Auditory Perception and Spatial (3D) Auditory Systems

Bill Kapralos

Michael R. M. Jenkin

Evangelos Milios

Technical Report CS-2003-07

July 20, 2003

Department of Computer Science

4700 Keele Street North York, Ontario M3J 1P3 Canada

Auditory Perception and Spatial (3D) Auditory Systems[4]

B. Kapralos[1,3], M. Jenkin[1,3] and E. Milios[2,3]

[1]Dept. of Computer Science, York University, Toronto, ON, Canada. M3J 1P3
[2]Faculty of Computer Science, Dalhousie University, Halifax, NS, Canada. B3H 1W5
[3]Centre for Vision Research, York University, Toronto, ON, Canada. M3J 1P3

{billk, jenkin}@cs.yorku.ca eem@cs.dal.ca

Abstract

In order to enable the user of a virtual reality system to be fully immersed in the virtual environment, the user must be presented with believable sensory input. Although the majority of virtual environments place the emphasis on visual cues, replicating the complex interactions of sound within an environment will benefit the level of immersion and hence the user's sense of presence. Three dimensional (spatial) sound systems allow a listener to perceive the position of sound sources, and the effect of the interaction of sound sources with the acoustic structure of the environment. This paper reviews the relevant biological and technical literature relevant to the generation of accurate acoustic displays for virtual environments, beginning with an introduction to the process of auditory perception in humans. This paper then critically examines common methods and techniques that have been used in the past as well as methods and techniques which are currently being used to generate spatial sound. In the process of doing so, the limitations, drawbacks, advantages and disadvantages associated with these techniques are also presented.

[4]The financial support of NSERC (Natural Sciences and Engineering Research Council of Canada), CRESTech (Centre for Research in Earth and Space Technology) and IRIS (Institute for Robotics and Intelligent Systems), is gratefully acknowledged.

Contents

Chapter 1

Introduction

The sounds we hear provide us with detailed information about our surroundings and can assist us in determining both the distance and direction to objects, at times, very accurately [159]. This ability is extremely beneficial for both humans and a variety of other species and in many situations, is crucial for survival. We can hear a sound in the dark where we may not necessarily make use of vision (sight) and in contrast to the limited visual field of view, the auditory system is omni-directional, allowing us to hear sounds reaching us from any position in three dimensional space. Given this omni-directional aspect, hearing serves to guide our visual senses, or to quote Cohen and Wenzel [30], "the function of the ears is to point the eyes". Hearing, or *audition* also serves to guide the more "finely tuned" visual attention system thereby easing the burden of the visual system [137].

Although sound is a critical cue to perceiving our environment, it is often overlooked in immersive virtual environments, where, historically, emphasis has been placed on the visual senses [30, 25]. The spatial audio cues present in many virtual environments are rather poor and do not necessarily reflect natural cues despite the fact that natural (spatial) sound cues can allow a user to orient themselves in a virtual environment. In addition, audio cues can add a "pleasing quality" to the simulation, add a better sense of "presence" or "immersion" and compensate for poor visual cues (graphics) [3, 137]. Furthermore, the virtual environments which actually employ spatial audio typically, assume a far field source acoustical model, emphasizing the direction (azimuth and elevation) to a sound source only, offering little, if any, sound source distance information [138, 108]. Despite the importance of distance discrimination in maintaining a sense or realism among the virtual sound sources [16], accurate sound source distance is often ignored in virtual audio

displays.

A three-dimensional (3D) (or spatial) audio system (or audio display) allows a listener to perceive the position of sound sources, emanating from a static number of stationary loudspeakers or a pair of headphones, as coming from arbitrary locations in three-dimensional space. Spatial sound technology goes far beyond traditional stereo and surround sound techniques by allowing a *virtual* sound source to have such attributes as left-right, back-forth and up-down [30]. The foundation of 3D audio rests on the ability to control the auditory signals arriving at the listener's ears such that these signals are perceptually equivalent to the signals the listener would receive in the environment being simulated [158]. When considering the design of any spatial sound system for use in a virtual environment, it is therefore necessary to consider issues related to human auditory perception [25].

In the natural environment, various acoustical cues, arising from the environment itself (e.g. source location, air propagation, reverberation etc.), as well as our own physical make-up (e.g. two ears physically spaced apart, notches and grooves of our pinnae etc.), allow us to localize sound sources. However, in a virtual environment, these cues may not necessarily be present and must therefore be simulated in order to reproduce (as closely as possible) the cues available under "natural" listening conditions. In order to localize a sound source, the human auditory system relies primarily on:

Interaural Time Difference (ITD): The difference in time between the arrival of the sound to each of the ears.

Interaural Level Difference (ILD[1]): The difference in sound pressure level (SPL) between the sound at both ears.

Head Related Transfer Functions (HRTFs): The complex interaction of a sound wave with the torso, shoulders, head and particularly the pinna (outer ear) of a listener. Essentially, the pinna of each ear filters every sound wave passing through it in some manner unique to the sound source position. Given these filtered signals, the brain estimates the exact 3D position of a sound source relative to the listener [34].

Reverberation: Reflections of the sound waves off of other objects in the environment (e.g. the walls of a room).

Interaction with Vision: We can determine the location of a sound source which we can see.

ITD and ILD cues are known as *binaural* cues, since they result from a comparison of the signals received at each ear. Monaural cues, such as HRTFs, result from the signal received at each ear independently, without any comparisons. Three-dimensional audio

systems spatialize a sound source by simulating some (or all) of the cues listed above. Although systems incorporating ITD and/or ILD cues only are fairly simple to model and implement, they generally produce poor results, providing limited sound spatialization capabilities. For example, a listener may not be able to disambiguate between a sound source directly in front or in back of them or from a sound source directly above or below them. As with human hearing in a natural setting and as will be described further in Section 1.2.2, the ability for a user in a virtual environment to spatialize a sound source and eliminate (or greatly reduce) such confusions can be greatly improved by incorporating HRTFs into the system.

The purpose of this report is to present an overview of some of the methods and techniques employed by 3D (spatial) sound systems in order to position virtual sound sources arbitrarily in three dimensional space. A review describing the biological and technical literature relevant to the generation of accurate acoustic displays for immersive projective virtual environments, is also provided. Common methods and techniques which have been used in the past as well as methods and techniques currently being used to generate spatial sound are critically examined. In the process of doing so, the limitations, drawbacks (and potential solutions to these drawbacks), advantages and disadvantages associated with these techniques, as well as their ability to be used in an immersive virtual reality system, are presented. Prior to discussing 3D sound technologies, however, given the importance of understanding the perception of sound and human sound localization, this report begins with a brief introduction on the physical attributes of sound and how these attributes can be measured, followed by an elaborate discussion of primary human auditory localization cues. Several auditory phenomena such as the precedence effect and auditory distance perception will be introduced. Chapter 2 presents a brief history of recording techniques, beginning with an introduction of monaural based systems, two-channel stereo based systems, and surround sound systems such as Quadraphonics and Ambisonics. Chapter 3 focuses on the simulation of human auditory localization cues for a spatial auditory display. In particular, this chapter describes techniques and models in order to re-create the cues available in our "natural" listening environment, including models to predict the ITD and methods which enable the measurement of Head Related Transfer Functions (HRTFs) for a specific position and techniques to model reverberation. Several techniques for conveying sound in a virtual environment using loudspeakers and the problems and drawbacks associated with these techniques is presented in Chapter 4, including transaural audio and amplitude panning. In addition, this chapter discusses some of the issues related to the presentation of spatial audio using headphones such as *in-head-localization* (IHL) and the *externalization* of sound. Finally, a summary and concluding remarks are presented in Chapter 5.

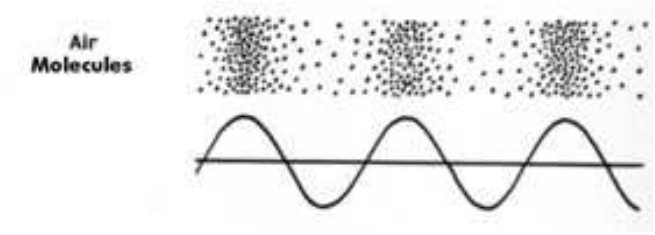


Figure 1.1: Sound waves consist of alternating regions of compression and rarefaction (e.g. "back and forth" motion) of the air molecules (top), corresponding to the "high" and "low" points of a "sine wave" (bottom).

1.1 What Exactly is Sound?

Sound results from the rapid variations in air pressure caused from the vibrations of an object (e.g. a vibrating guitar string, human vocal chords etc.) or an object in motion [102]. As shown in Figure 1.1, sound waves consist of alternating regions of compression and rarefaction (e.g. "back and forth" motion) of the molecules comprising the medium [167] (typically air although sound can also propagate through other mediums such as water or steel). The molecules themselves do not move with the wave but rather oscillate about some position. The wave itself propagates through the interaction of molecules in the medium. Considering a sound propagating through air, the air molecules surrounding the vibrating object will be compressed during forward motion and expanded during the object's backward movement. As these molecules are displaced, they will also "push or pull" the molecules neighboring them, causing these neighboring molecules to also be displaced from their resting position. This forward and backward movement of the molecules propagates throughout the entire medium, with each molecule displacing its neighbors. Sound waves may propagate in an *omni-directional* manner whereby the wave propagation is independent of direction (e.g. equal in all directions), or it may exhibit *directional* properties leading to wave propagation in a particular direction only.

Perception of sound begins with the arrival of this varying sound pressure at our ear drums. Through the actions of the eardrum these oscillating ("mechanical") variations of air pressure are passed through to the middle ear and converted (transduced) into electrical signals in the inner ear and ultimately coded into a pattern of neuronal spikes which are interpreted by the brain (a complete discussion of the physiology of the ear is beyond the scope of this report - see [102, 17] for greater details). However, the pattern of sound pressure variations arriving at our ears may not necessarily be identical to the pressure variations originally generated by the vibrating object. In order to propagate, sound waves require a medium (e.g. they are *mechanical* waves and therefore cannot travel in a vacuum). However, any medium (e.g. liquid, gas or solid), will affect the waves traveling through

it in some manner. As a wave propagates, a portion of it is absorbed by the medium, modifying the sound spectrum in some manner. The amount of absorption of a sound wave as it propagates through the medium is affected by the characteristics of the medium itself, including (when the medium is air for example) temperature, and humidity level (see Harris [66] for a detailed description on the effect of relative humidity and absorption of sound in air) and the distance the wave has traveled [74]. In addition, absorption of sound waves is also a function of frequency, where the higher frequency components are absorbed more readily than the lower frequency components. Furthermore, typical listening environments are *echoic*, as opposed to *anechoic*, whereby reflections and refractions result when the sound waves encounter any number of obstacles/objects in the environment (e.g. walls of a room), on their path from the source to the listener. In an anechoic environment there are (ideally) no reflections and the listener will receive the sound on the direct path from the sound source only (e.g. no reflections will arrive). Anechoic settings occur rarely in nature. A large open space or the top of a mountain summit does however approach anechoic [16]. Anechoic chambers are artificially created anechoic environments. They are rooms where the walls, floor and ceiling are covered with sound absorbing material to prevent any reflections of sound waves which may encounter any surfaces.

A very simple type of sound wave is a sinusoid (sine wave), illustrated in Figure 1.2. Sinusoids are also known as *tones* or *pure tones* and actually result in simple auditory responses, producing a very "clean" sound [102]. Mathematically, a sinusoid $x(t)$ can be described as:

$$x(t) = Acos(2\pi f_o t + \phi) \tag{1.1}$$

$$p = \frac{1}{f_o} \tag{1.2}$$

where, referring to Figure 1.2, A is the *amplitude* or *intensity* of the sine wave (e.g. amount of variation about the mean), f_o is the *frequency*, representing the number of "cycles" per second or in other words, the number of times each second the sinusoid repeats itself, measured in Hertz (Hz) and ϕ is the *phase* or relative starting time (generally important only when considering more than one sinusoid). The time taken for one complete cycle of the wave is known as the *period* p and can be obtained by taking the reciprocal of the frequency. Waves exhibiting this periodic property are known as *periodic* and periodicity is certainly not specific to sinusoidal waveforms as non-sinusoidal waveforms can also can be periodic.

Although sinusoids are very simple to analyze, they are not typically encountered in normal listening situations. Rather, the sounds we hear under normal listening conditions

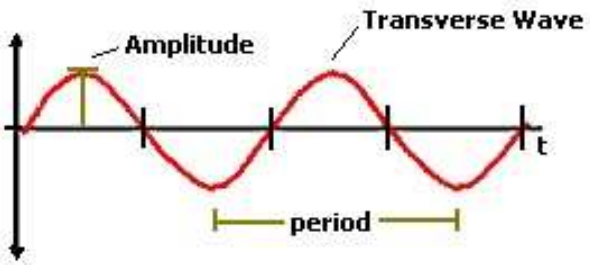


Figure 1.2: A sinusoid (sine) wave is a very simple sound wave. Taken from [106].

are much more complex and may not even be periodic. These complex waveforms can be "broken-down" into a series of sinusoids, each with its own frequency, amplitude and phase, using Fourier analysis [110]. A complex tone (periodic and non-periodic) can be described as the superposition of a number of sinusoids, where the frequency of each sinusoid is an integral multiple of the *fundamental frequency*, the frequency of the lowest common "fundamental" component which may not necessarily be present [102]. Frequencies other than the fundamental are known as the *harmonics*, where the first harmonic is the first multiple of the fundamental, the second harmonic is the second multiple of the fundamental and so on. For example, a square wave consists of a fundamental frequency sinusoid and the superposition of the odd harmonics of the fundamental (e.g. if the fundamental is 100Hz, the odd harmonics are 300Hz, 500Hz, 700Hz etc.). The amplitude of each harmonic is equal to the amplitude of the fundamental scaled by the inverse of the harmonic index (see Figure 1.3). Mathematically, a square wave $x(t)$ is represented as follows:

$$x(t) = \sum_{K=1,3,5,\ldots}^{\infty} \frac{1}{K} \sin(2\pi K f) \tag{1.3}$$

where, f is the fundamental frequency and K is the harmonic index (a square wave contains an infinite number of odd-harmonics). A discussion of Fourier analysis is beyond the scope of this report, however, an excellent mathematical description is provided by Oppenheim et. al. [110].

Finally, the range of frequencies to which humans are sensitive (e.g. can hear) is restricted to the range of 20Hz to 20kHz for a young healthy adult [25]. Frequencies below 20Hz are known as *subsonic* and can at times be felt rather than heard, while frequencies above 20kHz are known as supersonic [16].

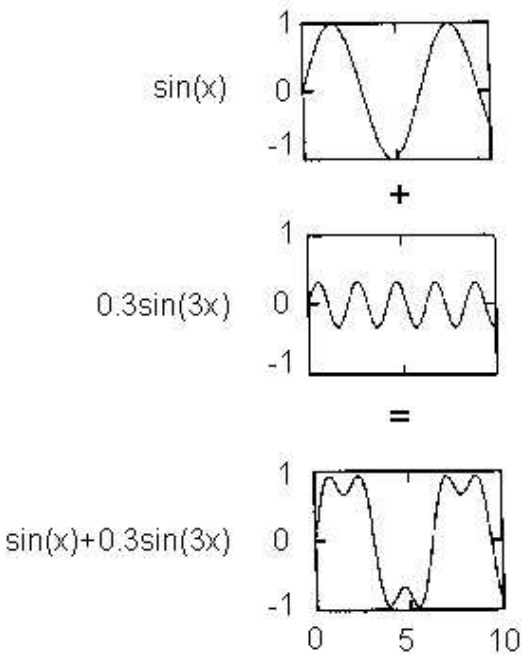


Figure 1.3: A square wave consists of a fundamental frequency sinusoid and the superposition of the odd harmonics of the fundamental. The amplitude of each harmonic is equal to the amplitude of the fundamental scaled by the inverse of the harmonic index. In this example, the superposition of a sine wave of frequency "x" (top diagram) along with its first odd harmonic (middle diagram) of "3x" suitably scaled (e.g. by $\frac{1}{3}$), produces the square wave approximation shown in the bottom diagram. As more scaled multiples of the fundamental are added, the superposition approaches the ideal square wave.

1.1.1 Measuring Sound

As described in Section 1.1, sound results from the variation of pressure arising when the molecules in the medium of propagation are compressed and expanded due to a vibrating object. Intensity is usually used to specify the magnitude of these variations (the compressions and expansions of the medium of propagation) and is defined as the sound energy transmitted each second through a unit area in a sound field [102].

The range of intensity levels that the human auditory system is sensitive to is very large, and therefore, rather than giving direct intensity measures, a logarithmic scale is used instead. Given this logarithmic scale, the measures are therefore known as *levels* and specified as a ratio with respect to some reference intensity measure [102]:

$$S_L = 10 \times log_{10}\left(\frac{I_1}{I_0}\right) \tag{1.4}$$

where, S_L is the number of Decibels (dB) corresponding to the ratio of intensities between I_1 and the reference intensity I_0. With a decibel scale, a 3dB increase in the intensity ratio corresponds to a doubling of the ratio of intensities.

Although the sound level ratio between two intensities can be determined, there may be times where a single measure of intensity is required. To allow for such a situation, a standard reference intensity level is used. The standard reference level chosen is the threshold of human hearing for a 1000Hz tone and is equal to 10^{-12}W/m^2 (Watts per meter square) or 20μPa (micropascals) when considering pressures [67]. Intensity levels given relative to this particular reference level are known as a *sound pressure level* (SPL). As an example, a sound level of 3dB SPL represents an intensity twice that of the reference level, while a sound level of 0dB SPL represents an intensity equal to the reference level. Finally, intensity ratios can also be given as pressure ratios as well since there is a relationship between intensity and pressure (e.g. intensity is proportional to the square of pressure) [67]:

$$\begin{aligned} S_L &= 10 \times log_{10}\left(\frac{I_1}{I_2}\right) && (1.5)\\ &= 10 \times log_{10}\left(\frac{P_1}{P_2}\right)^2 && (1.6)\\ &= 20 \times log_{10}\left(\frac{P_1}{P_2}\right) && (1.7) \end{aligned}$$

where P_1 and P_2 are the two pressure measurements in Pa (Pascals). As with intensity level, pressure level can also be given relative to the standard measurement, where the standard pressure measurement is 20μPa (e.g. $P_0 = 20\mu$Pa).

1.1.2 Near Field vs. Far Field

In physical acoustics, when describing the distance to a sound source, a distinction is made between a sound source in the *near field* and in the *far field*. When the distance to the sound source is "very large", the sound source is said to be in the far field and the sound waves reaching a listener are assumed to be *planar*. On the other hand, the sound waves reaching the listener from a sound source that is "very close" are not planar but rather are spherical in nature and therefore curved with respect to the listener's head.

The notion of a "very large" source distance or of a sound source "very close" to the listener is not very clear. Brungart and Rabinowitz [21], define the near field as "the region of space surrounding the listener within a fraction of a wavelength away from the sound source". Using this definition, the designation of a near field vs. a far field sound source is frequency dependent given the inverse relationship between frequency and wavelength. However, for practical considerations, assuming propagation in the air, when the distance to a sound source is greater than approximately one meter, a far field source is assumed [21] and the propagating waves are approximated by planar waves (for propagation underwater, you must multiply by a factor of four). As a result, *binaural* localization cues (e.g. cues involving a comparison between the signals arriving at both ears) are assumed to be independent of source distance (e.g. source distance can be ignored). Generally, a sound source within one meter of an observer is considered to be in the "near field". Given the spherical nature of the sound waves in the near field, ILD cues as well as monaural spectral cues (HRTFs) are very dependent on sound source distance and unlike the planar waves, these spherical waves are influenced greatly by such factors as head size and pinnae structure [21].

1.1.3 Coordinate System

In any audio environment, the position of a sound source is given relative to some reference point. In a single user system, typically the listener is chosen as the reference point and sound source positions are given relative to the listener. However, with systems supporting multiple users, some arbitrary point may be used instead. Various coordinate systems exist. In the "head centered rectangular system", the center of the head defines the origin, with positive x-axis (also known as the *interaural axis*) going through the right ear, positive y-axis pointing directly in front of the head while positive z-axis points directly upwards

(vertically). In this coordinate system, the axes form three planes (see Figure 1.4). The y-z axis form the *median* (or *sagittal*) plane whereby any point on this plane is equidistant from the left and right ears. The x-y plane is known as the *horizontal* plane and is level with the listener's ears and finally, the x-z plane is referred to as the *frontal* plane.

Rather than specifying individual x, y, z axis components, a "spherical coordinate system", in which coordinates are specified with an *azimuth*, *elevation* and *range*, may be used instead. In the "single pole" spherical system (see Figure 1.5(a)), the center of the head defines the origin while azimuth (θ) and elevation (ϕ) are specified by lines of latitude and longitude respectively [25]. An azimuth angle of 0^o is directly in front (e.g. median plane) while an angle of -90^o is directly to the right (e.g. moving clockwise from 0^o results in negative azimuth angles). The horizontal plane is at an elevation of 0^o and moving upwards from this point, elevation increases positively, with $+90^o$ directly on top of the head. Range specifies the distance between the origin (center of the head) to the point of interest. The single pole system is intuitive and the most widely used coordinate system. However, it does have its problems. Most importantly, the length of an arc length (semi-circle) between two angles of azimuth is dependent on elevation. For example, the arc length between 0^o and 90^o azimuth at an elevation of 0^o is greater than the same arc at an elevation of 75^o. As will be described in Section 3.3.3, this dependence of arc-length and elevation may be problematic when measuring HRTFs.

Another spherical coordinate system is the "double pole" system (Figure 1.5(b)). In the double pole system, elevation is specified in the same way as in the single pole system however, azimuth is given as a series of rings which are parallel to the "midline" (the z-axis) and centered at the poles at each interaural axis [25]. In this system, the arc length between two angles of azimuth is independent of elevation however, this system is not as intuitive as the single pole system and is therefore not widely used.

For the remainder of this report, unless specified otherwise, all positions are specified with respect to the single pole polar system.

1.2 Sound Localization

In this section, human sound localization will be introduced, beginning with the duplex theory formulated in the early 1900s followed by the head related transfer functions (HRTFs). In addition, several other sound localization cues will also be introduced, including reverberation, precedence effect and head movements. Finally, a description of auditory distance perception will also be provided.

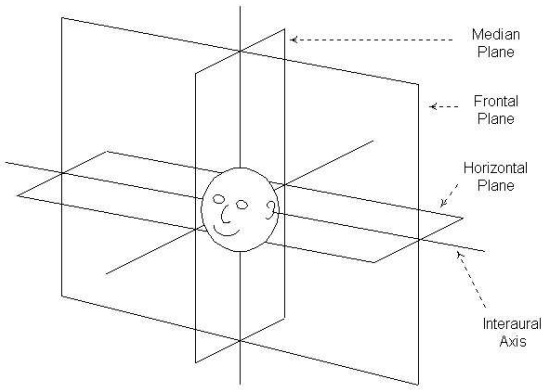
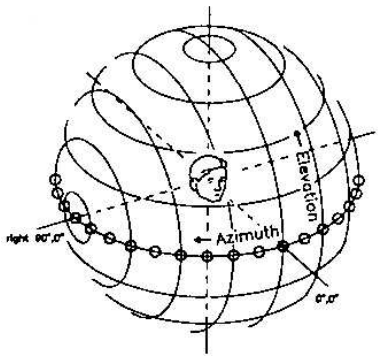
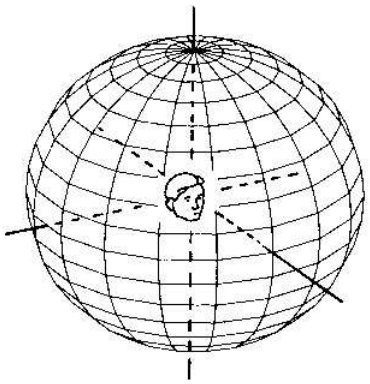


Figure 1.4: Coordinate system: Three planes of interest. Reprinted from [82].



(a) Single pole polar coordinate system.



(b) Double pole polar coordinate system.

Figure 1.5: Coordinate system: Single (a) and double (b) pole coordinate systems. Reprinted from [25].

1.2.1 Duplex Theory

The *duplex theory* formulated by Lord Raleigh [78] is a theory of human sound localization based on the two binaural cues, interaural time delay (ITD) and interaural level difference (ILD) and on the assumption that the head is spherical with no external ears (pinnae). These two cues arise from the fact that the two ears do not share the same position in space but are rather separated by the (rather large) head. Given this separation, unless the sound source lies on the median plane, the distance traveled by the sound waves emanating from the sound source to the listener's left and right ear will differ. This will cause the sound to reach the *ipsilateral* ear (the ear closest to the sound source) prior to reaching the *contralateral* ear (the ear farthest from the sound source). The difference between the onset of non-continuous (transient) sounds or phase of more continuous sounds [25] at both ears is known as the interaural time delay (ITD). Similarly, given the separation of the ears by the head, when the wavelengths of a sound are short relative to the size of the head, the head will act as an "acoustical shadow", attenuating the sound pressure level of the waves reaching the contralateral ear [164]. This difference in level between the waves reaching the ipsilateral and contralateral ears is known as the interaural level difference (ILD).

When the sound source lies on the median plane, the distance from the sound source to the left and right ear will be the same therefore causing the sound to reach each of the ears at the same time. In addition, the sound pressure level of the sound at both ears will also be the same. As a result, both the ITD and ILD will be (near) zero. As the source moves to the right or left ITD and ILD cues will increase until the source is directly to the right or left of the listener respectively (e.g. $\pm 90^o$ azimuth). Similarly, when the sound source is directly behind the listener, both ITD and ILD will be (near) zero and as the sound moves to the right or left, ITD and ILD cues will increase until the sound source is directly to the left or right of the listener.

"Separation" of ITD and ILD Cues

Although the duplex theory incorporates both ITD and ILD cues, they do not necessarily operate together. ITDs are prevalent primarily for low frequencies, less than approximately 1500Hz [17], where the wavelengths of the arriving sound are long relative to the diameter of the head and the phase of the sounds reaching the ears can be determined without ambiguity. For wavelengths smaller than the diameter of the head, the difference in distance may be greater than one wavelength, leading to an ambiguous situation (e.g. aliasing), where the difference does not correspond to a unique location [27].

For low frequency sounds in which the ITD cues are prevalent and the waves are greater than the diameter of the head, the sound waves experience *diffraction* whereby, they are not blocked by the head but rather they "bend" around the head to reach the contralateral ear. As a result, ILD cues for these low frequency sounds will be very small (although they

can at times be as large as 5dB [164]). However, for frequencies greater than approximately 1500Hz, where the wavelengths are smaller than the head, the wavelengths are too small to bend around the head and are therefore blocked by the head (e.g. "shadowed" by the head). As a result, a decrease in the energy of the sound reaching the contralateral ear will result and hence the ILD cue.

Shortcomings of the Duplex Theory of Sound Localization

The duplex theory can explain localization of a sound source in the azimuthal plane, where a sound is perceived to be closer and louder to the ear in which the sound first arrives. However, the duplex theory alone is incomplete as it cannot account for many aspects of human auditory localization [25]. We are capable of localizing a sound source even with a single ear as evidenced by listeners who are deaf in one ear [142]. Furthermore, ITD and ILD cues are not unique. According to the duplex theory, for sound sources located on the median plane (e.g. $\theta = 0^o$ or $\theta = 180^o$), or directly above or below the listener (e.g. $\phi = 0^o$ or $\phi = 180^o$), both ITD and ILD cues are (nearly) zero, presenting an ambiguous situation. Indeed, as shown in Figure 1.9, a sound source positioned anywhere on the surface of a cone (the *cone of confusion*), centered on the interaural axis will have identical ITD values [82]. Strictly speaking, the cone of confusion as well as ITD and ILD values of zero occur only in theory with the assumption of a perfect spherical head without the external ears (pinnae). In reality, of course, our head is not completely spherical and we certainly cannot disregard the effects of the pinnae (as discussed in Section 1.2.2). As a result, ITD and ILD cues are never really zero and will differ slightly even when the source is directly in front or directly in back of us. More generally, when the ITD or ILD cues are similar for two different locations, an ambiguous situation can potentially arise without the presence of any other cues [16].

Under normal listening conditions, humans are capable of resolving ambiguous situations such as the front-back confusions, leading many researchers to believe the duplex theory is incomplete. Although it does have its shortcomings, the duplex theory remained the dominant theory of human auditory localization for about half a century after its introduction in the beginning of the 20th century. However, as described by Wightman and Kistler [164], the next "revolution" in the study of human auditory localization occurred with the theories published by Batteau [8] in 1967, on the filtering effects introduced by the pinnae (external ear). These monaural filtering effects have come to be known as head related transfer functions (HRTFs) and allow us to overcome the localization limitations inherent using ITD and ILD cues alone. Greater details regarding the head related transfer functions and their importance to sound localization are provided in the following section.

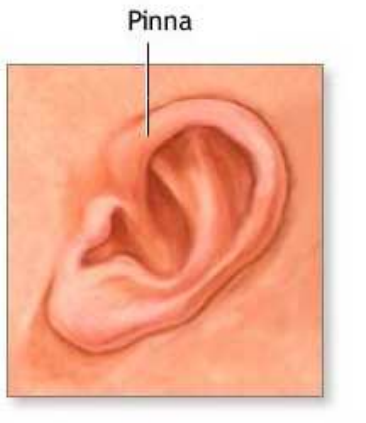


Figure 1.6: Diagram of the human pinna. After [95].

1.2.2 Head Related Transfer Function (HRTF)

The filtering of the sound source spectrum caused by the complex interactions of the sound waves with the head, shoulders, torso and particularly the outer ear (pinna or *auricle*) prior to reaching the ear drum (in addition to the interaural time delays and level differences), are collectively known as the *head related transfer functions* (HRTFs). The physical structure of each person's pinna consists of a series of grooves and notches and varies widely amongst individuals (an illustration of a typical person's pinna is provided in Figure 1.6). These asymmetrical grooves and notches accentuate or suppress the mid and high frequency energy content of the sound spectrum to a certain degree, depending very much on both the location and frequency content of the sound source. The multiple reflections of the sound waves off the grooves and notches of the pinnae lead to very small time delays, in the order of 0 - 300μs, once again depending on the source location [7]. Essentially, the HRTF modifies the spectrum and timing of a sound signal reaching the ears in a location dependent manner which is recognized by the listener and used as a localization cue [16].

Mathematically, according to Zotkin et. al. [171], the left and right ear HRTFs (H_L and H_R respectively), can be defined as the ratio between the sound pressure level (SPL) present at the eardrum of the left and right ears, $\Phi_L(\omega, \theta, \phi, d)$ and $\Phi_R(\omega, \theta, \phi, d)$, and the *free field* SPL at the position corresponding to the center of the head but with the head absent $\Phi_f(\omega)$:

$$H_L = \frac{\Phi_L(\omega, \theta, \phi, d)}{\Phi_f(\omega)}, \qquad H_R = \frac{\Phi_R(\omega, \theta, \phi, d)}{\Phi_f(\omega)} \tag{1.8}$$

where ω is the angular frequency, θ and ϕ are the azimuth and elevation angles and d

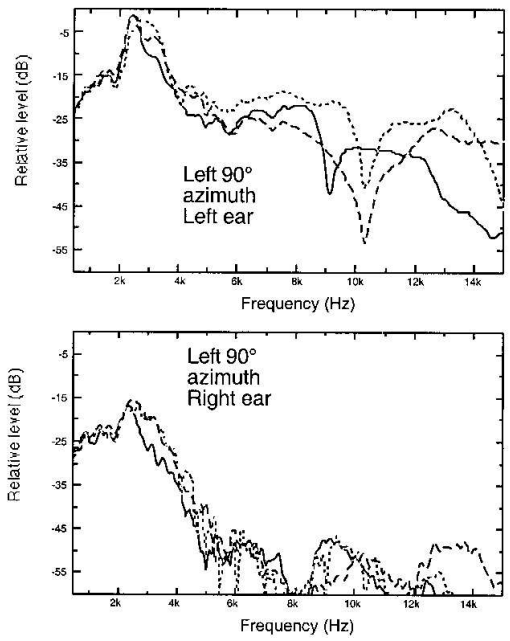


Figure 1.7: Example left and right ear HRTF measurements for a source at azimuth $\theta = 90^o$ and elevation $\phi = 0^o$. Reprinted from [16].

is the distance from the listener to the sound source (measured from the center of the listener's head). Example HRTFs from three individuals, as measured by Wightman and Kistler [16] are shown in Figures 1.7 and 1.8. (See Section 3.3 for greater details regarding the measurement of HRTFs.) Figure 1.7 top and bottom illustrate the resulting left and right ear HRTFs respectively, for a sound source located at $\theta = 90^o$ and $\phi = 0^o$ (e.g. directly to the left of the listener), The left and right HRTFs for a sound source located at $\theta = 0^o$ and $\phi = 36^o$ are shown in Figure 1.8 (top and bottom respectively). Examination of each plot reveals several differences. The inter-subject differences in each plot are clearly evident (the HRTF for each individual of each plot are denoted by the three different line styles: non-dashed, small dashes and large dashes), especially for higher frequencies i. e. frequencies greater than approximately 5kHz.

HRTFs can provide information used to judge vertical directions and to disambiguate front-back confusions [102]. Many studies have been performed in order to investigate the filtering effects of the pinnae. In several studies, when portions of the outer ear were occluded (filled with plasticine for example) [48, 109], an increased number of front-back confusions and a decrease in elevation accuracy occurred. The filtering actually performed on the source spectrum is dependent on the source frequency content as well. Studies have shown that the number of front-back ambiguities increases and localization accuracy

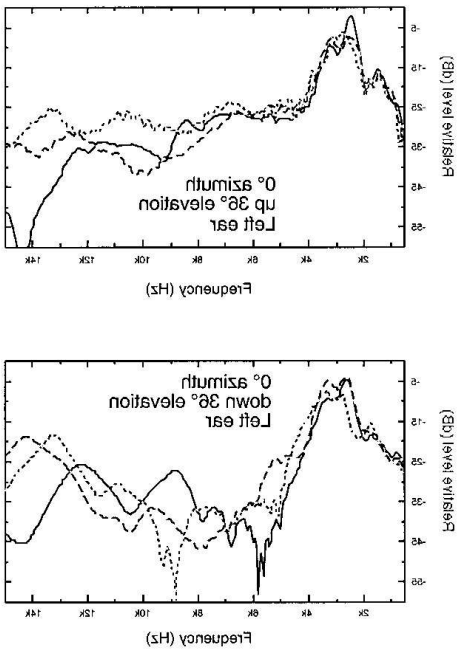


Figure 1.8: Example left and right ear HRTF measurements for a source at azimuth $\theta = 0^o$ and elevation $\phi = 36^o$. Reprinted from [16].

decreases as the bandwidth of the source is decreased [25, 17, 100], leading Carlile [25], to believe that, to allow for accurate source localization, a source containing a wide range of frequencies is required.

Greater detail regarding both the measurement and use of HRTFs in a spatial auditory display as well as the problems associated with their use, is provided in Section 3.3.

1.2.3 Reverberation

Various factors affect a propagating sound wave before it reaches the listener (receiver). The condition of the air itself (e.g. humidity level, heat etc.) may have an effect on the propagating waves (see Section 1.2.6 for further details on how the medium affects a propagating sound). In addition, the sound waves may encounter any number of objects and obstructions both on the path from the source to the listener and after reaching the listener (e.g. the listener will not solely receive and completely absorb the sound waves, but rather, a portion of the wave Γ may continue to propagate). When a sound wave encounters an object, the object itself may absorb a portion of the wave while the remainder is reflected in some other direction. In other words, typical environments are rarely anechoic, except,

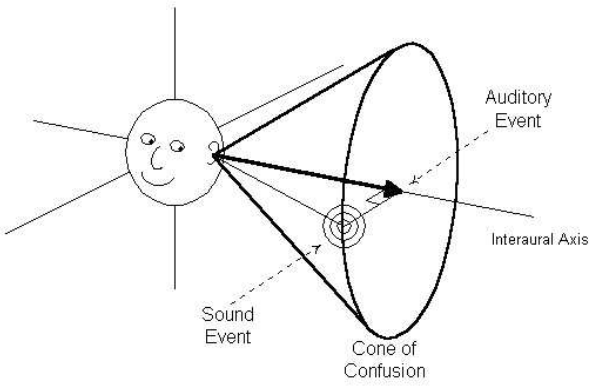


Figure 1.9: Cone of confusion. A sound source positioned anywhere on the surface of the cone will produce an identical ITD value. After [82]

as previously described, in certain infrequent situations such as within a large open area with snow covered ground or on a mountain summit [16]. As shown in Figure 1.10 in a typical listening environment, sound waves emitted by the source reach the listener both directly, via the straight line path between the source and receiver (assuming there is such a path) and indirectly as reflections (e.g. echoes) from any walls, floor, ceiling or any other obstacles and obstructions. This collection of reflected waves, which may consist of several thousands, reflecting from the various surfaces within a space, is known as *reverberation* [29].

The collection of reflected sound reaching the listener varies as a function of the geometry of the room relative to the listener [25], as well as the material of the room, the source spectrum (e.g. frequency components) and is rather irregular [53]. As will be described in Section 1.2.6, reverberation can also be used as a cue to source distance estimation, and can also provide information with regards to the physical "make-up" of a room (e.g. size, types of materials on the walls, floor, ceiling). Reverberation can also add a pleasing "lively" aspect to voice and music [159], making it attractive to the music recording and entertainment industry [169]. Radio and home theater manufacturers have also taken advantage of the benefits reverberation has to offer. Many radios, sound systems and home theater systems include DSP technology offering various reverberation settings. Greater details regarding the characteristics of reverberation are provided in the following section.

Characteristics of Reverberation

Reverberation results from the reflections of the sound waves. Prior to reaching the listener, a sound wave may be reflected multiple times from different surfaces. The number of times

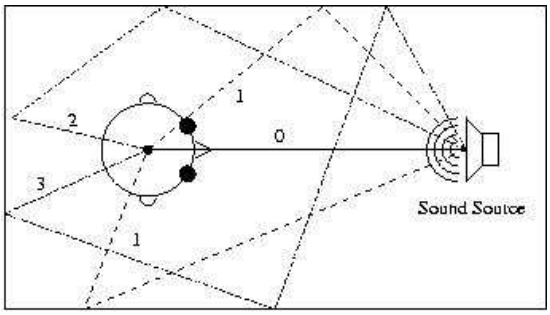


Figure 1.10: Direct and reflected sound waves reaching a listener. In addition to the sound waves reaching a listener via a "straight line path" directly from the sound source, reflected sound will also reach a listener. The number of times a wave is reflected before reaching a the listener is know as its *order*. The wave order in this example is provided next to the reflected wave. The direct sound has an order of zero. In a typical scenario, the number of reflected waves may reach several thousands.

a wave is reflected is denoted by its *order* (e.g. a reflection of order n indicates the wave has been reflected n times). In many situations, a higher reflection order, indicates a reduction in the intensity level due to the absorption by the reflecting surfaces and the inverse square law characteristics of propagating waves [141]. An example illustrating the order of reflected waves is provided in Figure 1.10.

In addition to the reflection order, reverberation can be broken down into two categories: *early* and *late* reflections. Reflections of order one, resulting from the room boundaries (e.g. walls, floor and ceiling), are known as early reflections and typically arrive within 80ms of the direct sound. Reflections arriving after 80ms and with reflection orders greater than one are known as late reflections. Late reflections, arising from "reflected reflections" from one surface to another, are assumed to arrive equally from all directions (e.g. diffuse) and can be described statistically as exponential decaying noise [53]. A graphical illustration of the concepts described above is provided in Figure 1.11, where the theoretical "room impulse response" (see Section 3.4.1) is shown.

Other parameters used to describe reverberation include *reverberation time* and *reverberation distance*. A definition of each parameter, according to Garas [53], is provided in the following sections.

Reverberation Time and Reverberation Distance

Reverberation time T_{60} can be defined as the time required for the sound pressure level (SPL) to be attenuated by 60dB (e.g. by a factor of one million), independent of the intensity of the sound after a steady state sound is turned off and can be approximated by [53]:

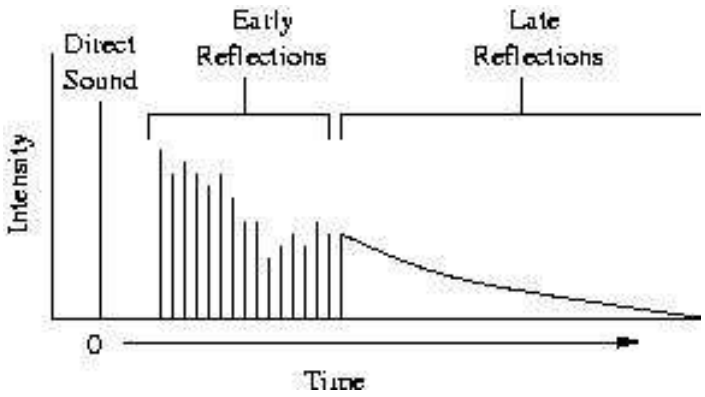


Figure 1.11: Theoretical room impulse response. In addition to the sound waves reaching a listener by traveling from the sound source to the listener directly, indirect sound waves reflected from the walls, floor or other objects in the environment will also reach the listener, albeit after the direct waves. Early reflection will occur within 80ms of the arrival of the direct sound. Reflections arriving after 80ms can be considered diffuse and can be described as exponentially decreasing noise.

$$T_{60} \approx \frac{V}{6 \times \beta \times S} \tag{1.9}$$

where V is the volume of the room (in m^3), β is the (frequency dependent) average absorption coefficient of the room boundaries and S is the sum of the surface areas of the room in m^2.

Reverberation time, as given, is rather arbitrary and depends on the characteristics of the enclosure, including the material of the walls, floor and ceiling, number and type of objects in the room etc. Depending on the level of the background noise, it may be the case that reflections arriving after T_{60} are still considerably audible [16]. However, the choice of 60dB was made by considering a good "music making area", such as a concert hall. In such a situation, the loudest level reached for most orchestral music is typically 100dB (SPL), while the level of background noise is around 40dB. As a result, a reverberation time of 60dB can be seen as the time required for the loudest sounds of an orchestra to be reduced to the level of the background noise.

Reverberation time is highly affected by the reflective surfaces encountered by the propagating waves. When a surface is highly reflective, very little energy is absorbed by the surface (e.g. the reflected wave contains most of its energy) leading to an increase in

the reverberation time. In contrast, highly absorbing materials will absorb much of the energy of a wave striking it, greatly reducing the energy in the reflected portion thereby reducing the reverberation time.

Late reflections can be considered diffuse, arriving any time after 80ms of the direct sound. However, as the distance between the source and listener d_s increases, the intensity (loudness) of the direct sound L_{direct} will decrease until the level of the direct sound equals the level of the reverberation L_{reverb}. Reverberation distance d_{reverb} is defined as the distance such that $L_{direct} = L_{reverb}$ and is given by the following expression [53]:

$$\begin{aligned} d_{reverb} &= 0.25 \times \sqrt{\frac{\beta \times S}{\pi}} \\ &= 0.006 \times \sqrt{\frac{V}{T_{60}}} \end{aligned} \tag{1.10}$$

1.2.4 Precedence Effect

In a typical listening situation, the listener receives the direct sound emitted by the sound source as well as delayed and attenuated versions of the direct sound resulting from the reflection of the sound with objects in the environment. The reflected sounds reaching the listener may emanate from any direction in the environment, potentially creating a false impression of a sound source at the location of reflection. However, this is certainly not the case as the auditory system can clearly localize a sound source in the presence of multiple reflections (reverberation). The ability of the auditory system to "combine" both the direct as well as reflected sounds such that they are heard as a single "entity" and localized in the direction corresponding to the direct sound has been termed the *precedence effect* by Wallach et. al. over fifty years ago [157] (also known as the *Haas effect* and the *law of first waveforms*). The precedence effect allows us to localize a sound source in the presence of reverberation, even when the energy of the reverberant sound is greater than that of the direct sound [102, 68].

Since the work of Wallach et. al. others have conducted various experiments to investigate the precedence effect. As described by Grantham [62], typically, these experiments include a listener and two loudspeakers, placed at differing locations, in an anechoic environment. One loudspeaker is used to provide the direct sound while the other provides a delayed and appropriately attenuated version of the direct sound in order to simulate a reflection. Such studies indicate the following:

- When the reflection and direct sound are presented simultaneously (e.g. a delay of

zero), a single sound source (virtual source) is perceived at a location about half way between the two loudspeakers.

- As the time delay between the direct sound and the echo is increased from 0 to 1ms the location of the perceived sound source moves towards the "direct sound loudspeaker" (this is known as *summing localization* [17]).
- When the delay is between 1 and 30ms the sound source is correctly localized (e.g. coming from the direct sound loudspeaker) without being affected by the reflected sound.
- When the delay exceeds approximately 30 to 35ms, the direct sound is correctly localized however, the delayed sound is also localized at the position of the "reflection loudspeaker".

The experiments show how we are capable of correctly localizing a sound source in the presence of reverberation provided the reflections arrive within a short period after receiving the direct sound.

1.2.5 Head Movements

ITD and ILD cues alone are not unique and may therefore result in ambiguous localization judgments, as evidenced by the cone of confusion. As previously described, other cues, most notably the filtering described by HRTFs, are used in order to resolve such ambiguities. Furthermore, in any normal listening environment, we are not stationary but are rather free to move about. In particular, we can move our heads, from side-to-side, up and down or in any other manner. These head movements are a very important and natural component of sound localization which can greatly reduce front-back confusions and can increase sound localization accuracy [156, 152, 165]. Head movements result in a change of position between the sound source and the listener, leading to changes in the ITD and ILD cues. According to Begault [14], we are capable of integrating these changes as they occur over time in order to resolve ambiguous situations such as front-back confusions. Referring to Figures 1.12 and 1.13, the following example illustrates how a simple head movement can be used to overcome an ambiguous front-back situation. Consider a sound source directly in front of a listener (e.g. $\theta = 0^o$). In such a situation, both ITD and ILD will be negligible and the listener will not be able to determine whether the sound source is directly in front or in back of them (assuming ITD and ILD cues are the only available cues and the listener of course cannot see the sound source). Now suppose the listener rotates his/her head to the left by a certain amount. Since the head has rotated, the ears

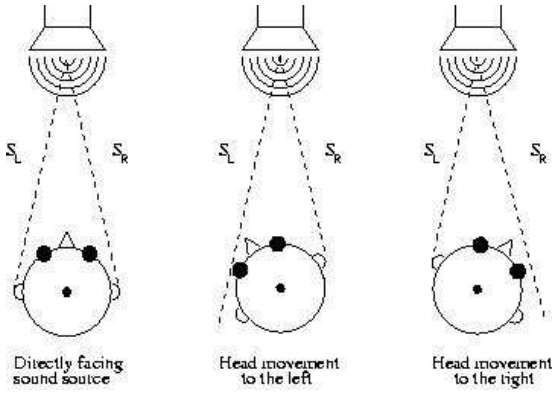


Figure 1.12: Head movements to resolve front-back ambiguities. When the sound source is directly in front of the listener, the path length to the left and right ears (S_L and S_R respectively), is the same. A head movement to the left will increase the distance between the left ear and the sound source S_L, while a head movement to the right will increase the distance between the right ear and the sound source S_R.

have moved from their initial position to some new position. Although the sound source has not actually moved from its initial location, relative to the listener, the sound source position has changed relative to the listener (e.g. whether the listener moves or the sound source moves, there is a relative change between them). Now, the sound source is no longer on the median plane but, is rather closer to the right ear. As a result, ITD and ILD have now increased. A similar situation arises when the listener's head is rotated to the right. However, in this case, the sound will be closer to the left ear.

Since the source was directly in front of the listener prior to the head movement, a head rotation to the left will bring the sound source closer to the right ear, while a head rotation to the right would bring the source closer to the left ear. However, had the source been directly in back of the listener, as shown in Figure 1.13, a head movement to the left would bring the sound source closer to the left ear while a head rotation to the right would result in the source being closer to the right ear. As a result, head movements to the left or right bring the source closer to one ear and which ear is actually closer to the sound source depends on whether the sound source is directly in front or in back of the listener, thus, eliminating any ambiguities. Finally, when head movements produce no change in ITD or ILD cues, the sound source lies directly above or below the listener.

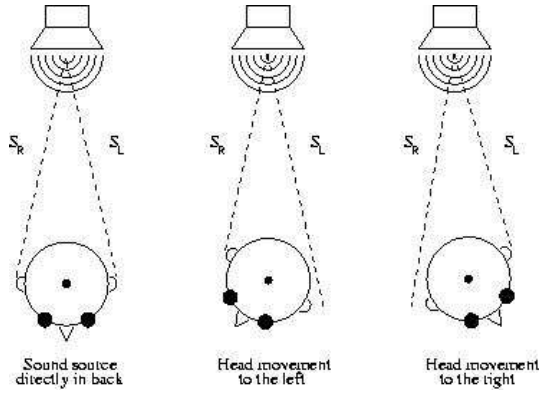


Figure 1.13: Head movements to resolve front-back ambiguities. When the sound source is directly in back of the listener, the path length to the left and right ears (S_L and S_R respectively), is the same. A head movement to the left will increase the distance between the right ear and the sound source S_R, while a head movement to the right will increase the distance between the left ear and the sound source S_L.

1.2.6 Auditory Distance Perception

The auditory system is capable of localizing a sound source using a variety of cues, including interaural time and intensity (level) differences and the head related transfer functions (HRTFs). However, when the sound source is in the far field (e.g. source distance greater than about 1m), these cues offer, primarily, directional information (e.g. azimuth and elevation), providing few, if any, cues to source distance.

Vision can be used to determine the distance to a sound source when the sound source is within our visual field of view but, the extent of the visual field of view is limited and is of little use in a very dark environment, when the sound source is not within the visual field of view or when a person is visually impaired. With the use of vision, auditory distance discrimination may not be as important however, auditory distance discrimination is of great importance when we cannot make use of visual cues such as in the dark, when the sound source is not within the visual field of view, or if a person is visually impaired. In such a situation, auditory information can be of critical importance and can be used to determine the distance to a sound source or the distance to some object in our environment which we cannot see. For example, by emitting certain sounds (e.g. "clicking" or "hissing" sounds from the mouth or tapping a cane on the ground), the visually impaired are capable of estimating the distance to an object(s) using the direct relationship between the distance to the object and the time taken for the reflected sound waves to reach the observer (see Section 1.2.3 for greater details). The greater the distance between an observer and a sound source, the greater the time required before the arrival of any reflections. This concept has been used in various "environmental" and navigational aids for the visually

impaired. For example, Milios et. al. [101] developed a device which converts a stream of range measurements obtained with a laser range finder into an auditory signal whose frequency and/or amplitude varies as a function of range, in order to provide the visually impaired user a greater sense of "spatial perception".

The ability to locate objects using reflected sounds (echoes) was termed *echolocation* by David Griffin [64] and has been used by the visually impaired to avoid obstacles which may be in their way [128]. Despite contrary belief, echolocation is certainly not exclusive to the visually impaired. Studies have shown that both the visually impaired as well as sighted blindfolded subjects are capable of employing echolocation to estimate the distance to, width and even material decomposition of objects [149, 16]. Similarly, various animals, most notable bats and dolphins, utilize echolocation to navigate and search for food [154]. Bats are very proficient with echolocation and can, using echolocation, determine the size, shape and texture of tiny insects.

Various researchers have examined the perception of auditory distance by humans (e.g. [52, 155, 32, 98, 16, 170, 107]) however, very little is known due to the inherent difficulties associated with the sound stimulus used in auditory distance experiments [170]. According to Coleman [32] and Mershon and King [99], the following auditory distance cues may potentially play a role in the perception of the distance to a sound source when both the observer and the sound source are stationary:

1. Intensity (sound level) of the sound waves emitted by the source.
2. Reverberation (direct-to-reverberant energy).
3. Frequency spectrum of the sound waves emitted by the sound source.
4. Binaural differences (e.g. ITD and ILD).
5. Type of stimulus used (e.g. familiarity with the sound source).

Source intensity (sound level) and reverberation (direct-to-reverberant energy) are believed to be the most effective cues [53], however, any number of these cues may be present and certain cues may dominate depending on the listening environment. As a result, auditory distance perception may be influenced by such factors as the user's familiarity with the room as well as the stimulus and the distance estimation process actually employed by a listener must adapt to the cues which may be available in each situation. In addition, changes in these cues may not necessarily be due to a change in distance between the listener and the source, but rather, may result from changes in the spectrum emitted by the source (e.g. the source power is reduced), or changes to the source spectrum due to changes in the environment, thereby further complicating matters, leading to poor judgments in

source distance estimation [170]. For example, as described in Section 1.2.6, as source distance is increased, the intensity of the sound received by the listener decreases. However, sound source intensity of the sound waves received by the listener may also decrease without an increase in source distance, but rather with a decline in source intensity. In such an ambiguous situation, the user may not necessarily be able to discriminate between the two scenarios. Fortunately, as described below, the presence of other distance cues may assist the listener in making the correct judgment.

Given these considerations, it appears that auditory distance studies should be conducted in normal, reverberant environments. Contrary to this, most earlier studies were conducted in anechoic environments [108] thereby limiting the cues presented to the listeners. Researchers have observed the importance of all the cues listed above (especially reverberation) required for accurate source distance estimation and there have been several studies conducted in normal reverberant environments. For example, a series of auditory distance experiments were conducted by Nielson [108] in "normal" reverberant rooms. In addition, to reduce the potential for erroneous results, rather than varying a single signal parameter, several parameters were varied in each of these experiments (e.g. source distance, angle between source and listener, source loudness, echoic and anechoic environments) to ensure the subjects did not "learn to use a single factor to achieve certain dynamics in the responses". Using virtual acoustics to provide accurate measurement and control of multiple auditory distance cues presented to the listener, Zahorik [170] examined auditory distance perception by listeners in a normal environment (a 264 seating capacity "complex shaped" room).

Source distance cues can be divided into two categories, *exocentric* and *egocentric*. Exocentric or relative cues provide information with respect to the *relative* distance between two sounds whereas egocentric or absolute cues provide information about the actual *absolute* distance between the listener and the sound source. Consider a sound source and a listener in a room where the listener cannot see and does not have any prior information regarding source position or distance. Now further imagine the source distance is doubled. Using the decrease in sound intensity between the sound source at the initial position and the sound source at the new position to determine that the distance has increased, is an example of an exocentric cue. On the other hand, when the listener uses the ratio of direct-to-reverberant levels to determine the source is five feet away from him or her is an example of an egocentric cue.

Greater details regarding the auditory distance cues listed above as well as their classification as either exocentric or egocentric are provided in the following sections, while a discussion on how these cues can be incorporated into an auditory display is provided in Section 3.5.

Intensity Cues and Loudness

The importance of sound intensity (level) as a cue to source distance has been known for many years. According to Zahorik [169], in 1892 Thompson observed that the intensity of the sound reaching the listener is the primary cue to source distance. Furthermore, given its relatively simple physical properties, it has also been the most studied auditory distance cue. This section will describe how sound intensity is used by humans as a cue to source distance and will also provide details of why it is an insufficient cue when used alone.

Consider a far field sound source and a listener placed in an anechoic environment. Furthermore, assume the listener's head is a perfect sphere with no external ears (pinnae). In such a situation, where the only auditory cue available is intensity of the emitted sound (a measure of energy propagated from the source per unit area and time), will be attenuated as the source distance s_d is increased following the *inverse square law* $1/s_d^2$, where the loss of intensity L_{loss} (in dB), due to increasing source distance, is given by Coleman [32] as:

$$L_{loss} \quad = \quad 20 \times \log_{10}\left(\frac{s_d}{s_0}\right) \tag{1.11}$$

where s_0 is the original source distance (e.g. reference distance). In other words, for each doubling of source distance, the intensity (level) of the sound waves reaching the listener will be decreased by 3dB. Such a model (although as described below, is certainly not completely correct) has been used in most 3D audio displays to convey source distance information to the users, without requiring any absolute (reference) sound pressure level [16, 19]

Reverberation as a Distance Cue

The inverse square $1/s_d^2$ attenuation of intensity of a propagating sound wave is valid under very restricted conditions. In particular, it assumes that the propagating waves reach the listener (receiver) directly without encountering any obstructions on their direct path from the source to the receiver or without any modifications due to environmental conditions. Furthermore, the majority of the experiments examining the relationship between distance and loudness were also conducted under intensity controlled conditions, taking place in anechoic chambers for example, where reflections of the propagating waves have a minimal (if not negligible) effect. However, in normal "everyday" listening situations, these restricted conditions rarely occur.

Reverberation can provide a cue to absolute source distance estimation, regardless of source intensity level due to changes in ratio of the direct-to-reverberant sound energy level

as a function of source distance. In particular, as source distance is increased, the level of the sound reaching a listener directly will decrease leading to a reduction in the direct-to-reverberant ratio. Greater detail regarding this phenomenon as well as the drawbacks associated with the inclusion of reverberation in an auditory display is provided in Section 3.5.2.

Spectral Content of the Sound as A Distance Cue

The majority of sounds encountered in the "real world" are comprised of many different frequency components and may contain components from the entire audible frequency range. It has been known for some time that the frequency spectrum of a sound source varies with respect to source distance due to absorption effects by the medium [32, 16, 108, 107]. In particular, there is a greater attenuation of the higher frequency components as source distance is increased. The spectral content of the received sound provides a relative distance cue only, unless the listener has prior information regarding the sound source. As with the loudness cue to source distance, allowing a listener to familiarize themselves with the sound source and the environment, improves the accuracy of source distance judgments [32]. The environmental conditions, including atmospheric conditions (the medium the sound must travel through), and any objects (reflective surfaces) the sound waves may encounter in the environment will affect any source distance estimation [16, 170].

Binaural Cues

As described in Section 1.2.1, ILD cues provide source localization information for frequencies greater than about 1500Hz. When the distance to the source is greater than approximately one meter, a far field source can be assumed (e.g. planar waves reaching the listener) and the ILD cues are distance independent. However, when the sound source is in the near field (e.g. within one meter of the listener), the waves reaching the listener can not be assumed to be planar. In this situation, the waves are spherical and the ILD cues are, in addition to direction, dependent on source distance [155, 138]. Confirmation of this can be seen in Figure 1.14 which is reprinted from the study performed by Brungart and Rabinowitz [20]. In this particular study, they investigated the distance dependence of HRTFs for source distances ranging from 0.12m to 1.0m through calculations using a rigid spherical head model and by actually measuring the response using a KEMAR dummy head. The graph shows the measured ILD for both a 500Hz (middle plot) and 3kHz (top plot) tone source as a function of distance (from 0m to 1.0m), for the source azimuths of 15, 45 and 90 degrees. The dependence of distance is clearly evident, as the ILD cues noticeably decrease as distance increases. In addition, this dependence to source distance

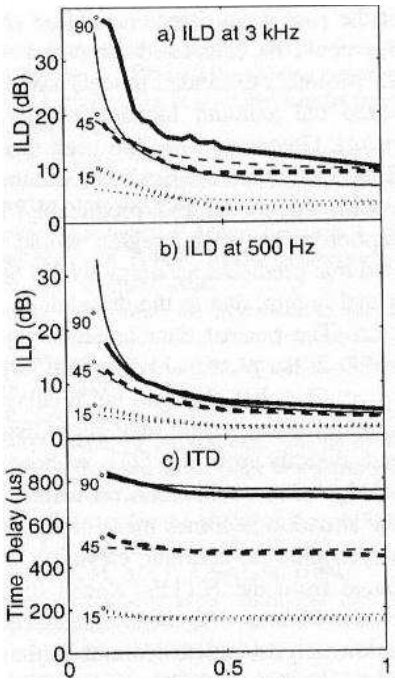


Figure 1.14: ILD as a function of distance (0m to 1.0m for two pure tones: 3kHz (top plot) and 500Hz (middle plot) for source azimuths of 15^o, 45^o and 90^o. Bottom plot illustrates the dependence of ITD on source distance for a source also at azimuths of 15^o, 45^o and 90^o. Reprinted from [22].

is greatest when the source is close to or on the interaural axis, reaching values as high as 20-30dB and decreases as the source moves towards the median plane. Also included is the graph illustrating the dependence of distance with respect to ITD (bottom plot). As shown, the ITD is less dependent on distance as opposed to the ILD cue.

Familiarity with the Sound Source and the Environment

A listener's prior experience with a particular sound source and the environment (e.g. the source transmission path) can greatly affect a listener's ability to determine the source distance, especially when the sound level is the only available cue. As previously described, without any prior information regarding the sound source and environment, the intensity (loudness) of a sound can only provide relative source distance cues. However, given prior experience, sound level can be used to determine the absolute distance to a sound source [138] and overall distance judgments may be improved [170, 107], especially when the sound is speech [16]. Prior information about a sound source or environment allows a listener to use their previous experiences and knowledge to provide a more accurate distance estimate

or to overcome ambiguous situations. For example from a very young age, we engage in conversations with others. For normal listeners, speech has become an important aspect of life as it allows us to communicate with others and express our thoughts. As a result, we have become familiar with the characteristics of speech (e.g. how loud a whisper or shouting may be and who is speaking) and are capable of accurately judging the distance to a live talker under normal conditions, especially when the distances are within a few feet [54, 22].

In many of the studies examining the relation between intensity (loudness) and source distance (as well as many other auditory studies), a single tone stimulus was employed. However, pure tones are not "particularly ecological" as most sounds in our environment contain various spectral components, generally leading to a decrease in localization accuracy [25]. In addition, as previously described, many of the earlier source distance studies were performed in restricted environments (e.g. anechoic chambers) in the absence of other distance cues, leading to a further reduction in localization accuracy. Familiarity with the sound source may have also affected the outcome of such experiments and even without prior knowledge of the listening environment and sound source, after repeated trials, knowledge of both may have been acquired by the subjects.

Finally, in a reverberant environment, source distance is also affected by the background noise [97]. In the presence of background noise, we tend to underestimate the distance to a sound source. This is probably due to the fact that since noise masks part of the weaker indirect portion of the sound reaching the listener, we cannot detect the entire extent of the reverberation [107].

Chapter 2

Recording Techniques

Since the introduction of the telephone in the late 1800s and the radio in the early 1900s, there have been many developments and improvements to technologies for presenting sounds to a listener in such a manner that the original sound field is reproduced. The exact reproduction of a sound field, including all spatial cues (e.g. reverberation, HRTFs, ITD and ILD), as they would occur in a "natural seeting", is certainly the goal of most 3D sound technologies. Current technologies also realize the importance of human psychoacoustics and employ many of the human auditory localization cues, including HRTFs. However, this was not the case in the "early years" of audio technology. In fact, it wasn't until the mid to late 20th century that an understanding of human auditory localization started to emerge (e.g. see the work of Batteau related to HRTFs [8], [7]). This "modern" approach to 3D sound, which employs human auditory localization cues, is fairly new, dating back approximately 20 to 25 years. Prior to this, many techniques involved recording a sound field (e.g. concert, musical performance etc.) using one or more microphones and then playing back the recorded sound using one or more loudspeakers or a pair of headphones. These techniques include monaural, stereo and surround sound and resulted in a large part due to the demanding needs of the entertainment industry (e.g. movie theaters, record companies etc.). Although such techniques are not "true" 3D sound technologies, they have paved the way for the many modern 3D sound technologies currently available. This chapter examines several of these recording techniques in greater detail, beginning with monaural, two-channel stereo and binaural procedures followed by multi-channel surround sound systems.

Since the techniques described in this chapter rely on the recording of a sound field using

one or more microphones, before discussing any of these techniques, a brief introduction outlining the operation and characteristics of microphones is provided.

2.1 Listener Sweet Spot

A problem common to all loudspeaker based audio systems, regardless the configuration or the technique used (e.g. recording technique or "true" 3D sound technology), is the fact that the intended auditory effect is restricted to a small region of space known as the *listener sweet spot*. The listener must therefore be placed in a specific location relative to the loudspeakers in order to achieve the desired effect. Any movements by the listener even small head movements away from the sweet spot, quickly degrades the intended effect. The size and span of the sweet spot as well as the degradation of the intended effect when the listener moves out of the sweet spot, is dependent on the technique used, the range of directions to be produced and the listening conditions (e.g. number of loudspeakers, loudspeaker layout and the directivity characteristics of the loudspeakers and the listening room) [80]. For example, in a two-channel stereo configuration, the listener should be positioned such that they form an equilateral triangle with the two loudspeakers [140]. Back and forth movements by the listener result in only a slight degradation of the intended stereo effect, as the distance to each loudspeaker remains the same. However, the intended effect may be greatly affected with sideways movements since in this case the distance between the listener and each loudspeaker differs.

2.2 Microphones

A microphone is a *transducer*, with the sole purpose of converting variations in air pressure into corresponding variations in electrical current (or electrical voltage) [104]. Each microphone contains a small component called the *diaphragm* which outputs a varying electrical current (or voltage). The propagating sound waves reaching the diaphragm cause it to vibrate and the rate of these vibrations determines the current (or voltage) output of the diaphragm (e.g. the greater the rate of vibrations, the greater the output produced by the diaphragm). The electrical output of the diaphragm can then be processed and used as desired (e.g. converted to a digital signal through an analog to digital converter, amplified etc.).

There are many different classifications of microphones, and typically, the classification is based on the operation of the diaphragm. The *dynamic* microphone relies on electro-

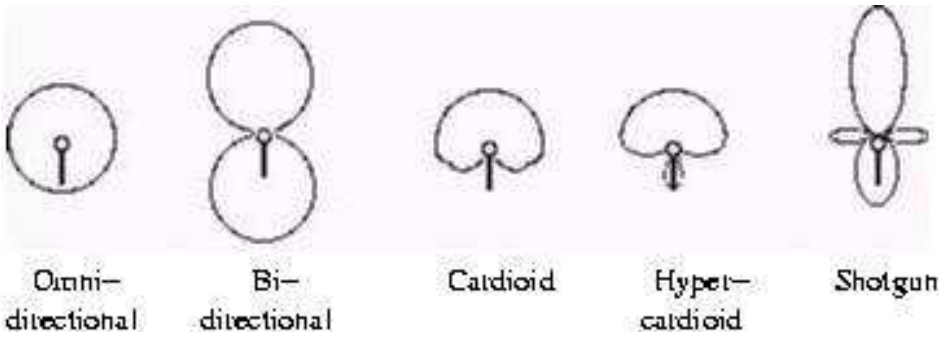


Figure 2.1: Microphone polar patterns.

magnetism. The diaphragm is attached to a coil of wire, which vibrates as the diaphragm vibrates. As the coil vibrates, its position relative to a magnet changes resulting in a varying current flow through the coil [129]. Given its high sensitivity, high frequency, response, low amplitude pick-up and its acoustically natural sound output, the *condenser* microphone (sometimes known as the *capacitor* microphone) is one of the most widely used microphones available [129]. The diaphragm of a condenser microphone is metal and forms one plate of a capacitor. Another metal disk positioned close to the diaphragm, acts as a "backplate" (the other plate of the capacitor). A steady D.C. voltage is applied to either the diaphragm or the backplate. As the diaphragm vibrates (due to the propagating sound waves), the distance between the diaphragm and the backplate changes. The change in distance leads to a change in capacitance which ultimately results in a change of electrical current output. Other types of microphones exist, however, they will not be discussed here. Greater details can be found in [42, 129].

Microphone Directivity Patterns

An important property of a microphone is its *directivity* or *polar* pattern. The directivity pattern refers to the direction(s) in which the microphone is sensitive. Ideally, the microphone will respond only to sounds which are propagating in the directions of the microphone's directivity pattern. Various types of polar patterns are available, and type of polar pattern used is typically determined by the application. Several of the more popular polar patterns are illustrated in Figure 2.1.

A microphone with an *omni-directional* polar pattern (known as an omni-directional microphone), will (ideally) respond equally to sounds coming from all directions (e.g. 360^o) and as a result, are often used to record ambient or background sounds. The *bi-directional* or "figure of eight" polar pattern allows the microphone to respond to sounds coming from in front or in back of the diaphragm and to be less sensitive to sounds that approach

at right angles. The *cardioid* microphone has a heart shaped polar pattern (hence the name "cardioid") and is sensitive to sounds coming from in front of the diaphragm while rejecting sounds from the back. This makes it useful in recording musical performances such as concerts, where only the performance is of interest and not any sound coming from the audience.

Although in theory the microphone is sensitive only to the directions defined by its polar pattern, in practise this is certainly not the case. The microphone will respond to some degree to sounds coming from directions other than those defined by its polar pattern, as such sounds are attenuated and not completely rejected. Finally, in addition to direction, a microphone may also be more sensitive to certain frequencies.

2.3 Monaural Systems

A monaural recording is made using a single microphone and typically conveyed to the listener through a single loudspeaker, although the signal can be output through multiple loudspeakers regardless of their placement or how many. It was the first method used to convey sound in films and remained a standard in the film industry for over fifty years and still is a standard for AM radio [61]. Since a single microphone is used, binaural cues cannot be captured, thereby giving the listener the impression that all sounds are coming from a single location (e.g. it is *unidimensional*). In addition, it is very difficult to convey any ambiance (e.g. a certain mood created by some environment), and any ambiance present is of poor quality, given that in a monaural recording, the signal of interest and the background noise essentially sound the same and therefore, it may be difficult to differentiate between the two. Finally monaural sounds do not contain any directional information making them impractical for use in a 3D auditory display.

Despite these shortcomings, monaural systems are still relevant today. They may not be used by the film and music industry anymore, or in spatial auditory displays however, monaural systems are very good for conveying "alert" (warning) sounds and provide good speech intelligibility [61]. In fact, the telephone, going back from its introduction in 1876 by Alexander Graham Bell, to the present continues to employ a monaural system to convey speech between participants.

2.4 Stereophonic Techniques

Stereophonic or *stereo* has become synonymous with two-channel audio. However, the word

stereophonic, derived from Greek for "solid sound" actually refers to the construction of believable, solid, stable sound "images" regardless of the number of channels used [122]. It can refer to any number of channels, including two, three, four, five and even six (the popular Dolby 5.1 surround sound format employs six channels). In fact, stereo and surround sound actually refer to the same thing (surround systems are covered in Section 2.6). Regardless of the number of channels actually used, the purpose of stereo is to provide the listener with a real life impression of a sound event [37]. However, given the widespread association between stereo and two channels, for the purpose of this report, unless otherwise stated, *stereo* will refer to two-channel audio whereby sounds are output with two loudspeakers (or headphones).

The first documented use of a stereo system was by Clement Ader, a French designer, in Paris in 1881 (see [148]). However, despite the potential this technique had to offer, very little attention was paid to it until the early 1930s. In the United Kingdom, Alan Blumlein, a researcher working for EMI Corporation, developed a stereo recording system based on *coincident* microphones while in the U.S, a team of researchers, led by Harvy Fletcher developed their own stereo techniques based on *spaced* microphones. It wasn't for many years later until either of these techniques were used for commercial purposes. Stereo has been (and still is) the main method for the playback of recorded music since the introduction of vinyl stereo records in 1958, FM radio in 1961 and stereo television in 1986 [162].

The typical stereo listening setup is illustrated in Figure 2.2, where the listener is typically positioned between the two loudspeakers. For optimal listening conditions, the listener, along with the two loudspeakers should form an equilateral triangle, where the separation between the two loudspeakers forms the base "b", the offset angle (α) is equal to 60^o and the listener is positioned on the vertex of the triangle [72].

As with human sound localization in a natural setting, stereo systems utilize sound level and (or) timing differences (ILD and ITD) to simulate a sound event between the two loudspeakers. Stereo can provide the listener with a "sense of depth", allowing them to perceive the presence of a particular auditory environment in which sounds may be localized, extending beyond the two loudspeakers [72]. Various stereo recording techniques have been developed and experimented with over the years, however, the following three methods are the most widely used [122]:

Artificial Techniques: Stereo *images* (or *phantom sources* or *virtual sources*) are produced by artificially adjusting the intensity and (or) time delays between the monaural signal delivered to the left (L) and right (R) channels (loudspeaker outputs).

Coincident Microphone Techniques: The sound event is recorded by two directional microphones whose capsules point in different directions but are placed as physically

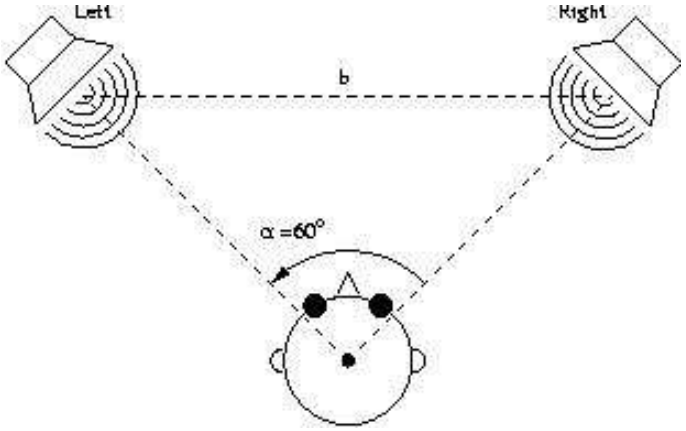


Figure 2.2: Ideal stereo configuration. The listener and two loudspeakers form an equilateral triangle, where the separation between the two loudspeakers forms the base b, the offset angle α is equal to 60^o and the listener is positioned on the vertex of the triangle.

close to each other as possible. Since they point in different directions, the capsules will be offset by a certain angle. Such a configuration will (ideally) eliminate any timing differences between the two recorded signals while capturing any intensity differences.

Spaced Microphone techniques: Two or more identical microphones spaced some distance apart from each other are used to capture the sound event. Timing differences between the sound at each microphone are captured and conveyed to the listener during playback.

Details regarding each of these three stereo techniques are provided in the following sections.

2.4.1 Artificial Stereo

In this technique, the difference in sound level and/or the time delay between the signal fed to the left and right loudspeakers is adjusted in order to position the virtual sound source somewhere between the two loudspeakers. This method takes advantage of the ITD and ILD cues employed by a listener to localize a sound source in a natural setting. As described in the following sections, the positioning of a stereo sound between loudspeakers using level and time delay adjustments are referred to as *intensity* (or *amplitude*) and *time panning* respectively [55] given that the position of the virtual source can be *panned* across the space between the two loudspeakers.

Stereo Time Panning

As described in Section 1.2.1, unless the sound source is directly in front of the listener, it will be closer to one ear (*ipsilateral* ear) and therefore arrive at this ear first, leading to the ITD cue. In the artificial stereo technique, this cue can be simulated by simply sending to the contralateral ear, a delayed version of the signal sent to the ipsilateral ear. For example, when the desired sound source position is to the left of the listener of a stereo setup, the right ear will receive a delayed version of the signal sent to the left ear. The amount of delay determines the position of the virtual source and therefore, by allowing for a variable time delay, the virtual source may be positioned between the two loudspeakers. The time delay actually required to position the sound source to either the left or right loudspeaker is rather small. According to Hugonnet and Walder [72], experiments indicate a delay of between 0.8ms to 1.4ms. Given this short range of delays required to position the virtual source to either of the loudspeakers, the effect produced by this technique quickly degrades with even small listener movements, especially side-to-side movements. Movements of a few feet may lead to time delays which are much greater than the small amount described above [55]. In other words, the extent of the sweet spot is very small.

Stereo Intensity Panning

Intensity panning is similar to time panning however, instead of adjusting the difference in the time of arrival between the signal delivered to the left and right loudspeakers, the difference in level (intensity) is adjusted instead to position the virtual source anywhere between the two loudspeakers. Generally, level differences of approximately 12dB to 16dB are sufficient to position the sound source to either of the loudspeakers, although the exact value may depend on the individual, the listening environment and equipment [122]. Stereo intensity panning is actually a more effective technique then stereo time panning. It is more robust, given the greater dynamic operating range of level differences as opposed to time delays. Furthermore, it is fairly consistent with different signal types and is less prone to errors when the listener moves away from the "sweet spot" (e.g. the listener is "off axis"), unlike the case with time delay adjustments [55]. Finally, as with time delay adjustments, when the level difference between the two loudspeaker signals is zero, the source appears to be directly in front of the listener. Once it reaches a maximum value, the virtual source will continue to emanate from the loudspeaker corresponding to the ipsilateral ear, even when the level difference is increased beyond the maximum.

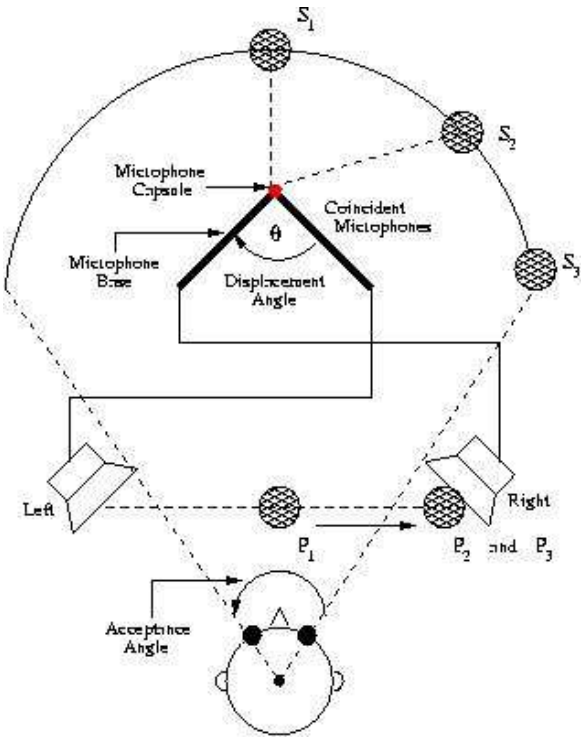


Figure 2.3: Coincident stereo microphone techniques. One microphone capsule is placed on top of the other, offset by θ degrees. Placing capsules on top of each other ensures there is no offset between them with respect to the horizontal plane, resulting in a minimal time delay between the recorded signals.

2.4.2 Coincident Microphone Techniques

Coincident microphone techniques, developed by Blumlein in the early 1930s, involve the recording of a sound event using two directional microphones, spaced as close as physically possible in order to avoid (or at least greatly reduce) any timing differences between the two recorded signals. As shown in Figure 2.3, typically, one microphone capsule (e.g. the part of the microphone housing the diaphragm) is placed on top of the other, displaced by an angle of θ degrees (the *displacement angle*). Placing the microphone capsules one on top of the other ensures there is no offset between both microphones with respect to the horizontal plane (e.g. horizontal plane displacement is zero). The lack of "horizontal plane displacement" between the microphone capsules ensures the time delay is zero when considering sound positions on the horizontal plane, the plane of interest for stereo recordings.

Coincident microphone techniques rely on intensity differences between the two recorded signals arising from the polar pattern of each microphone. The choice of microphone po-

lar pattern actually used (e.g. the type of microphone), determines the range of allowable angles spanned by the virtual source in front of the microphones, or in other words, the *acceptance* angle, as illustrated in Figure 2.3. In addition, the choice of polar pattern used for each microphone may differ and certain combinations of polar patterns may produce more favorable results for certain listening scenarios.

Referring to Figure 2.3, when the source lies directly in front of the two microphones, at source position S_1, the level of the sound recorded by each microphone will be the same (e.g. the sound must travel an equal distance to reach the right and left ear) and therefore, the difference in level will be zero. Without any difference in level, during playback, the listener will perceive a source directly in front of them. As the source is moved to the right (left), the level of the signal intended for the right (left) loudspeaker will be greater, leading to an increase in the level (intensity) difference. When played back, the listener will perceive a source to the right (left). As previously described, a level difference L_d between approximately 12 to 16dB (depending on the displacement angle between the microphone capsules) is required to place the sound source completely to the right (left). For example, when the sound source is positioned at S_2 (the rightmost position within the acceptance angle), the difference in level between the left and right microphone signals will reach its maximum, with the level of the signal being fed to the right loudspeaker being greater and the virtual source will come from the right loudspeaker. Furthermore, the virtual source will continue to come from the right loudspeaker even if the level difference is greater than L_d [72], which for example, will occur if the desired position of the virtual source is at S_3. Two common coincident microphone approaches are known as the "XY" and "MS" techniques. A brief description of each technique is provided in the following sections. Finally, regardless of the actual technique used, according to Theile [151], coincident microphone techniques lack any sense of depth and space as the signals of each of the two channels lack the interaural correlation available naturally. This problem can be overcome however using the *sphere* microphone. This microphone captures the interaural differences naturally available and produces favorable results with respect to spatial perspective, localization accuracy and overall quality [151].

XY Coincident Microphone Technique

In this technique, the two microphones must have the same directivity (polar) pattern, which is usually cardioid or bi-directional ("figure of eight"). The microphone pointing to the right is used to record the sounds intended for the right (R) loudspeaker while the microphone pointing to the left records the sounds intended for the left (L) loudspeaker. Using this technique, a monaural signal S_m can be obtained as follows:

$$S_m = X + Y \tag{2.1}$$

where, X and Y are the left and right microphone channels respectively. When played back, XY recorded sounds lack the sense of depth and perspective but can allow for clear sound source localization [72].

Mid and Side (MS) Microphone Technique

In this technique, one microphone (the "M" microphone), having any polar pattern, including omni-directional, cardioid or figure of eight, faces forward, capturing the sound coming directly in front of the sound event (e.g. orchestra, performance etc.). The other microphone (the "S" microphone), has a bi-directional polar pattern and faces sideways, perpendicular to the M microphone in order to capture the sounds coming from the side of the sound event in addition to a large amount of reverberation.

Once the M and S signals have been recorded, the left (L) and right (R) channel signals which will be played back to the listener via the left and right loudspeakers respectively, are formed as follows:

$$L = M + S \tag{2.2}$$
$$R = M - S \tag{2.3}$$

The MS technique offers several advantages over the XY stereo recording technique and is widely used for television sound recording [122]. Varying the ratio of the mid (M) and side (S) signals allows for one to modify the useful acceptance angle without modifying the configuration of the microphones in any way [72]. This is particularly advantageous in situations were the configuration cannot be physically adjusted, such as during a live performance or a concert [5]. Furthermore, in practise, the MS technique is less error prone over the XY technique leading to greater recording fidelity [71].

Equivalence Between XY and MS Techniques

Theoretically, the XY and MS techniques can be considered equivalent and a transformation between them can be performed assuming ideal microphone characteristics [71]. However, in practice, microphones are certainly not perfect and do not exhibit ideal directivity patterns (e.g. the real polar pattern differs from the mathematical ideal one) making the practicality of such a transformation very limited.

2.4.3 Spaced Microphone Techniques

Spaced microphone techniques rely primarily on time panning in order to position the sound source anywhere between the two loudspeakers. The simplest spaced microphone technique is shown in Figure 2.4. Two microphones, typically omni-directional, spaced a few centimeters apart with a displacement angle of zero degrees, face the sound source. When the distance between the sound source and each of the microphones is the same (e.g. position S_1 in Figure 2.4), each microphone will record the same sound signal without any delay. However, as the source moves towards the left or right, the time delay Δt between the signal increases, reaching a maximum when the source has moved to its maximum allowable leftmost or rightmost position. As with the artificial stereo technique described previously, the maximum time delay is approximately 0.8 to 1.4ms. Furthermore, when the time delay does reach this amount, as with the coincident microphone technique, the signal will be output from the left or right loudspeaker, even if the delay is increased further.

The spacing between the two microphones determines the size of the active-arc (acceptance angle). Microphone spacing is limited to between approximately 25 and 50cm which corresponds to acceptance angles of 80^o to 130^o respectively [72], however, there are generally no other rules for spacing the microphones and usually, it is a matter of trying various distances until one producing favorable results is found [123].

2.4.4 Combining Coincident and Spaced Microphone Techniques

The spaced and coincident stereo recording techniques can be combined to take advantage of both time delay and intensity differences. In such a situation, the two microphones are spaced apart by some distance but are also displaced by an angle θ. When the source is moved towards either loudspeaker, both the time delay and level differences between the signals of the left and right microphones will increase. Various combination systems have been developed. For example, as shown in Figure 2.5 the ORTF technique developed by the *Office de Radiodiffusion Television Francaise* (hence the acronym ORTF), uses two

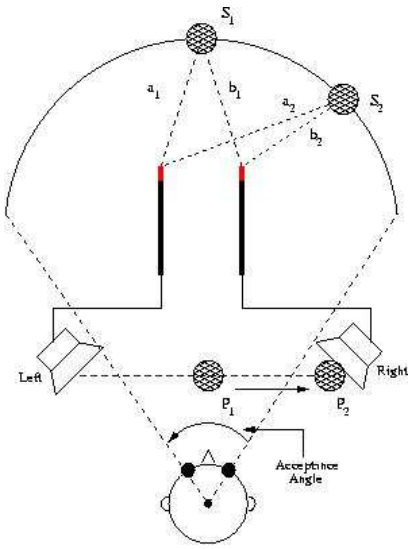


Figure 2.4: Spaced stereo microphone techniques rely primarily on time panning only in order to position the sound source anywhere between the two loudspeakers. Two microphones, typically omni-directional are spaced a few centimeters apart with a displacement angle of zero degrees, facing the sound source.

cardioid microphones which are separated by a distance of 17cm and a displacement angle (θ) of 110^o. Similarly, the NOS technique employs two cardioid microphones spaced 30cm apart and displaced by 90^o. Further details regarding these techniques as well as other combination techniques may be found in [42, 72, 122, 123].

2.5 Binaural Audio

Given a particular listening environment to be simulated, with the sound source and listener each at some particular position, *binaural audio* can be defined as the reproduction of the acoustic signals that would naturally be present in this particular situation. Ideally, the reproduced acoustical signals and the acoustical signals present in the natural listening condition will be identical, such that when presented to the listener, the listener may use any of the naturally available cues in order to perceive the sound as emanating from the desired position. The reproduced acoustical signals can be obtained using *binaural recording techniques* or *binaural synthesis*. Greater detail regarding binaural recording techniques are provided in the following sections. Binaural synthesis is described in Section 3.2.

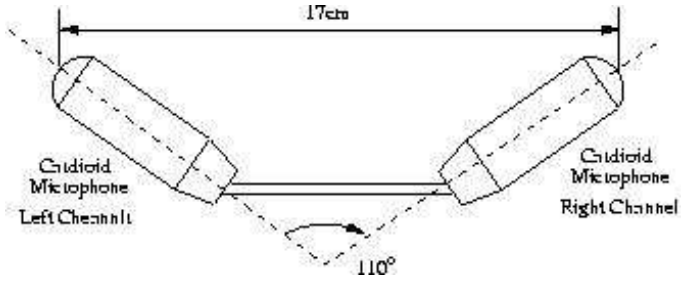


Figure 2.5: Near coincident (ORTF) stereo microphone technique. Two cardioid microphones are separated by a distance of 17cm and displaced by 110^o. This technique is able to capture both time delay and intensity differences between the signals arriving at each microphone.

2.5.1 Binaural Recording Techniques

As illustrated in Figure 2.6, in this technique, small microphones are placed typically at the entrance of the ear canals of a person or an anthropomorphic dummy head to separately record the sound at the left and right ears as it occurs in the natural environment. Since the microphones are placed at the entrance of the ear canals, the recorded sound will include any environmental modifications (e.g. reverberation, air absorption, attenuation etc.) encountered by the sound on its path to the corresponding ear and HRTF filtering. Once the sounds have been recorded, playing back the left and right recorded signals to the left and right ears respectively will reproduce the listening situation in which the recordings were made, and ideally, lead to the perception of a sound emanating from the original sound source position. All the audio environmental cues, including reverberation, air attenuation, source distance etc. and any binaural cues such as ITD and ILD will be present.

Binaural recordings are capable of producing very realistic results, allowing for a strong perception of a sound source at some specific location (e.g. of "being there in real life"), even when using standard audio equipment. However, binaural recordings do have their problems. In particular, the recorded signals are specific to the environmental setting in which they were made, as well as the source position. Therefore, only this specific listening situation can be reproduced during playback. Any changes in the position of the sound source, listener or the environment (e.g. the introduction of new objects in the path between source and listener) requires a new pair of recordings to be made. Furthermore, as described further in Section 1.2.2, each person's pinnae differ, leading to spectral filtering of the sounds present at each ear specific to each individual. Since the filtering effects of the "dummy head" differ from the filtering effects of the listener during playback, a degradation of the desired effect results. Finally, for optimal results, the recorded signals must be played back to a listener's ears in isolation, ensuring the left and right signals arrive only to the

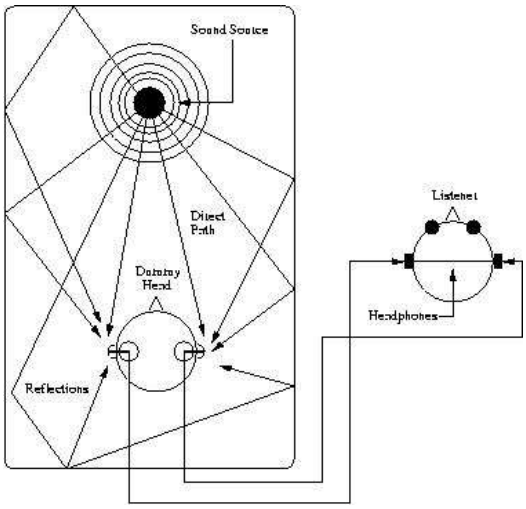


Figure 2.6: Binaural recording technique. Small microphones are placed in the ear canal of a "dummy" head (or person) in order to record a sound event in a particular environment. The recordings will capture any auditory localization cues present such as ITD, ILD, reverberation and HRTF filtering. When the recording is presented to the listener, the listener will perceive the sound as emanating from the original position and environment in which the recordings were made.

left and right ears respectively. As a result, the recorded signals are usually played back over headphones as opposed to loudspeakers, to avoid any *crosstalk*. Headphones have their share of problems (see Section 4.1) and there may be times where loudspeaker output is desired. Binaural recordings can be presented over loudspeakers provided some method of crosstalk cancellation is used to remove the crosstalk signals arising from the fact that a portion of the signal emanating from the left (right) loudspeaker will reach the right (left) ear (see Section 4.2.1 on crosstalk and crosstalk cancellation). Given the elimination of this crosstalk and assuming the listener is placed in the appropriate position (e.g. in the "sweet spot" which is typically symmetrically between the loudspeakers), binaural recordings heard over loudspeakers can provide good results.

2.6 Surround Sound

Stereo may involve any number of channels and is certainly not restricted to two. However, when introduced to home consumers in 1958 only two channels were actually used. The use of two channels was based solely on the fact that two channels was all that could

physically be placed on phonograph records, the main recording medium at that time [73]. Surround systems can consist of any number of loudspeakers "surrounding" a listener [55] in order to provide them a greater sense of realism, making them feel as if they are in attendance (live) at the music performance (concert etc.) or when watching a movie, as if they were part of the "action". Surround sound systems allow the listener to hear sounds coming from all around them, not only in front of them as with traditional stereo setups. The majority of people associate surround sound with "something being added to two-channel stereo", including the addition of more loudspeakers. The surround systems described in the following sections do contain more than two loudspeakers and require that the loudspeakers be physically placed around the listener in some particular configuration.

Despite the widespread use and popularity of stereo in the home consumer market, various "more than two" channel stereo systems are also being investigated and developed, especially for the movie theater market. Research conducted by Fletcher's group at the Bell Laboratories, involved the use of a large number of microphones and loudspeakers to record and playback a sound event respectively. One of the earliest systems, the "Wall of Sound" developed by Fletcher used an array of up to 80 microphones, mounted horizontally across the front of an orchestra [122]. The playback of sound over an equal number of loudspeakers produced very accurate and pleasing results. Such a large number of microphones resulted in a large sweet spot, providing the listener greater freedom to move about in the environment [42]. However, the use of such a large number of microphones and loudspeakers was clearly impractical and so, the number of microphones and loudspeakers were reduced to three. Three channels allow for greater precision in positioning the virtual source than when using two channels alone [151]. The use of three channel produces more favorable results when compared to two-channel stereo and is still in widespread use today to produce frontal sound in movie theaters [123].

Although favorable results were achieved with the three microphones and loudspeaker setup, such a method was not very popular amongst recording companies who could not physically "cut" more than two channels into a vinyl record. Given this two-channel recording restriction, the majority of consumer audio products supported (and still support), two-channel stereo only, making the use of a third recording channel impractical except in very restricted situations. The three recorded channels can be transformed into a two-channel stereo compatible signal by mixing the signal recorded with the center microphone into the signals intended for the left and right loudspeaker [42].

The first public use of a true surround system was by Walt Disney, in his animated movie *Fantasia* released in 1940, which combined animated "mini-features" with popular orchestral music [124] and included as many as eight separate music and effects channels. Movie theaters at this time did not have the equipment to support eight channels of audio and as a result, the film was toured throughout the country with its own technical crew and reproduction system.

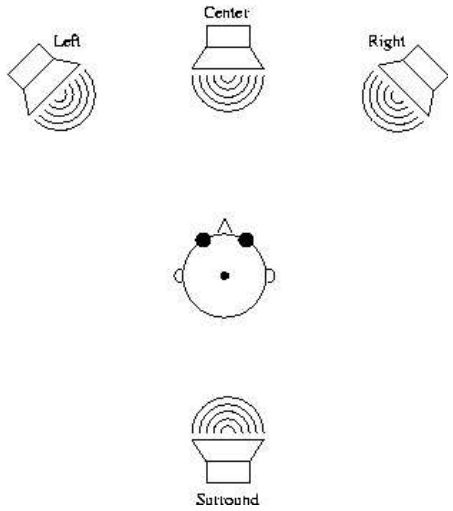


Figure 2.7: Three loudspeaker surround sound system configuration. Obtained by adding a loudspeaker in between the left and right stereo loudspeakers.

Despite the widespread popularity of stereo (two-channel) systems in the home consumer market, two channels were insufficient for move theaters. Given the large width of movie screens, localization of sound for any viewers which happened to be seated towards one side of the screen (e.g. away from the center), was very poor [73]. In an attempt to overcome this limitation, during the 1950s and 1960s several multi-channel cinema audio systems were experimented with. In one technique, (shown in Figure 2.7), a loudspeaker (center channel) was added in between the left and right stereo channels, producing a three channel surround system. The center channel could now be a substitute for the right or left channel loudspeaker for persons seated towards the left or right side of the movie screen respectively.

Several other competing systems were developed as well, including systems with up to six channels [73]. For example, 70mm movie prints with magnetic audio tracks provided a total of six channels, five placed across the front of the screen and one channel, the *surround* channel, was placed towards the side and rear of the theater. None of these early multi-channel systems lasted very long for several reasons, including the fact that the early film audio tracks were of poor quality and very noisy.

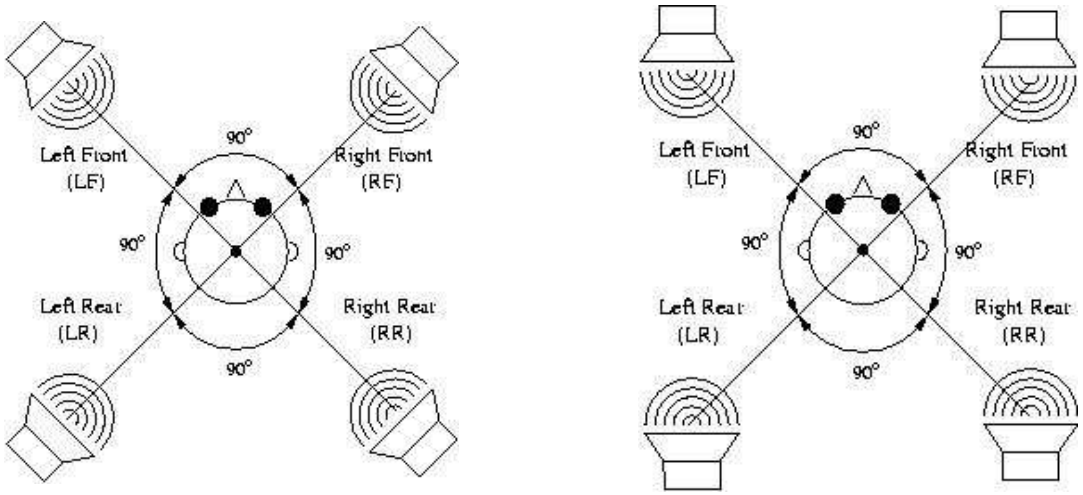
The following sections provide greater details regarding several of the more popular surround sound system formats including Quadraphonics, Ambisonics and several systems developed Dolby Laboratories, including Dolby Stereo and Dolby Digital.

2.6.1 Quadraphonic

The Quadraphonic (also known as 'Quadrisonics"), or "Quad" system was the first "surround system" to be introduced to consumers. The first demonstration of a Quadraphonic system was by Vanguard records in 1969 and the technology was made publicly available in the early 1970s. Quad systems were developed to improve the limitations associated with monaural and stereo recorded sound. Although monaural and stereo systems did provide the listener with the impression of looking towards a performance (a sound source), they did not provide the listener with the sense (illusion) of "actually being there", live, while the performance was taking place. In order to accomplish this, sounds would have to reach the listener from any direction in three dimensional space, something clearly monaural and stereo systems were incapable of achieving. In a stereo setup the listener faces the two loudspeakers and hears sounds coming from in front of them, or as mentioned, they "look towards the performance". The principle behind quad systems was simple: add another two loudspeakers behind the listener in a traditional stereo setup to allow sounds emanating in the rear to reach the listener as well. Quad systems consist of four loudspeakers, two in front of the listener, left-front (LF) and right-front (RF) and two in back of the listener, left-rear (LR) and right-rear (RR). Although no standard was developed for the actual placement of the loudspeakers, they were typically placed at the four corners of a listening area, either facing inwards towards the listening area as shown in Figure 2.8a or as shown in Figure 2.8b, the two rear loudspeakers could face the two front loudspeakers [38]. In either setup, the angle of separation between each of the loudspeakers is 90^o, equally dividing the entire 360^o space surrounding a listener. Quad systems were intended to allow for the perception of sound emanating from any direction on the plane in which the four loudspeakers were placed (all loudspeakers are placed on the same plane). Each of the loudspeakers received a signal which was previously recorded from a microphone element, intended to capture sounds emanating from the direction corresponding to the position of the loudspeakers. The following section provides greater detail regarding the quad microphone system.

Quadraphonic Microphone

Quadraphonic recordings are made by capturing the sound with four microphone elements, (typically packaged together in a single housing) and played back over four loudspeakers. Usually, these microphone housings can be further divided into two units: the upper and lower capsules. The upper capsule consists of two microphone elements that are used to capture sound from the left-front (LF) and the right-rear (RR). The two microphone elements of the lower capsule are used to capture sounds coming from the right-front (RF) and left-rear (LR). Each of the four elements contains its own separate channel, hence the



(a) Quadraphonic Setup.

(b) Quadraphonic Setup.

Figure 2.8: Quadraphonic loudspeaker configuration. (a) Four loudspeakers at each corner of the listening area facing inwards towards the listener. (b) The two front loudspeakers facing the two rear loudspeakers.

requirement of four loudspeakers during playback.

Encoding and Decoding of Quadraphonic Sound

Quadraphonic recordings resulted in four separate channels of audio information, one channel for each microphone and loudspeaker pair. As a result, four channels had to be transferred from the recording process to the playback process. Although several storage mediums supporting four channels were available, including quad reel-to-reel tape and eight-track cartridges [124], given the widespread use and success of stereo, the majority of consumer equipment, including vinyl records supported two-channel stereo only [38]. Since it would be very difficult to convince users to purchase new dedicated quad equipment, methods were developed to allow the four channels of Quadraphonic recorded information to utilize the existing two-channel transmission medium. The technique used to *encode* four channels of information into two channels, transmit and then decode the two channels back into four channels for Quadraphonic playback is referred to as *matrixing*, and the set of equations to perform the task is referred to as the *matrix*. By encoding the four channels of audio information into two channels, quad recordings were "backward compatible", allowing them to be played back using standard two-channel record players, instead of requiring new, dedicated equipment [124]. Quad matrixing was also known as the "4-2-4" method, denoting the encoding of the four original channels into two channels for

storage and transmission and then the decoding back into four channels during playback. Matrixing is not restricted to four and two channels only, neither is it specific to Quad systems however. Matrixing can involve the encoding and decoding between any number of channels. For example, a 5-2-5 system encodes five original channels into two channels and later reconstructs the five channels once again. The first 4-2-4 matrix was proposed by Scheiber in 1970 [131, 132]. The encoding and decoding equations comprising the Scheiber matrix are provided in the following equations:

$$
\begin{aligned}
L_{stereo} &= 0.924 \times LF + 0.924 \times LR + 0.383 \times RF - 0.383 \times RR && (2.4)\\
R_{stereo} &= 0.924 \times RF + 0.924 \times RR + 0.383 \times LF - 0.383 \times LR && (2.5)\\
LF &= 0.924 \times L_{stereo} + 0.383 \times R_{stereo} && (2.6)\\
RF &= 0.383 \times L_{stereo} + 0.949 \times R_{stereo} && (2.7)\\
LR &= 0.924 \times L_{stereo} - 0.383 \times R_{stereo} && (2.8)\\
RR &= 0.383 \times L_{stereo} + 0.924 \times R_{stereo} && (2.9)
\end{aligned}
$$

where L_{stereo} and R_{stereo} are the left and right stereo signals respectively. By examining the equations of the Scheiber matrix, it can be seen that each channel will contain a component at -3dB from the channels adjacent to it. For example, the right-front channel will contain a -3dB component from the front-left and rear-right channels. Furthermore, the diagonal signals between front-right and rear-left and between the front-left and rear-right, will be canceled.

The Scheiber matrix was never actually developed into a commercial product, however it formed the basis for the first commercially available Quadraphonic encoding/decoding system released in 1972 under the name of SQ for Surround Quadraphonic by CBS. This was followed by the release of the QS matrix system developed by Sansui Corporation, which was however incompatible with SQ.

Problems with Quadraphonic Systems

Quadraphonics was never really a "hit" with consumers and lasted for a short time only. The first Quadraphonic recordings were released in the early 1970s (open reel tape) while the last encoded recordings were released in 1980. Given the widespread use of stereo equipment, very few consumers rushed out to purchase additional new and expensive equipment to support Quadraphonics on their existing systems. Furthermore the different record com-

panies and stereo equipment manufacturers each supported different incompatible encoding and decoding schemes, creating much confusion amongst consumers.

In addition to the non-technical issues described above, there were other serious technical issues associated with Quadraphonic systems which inevitably led to their downfall. Most importantly, despite the promise of full 360^o localization on the azimuthal plane (e.g. the ability to convey 3D sound), Quadraphonic systems were inaccurate and non-realistic in presenting a 3D sound source [38]. With respect to matrixing, once encoded, the original signals can never be completely reconstructed as information will always be lost in the process, resulting in undesirable effects [124]. As with any loudspeaker auditory display, *crosstalk* (see Section 4.2.1 for greater details regarding crosstalk) also degrades the performance and effects of the resulting playback sound. In a Quadraphonic setup, the sweet spot is located equidistant from all four loudspeakers (e.g. in the center of the listening area) and is rather narrow, as even small head movements by the listener result in dramatic changes in the desired effect. Despite its shortcomings and lack of interest by consumers, Quadraphonics paved the way for the surround systems currently available.

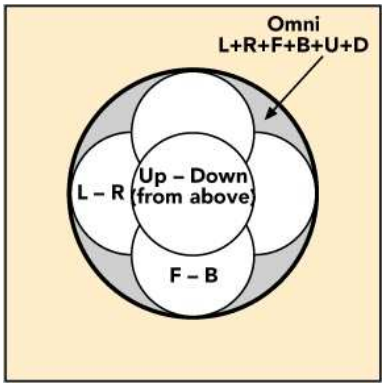
2.6.2 Ambisonics

Ambisonics is a high resolution surround sound system developed primarily from the work of British researchers Michael A. Gerzon and Peter Fellgett in the late 1960s and early 1970s, building on Blumlein's earlier work of stereo recording and playback. Ambisonics was primarily developed to overcome the major difficulties associated with the Quadraphonic systems available at that time. It was created to allow a recorded musical performance to be played back in a typical living room such that the original sound and environment in which the performance took place would be recreated [46] or in other words, it was conceived as a system capable of recreating accurate 3D sound from original recordings [125]. Furthermore, Ambisonics is capable of encoding (and then decoding) sound sources from any direction in space, including vertically [45]. The following sections provide greater details regarding both the recording and playback of an Ambisonic system.

Recording Stage

Various Ambisonic microphones can be used to capture the sound field. However, regardless of the actual microphone used, the principles are the same. The microphone components are arranged in such a manner such that they simulate a single omni-directional capsule along with three "figure-of-eight" capsules, where one figure-of-eight capsule is pointing left-right, the other front-back and the third one up-down. An illustration of a very popular

(a) Soundfield microphone.



(b) Soundfield directivity pattern.

Figure 2.9: Soundfield microphone (a) and its directivity (polar) pattern (b). Reprinted from [45].

Ambisonic B-Format microphone, the *sound field*, [47] is illustrated in Figure 2.9(a). As shown, four cardioid capsules are arranged in a tetrahedral array to provide the pattern described above. A graphical illustration of the polar pattern is provided in Figure 2.9(b).

The Ambisonic recording phase produces a four-channel signal (W, X, Y and Z components), collectively known as "B-Format". The W component is the monaural signal arising from the omni-directional capsule, while the X, Y and Z components result from the three figure of eight capsules. The three figure of eight capsules are used to determine the direction of the arriving sound while the omni-directional capsule provides an overall level reference. Mathematically, the four components are encoded as follows:

$$W = Left + Right + Front + Back + Up + Down \tag{2.10}$$
$$X = Front - Back \tag{2.11}$$
$$Y = Left - Right \tag{2.12}$$
$$Z = Up - Down \tag{2.13}$$

With the use of the Soundfield microphone, the produced B-Format signal is processed to provide a flat frequency response for all directions of incidence, a quality not shared

by conventional microphones [125]. In addition to surround sound capability, "B-Format" allows for the encoding of height information as well, which is usually not included in other surround sound systems, despite the fact that height information does improve the realism [46].

Playback Stage

The decoding of Ambisonic signals offers considerable flexibility, allowing for any number of loudspeaker configurations, depending on the desired output. A minimum of four loudspeakers is required to allow for horizontal localization (*planar* surround), eight loudspeakers can, in addition to planar surround, provide height information as well, thereby permitting 3D localization (e.g. *periphony* or *full sphere* surround), while twelve loudspeakers can be used in a large room such as a movie hall or auditorium. Furthermore, Ambisonics provides greater freedom with respect to loudspeaker placement for the playback of a sound field (or sound event). Loudspeakers can be placed in any rectangular configuration as long as the ratio of length vs. width of the rectangle does not exceed $2:1$ [45]. The listener simply tells the decoder where the loudspeakers are located. Contrary to other recording techniques, such as Quadraphonic, the playback loudspeakers certainly do not have to be configured such that they correspond to the position of the sound sources during the recording stage. In other words, there is no "one-to-one" mapping between the recording microphones and the playback loudspeakers such that a playback loudspeaker will simply output the sound received by its corresponding microphone, leading to a very small "sweet spot". As a result, with Ambisonics, the sweet spot is much bigger, allowing for the "surround effect to be more pronounced and stable" over a wider listening area [46]. Furthermore, it also permits for the adjustment of the sweet spot, allowing the listener's position to be taken into account [24].

As described above, each of the four channels of the B-Format signals contain a combination of sum and difference signals, making it impractical for playback over standard monaural or stereo systems. However, given the widespread availability of such equipment, the *UHJ hierarchy* encoding system was developed to allow for the playback of Ambisonics encoded sounds over existing monaural and stereo equipment. With respect to encoding, UHJ carries the same information as the B-Format signal, however it offers more flexibility when it comes to playback, allowing for several configurations depending on the availability of loudspeakers and equipment. By ignoring the fourth channel, the UHJ decoder allows three loudspeakers to provide a high resolution horizontal surround signal, while two channels provide "very effective" but less accurate horizontal surround. As with B-Format, with the availability of four channels, 3D (periphonic) sound can be reproduced. Finally, the UHJ decoder may be bypassed, allowing a two-channel UHJ signal to be treated as a standard monaural or stereo signal [46]. The UHJ components are referred to as L, R,

T and Q. The L and R signals are used for stereo playback and are derived from the W, X and Y components of the B-Format signal. The T component is used to permit the complete reproduction of the W, X and Y B-Format signals and when four channels are available, the Q component is mapped to the Z component of the B-Format signal (e.g. provides height information) [125].

In addition to the benefits associated with Ambisonic encoded sound, as described by Gardner [55], there are also several drawbacks associated with it. Although the Ambisonic sweet spot is wider than most other systems, it does have a sweet spot, thereby limiting its use. Furthermore, there is a distinct timbral artifact as the listeners move their head near the sweet spot due to the fact that all speakers reproduce the omni-directional component. In addition, even when the original direction to the sound source corresponds to the direction of one of the playback loudspeakers, rather than have all the sound come from this one loudspeaker, the sound will be reproduced (erroneously) from more than one loudspeaker.

Finally, although Ambisonics may not be the standard surround sound format, according to Elen [46], Ambisonics is "still alive" and has simply taken the "back seat" to Dolby Surround and is still, presently, being used to create recordings. In addition, thanks to the scheme developed by Gerzon and Barton [60] Ambisonics can also (in theory) be encoded onto a DVD audio disk. Of course current DVD players support the Dolby Digital surround sound format and therefore adding an Ambisonic decoder to a player would raise the price of the player, something manufacturers are currently not willing to do.

2.6.3 Dolby Stereo

In the early 1970s, Dolby Laboratories introduced "Dolby A" noise reduction in an attempt to improve the low quality sound in films at the time, thus bringing Dolby into the cinema industry. In addition to noise reduction, Dolby was looking at other means of improving the sound quality heard in cinema films. In 1978 Dolby introduced *Dolby Stereo* (also referred to as *Dolby Surround* or *Dolby MP*) for 35mm films, based on optical sound-track technology, a technology used to place monaural sound on film since the 1930s. As shown in Figure 2.10, Dolby Stereo involved the use of four loudspeakers and basically changed the Quadraphonic rectangular configuration consisting of front-left, front-right, rear-left and rear-right into a diamond shape where the speakers were now left, center, right and surround (L, C, R and S respectively) [85], or in other words, three frontal loudspeakers and one "surround" loudspeaker.

The surround sound system consisting of three frontal loudspeakers, as described in Section 2.6, provided good localization for frontal sounds and for people seated to the

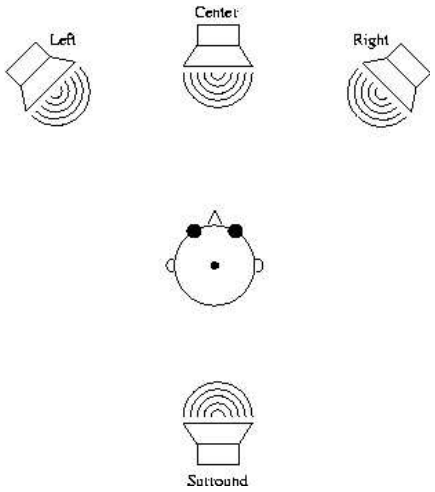


Figure 2.10: Dolby Stereo loudspeaker setup. The original Dolby Stereo format had four loudspeakers (front-left, center, front-right and surround) arranged in a diamond shape around the listener.

left or right of the screen. The surround channel was used to provide greater overall audio realism (e.g. "surround the listener"). Its purpose was to deliver background sounds in order to convey environmental context such as reverberation and other spatial sound effects [16].

Given that many theaters were only equipped for monaural or stereo playback, they could not support Dolby stereo. Building on and improving the matrixing techniques introduced for Quadraphonic systems, Dolby introduced their own encoder/decoder pair allowing their Dolby Stereo format to be encoded into the traditional two-channel stereo format and to be played back as stereo or monaural or with suitable equipment available, decoded back into a four-channel format. The following sections provide greater detail regarding the recording of Dolby Stereo sound as well as Dolby Stereo matrixing (encoding and decoding stages).

Dolby Surround Recordings

Quadraphonic and Ambisonic recordings were made using specific microphones capable of detecting sounds from more than one direction. This is not the only method available to create surround sound recordings. Surround recordings (such as Dolby Stereo and the other Dolby methods, described in the following sections), can also be produced by either recording or creating synthetic versions of (e.g. using a computer) each of the desired sounds (e.g. dialogue, special effects etc.) independently (possibly at different locations)

and then *mixing* each of the sounds in a *mixing studio* (e.g. assigning the sounds to the channels).

Encoding Stage

Encoding the four-channel Dolby Stereo format into a two-channel stereo format was accomplished using the Dolby MP (Motion Picture) matrix encoder. A graphical illustration outlining the operation of the encoder is provided in Figure 2.11. The four Dolby Stereo signals (L, C, R and S) are input and a two-channel stereo signal (or "total signal") L_t and R_t is output. The left (L) and right (R) Dolby Stereo signals are fed directly into the left and right stereo outputs respectively without any modification. The center channel (C) is divided equally but with a 3dB decrease in level, between the left and right stereo outputs. Finally, the surround channel signal (S) is also divided equally to the left and right stereo outputs. However, prior to doing so, the following operations are performed on the surround channel signal:

1. Bandpass filtering to allow frequencies in the range of 100Hz to 7kHz only.
2. Encoding with Dolby B noise reduction.
3. Signal is split into two parts: part one is phase shifted by $+90^o$ and added to the L_t while the other part is phase shifted by -90^o and added to R_t.

Since the left and right (L, R) Dolby signals are fed directly to the corresponding stereo channels and remain completely independent, there is no loss of separation between the left and right stereo outputs and between the center and surround signals [39].

Once the four-channel Dolby Stereo signal has been encoded, it may be stored, transported or played back on any two-channel supported equipment. However, in order to take advantage of the four-channel configuration in which the particular sound event was recorded and intended to be heard, the two-channel encoded stereo signal must be decoded to retrieve the original four channels of information. The following section describes the decoding process in greater detail.

Decoding Stage

The four channels (L, C, R, S) may be recovered from the previously encoded left and right signals L_t and R_t respectively, by essentially reversing the encoding process. The simplest form of the decoding process is outlined in Figure 2.12 and is referred to as *passive* decoding. The left and right encoded signals (L and R), are assigned (without modification) the left

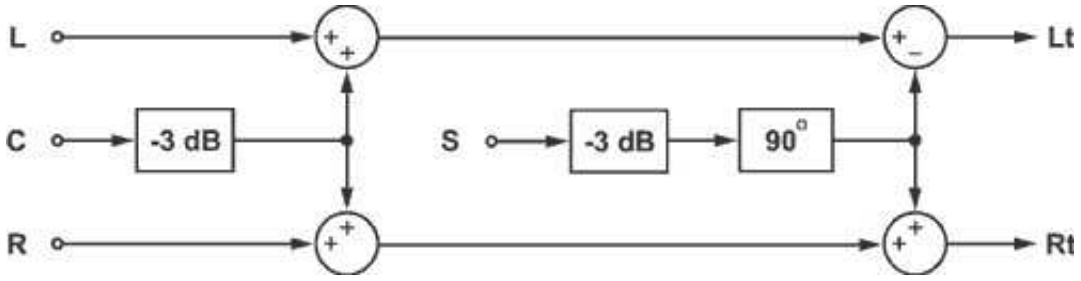


Figure 2.11: Dolby Stereo encoding process. The four channels of the Dolby Stereo format are encoded as two-channels to allow compatibility with existing two-channel stereo equipment. From [39].

and right encoded signals (L_t and R_t) respectively. The encoded signals also contain the center and surround channels, C and S respectively. The surround signal S is recovered by taking the difference d_t between L_t and R_t and performing the following steps to limit any crosstalk between the front and rear loudspeakers:

1. Low pass filter the difference signal d_t to avoid any aliasing
2. Delay d_t by about 15 to 20ms
3. Low pass filter d_t with a cut-off frequency of 7kHz
4. Apply Dolby noise reduction to the difference signal

During playback of the decoded signal, in the absence of a center channel (which is the case with most passive decoder systems [39]), a "virtual source" will be formed between the left and right loudspeakers and as such, no operations are required to recover the center channel.

Given that the encoded stereo signals are passed unmodified and assigned to the left and right decoded signals (L and R respectively), they will also contain the encoded surround signal which is output during playback (e.g. no process is taken to eliminate the surround signal from L and R or from L_t and R_t prior to assigning them to L and R). However, this signal will be heard out of phase (e.g. recall during the encoding phase the surround signal going to the left and right two-channel signals are 180^o out of phase) leading to a diffuse sound [39].

2.6.4 Dolby Pro Logic

Dolby Pro Logic, Dolby's second generation of surround sound was introduced following the original Dolby Stereo system. It was created to balance the improvements seen with respect

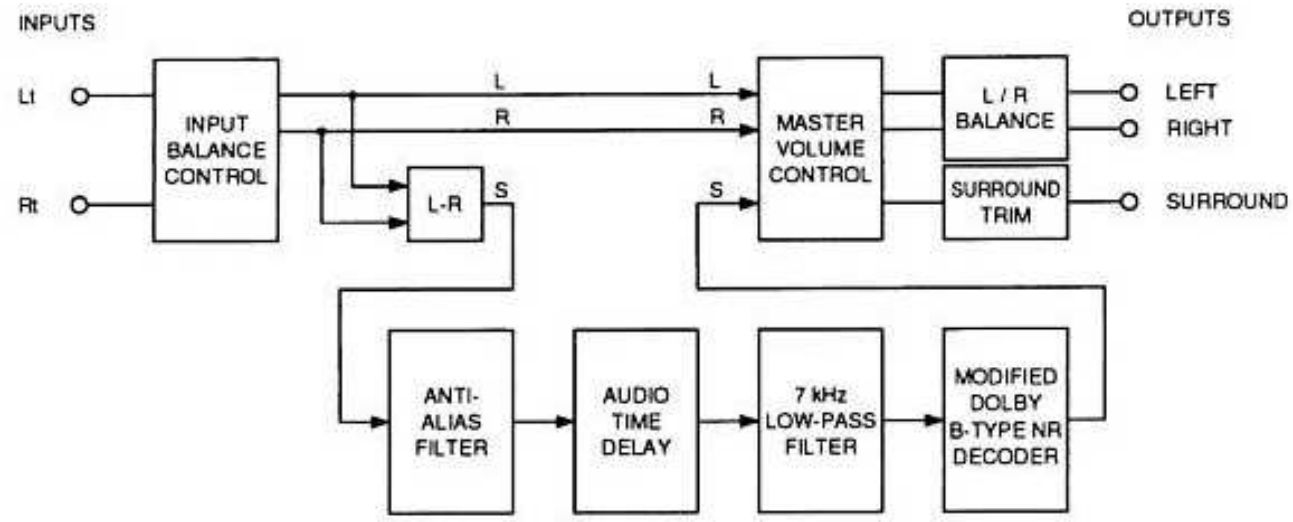


Figure 2.12: Dolby Stereo decoding process. Obtaining the original Dolby Stereo signals form the previously encoded two-channel stereo signals Lt and R_t. Reprinted from [39].

to the video presentation of home movies, such as increased video resolution and larger screens. As with Dolby Surround, Pro Logic consists of four channels, left (L), center (C), right (R) and surround (S) and in fact, uses the identical encoding matrix. However, Pro Logic differs with respect to the decoding stage only, where it employs "active decoding" as opposed to *passive* decoding used in the original Dolby Surround system which employed a simple difference operation. Active decoding allows Dolby Pro Logic to maintain a high separation (30dB) between all output channels and not only the front channels as in the original Dolby surround decoding stage.

2.6.5 Dolby Digital

Dolby Digital, the newest "sound innovation" from Dolby, was first introduced in 1992 with the film *Batman Returns* and made its way to home consumers in 1995 on multi-channel Laser Disc (LD) format. It is based on the Dolby AC-3 method (a method of storing and transmitting multi-channel audio in a fraction of the space needed for standard audio signals) and allows for high flexibility with respect to several operating parameters (e.g. bit rate, number of channels). As with analog Dolby Stereo systems, Dolby Digital includes three front speakers, the left, center and right channels (L, C, R respectively) however, rather than a single surround channel, as with Dolby Stereo, Dolby Digital includes one or more independent surround channels, on each side of the listener in addition to a sub-woofer used for the playback of low frequency effects (LFE) (this configuration is known

as 5.1 and is described in greater detail in the following section). The two independent surround channels allow for "true stereo surround effects" leading to a greater sense of depth, localization and overall realism [90].

Dolby Digital employs Dolby noise reduction to reduce noise levels when no audio signal is present. In addition, this format also takes advantage of human listening, in particular human *auditory masking*, whereby one sound may be made indistinguishable (non-audible) in the presence of another sound [102]. According to Dolby [90], "Dolby Digital separates the frequency spectrum of each channel into narrow frequency bands of different sizes optimized with respect to the frequency selectivity of human hearing. This allows for "sharp" filtering of any present noise, ensuring the frequency spectrum of the noise is close to the frequency spectrum of the signal being coded to take advantage of audio masking, leading to an overall reduction in noise thereby providing higher quality audio delivery".

Being a digital format, signals are represented using bits. However, instead of representing each sample of a particular signal with a static number of bits (as done with standard compact discs), Dolby Digital employs *perceptual coding*, whereby bits are distributed to each frequency band as needed, ensuring a proper number of bits are used to code each signal and the noise is properly masked. Channels with a wider frequency spectrum can be allocated a greater number of bits. In contrast, the audio coding used for compact disc (CD) format, requires a fixed number of bits per sample (16 bits), regardless of the frequency bandwidth. Given the 48kHz sampling rate used in the CD audio coding, the amount of data produced is too large to transmit and store even the standard two-channel stereo configuration, let alone multi channel audio with, for example, six channels.

Dolby Digital is the standard multi channel surround sound format used in digital TV broadcasts in the United States, digital cable and satellite transmissions. Furthermore, it is the standard audio format for DVD in countries which employ the NTSC television standard [90]. In addition, Dolby Digital offers great flexibility with respect to decoding, allowing the signal to be decoded depending on the listener's preference, budget and listening space [90]. This permits a Dolby encoded soundtrack to be heard on a monaural, two-channel stereo, four-channel Dolby Surround or Dolby Digital Surround configurations.

Dolby 5.1

To many, the term "5.1" has become synonymous for and is often used to define surround sound systems [125]. However, 5.1 simply refers to one of the many possible (although the most widely used and most famous), surround sound system loudspeaker configurations. It certainly does not define surround sound. The 5.1 configuration consists of six discrete

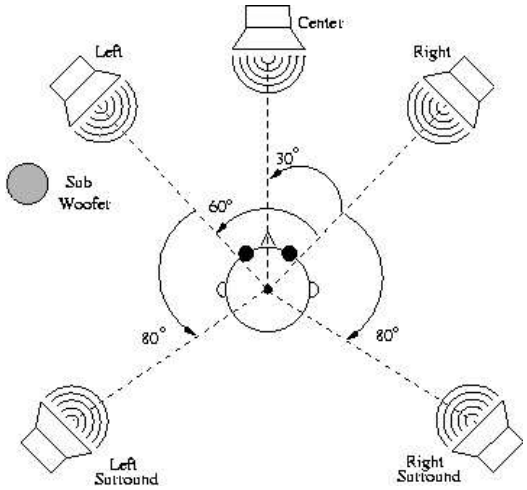
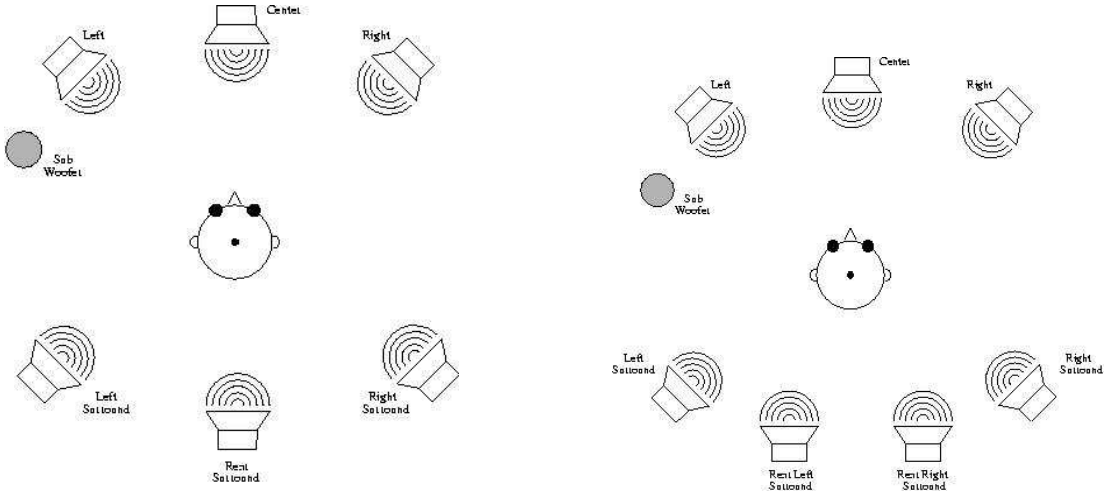


Figure 2.13: Dolby Digital surround 5.1 loudspeaker configurations according to the ITU-R BS 775-1 recommended specification.

channels, was defined in 1987 and became commercially available in 1993 by Dolby Laboratories after several studies by film industry groups found a six speaker configuration produced "satisfying results" in a cinema [73]. The 5.1 configuration (see Figure 2.13) consists of five discrete full bandwidth channels, (hence the 5 in "5.1"), left (L), center (C), right (R), left surround (LS) and right surround (RS), each capable of conveying signals in the range of 20Hz to 20kHz and a sixth, low frequency channel drives a sub-woofer in order to convey low frequency effects (LFE) such as explosions and operates in the frequency range of approximately 5Hz to 120Hz. Since this LFE channel requires a fraction of the full range channel bandwidth, it is known as the ".1" channel and, hence, when combined with the five full range channels, we have 5.1. Placement of the five, full range loudspeakers, according to the International Telecommunications Union specification (ITU-R BS 775-1) is illustrated in Figure 2.13. Since humans cannot localize such low frequency sounds, the sub-woofer may be placed anywhere in the room [92].

Dolby Digital Surround EX

Dolby EX extends Dolby 5.1 by providing an additional surround loudspeaker, placed directly in back of the listener between the existing left and right surround channels, as shown in Figure 2.14(a). The rear surround channel is encoded onto the left and right surround channels as it does not contain its own (discrete) channel. This ensures backward compatibility with the 5.1 format. The addition of this rear surround channel provides greater localization over the three surround channels, allowing for better effects,



(a) Dolby Digital 6.1.

(b) Dolby Digital 7.1.

Figure 2.14: Dolby Digital Surround 6.1 and 7.1 loudspeaker configurations. (a) The 6.1 loudspeaker configuration is obtained by adding a loudspeaker in back of the listener, between the two surround channels of the 5.1 configuration. (b) The 7.1 loudspeaker configuration is obtained by adding two loudspeakers in back of the listener, between the two surround channels of the 5.1 configuration.

in both home and cinema settings. This configuration is known as 6.1. Adding yet another rear surround speaker yields a 7.1 system. As shown in Figure 2.14(b), the additional two loudspeakers are placed between the two original surround loudspeakers of the 5.1 configuration.

2.6.6 Digital Theater Systems (DTS) Digital Surround

DTS Digital Surround is a 5.1 channel surround format. It is very similar to Dolby Digital, consisting of up to five full bandwidth loudspeakers (front left (L), front center (C), front right (R), surround left (LS) and surround right (RS)) and a sixth low frequency effects channel (LFE). As with Dolby, it also employs perceptual coding, using the characteristics of human hearing to reduce noise in order to produce a high quality audio output and to also reduce the amount of data necessary to both transmit and store 5.1 channels of information.

The main difference between DTS and Dolby Digital is with respect to the supported encoding data rates. DTS supports a much higher data rate (1.5Mbit/s), almost four times the rate of Dolby Digital (448kbit/s). DTS also employs less audio compression than Dolby Digital. This, along with the higher data rates, leads to, according to many home

theater enthusiasts and industry experts, "superior sound quality and clarity, far greater than Dolby Digital". However, DTS is certainly not as popular as Dolby Digital and the available soundtracks and movie titles supporting this format is actually much smaller than its counterpart.

The DTS format has also been extended to allow an additional surround channel. DTS Extended Surround (DTS ES) is a 6.1 format, which is similar to Dolby 6.1, includes an additional channel for a surround loudspeaker placed directly in back of the listener. There are two version of DTS Extended Surround. The *DTS Extended Surround Matrix* is simply a 5.1 channel format with the rear surround channel being encoded into the left and right surround channels (e.g. there is no independent rear surround channel). The other format, *DTS Extended Discrete* 6.1, allocates an independent (discrete) channel for the rear surround loudspeaker. This format allows for greater sound localization over the three surround channels. Finally, a DTS ES soundtrack can be played back on a DTS 5.1 setup (e.g. it is backward compatible). The rear surround is simply ignored by the 5.1 format decoder. renewcommand2.01.0

Chapter 3

Simulating Audio in a Virtual Environment

This chapter introduces several techniques used to simulate or recreate, the audio localization cues presented in Section 1.2, to allow for 3D sound in a virtual environment or spatial auditory display. The chapter begins with a discussion of the modeling of ITD cues using a spherical head model as well as an anthropomorphic manikin, followed by a description of binaural synthesis, including an in depth discussion on the techniques available for the measurement of the Head Related Transfer Functions (HRTFs). Finally, several of the more common methods used to include reverberation cues and model a room's acoustics are described.

3.1 Modeling the ITD

A model to predict the ITD for a sound source located on the horizontal plane was presented by Woodworth [168]. This model assumes a spherical head without any external ears and a sound source at an infinite distance away located on the horizontal plane. Given these two assumptions, the ITD τ_{delay}, can be calculated as follows:

$$\tau_{delay} = \frac{a}{c}(\theta + \sin\theta) \tag{3.1}$$

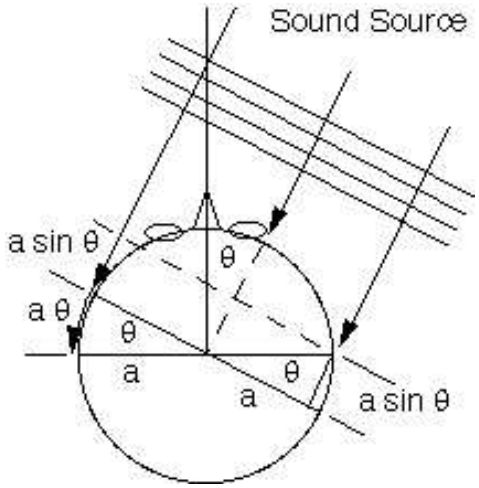


Figure 3.1: Woodworth's prediction of the ITD based on a spherical head model. Reprinted from [1].

where, as illustrated in Figure 3.1, θ is the azimuth angle of the sound source, c is the speed of sound and a is the radius of the sphere representing the head. This formula is valid for angular frequencies (e.g. frequencies expressed in radians per second), which are greater than a/c and is actually very close to the exact theoretical solution even when the source is near the sphere as opposed to an infinite distance away [41]. According to Kuhn [86] however, this model is applicable for "steady state" high frequency signals and clicks only.

In an attempt to overcome some of the limitations associated with Woodwroth's model, Kuhn proposed another model for the prediction of ITDs on the azimuthal plane [86]. As opposed to Woodworth's model, this model is based on ITD measurements made with an anthropomorphic manikin comprised of a head and torso and is also frequency dependent. For low frequencies, the ITD τ_{low} is modeled by the following formula:

$$\tau_{low} = \frac{3a}{c} \sin\theta \tag{3.2}$$

where a is the radius of the head, c is the speed of sound and θ is the azimuth angle of the sound source. This formula is valid provided $\frac{2\pi af}{c} \ll 1$. For higher frequency sounds, where $\frac{2\pi af}{c} \gg 1$, the ITD τ_{high} is modeled by the following formula:

$$\tau_{high} = \frac{2a}{c} \sin\theta \tag{3.3}$$

Although Kuhn's model accounts for both low and high frequency sounds, as with

Woodworth's model, it also restricted to sounds on the azimuthal plane as it does not account for sound source elevation. In fact, very few studies have been conducted to investigate the dependence of elevation on the ITD. However, the few results available do indicate an inverse relationship between ITD and elevation whereby as elevation increases, the ITD decreases [74]. A spherical head model to predict ITD values and also account for elevation was presented by Larcher and Jot [91]. In this formulation, the ITD τ_{delay} is calculated as:

$$\tau_{delay} = \tau_{contralateral} - \tau_{ipsilateral} \tag{3.4}$$

$$\tau_{delay} = \frac{r}{2c}(\sin^{-1}(\cos\phi\sin\theta) + \cos\phi\sin\theta) \tag{3.5}$$

where $\tau_{contralateral}$ and $\tau_{ipsilateral}$ is the time required for the sound to reach the contralateral and ipsilateral ears respectively, θ and ϕ are the angles of azimuth and elevation respectively, r is the radius of the head and finally, c is the speed of sound in air.

Duda et. al. [41] present a model for predicting ITDs based on an ellipsoidal model of the head with offset ears which produces correct variations of ITDs with respect to both azimuth and elevation. The ellipsoidal model does however require values for five parameters, measured from the listener's head.

Although three-dimensional (spatial) sound systems incorporating interaural (ITD and/or ILD) cues only are fairly simple to model and implement, they generally produce poor results, providing limited sound spatialization capabilities (usually restricting sound source localization to the horizontal plane). Furthermore, ITD cues are one of many potential cues available to us in our natural environment. As with human hearing in a natural setting, and as described in future sections, localization improvements can be made by incorporating HRTFs into the system.

3.2 Binaural Synthesis

Rather than recording the signal present at the ears for a particular listening situation as done with binaural recordings (as described in section 2.5.1), binaural synthesis imitates the binaural recording process by processing (filtering) a monaural sound source with a pair of left and right ear HRTFs corresponding to the desired position $\vec{p}$. The HRTFs may be obtained by either physically measuring them directly from a human (or "anthropomorphic

dummy"), or by modeling them.

Theoretically, HRTF filters can be determined by solving the wave equation taking into consideration the interaction of the wave with the head, torso and pinnae, however, such an approach is beyond current computational and analytical reach [40]. Approximations to this solution have been proposed and usually involve simplifying the problem by ignoring the pinnae and torso altogether and assuming a spherical head. The use of a spherical head model does of course ignore the filtering effects introduced by the pinnae, despite the fact that the interaction of the sound wave with the pinnae is the major contributor to the HRTF. As a result, such approximations lead to decreased performance when employed in any 3D sound display. The model introduced by Lord Rayleigh (as described in Section 1.2.1), almost 100 years ago can be considered as such an approximation. Other more complicated models have also been proposed (e.g. see [81, 65, 93]), however, according to Duda [40], there are four major problems associated with modeling of HRTFs:

1. Approximating the effect of wave propagation and diffraction using simple, low order filters is difficult.

2. Complicated relationship between azimuth, elevation and distance on the HRTF.

3. Difficult to measure the accuracy of the approximation due to a lack of any quantitative criteria for doing so.

4. Large variation amongst the HRTFs of individuals

Given the above problems associated with modeling HRTFs, most practical systems employing HRTF cues utilize measured HRTFs and as such, for the remainder of this report, emphasis will be placed on measured HRTFs. The measured HRTFs typically form the coefficients of a FIR filter. The signal delivered to the left and right ears is obtained by filtering (typically through convolution), the monaural sound with the coefficients corresponding to the measured left and right ear HRTFs response respectively. When the filtered sounds are presented to the user either through headphones or loudspeakers, it provides the impression of a sound source at the desired position. Measurement of HRTFs and the limitations and problems associated with such a procedure are described in Section 3.3. Convolution is unfortunately an extremely computationally expensive technique, especially when computed directly using FIR filters, thus limiting the performance of any real time 3D audio system. Improvements can be made using *block* methods which are based on the Fast Fourier Transform (see [147]). Block methods unfortunately introduce a significant amount of delay to the system, once again limiting their "real-time" usefulness. An efficient convolution method using a combination of both direct form and block convolution which does not introduce any delay, has been developed by Gardner [56]. Given the efficiency of this method, it can be used to allow for real-time auralization.

3.3 HRTF Measurement

The HRTF is typically modeled as linear time invariant (LTI) systems [27]. As a result, a common technique used to measure an individual's left and right ear HRTF for a sound source at a position $\vec{p}$ relative to the user, is to output an "excitation" signal $s(n)$ with known spectrum from a loudspeaker placed at position $\vec{p}$ and measure the resulting impulse response h_L and h_R using small probe microphones inserted in the vicinity of the individual's left and right ear canals respectively [16]. Various excitation signals can be used, including an impulse such as a clicking sound, white Gaussian noise, stepped sines, maximum length sequences (MLS) and sweeps (Muller and Massarani [94] provide a review of the measurement of transfer functions using various excitation signals including an in-depth coverage using sweeps). The spectrum of the excitation signal should contain all frequencies of interest (e.g. all frequencies which could be used by the spatial sound system) and since there will be noise introduced by the measurement process itself, the excitation signal should also contain a large amount of energy to ensure a high Signal-to-Noise (SNR) ratio, typically greater than 90dB [94]. The responses h_L and h_R as measured at each ear are in the time domain. The time domain representation of the HRTF is known as the Head Related Impulse Response (HRIR) [53]. Applying the Discrete Fourier Transform (DFT) to the time domain impulse responses h_L and h_R results in the left $H_L(f, \theta, \phi, d)$ and right $H_R(f, \theta, \phi, d)$ ear HRTF respectively, where, as described in Section 1.2.2, θ and ϕ are the azimuth and elevation of the sound source respectively, f is the frequency and d is the distance to the sound source.

A particular HRTF is specified by four parameters, azimuth (θ), elevation (ϕ), frequency (f) and distance (d). As described in Section 1.1.2, distances greater than approximately $1m$ (e.g. a far field acoustical model), have a minimal effect on the measured response [55] and therefore, provided the distance to the source is at least one meter, distance is typically ignored. When the sound source distance is closer than about one meter (e.g. the source is in the near field), the HRTF is dependent on the distance to the sound source and therefore, source distance cannot be ignored [21].

To allow for greatest flexibility, robustness and accuracy, the measured HRTF response should not include any reverberation reaching the ears that may may result from reflections of the impulsive sound off of any surfaces (e.g. walls, floor, ceiling and any other objects) in the environment in which the measurements are made. As a result, in order to minimize the effect of reverberation HRTF measurements are usually obtained in an anechoic chamber. HRTF measurements can be made in a "normal" reverberant room. In such a situation, the HRTF will include reverberation effects (e.g. binaural synthesis including reverberation). This will limit the auditory display to the reverberation pattern of this one particular room (the room in which the measurements took place), thereby allowing the simulation of this one room only, at the time the emasurements were made. Furthermore, throughout

the HRTF measurement process, the individual should remain motionless as even small head or body movements can degrade the measured response (e.g. the measured response may no longer correspond to the desired position). However, since it is difficult to have a human subject stay completely motionless during the measurements process (as described in Section 3.3.2), an anthropomorphic "dummy head" is usually used instead.

There is a large variation between the measured HRTFs across different subjects, which, according to Carlile [25], results from a number of factors, including the following:

Variation of Each Person's Pinnae: The physical make-up of each person's pinnae differs, leading to differences in the filtering effects introduced by the pinnae and therefore the measured HRTFs.

Differences in the Measurement Procedures: Currently there is no single standard approach for measuring the HRTFs [16] and as a result, the procedure itself varies widely. A major concern relates to the position within the outer ear at which the HRTF should be measured. Although measurements of the response can be recorded anywhere within the ear canal, one should place the microphone as close as possible to the eardrum to avoid the reflections of the incoming sound by the eardrum itself [83].

Perturbations of the Soundfield by the Measuring Instruments: The microphone used to measure the response may itself interfere with the measurement. Despite the fact that the microphones used are rather small (e.g. less than 5mm in diameter), they can perturb the soundfield, especially at high frequencies [25].

Variations in the Relative Position of the Head: HRTFs may be very sensitive to variations in the subject's head position, where even small head movement during the measurement procedure can result in a large variation in the HRTF measurements within one subject.

Another problem with measuring HRTFs is the fact that the measured HRTF is valid for one particular position only. The process should, ideally, be repeated for every possible position in three-dimensional space, as the HRTF at each unique position is itself unique. This is impractical as the number of unique positions to be simulated by a 3D audio system is potentially very large and the task of actually collecting HRTF measurements is both tedious and time consuming. For practical considerations, HRTFs are typically sampled at a number of discrete positions around the individual. This sampling in turn results in further problems. Since the HRTFs are collected at discrete positions, there will surely be positions which cannot be accurately simulated as there is no corresponding measured HRTF impulse response. Various techniques have been developed to deal with such non-sampled positions including simply using the HRTF corresponding to the position closest

to the intended (target) position or, as described in the following section, interpolating between HRTF measurements.

3.3.1 Interpolation of HRTFs

The simplest interpolation technique is linear interpolation whereby the desired HRTF is obtained by taking a linear average of neighboring HRTFs. This technique results in HRTFs which are acoustically different when compared to the actual measured HRTF of the desired target location [87]. According to Wenzel et. al. [166], localization accuracy is not affected by linear interpolation of non-individualized HRTFs even with a large interval separating the sampled HRTF measurements. They believe that despite the error associated with interpolation of HRTFs, this error is smaller relative to the error associated with the use of non-individualized HRTFs as opposed to individualized HRTFs.

Various other interpolation techniques can also be used, such as the more complex spline interpolation techniques, used in various other fields, including computer graphics [69]. Regardless the interpolation technique actually used, some method is needed to handle the fact that it is clearly impractical to measure and store HRTF responses for each location in space relative to the listener.

3.3.2 The Use of Non-individualized ("Generic") HRTFs

The pinnae of individuals differ with respect to size, shape and general make-up, leading to differences in the filtering of the sound source spectrum, particularly at higher frequencies amongst individuals. Regardless the individual, the higher frequencies are attenuated by a greater amount when the sound source is to the rear of listener as opposed to the front of the listener and in the 5kHz to 10kHz frequency range, the HRTFs of individuals can differ by as much as 28dB [163]. This high frequency filtering is an important cue to source elevation perception and in resolving front-back ambiguities [126, 127, 100, 160, 16]. The unique filtering effects performed by each person's pinnae results in a differing set of HRTFs, where the differences are large enough to warrant the use of *individualized* HRTF measurements in a spatial sound system [25]. Best results are achieved when an individuals own HRTFs are used [160].

Despite the benefits which may be offered to a listener through the use of individualized HRTFs, the process of collecting a set of an individualized HRTFs is an extremely difficult, time consuming, tedious and delicate process requiring the use of special equipment and environments, such as an anechoic chamber. Furthermore, although researchers are actively pursuing methods and techniques to accurately measure and gather HRTF responses, currently, there is no single scientifically accurate method for doing so [16]. It

is therefore very impractical to use individualized HRTFs and as a result, generalized (or generic) *non-individualized* HRTFs are used instead. As will be described in the following sections, non-individiualized HRTFs can be obtained using a variety of methods such as measuring the HRTFs of an anthropomorphic "dummy" head, or of an above average human localizer or averaging the HRTFs measured from several different individuals (and/or "dummy heads"). However, studies indicate that these non-individualized HRTFs reduce localization accuracy, especially with respect to elevation. Wenzel et. al. have performed various studies examining the effect of non-individualized HRTFs and have determined that the use of non-individualized HRTFs resulted in a degradation of the subjects' ability to determine the elevation of a sound source [160, 161]. Similarly, studies performed by Begault and Wenzel [15], in which subjects localized a speech stimuli as opposed to broadband noise, (as used in easlier studies), indicate a decrease in elevation judgments as well [25].

In addition to the filtering effects introduced by the pinnae, HRTFs are also affected by the head, torso and shoulders of the individual, leading to further degradations when using non-individualized HRTFs. Regardless of the method used to obtain the set of non-individualized HRTFs, the performance of the auditory display will be greatly reduced when the size of the listener's head differs greatly from the size of the head used to obtain the HRTF measurements (dummy head or person) [82].

HRTF Measurements Obtained with an Anthropomorphic Dummy Head

In order to eliminate the possibility of errors in the collected HRTFs due to subject head or body movements and to overcome the fact that it is a long and tedious process for any human subject participant, rather than using human subjects to collect HRTF measurements, an anthropomorphic manikin can be used instead. Begault [16] provides a description of several dummy heads, including the popular and widely used, Knowles Electronics KEMAR (Knowles Electronic Mannequin for Acoustic Research) standard anthropomorphic dummy head. The KEMAR consists of a head, torso and pinnae (see Figure 3.2), obtained from human median measurements [23] and contains removable pinnae to allow for the use of different pinnae models.

HRTF Measurements From an Above Average Localizer

Given individual variation of the pinnae filtering effects, it seems intuitive that there exists individual variation amongst the localization ability and accuracy. There are people who are "good localizers", capable of producing accurate azimuth and elevation localization results while others are "poor localizers", who demonstrate little localization ability. It may then seem plausible that non-individualized HRTFs obtained from good localizers may improve the localization accuracy of average or poor localizers. Evidence does suggest this

Figure 3.2: KEMAR Mannequin. The KEMAR (Knowles Electronic Mannequin for Acoustic Research) is often used to obtain HRTF measurements. Reprinted from [82].

is the case. A study by Wenzel et. al. [161] examined whether HRTFs obtained from a good localizer could improve the localization accuracy of real sound sources for a poor localizer. The HRTFs of both the good and poor localizers were measured. When synthesizing the sound source with their own (individualized) HRTFs, as expected, the good localizers produced accurate localization results while the poor localizers showed poor localization results. Similarly, when presenting the sound source to the good localizers using HRTFs obtained from other good localizers, accuracy decreased only slightly. When presenting the sound source to the good localizers using HRTFs from a poor localizer, the localization accuracy decreased substantially. However, localization accuracy was improved for the poor localizers when using HRTFs obtained from good localizers. Similar results were found in a more comprehensive study also performed by Wenzel et. al. [160], investigating the two-dimensional localization accuracy of 16 inexperienced localizers using non-individualized HRTFs obtained from a "good" localizer. For 14 of the 16 subjects, localization accuracy of virtual sources presented over headphones using the non-individualized HRTFs was comparable to the localization of a real sound source presented without headphones and any HRTF processing (e.g. sound source in the free-field). Furthermore, their results also suggest that the use of non-individualized HRTF measurements results primarily in an increase in the front-back confusions.

HRTFs Obtained by Averaging the Response of Several Listeners

In this method, a non-individualized HRTF dataset is obtained by averaging the Fourier domain representation of the HRTF measurements of several human and/or anthropomor-

phic manikin. The motivation behind this method is to remove (through the averaging process) any distinct spectral features of any one individual's HRTF response [16]. One drawback of this technique is that the time consuming, tedious and delicate process of collecting the HRTF responses must be repeated for every one of the subjects included in the averaged dataset.

Rather than restricting an auditory display to a single dataset of HRTF measurements, several datasets, obtained using either of the methods previously described (e.g. averaged, good localizer or anthropomorphic dummy), can be used and for each user of the display, a single HRTF dataset is chosen based on some criteria to allow for maximum accuracy. This is the approach taken by the 3D sound system of Chong et. al. [150], in which sounds are synthesized using one of several available HRTF datasets, depending on the user. Prior to using the sound system, a listener test is presented to the user. The test evaluates the user's localization performance using each of the six datasets. The dataset resulting in the highest accuracy is chosen and will be used by the system to synthesize a sound source(s) for this particular user.

3.3.3 Available HRTF Datasets

Given the difficult and tedious task associated with the measurement of HRTFs, very few HRTF datasets exist. Furthermore, given the potential expense associated with collecting a dataset of HRTFs, researchers and companies who take the initiative to actually collect a set of HRTFs, are often reluctant to share them with others. Fortunately, however, several HRTF datasets have been made freely available to the research community. The following sections provide greater details regarding three such HRTF sets, the dataset from MIT's Media Laboratory of Perceptual Computing, measured by Gardner et. al. [59], the CIPIC HRTF dataset measured by Algazi et. al. [1] and the LISTEN HRTF dataset measured as part of the LISTEN project [76].

MIT KEMAR HRTF Measurements

This set of "raw", un-processed HRTFs were measured using the anthropomorphic dummy KEMAR. The KEMAR was equipped with two different pinnae (each of the two ears had its own, different, pinna model), and as described below, this allowed for the simultaneous measurement of two HRTF sets, one set for each corresponding pinnae model.

The KEMAR was mounted on an electronically controlled turntable capable of being rotated 360^o and placed in an anechoic chamber, at a distance of 1.4m from the sound source (a Realistic Optimus Pro 7 loudspeaker). The loudspeaker itself was positioned on an electronically controlled "boom", allowing it to be positioned at any elevation relative to the

Elevation	Total Measurements	Azimuth Increment
-40	56	6.43
-30	60	6.00
-20	72	5.00
-10	72	5.00
0	72	5.00
10	72	5.00
20	72	5.00
30	60	6.00
40	56	6.43
50	45	8.00
60	36	10.00
70	24	15.00
80	12	30.00
90	1	-

Table 3.1: Resolution of the KEMAR HRTF measurements. Each row lists the number of azimuth samples obtained at the corresponding elevation.

KEMAR. By placing the loudspeaker at some specific elevation and azimuth with respect to the KEMAR, the HRTF measurement corresponding to that particular position was obtained by outputting a sound through the loudspeaker and recording the sound with a probe microphone in each of the ear canals of the KEMAR. In total, 710 measurements were sampled, one elevation at a time, by moving the loudspeaker to some particular elevation, from -40^o to 90^o (in 10^o increments) and rotating the KEMAR a total of 360^o, in equal increments for each elevation. The increment size was chosen to "maintain approximately 5^o great-circle increments" [59]. Table 3.1 illustrates the azimuth increments for each elevation. The impulse response measured at each ear contains a total of $16,383$ samples, sampled at a rate of $44,100$Hz. and stored as 16-bit integers.

Although for each HRTF measurement the measured response was $16,383$ samples long, not all samples are included in the dataset. Rather, each response has been reduced to 512 samples. Since sound does not travel instantaneously, it does take (a small) amount of time for the sound to travel from the speaker to the ear. Also, there is an additional delay of 50 samples introduced by the measurement system. As a result, the first 200 samples were discarded to account for this (e.g. assuming a speed of $345m/s$ for sound waves traveling through air, the time in number of samples to reach the ear is $\frac{1.4m}{344m/s} = 180$ samples). Similarly, the last $15,671$ samples have been discarded to avoid corruption of

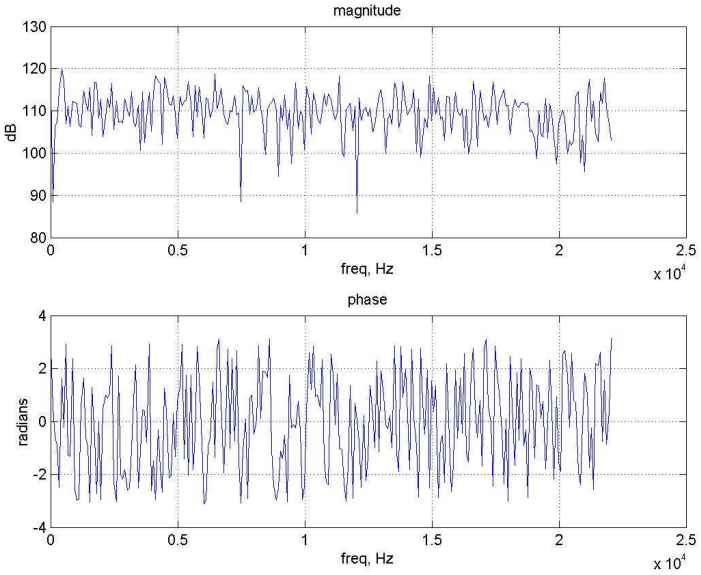


Figure 3.3: MIT KEMAR HRTF measurement for the sound source positioned at 0^o elevation and 0^o azimuth. Magnitude, in dB (top) and phase (bottom).

the measurement with respect to reverberation caused by reflection of the sound waves with other objects in the anechoic chamber, including the KEMAR itself, the boom and the turntable. A sample of the magnitude (in dB) and phase of two HRTF measurements are illustrated in Figures 3.3 and 3.4. In particular, the response for the sound source at elevation 0^o and azimuth 0^o is shown in Figure 3.3, while the response for the sound source positioned at elevation 40^o and azimuth 90^o is illustrated in Figure 3.4.

In addition to the complex interactions between the sound waves and the KEMAR (pinnae, torso etc.) the impulse response contains the response of the measurement system as well (speaker, amplifiers, environment etc.) and may produce poor results when used in a spatial display. However, as described in Section 3.3.4, the HRTFs may be equalized to compensate for these unwanted effects.

The KEMAR was fitted with a different pinnae model for each ear and the response was measured simultaneously at each ear. As a result, the responses lack any ITD cues. If required, ITD cues must be added by the system (see Section 3.1). Details regarding how to actually access a desired HRTF response from the dataset (e.g. file/directory names etc.) may be found in [59].

Finally, in addition to the HRTF measurements, the responses of the loudspeaker, headphones, microphones and measurement system (e.g. electronics equipment) are also

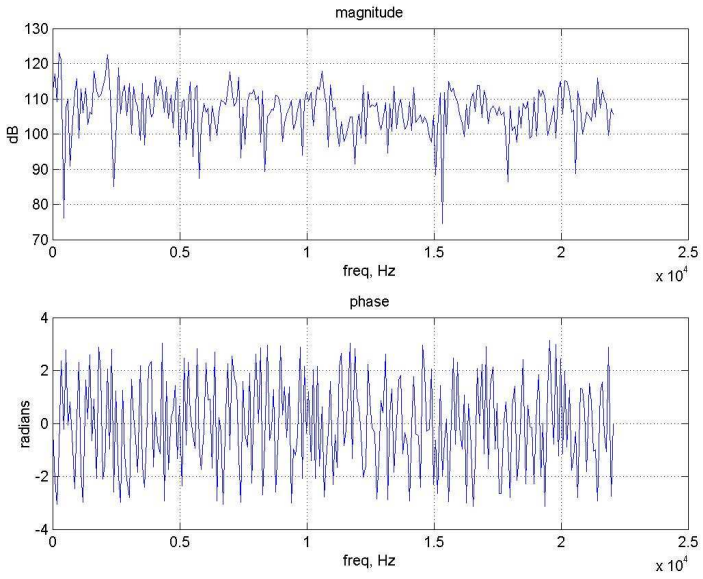


Figure 3.4: MIT KEMAR HRTF measurement for the sound source positioned at 40^o elevation and 90^o azimuth. Magnitude, in dBs (top) and phase (bottom).

included separately and as described in Section 3.3.4, can be used to "equalize" each HRTF measurement (e.g. remove the measurement system response from the HRTF measurement).

The CIPIC HRTF Database

The publicly available CIPIC HRTF database [1] consists of 45 individual HRTF datasets obtained from 43 different human subjects (27 men and 16 women) and a KEMAR mannequin (with two different pinnae models). For each subject, a total of $1,250$ measurements were taken at each ear (25 different azimuths and 50 different elevations). For this dataset, an "interaural-polar coordinate" system was used as opposed to the "vertical coordinate system" introduced in Section 1.1.3. As shown in Figures 3.5a,b, in the interaural polar coordinate system the origin is defined at the center of the head between the ears. As with the single and double polar coordinate systems, sound source locations are given by specifying azimuth and elevation angles (θ and ϕ respectively) in addition to range r. However, in this coordinate system, azimuth measures the angle between the vertical median plane and a vector to the sound source while elevation measures the angle from the horizontal plane to a plane through the source and the x axis (the interaural axis) [1]. In this dataset, azimuth angles were at -80^o, -65^o, -55^o, from -45^o to 45^o in increments of 5^o, 55^o, 65^o and 80^o to $+80^o$. Elevation ranged from -45^o to $+230.625^o$, equally sam-

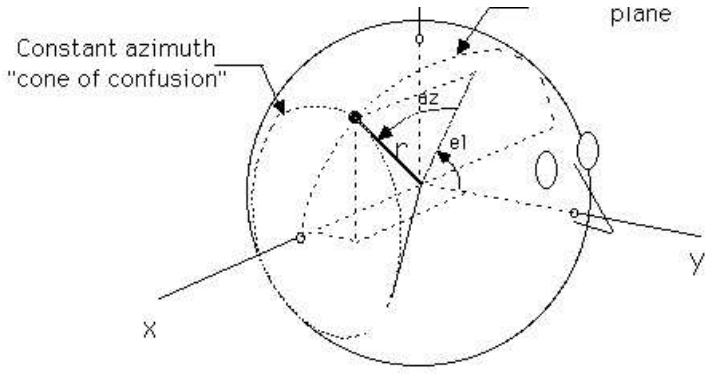


Figure 3.5: Interaural polar coordinate system used to obtain the CIPIC HRTF measurements. Reprinted from [1].

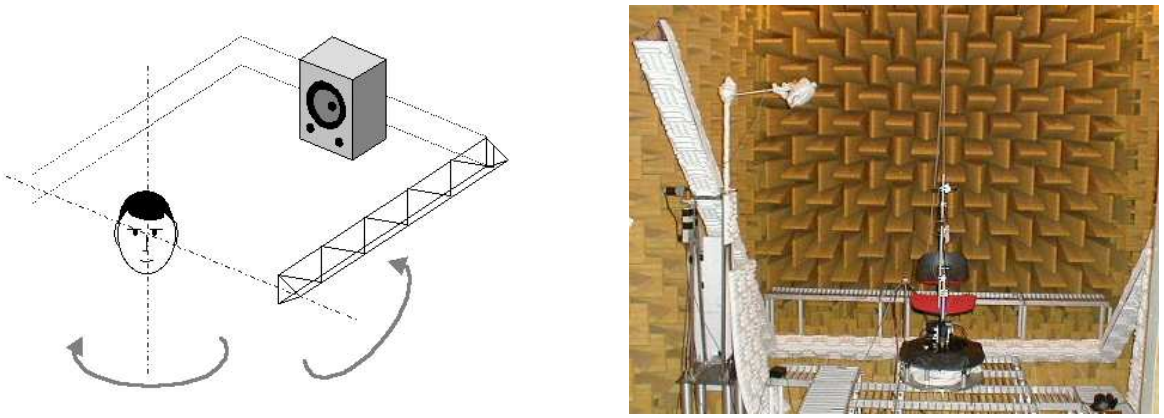
pled in 6.625^o increments. Each of the measurements contains 200 samples, sampled at a rate of 44.1kHz. In addition, the measurements were also equalized to account for the response of the measurement system as well as the microphone and loudspeaker used in the measurement phase (see Section 3.3.4 for more details regarding the equalization of HRTFs).

The microphone was placed just outside the ear canal entrance and as such, the response of the ear canal is not included. Finally, included with the database are detailed anthropomorphic measurements (e.g. head width, head height, shoulder width etc.) for each subject.

The LISTEN HRTF Database

This publicly available dataset currently consists of the HRTF measurements of 49 human subjects and was made available towards the end of 2002 by Ircam (*Institute de Recherche et Coordination Acoustique/Musique*) and AKG Acoustics (manufacturer of studio microphones and broadcast equipment), as part of the LISTEN project [44, 43]. The database is periodically updated with the addition of HRTF measurements from new subjects.

The measurements were made in an anechoic chamber. A graphical illustration of the anechoic chamber and the equipment set-up is available in Figure 3.6. The subject was seated on a "common office chair" which was itself mounted on a computer controlled turntable, capable of rotating 360^o (see Figure 3.6(b)). A single loudspeaker (TANNOY System 600 driven by a YAMAHA amplifier), mounted on U-shaped crane, was used to output the impulsive sound (see Figure 3.6(a)). The elevation of the loudspeaker was set by adjusting (via computer) the elevation of the crane. For each loudspeaker elevation,



(a) Loudspeaker, crane, subject.

(b) Actual set-up.

Figure 3.6: LISTEN HRTF measurement set-up. The loudspeaker is mounted to a computer controlled crane. (a) The elevation of the crane is then adjusted to the desired elevation and for each elevation, the subject is rotated to appropriate azimuth and the measurement is taken. (b) Actual photograph of the set-up. Reprinted from [76].

the chair was rotated to the appropriate azimuth angle and a pair (left and right ear) of HRTF measurements were made. As with the MIT KEMAR dataset previously described, the number of azimuth settings varied depending on the elevation. Table 3.2 summarizes the azimuth increments for each elevation.

For each subject, the HRTF measurements of 187 discrete locations were made. The response was measured using a pair of very small (e.g. 2.54mm diameter with a height of 2.54mm) blocked-meatus microphones (Knowles FG3329). The microphones were held firmly in place with silicon putty. Blocked-meatus microphones are meant to be inserted into the ear canal and although they do not capture any ear canal resonance, according to, Brown and Duda, it is generally believed that they capture the direction dependent components of the HRTF [18].

For each subject, the "raw", non-equalized data as well as the equalized data, both in Microsoft^{TM} WAV and Matlab^{TM} formats, are available. Non-equalized measurements consist of the entire HRTF measurement, a total of 8192 samples (e.g. 0.186s duration) and include the response of the measurement system and equipment (see Section 3.3.4). The equalized dataset was obtained by removing a portion of the start/end of the measurement to avoid any propagation delays introduced by reverberation and equalizing the measurement using diffuse field equalization (see Section 3.3.4). The equalized measurements are 512 samples long (e.g. 0.012s duration). Included with the dataset is an "information database" which provides information related to the subject, the measurement environment and the measurement system and equipment (e.g. subject's age, the dimensions of

Elevation	Total Measurements	Azimuth Increment
-45	24	15
-30	24	15
-15	24	15
0	24	15
15	24	15
30	24	15
45	24	15
60	12	30
75	6	60
90	1	360

Table 3.2: Resolution of the LISTEN HRTF measurements. Each row lists the number of azimuth samples obtained at the corresponding elevation.

the anechoic room the measurements took place, the distance between the subject and the loudspeaker etc.).

3.3.4 Equalization of the HRTF Impulse Response

In addition to containing the actual impulse response $h_{actual}[n]$ due to the head, pinnae, torso and shoulders, the measured HRTFs $h_{measured}[n]$, include the impulse response $m[n]$ due to the loudspeaker, headphones and electronic measurement system [55]. Mathematically,

$$\begin{aligned} h_{measured}[n] &= h_{actual}[n] * m[n] \quad or \\ H_{measured}(e^{j\omega}) &= H_{actual}(e^{j\omega}) \times M(e^{j\omega}) \end{aligned} \tag{3.6}$$

where $X(e^{j\omega})$ is the Discrete Fourier transform of the finite signal $x[n]$ and "$*$" denotes convolution. For a spatial auditory system incorporating HRTFs, it is $h_{actual}[n]$ which is desired and not a response which has been modified in any way, including the introduction of $m[n]$, as such modifications will negatively affect the performance of the system. Various *equalization* methods have been developed in order to compensate for (remove) the response of the measurement and playback systems. These methods typically involve "filtering" the HRTF measurements with the inverse of a filter $H_{filter}(e^{j\omega})$ which includes the un-wanted

components, including the measurement system response. In the frequency domain, the filtering can be performed by multiplying each HRTF response $H_{measured}(e^{j\omega})$ by the inverse of $H_{filter}(e^{j\omega})$:

$$H_{actual}(e^{j\omega}) = H_{measured}(e^{j\omega}) \times \frac{1}{H_{filter}(e^{j\omega})} \tag{3.7}$$

Below is a summary of three methods which can be used to obtain the filter $H_{filter}(e^{j\omega})$ [55]:

Measurement Equalization: HRTFs are equalized with respect to the response of the measurement system, which is obtained by measuring the response at the position corresponding to the center of the head without the head present. The response should be measured using the same microphone and equipment which are used to obtain the actual HRTF measurements. Since an HRTF measurement ideally represents the interaction between a sound and the head torso and pinnae, measuring the response without the head (real person or "dummy" head) present should provide the response of the measurement equipment, including the loudspeaker, amplifiers, A/D cards and microphone (i.e. $m[n]$ as described above) which is present in all measurements of the dataset.

Free-Field Equalization: The HRTFs are equalized with respect to one of the measured HRTFs. Since the measured HRTF contains the response of the measurement system, when equalizing with respect to one measurement, the unwanted measurement response will be removed while the directional components will (ideally) remain intact.

Diffuse-Field Equalization: Equalize each of the HRTFs with respect to the diffuse field average. The diffuse field average, $H_{DF}(e^{j\omega})$ is obtained by averaging the power of the HRTFs measured at all locations. Mathematically, the magnitude of the diffuse field average response is obtained as follows (phase is ignored) [55]:

$$H_{filter}(e^{j\omega}) = |H_{DF}(e^{j\omega})| = \sqrt{\frac{1}{N}\sum_{i=1}^{N}|H_i(e^{j\omega})|^2} \tag{3.8}$$

where, N is the total number of HRTFs measured and $H_i(e^{j\omega})$ is the HRTF measured at location i. Since the diffuse field signal is an average of all the measured HRTFs, it will contain components common to all the HRTFs, including the response of the

measurement system. When equalizing a measured HRTF with this average (e.g. by multiplying the HRTF by the inverse of H_{DF}), all "common" components will be eliminated, leaving only the (desired) directional components which are specific to the position of interest.

3.4 Modeling of Reverberation and Room Acoustics

Given the benefits reverberation has to offer, incorporating reverberation into a virtual auditory display seems obvious. Indeed, adding reverberation to an auditory display can be advantageous for several reasons. For example, it will improve distance estimation accuracy and create a more realistic sounding environment [139]. Furthermore, as described by Begault [16], without reverberation, the auditory display can only output sounds as in an anechoic environment, thereby lacking any realism. In addition, as described in Section 4.1, reverberation also allows for the externalization of a sound source presented over headphones. Despite the benefits reverberation offers, it also has its share of drawbacks. Most importantly, the reflections reaching a listener will vary depending on the geometry of the room, the material composition of the walls, ceiling and floor, objects present in the room and the listeners position in the room. However, exactly imitating these complex interactions of the reflected waves, especially with systems in which any of the room parameters may be updated in real time, is extremely computationally intensive, making the simulation of a "true" realistic reverberant environment impossible, with the computing and DSP technology currently available [139].

Two basic techniques are available to enable the inclusion of reverberation in a virtual auditory display. With *auralization* techniques, the desired listening environment is recreated by determining the reflection patterns of any sound waves in the environment, using either physical or mathematical modeling. Rather than relying on such models, reverberation can also be added using *artificial* techniques. These techniques are not necessarily concerned with recreating the exact reflections of any sound waves in the environment. Rather, they "artificially" recreate reverberation by simply presenting the listener with delayed and attenuated versions of a sound source, where the delays and attenuation factors do not necessarily reflect the physical properties of the environment being simulated. These factors are chosen by "trial and error", adjusting these settings until a desirable effect is achieved [35]. Although such techniques do not necessarily accurately model the early reverberation patterns of a room, they are capable of providing convincing late reverberation effects [51]. Greater details regarding artificial reverberation techniques can be found in [35, 51, 134, 133, 79, 57]. The following section examines auralization techniques in greater details.

3.4.1 Auralization

According to Kleiner et. al. [84], auralization is defined as "the process of rendering audible, by physical or mathematical modeling, the sound field of a source in space, in such a way as to simulate the binaural listening experience at a given position in the modeled space". The goal of auralization is to recreate a particular listening environment by determining the reflection patterns of sound waves which emanate from a sound source(s) as they propagate through the environment. This is accomplished by computing the *binauaral room impulse response* (BRIR), or in other words, the response of the "real" room.

In a manner similar to the measurement of HRTFs, the response of a "real" room can be measured by outputting a sound with known characteristics through a loudspeaker positioned somewhere in the room and measuring the response with a microphone positioned elsewhere in the room. The microphone captures the direct sound emitted by the source as well as any reflections (both early and late) which may arise, or in other words, it can capture the room acoustics for that particular sound source and listener (microphone) configuration. The measured response is known as the Binaural Room Impulse Response (BRIR), and captures the reflection properties, sound attenuation and absorption properties of a particular room configuration. As with HRTF measurements, the BRIR can then be used to filter a monaural sound source and when this processed sound is presented to the listener, the original acoustic environment is reproduced. Figure 3.7 provides a graphical illustration of an actual impulse response measured in a "standard classroom" [137]. The process of measuring the BRIR is for one specific room configuration with the sound source and listener at some particular position and as a result, only this particular configuration can be re-created. Changes in the position of the sound source, listener, objects in the room or room configuration (e.g. introduction of new objects in the room), will potentially result in a change of reflection patterns reaching the listener and therefore, the BRIR may no longer be valid. As with HRTF measurements, the BRIR can be sampled at various locations of the room and during any changes in the virtual environment simulation, some form of look-up and interpolation can be used to determine the appropriate BRIR. Once the BRIR has been obtained, it is then used to filter a monaural sound and this filtered sound is presented to the listener. As with HRTFs, this filtering is accomplished using convolution in the time domain or multiplication in the frequency domain.

The BRIR for a particular environment can be obtained using either *acoustic scale modeling* or *computer modeling*. In the acoustic scale modeling technique, three dimensional scaled down actual material models of a particular environment are built and used to examine the acoustical properties of the real environment and ultimately measure the BRIR. These methods allow for the correct inclusion of all the room effects, including scattering and diffraction of the sound waves as they encounter surfaces in the environment rather than relying on mathematical approximations as done with the computer modeling

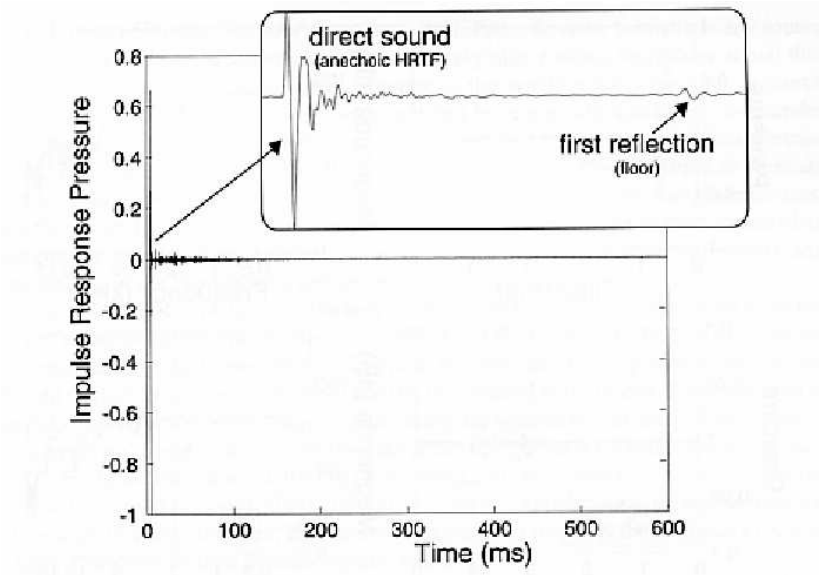


Figure 3.7: Binaural room impulse response measured in a "standard classroom". Right ear with a sound source positioned at 45^o azimuth, 0^o elevation and at a distance of 1m. Taken from [137].

techniques [84]. Measurement of the BRIR can also be made using a dummy head or a human listener. The measured BRIR in this case will include the HRTF response as well. This is in fact the process used to measure binaural recordings (see Section 2.5.1), which themselves are a form of BRIRs.

With computer modeling techniques, the BRIR is predicted (modeled) using some form of mathematical model and a computer. Approaches using this technique can be divided into two categories, *wave-based modeling* and *ray-based modeling* [130][1]. With wave-based methods, the objective is to solve the wave equation (also known as the *Helmholtz* equation) in order to completely recreate a particular sound field. An analytical solution to the wave equation however is rarely feasible [130], thereby limiting its use. Wave-based methods using numerical approximation, such as finite element methods (FEM), boundary element methods (BEM) and finite difference time domain methods (FDTD) can however be used [130]. Such methods are used to solve complex integral equations, by sub-dividing the domain of these complex functions into smaller units such that for each smaller unit, the function can then be approximated using simpler functions [12, 36]. In other words,

[1]Actually, Savioja [130] includes a third category called *statistical modeling*. However, described in [130], statistical modeling is primarily applied to "predict noise levels in coupled systems in which sound transmission by structures is an important factor" and hence not suitable for auralization.

numerical approximations such as FEM and BEM essentially project the original, complex function into a *finite function* space where now, the approximated function is characterized by a finite number of unknowns which can then be solved numerically [31]. Numerical approximation are widely used for a variety of engineering analysis tasks, including heat transfer formulations, and the popular global illumination, image rendering technique *radiosity* [70]. With respect to acoustical wave-based methods, numerical approximations sub-divide the boundaries of a room into smaller units (*elements*). Then, by assuming the pressure at each of these elements is a linear combination of a finite number of basis functions, the "boundary integral form" of the wave equation can be solved [51]. However, numerical approximations are computationally very expensive making them practical for simple, static environments and low frequency sounds only. In fact, according to Rabenstein et. al. [118] 4.2 Gigaflops (4.2 billion floating point operations each second), are required to simulate the propagation of 3kHz sound wave in a $100m^3$ room using such a method and the number of floating point operations (*flops*) increases linearly with the volume of the room and is proportional to the fourth power of the frequency of interest.

As with methods used in computer graphics to render a scene, in ray-based acoustical modeling, the propagation paths taken by the sound waves as they travel from the sound source to the receiver (listener), are found by following "rays" emitted by the source. While traveling in the environment, these rays may interact with any number of surfaces in the environment (e.g. reflected when they encounter a wall). Mathematical models are used to account for source emission patterns, atmospheric scattering reflections off surfaces (including taking into account absorption of a portion of the wave the surface itself), diffraction and absorption of the sound by the medium (air), which may occur as the sound waves interact with objects in the environment [51]. The BRIR is obtained by combining the (filtered) rays actually reaching the receiver. The ray-based methods are not completely valid as they completely ignore the wavelength of sound waves as well as any phenomena associated with it (e.g. diffraction) [88]. For example, the wavelength of very low frequency sounds can be large, and can actually "bend" around certain objects whose dimension happens to be smaller than the wavelength of the sound wave. These methods are therefore valid only when dealing with sound wavelengths which are smaller than the dimensions of the objects in the environment but larger than the roughness of these objects [74]. As a result, such methods are valid for higher frequencies sounds only. In addition, although ray-based methods are rather simple to implement, they can only accurately model the early portion of reverberation pattern as they do become computationally expensive as the number of reflections and diffractions increases [51]. Furthermore, as described by Kleiner et. al. [84], these methods can be rather complicated for all but very simple, theoretically ideal cases. Greater details regarding two of the more popular ray-based methods, (*image source* and *ray tracing*), are provided in the following sections.

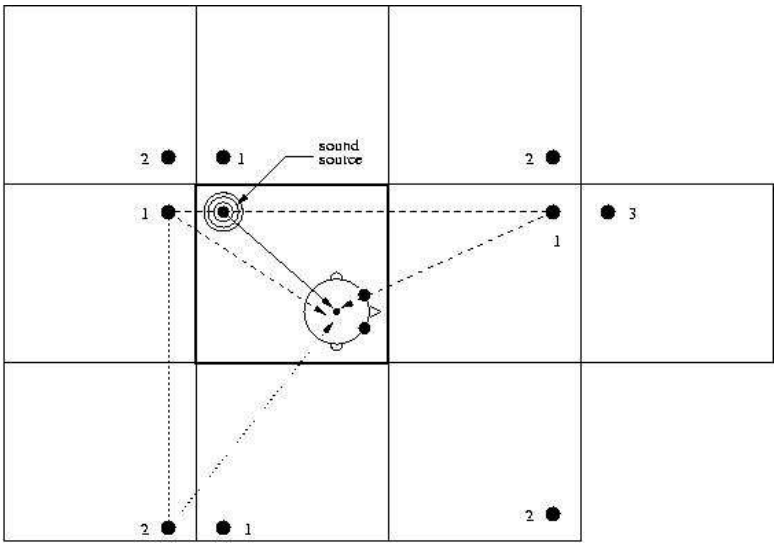


Figure 3.8: Image source method to determine low order specular reflections. The bold outlined rectangle represents the actual room with the listener and sound source. First order reflections are created by mirroring the sound source once (labeled with a "1"). Multiple order reflections are created by mirroring the first order reflections (labeled with a 2) and so on....

Image Source Method

The image source method [2] is used to determine the path followed by low order specular reflections (e.g. a reflection in which the angle 7of reflection is equal to the angle of incidence). A virtual sound source "copy" S_i of the original source S is created at a position obtained by *mirroring* the original sound source over each polygonal surface of a room [50]. Reflections up to any order can be produced by recursively repeating this procedure. A graphical illustration of the image source method is illustrated in Figure 3.8. For example, referring to Figure 3.8, after creating the first virtual sound source S_1 by mirroring the original source, a second order reflection can be created by treating S_1 as an "original" source and then mirroring it to create another virtual source S_2 and so on. For each virtual source, a "visibility" check is made to determine whether the virtual source is "visible" to the listener (the visibility check may be complex depending on the room being simulated). If the source is "visible" to the listener, it can be adjusted to account for the $1/r^2$ reduction of intensity of a propagating sound, absorption of the wave energy by the medium of propagation (e.g. air etc.) and added to the spatialization algorithm being used (e.g. it may be HRTF processed etc.). This of course will require maintaining and possibly updating information related to each virtual source, such as source distance and the position (elevation and azimuth) relative to the listener.

Changes in the environment (room) which may occur due to a variety of reasons, including movement of the original source or listener or the introduction of any objects/obstructions in the environment (room) may require the re-computation of all image sources as their visibility relative to the listener may change (e.g. one or more image sources which were previously visible may now become occluded and vice versa). If the listener or sound source are only rotated, then the visibility will not change and only azimuth and elevation angles may potentially need to be updated.

Although the image source method can find all specular reflections up to a certain order, it does have its shortcomings. Most importantly, as described by Funkhouser et. al. [50], it can only model specular reflections and its computational complexity is exponential with respect to the order of reflections (e.g. $O(n^r)$ virtual sources are created for a room with n surfaces with r reflections [50]). Given the potentially complex visibility checks which must be performed, the number of image sources which can be calculated is dependent on the processing power available.

Ray Tracing

As with the image source method, the ray tracing methods (see [69, 77]) find the paths between a sound source S and the listener. However, rather than mirroring the source, as shown in Figure 3.9, "rays" are emitted from the source in all directions and followed through the environment until some pre-defined number of them reach the listener. On their path from the sound source to the listener, the rays may encounter any number of surfaces (e.g. walls) or obstacles/obstructions. At this point, the rays are reflected once again (specular reflections are typically assumed, although diffuse reflection, diffraction and refraction can also be modeled). As with the image source method, the intensity of each reflection is reduced following the $1/r^2$ rule (or some variant of it), absorption of the wave energy by the medium of propagation (e.g. air), and the object it encounters.

As mentioned, rays are emitted from the source in all directions. In practise however, this is rarely the case. Having rays emitted from the source in all directions is clearly impractical computationally as it will lead to a large number of reflections which must be followed. Rather, a subset of rays is emitted instead. Various methods can be used to choose this subset, including Monte Carlo techniques which choose the paths followed by the rays randomly [50]. Ray tracing methods are well known and are widely used in computer graphics applications to render scenes, however, the ray tracing method has its share of advantages and disadvantages. Advantages include simplicity and manageable computational complexity, which increases sub-linearly with respect to the number of surfaces in the environment [50]. With respect to disadvantages however, given that a subset of the actual paths from the source to the listener are actually followed, certain paths may

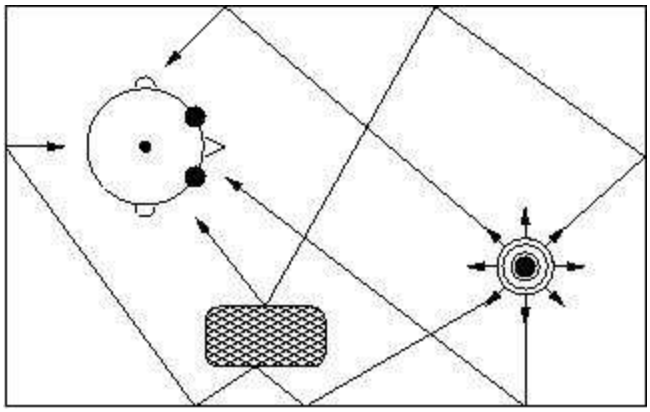


Figure 3.9: Ray tracing to determine the reflection paths of the sound waves traveling from the sound source to the listener.

be missed altogether. Furthermore, since only a portion of the actual paths will be sampled, aliasing and incorrect BRIRs will result and certain paths may be missed altogether [51, 118]. Although the number of paths sampled can be increased, doing so will lead to increased computational demands. Finally, these methods are typically used only when the receiver (listener) remains static. To allow for a moving receiver, the paths followed by the sound waves would need to be re-computed and depending on how fast the movements may be, once again, this would be computationally impractical.

3.5 Distance Simulation

This section examines the reproduction of the sound source distance cues presented in Chapter 1, along with any potential problems associated with their reproduction. The distance cues include intensity (loudness), reverberation (ratio of direct-to-reverberant sound levels reaching the listener), sound source spectral content and binaural cues. Since loudness and reverberation are the two most prominent distance cues and the simulation of reverberation was discussed in Section 3.4, emphasis will be placed on the simulation of loudness.

3.5.1 Loudness as a Distance Cue

Intensity (sound level), is an exocentric, relative distance cue. We don't necessarily need to know the distance to the original (reference) sound source position to make use of this cue.

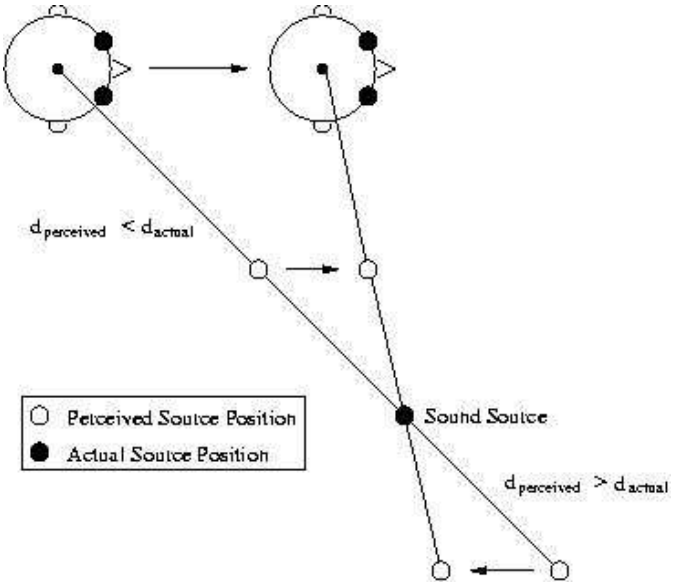


Figure 3.10: Potential problems arising from incorrect sound source distance estimation. Incorrect source distance judgments may lead to erroneous perception of a moving sound source when the listener is moving forward. Reprinted from [96].

This cue is rather simple to implement in a 3D sound system (e.g. simply scale the output presented to the loudspeakers/headphones by the inverse squared source distance) and is certainly intuitive. However, the inverse square reduction in intensity assumes a spherical head without any pinnae and an anechoic environment. Since our head is not a perfect sphere and our world is (generally) not anechoic, the inverse square law is not completely accurate. Although there is definitely an inverse squared relationship between sound source distance and sound intensity reaching the listener, there are other factors that influence the intensity of a sound reaching a listener in a "real world environment" and hence affect the $1/r^2$ loss model. Given these considerations, in a virtual environment, it may not necessarily suffice to simply use the inverse relationship between source distance and sound intensity described by Equation 1.11, as it may lead to errors. Speigle and Loomis [143] demonstrated that incorrect distance perception may lead to changes in appearance even when the correct directional information (e.g. azimuth and elevation of the source) is available. For example, referring to Figure 3.10 [96], consider a sound source at a distance d_{actual} from a listener (observer). Assume now that the perceived distance to the sound source $d_{perceived}$ is less than the actual distance. If the observer starts moving forwards, they may perceive the sound source as moving away from them. Similarly, if the perceived distance is greater than the actual distance, as the listener walks towards the source, they may perceive the source as moving towards them.

Although the inverse square law relates the intensity of sound waves to source distance, we perceive intensity as *loudness* [16, 169]. According to Moore [102], "loudness is defined as that attribute of auditory sensation in terms of which sounds can be ordered on a scale extending from quiet to loud". It is a quantity of auditory sensation corresponding most closely to the physical measure of sound [111]. Loudness is a subjective measure and therefore cannot be measured directly. In addition, it may not always be an accurate representation of intensity [102] as the loudness of pure tone sounds is frequency and bandwidth dependent [49, 120, 103].

Various studies examining loudness have been performed in order to understand and determine the relationship between loudness and intensity. The following sections provide greater details regarding the findings of these studies and the implications they may pose for a 3D sound system employing loudness cues to convey distance information.

Loudness Studies

"Loudness matching" experiments, where the listeners adjust the intensity of a pure tone so that it sounds as loud as a reference pure tone, can be used to create *equal loudness contours* [49, 120], describing the dependence of the loudness of pure tones. The measure of loudness level for a tone of any frequency t_f, is given in *phons*, and describes the sound level (in dB SPL), required for a 1000Hz reference tone t_{ref} to sound equally as loud. The equal loudness contours for loudness levels of 10 to 110 phons, as measured by Robinson and Dadson [120] are illustrated in Figure 3.11. As shown, generally, the lower frequency tones (e.g. below 1000Hz) are not as loud as the higher frequency tones, especially for smaller phon levels. For example, consider the 10 phone curve in Figure 3.11. As shown, the intensity of a 100Hz tone must be increased by about 40dB in order to sound equally loud as a 1000Hz tone. Also included with the equal loudness contours is the MAF curve, which describes the minimum audible threshold (e.g. below this level, the tone cannot be heard). As shown, the contours for all phon levels are similar in shape to the MAF however, as the phon level is increased, the curves become less steep.

The equal loudness contours illustrate the relationship between frequency and loudness of pure tones. However, there is no single equation which can describe the function between them [102]. The function has been approximated in various applications. For example, as described in [102], sound level meters provide an approximate measure of the intensity of complex sounds and have been designed to account for the equal loudness contours. These meters contain weighting networks, which provide a weight to the intensity of each component frequency according to the equal loudness contours. At low sound levels, the intensity of the higher frequency components contribute more to the overall sound level. As a result, a smaller weight is assigned to the intensity of the lower frequency components.

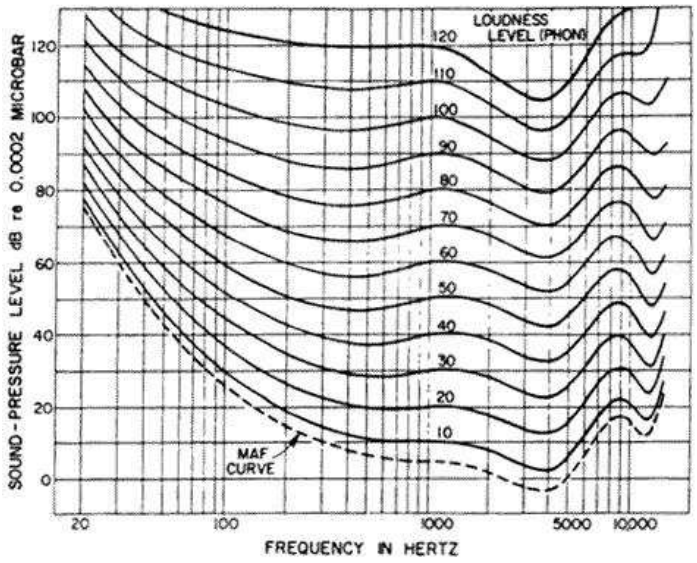


Figure 3.11: Robinson and Dadson free-field equal loudness contours. Reprinted from [153].

In addition to loudness matching experiments, *magnitude estimation* (where a number is assigned to sounds of different intensities) and *magnitude production* (a listener is given a number and must then adjust the intensity of a sound so that it matches the number), studies have given way to the development of "loudness scales" [102]. According to Stevens [145, 146], the loudness of a pure tone can be given according as follows:

$$L_t = kI_t^{0.33} \tag{3.9}$$

where I_t is the intensity of the pure tone, k is a constant which depends on the listener and on the units used, and L_t is loudness, measured in *sones*. In the sone scale (introduced by Stevens), one sone is defined as the loudness of a 1000Hz tone at 40dB SPL (sound pressure level) and loudness levels are given relative to it. For example, a sound with a loudness of two sones is twice as loud as the 1000Hz tone at 40dB. In this scale, a doubling of the source distance will result in a loudness decrease of 10dB as opposed to 6dB predicted by the inverse square law.

Finally, although loudness can be used as an effective cue to source distance estimation, when used alone, there is evidence suggesting the perceived distance is under estimated and may be insufficient (see [169]). Greater details regarding how the other distance cues (e.g.

reverberation, absorption of the sound by the medium) affect source distance estimation and how they affect the $1/r^2$ inverse square relationship between source distance and intensity (loudness), are described in the following sections.

3.5.2 Reverberation as a Distance Cue

Reverberation can used to provide absolute source distance estimation independent of overall sound source intensity [138, 25], due to the variation of the direct-to-reverberant sound energy level as a function of source distance [32, 155, 138, 29, 16, 108, 19]. In particular, as the source distance is increased, the ratio between the direct-to-reverberant levels

$$L_{ratio} = \frac{L_{direct}}{L_{reverb}} \tag{3.10}$$

will decrease. Referring to the definitions presented in Section 1.2.3, when the direct distance to the sound source d_{direct}, is less than the reverberant distance (e.g. $d_{direct} < d_{reverb}$) the intensity (the perceptual equivalent of intensity is *loudness*), of the direct sound will be greater than that of the reverberant sound (e.g. $L_{direct} > L_{reverb} \Rightarrow L_{ratio} > 1$). In contrast, when the reverberant distance is greater than the direct source distance ($d_{reverb} > d_{direct}$) the intensity of the reverberant sound will dominate (e.g. $L_{reverb} > L_{direct} \Rightarrow L_{ratio} < 1$). As described in Section 3.4, several methods are available to allow for the incorporation of reverberation cues into a 3D sound system. With an accurate reverberation model in place, the ratio direct-to-reverberation levels should be accounted for.

Although the effects of reverberation on source distance, have been known for some time [119, 144] according to Carlile [25], von Bekesy [155] was the first to demonstrate the affect of the ratio of direct-to-reverberant levels on the perception of source distance. However, in contrast to many other more recent studies, von Bekesy did not believe the ratio of direct-to-reverberant intensity levels represented a true perception in source distance, but rather, source distance was determined by other cues. He based this belief on the fact that in a "free space" (e.g. very large room or large open space) or in an anechoic environment, "the sensation of distance is even more distinct and of much greater extensiveness than elsewhere" [155]. Many studies performed after this have however demonstrated the effectiveness of reverberation (and in particular, the ratio between direct-to-reverberant intensity levels) as a cue to absolute source distance. Mershon has performed various studies examining the effect of the direct-to-reverberant ratio in source distance estimation [99, 98, 97]. These studies provide evidence that distance judgments are more accurate in the presence of reverberation than in an anechoic environment and as previously described,

that the ratio of direct-to-revereberant intensity is a cue to absolute source distance judgment.

Reverberation may be altered drastically with small changes to the objects in the environment themselves, changes in their positions, changes to the medium the sound is propagating in (typically air) or with the introduction of new objects in the environment. Although in general, the ratio between direct and reverberant sound decrease/increase as the source distance is increased/decreased, this may not necessarily always be true. Furthermore, although evidence indicates that reverberation does provide a cue to absolute source distance, studies also indicate reverberation can have negative affects as well. In particular, it leads to a decrease in directional localization accuracy in both real and virtual environments [16], and although this effect is of small magnitude, it is nevertheless measurable [139].

3.5.3 Source Spectral Content as a Distance Cue

As sounds travel through air, the sound waves are attenuated due to absorption by the medium itself, with higher frequencies being attenuated more. This attenuation of the high frequency components is also a function of source distance, where it manifests itself as an increasing low pass filter as the source distance is increased [74], providing a relative cue to source distance judgment whereby sounds with attenuated higher frequency components sound farther away [32, 170, 19, 159]. This high frequency attenuation cue is of particular importance for larger distances, greater than 15m and provides little information when the distance is small [19]. Ingard [75] however states that this attenuation is rather small, with a three to four dB loss for a 4kHz wave every 100m of propagation. Similarly, Begault [16] also states that this attenuation cue is rather weak when compared to the other cues (loudness, familiarity and reverberation). He describes how it may be difficult for a user of an auditory display to establish any "reference spectra" for this cue given the dynamic nature of the sound source location and spectra. In addition, in an indoor environment, the sound spectra reaching the listener's ear will also be fluctuating due to heating and air conditioning systems leading to further complications. This view is also shared by Brungart [19], who believes that although evidence suggests that a simple low pass filtering of a sound, where the cut-off frequency is inversely proportional to the distance, does increase the perceived source distance, the usefulness of this cue is limited given the large variation seen amongst subjects using this cue.

Bass et. al. have provided analytical expressions to predict the absorption of sound in air as a function of humidity, temperature, frequency and distance [6] which have been standardized by the ISO [74]. As presented by Bass et. al. and also given in [74], the attenuation $\alpha(f, T, h, p_a)$ (in dB per meter) of sound waves traveling through air at a

frequency "f", temperature "T" (in Kelvin), ambient sound pressure amplitude "p_a" and molar concentration of water vapor "h", can be described as follows:

$$\begin{aligned}\alpha(f,T,h,p_a) &= 8.686f^2\left(\left[1.84\times10^{-11}\left(\frac{p_a}{p_r}\right)^{-1}\left(\frac{T}{T_0}\right)^{\frac{1}{2}}\right]+\left(\frac{T}{T_0}\right)^{-\frac{5}{2}}\times\right.\\ &\quad\left(0.01275e^{-\frac{3352.0}{T}}\left[f_{ro}+\left(\frac{f^2}{f_{ro}}\right)\right]^{-1}+\right.\\ &\quad\left.\left.0.1068e^{-\frac{3352.0}{T}}\left[f_{rn}+\left(\frac{f^2}{f_{rn}}\right)\right]^{-1}\right)\right)\end{aligned}\tag{3.11}$$

where, $T_0 = 293.15$K is the reference air temperature in Kelvins and $p_r = 101.325$kPa is the reference ambient atmospheric pressure. The quantities f_{ro} and f_{rn} are the oxygen and nitrogen relaxation frequencies respectively and are given as follows:

$$f_{ro} = \frac{p_a}{p_r}\left(24+4.04\times10^4h\frac{0.02+h}{0.391+h}\right)\tag{3.12}$$

$$f_{rn} = \frac{p_a}{p_r}\left(\frac{T}{T_0}^{-\frac{1}{2}}\right)\left(9+280he^{-4.170\left[\left(\frac{T}{T_0}^{-\frac{1}{3}}-1\right)\right]}\right)\tag{3.13}$$

The above equations were realized by Huopaniemi using IIR filters. Figure 3.12 (as it appears in [74]), provides a graphical illustration of the magnitude response of the absorption of sound in air as a function of frequency for a temperature of 20^C with a humidity (e.g. water vapor concentration), of 20% ($h = 0.4615$) for several sound source distances (1m, 10m, 20m, 30m, 40m and 50m). The absorption of higher frequency components is clearly evident and cannot be ignored, especially when considering larger sound source distances [74].

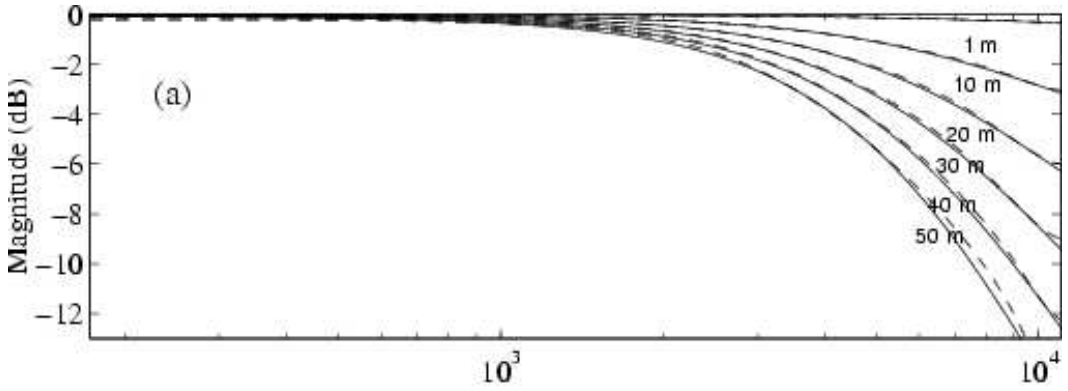


Figure 3.12: Graphical illustration of the absorption of sound in air as a function of frequency for a temperature of 20^C with a humidity (e.g. water vapor concentration), of 20% ($h = 0.4615$) for several sound source distances (1m, 10m, 20m, 30m, 40m and 50m). Dashed line represents IIR filter response while continuous line represents them ideal response. Reprinted form [74].

3.5.4 Binaural Cues

As described in Section 1.2.6, for near field sound sources, the ILD is highly dependent on source distance. However, as Brungart states [22], it may be unnecessary to include binaural distance cues in a virtual auditory display given these cues may be insignificant relative to the other distance cues such as reverberation and loudness. Furthermore, as expressed by Blauert [17] and described in [16], given the numerous conflicting data in the literature, the effect of binaural cues on source distance remains an unresolved issue.

3.5.5 Sound Source Familiarity

As previously described, familiarity of the sound source does improve source distance accuracy and localization in general. It therefore seems intuitive that the user is familiar with the sounds and the environmental context associated with them, which are presented in a virtual auditory environment [136]. The importance of sound source familiarity and its effect on a virtual auditory display is best summarized by Begault [16], who states "any reasonable implementation of distance cues into a 3D sound system will probably require an assessment of the cognitive associations for a given sound source". Unfortunately, it may be difficult to determine exactly which sounds are familiar to each user of the display, as this varies depending on each users prior experience. In addition, it may also be impractical to allow a user to become familiar with the characteristics of a sound source through repeated use of the auditory system. This will certainly limits the ability of the

auditory system to process arbitrary input stimulus in real time [19].

Chapter 4

Conveying Sound in a Virtual Environment

In an auditory display the audio output is conveyed to the listener either through loudspeakers or through headphones worn by the listener. Both headphones and loudspeakers each have their advantages as well as shortcomings and one or the other may produce more favorable results depending on the application. This chapter examines issues associated with headphone and loudspeaker displays. In doing so, two loudspeaker based 3D audio techniques known as *transaural audio* and *amplitude panning* will also be described.

4.1 Headphone Listening

Many 3D audio systems employing binaural techniques are "headphone based", conveying sounds to the users over headphones as opposed to loudspeakers. Headphone based systems offer several advantages over loudspeaker based systems. In particular, headphones provide a high level of channel separation thereby minimizing any *crosstalk*, arising when the signal intended for the left (or right) ear is also heard by the right (or left) ear. Headphones also isolate the listener from external sounds and reverberation which may be present in the environment [55], ensuring the acoustics of the listening room or the listener's position in the room, do not affect the listener's perception [74]. These factors make headphones the only means of delivering audio in various auditory displays, including displays intended for aircraft cockpits or multiple users [89], where loudspeakers are impractical and cannot be used.

Despite the potential benefits headphone based systems offer, they certainly do have their shortcomings and limitations as well. According to Kyriakakis [89], the four major

drawbacks of headphone based systems are as follows:

Use of Non-individualized HRTFs: The filtering performed by each person's HRTFs may differ significantly. However, as described in Section 3.3.2, for practical purposes, the dataset of HRTFs used in an auditory display is typically not obtained from the individual user but rather a non-individualized ("generic") set of measurements is used instead. The differences between the individual's and the generic HRTFs can lead to errors. The problems associated with the use of non-individualized HRTFs are not unique to headphone based systems and may also be present when using loudspeaker based (transaural) techniques.

Ambiguous Cues: Ambiguous situations arising when the sound source is positioned on the median plane or directly above or below the listener. The interaural cues are (nearly) zero in such a situation, leading to confusion between a sound source directly in front or directly behind a listener. The inclusion of individualized HRTF information helps to reduce such ambiguities. Similarly to the use of non-individualized HRTFs, ambiguous cues may also arise when using loudspeaker based systems.

Comfort Level: Headphones may be uncomfortable to wear and can be cumbersome.

Inside-the-head Localization (IHL): Sounds are not *externalized* (e.g. appear to be emanating beyond the listener) but rather, appear as if they are originating from inside the head.

Ambiguous cues (e.g. front-back reversals and the cone of confusion as introduced in Section 1.2.1), will arise, regardless of whether the system is headphone or loudspeaker based, when HRTF cues are not employed. As previously described, ambiguous cues can be greatly reduced, and hence system performance improved, with the incorporation of HRTF information. However, errors resulting from the use of non-individualized HRTFs may offset any improvements.

Greater details regarding the use of HRTFs (including non-individualized HRTFs) and the ambiguous cues present when relying solely on the duplex theory of sound localization were provided in Sections 1.2.2 and Section 1.2.1 respectively. The following sections elaborate further on the problems of comfortability and inside-the-head localization with respect to headphone based auditory systems.

4.1.1 Headphones and Comfort

In many situations, it may be inconvenient and impractical for the listener to use headphones. The use of headphones may limit a user's immersion with a virtual environment.

For example, in a six-sided virtual environment such as York University's IVY [121], the goal is to totally immerse the user in the virtual world. This is accomplished by surrounding the user with realistic visual imagery projected on the four walls, ceiling and floor and also providing them with the corresponding spatial auditory cues to accompany the visual information. However, the physical presence of the headphones over the listener's ears is a constant reminder that they are in a virtual environment.

After wearing headphones for an extended period of time, they can become very uncomfortable [89]. Furthermore, small movements of the headphones themselves, while being worn by the listener (which, may result if the listener repositions them over the ears), may affect the HRTF considerably [135] as it can change the position of the sound source relative to the listener.

4.1.2 Inside-the-Head Localization

Inside-the-head localization (IHL) refers to the lack of externalization of a sound source, resulting in the false impression that the sound is originating from inside the listener's head and can only move left and right inside the head along the interaural axis, being biased towards the rear of the head [82]. This is actually the main drawback associated with headphone displays and other than this problem, according to Begault [16], headphone displays are actually superior for conveying 3D audio. Although rare, IHL can also occur when listening to "external" sound sources in the real world, especially when the sounds are unfamiliar to the listener or when the sounds are obtained (recorded) in an anechoic environment [30].

IHL results from various factors including the lack of correct environmental context (e.g. lack of reverberation and HRTF information). IHL can be greatly reduced, if not eliminated, by ensuring the sounds delivered to the listener's ears reproduce the sound as it would be heard naturally or in other words, providing the listener with a "realistic spectral profile of the sound at each ear" [135]. Delivering the correct spectral profile of the sound to the ears is of course a difficult task and ultimately the goal of any 3D audio display. In any case, it will involve incorporating HRTF information into the auditory display. Although the externalization of a sound source is difficult to predict precisely, it does increase as the sound becomes more "natural" and contains localization cues, including individualized HRTFs, and reverberation cues, which are updated appropriately with any head movements, as they are in "normal" listening situations [16]. Of course, each of the cues mentioned above has its share of problems. As described in Section 3.3.2, the inclusion of individualized HRTFs is usually impractical given the difficult and time consuming task of measuring the individual's HRTFs. Although non-individualized or generalized HRTFs can be used instead, they result in reduced performance and listener localization accuracy.

Similarly, various methods exist to allow the inclusion of reverberation cues. However, as described in Section 3.4, these methods are certainly far from perfect and have their share of troubles including the fact that they are computationally expensive. In addition, as demonstrated in a study performed by Begault [13], although the addition of reverberation basically eliminated IHL, it resulted in a decrease in user localization accuracy. Head movements can also aid in the externalization of a sound source [30]. However, as with the inclusion of HRTFs and reverberation information, this is also a difficult task. It requires some method of tracking the position (and possibly orientation) of the user's head in order to account for any head movements which may require the updating of the sound source location.

Finally, as a final note regarding the use of HRTF based headphone systems, when the headphones are placed over the ears, an "acoustic cavity" forms between the the headphone and the ear (separate cavity for the left and right ears), resulting in an additional transfer function which can lead to further performance reductions [51]. As a result, the response of the headphones should also be measured and accounted for using the equalization methods described in Section 3.3.4.

4.2 Loudspeaker Displays

In the following sections, the two most common loudspeaker based 3D audio techniques will be introduced. The first method, *transaural audio*, allows for the presentation of binaural audio over loudspeakers as opposed to headphones. Although it overcomes many of the problems encountered when binaural audio is presented over headphones, as will be discussed, it also has its share of problems as well. The other technique to be discussed will be *amplitude panning*, where the desired spatial audio effect is achieved by scaling the intensity (amplitude) of up to N loudspeakers by some pre-defined weighting factor.

As with the recording techniques introduced in Chapter 2, the intended effect produced by each of the techniques described in this section is restricted to a small region of space.. In other words, these techniques also assume a listener sweet spot and deviation from this region will lead to serious degradations in system performance.

4.2.1 Transaural Audio

The presentation of the left and right binaural audio signals to the corresponding left and right ear using stereo loudspeakers is known as *transaural audio* [26]. Transaural audio can overcome some of the limitations inherent with headphone based binaural audio, such

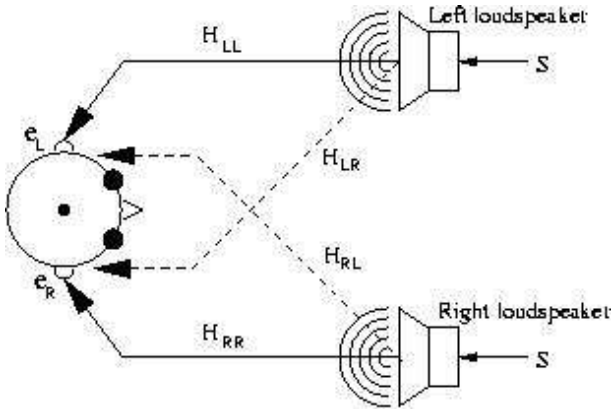


Figure 4.1: Crosstalk Defined. When using loudspeakers as opposed to headphones to convey sound to a user of a 3D sound system, in addition to the desired left loudspeaker signal H_{LL} reaching the left ear e_L, a delayed and attenuated portion of the right loudspeaker signal H_{RL} will also reach the left ear. A similar situation occurs with the signal reaching the right ear e_R, where in addition to the desired signal H_{RR} from the right loudspeaker, a delayed and attenuated portion of the left loudspeaker H_{LR} will also reach the right ear.

as IHL, however, transaural audio has its share of problems as well. Most importantly, the *crosstalk* signal arriving at each ear from the opposite loudspeaker must be removed. Consider the two-channel stereo loudspeaker set-up illustrated in Figure 4.1, where the listener is symmetrically placed between the left and right loudspeakers. In a virtual auditory display, the signal emitted from the left and right loudspeakers must be delivered to the left and right ears respectively only. However, as illustrated, this is certainly not the case. In a typical two loudspeaker (stereo) scenario, the signal received at the left and right ears (e_L and e_R respectively), is a linear combination of the signal output by the left and right loudspeakers, including any filtering effects introduced by the loudspeakers and the environment (e.g. the speaker frequency response, absorption of sound by the medium and head response) [55]. Ideally, the signal emitted by the left (right) loudspeaker should reach the left (right) ear only, in isolation. However, in addition to the desired signal coming from the left and right loudspeakers H_{LL} and H_{RR} respectively, a delayed and attenuated portion of the left loudspeaker signal H_{LR} will reach the right ear while a delayed and attenuated portion of the right loudspeaker signal H_{RL} will reach the left ear. This delayed signal reaching the left (right) ear from the right (left) loudspeakers is known as *crosstalk* and can greatly affect the "spectral balance" and interaural differences (ITD and ILD) [16], thereby limiting the effectiveness of a loudspeaker based system. Crosstalk should therefore be minimized or, ideally eliminated. The unwanted crosstalk signals can be removed using a technique known as crosstalk cancellation. Greater details regarding crosstalk cancellation are provided in the following section.

Crosstalk Cancellation

Crosstalk cancellation was first proposed by Bauer in 1961 [11] in order to allow for the delivery of binaural audio (see Section 2.5) using a pair of loudspeakers. Two years later, the first crosstalk canceller was actually implemented by Atal and Schroeder [4] in order to allow binaural recordings made in concert halls to be played back over loudspeakers [74]. Essentially, the basic idea behind the Atal and Schroeder crosstalk canceller involves adding a delayed and inverted version of the crosstalk signal to the opposite loudspeaker output. A delayed and inverted version of the crosstalk signal going from the right loudspeaker to the left ear H_{RL} would be added to the left loudspeaker output, while a delayed and inverted version of the crosstalk signal going from the left loudspeaker to the right ear H_{LR} would be added to the right loudspeaker output. Given that the inverted signals are 180^0 out of phase and delayed, theoretically, if the delay is chosen such that it equals the amount of time it takes for the crosstalk signal to reach the opposite ear, the crosstalk will be completely cancelled [16]. A frequency domain, matrix solution to the crosstalk cancellation method as proposed by Atal and Schroeder following the notation of Mouchtaris et. al. [105] is provided below. This solution assumes the listener is symmetrically positioned between the two loudspeakers (e.g. the listener and the two loudspeakers form an equilateral triangle) and a spherical head model without any external ears.

With a typical headphone based binaural audio system, in order to spatialize a sound to some particular location, the signal received at the left and right ears (e_L and e_R respectively), is obtained by processing (e.g. convolution in the time domain and multiplication in the frequency domain) a monaural sound with the measured HRTF of the left and right ear corresponding to the desired synthesis location (e.g. position of the virtual sound source), In matrix notation, the signal presented to the left and right ears is given as follows:

$$E = H_{hrtf} S$$

where,

$$E = \begin{bmatrix} e_L \\ e_R \end{bmatrix}, \qquad H_{hrtf} = \begin{bmatrix} H_L & 0 \\ 0 & H_R \end{bmatrix}, \qquad S = \begin{bmatrix} s \\ s \end{bmatrix} \tag{4.1}$$

such that S is the column vector representing the monaural (non-synthesized), sound source s (prior to filtering with the appropriate HRTF response, s is identical for both ears) and H_{hrtf} is the matrix containing the left and right ear HRTFs H_L and H_R respectively, corresponding to the desired synthesis location, which have been equalized to remove the response of the measurement system.

When using loudspeakers as opposed to headphones, a similar situation arises, a monau-

ral signal is processed with a pair of HRTFs (H_L and H_R) corresponding to desired synthesis location in order to obtain the binaural signals for the left and right ears. The left and right processed signals are then delivered to the listener through a pair of corresponding loudspeakers. However, as previously described, due to crosstalk, the left and right loudspeaker signals are not delivered exclusively to the left and right ears respectively. Rather, as shown in Figure 4.1 and as previously described, in addition to the signal from the right (left) loudspeaker reaching the right (left), a delayed and attenuated version of the left (right) loudspeaker signal will be delivered to the right (left) ear (e.g. crosstalk). In matrix notation, this can be represented a follows:

$$E = H_{tf} H_{hrtf} S \tag{4.2}$$

where,

$$H_{tf} = \begin{bmatrix} H_{LL} & H_{LR} \\ H_{RL} & H_{RR} \end{bmatrix} \tag{4.3}$$

is the 2×2 "acoustic transfer matrix" representing the transfer function from the loudspeakers to the two ears. The terms H_{LL} and H_{RR} are known as the *ipsilateral* terms and describe the transfer function from the left and right loudspeakers to the left and right ears respectively. H_{LR} and H_{RL} are known as the *contralateral* terms and describe the transfer function from the left and right loudspeakers to the right and left ears respectively (in other words, H_{LR} and H_{RL} are the crosstalk signals).

The desired output for a transaural system does not include any crosstalk but is rather the delivery of the left and right binaural (HRTF processed) signals to the corresponding ears. This can be accomplished by eliminating the acoustic transfer functions described by matrix H_{tf} using the "crosstalk canceller" C. Matrix C is essentially the inverse of H_{tf} (e.g. $C = H_{tf}^{-1}$), leading to the following:

$$\begin{bmatrix} e_L \\ e_R \end{bmatrix} = \begin{bmatrix} H_{LL} & H_{LR} \\ H_{RL} & H_{RR} \end{bmatrix} \begin{bmatrix} H_{LL} & H_{LR} \\ H_{RL} & H_{RR} \end{bmatrix}^{-1} \begin{bmatrix} H_L & 0 \\ 0 & H_R \end{bmatrix} \begin{bmatrix} s \\ s \end{bmatrix} \tag{4.4}$$

The inverse matrix C can be computed as:

$$C^{-1} = \frac{1}{H_{LL}H_{RR} - H_{RL}H_{LR}} \begin{bmatrix} H_{LL} & -H_{LR} \\ -HRL & H_{RR} \end{bmatrix} \tag{4.5}$$

where,

$$\frac{1}{H_{LL}H_{RR} - H_{RL}H_{LR}} \tag{4.6}$$

is the determinant of the matrix H_{tf}. Of course, C^{-1} is undefined when the determinant is equal to zero (e.g. matrix H_{tf} is singular) and therefore, in such a situation, the inverse matrix cannot be calculated.

In order to realize the crosstalk canceller as described in Equation 4.4, four transfer functions (H_{LL}, H_{RR}, H_{LR} and H_{RL}), must be obtained. However, when originally introduced, for simplicity, the listener's position was assumed to be symmetrical between the two loudspeakers (e.g. such that the listener and the two loudspeakers form an equilateral triangle) and furthermore, that the listener's head was also a perfect sphere. These two assumptions ensure the ipsilateral transfer functions for both the left and right loudspeakers are identical (e.g. transfer function from the left loudspeaker to the left ear is the same as the transfer function from the right loudspeaker to the right ear) and for simplicity, denoted by H_i. Furthermore, the two assumptions further result in identical contralateral transfer functions for the left and right loudspeakers (e.g. the transfer function from the left loudspeaker to the right ear is the same as the transfer function from the right loudspeaker to the left ear), which are denoted by H_c. With this simplification, Equation 4.4 described above can be re-stated as:

$$\begin{bmatrix} e_L \\ e_R \end{bmatrix} = \begin{bmatrix} H_i & H_c \\ H_c & H_i \end{bmatrix} \begin{bmatrix} H_i & H_c \\ H_c & H_i \end{bmatrix}^{-1} \begin{bmatrix} H_L & 0 \\ 0 & H_R \end{bmatrix} \begin{bmatrix} s \\ s \end{bmatrix} \tag{4.7}$$

where the inverse matrix now becomes:

$$C^{-1} = \frac{1}{H_iH_i - H_{cc}H_c} \begin{bmatrix} H_i & -H_c \\ -Hc & H_i \end{bmatrix} \tag{4.8}$$

After substituting Equation 4.8 into Equation 4.7, the following expression is obtained:

$$\begin{bmatrix} e_L \\ e_R \end{bmatrix} = \begin{bmatrix} H_i & H_c \\ H_c & H_i \end{bmatrix} \frac{1}{H_iH_i - H_cH_c} \begin{bmatrix} H_i & -H_c \\ -H_c & H_i \end{bmatrix} \begin{bmatrix} H_L & 0 \\ 0 & H_R \end{bmatrix} \begin{bmatrix} s \\ s \end{bmatrix} \tag{4.9}$$

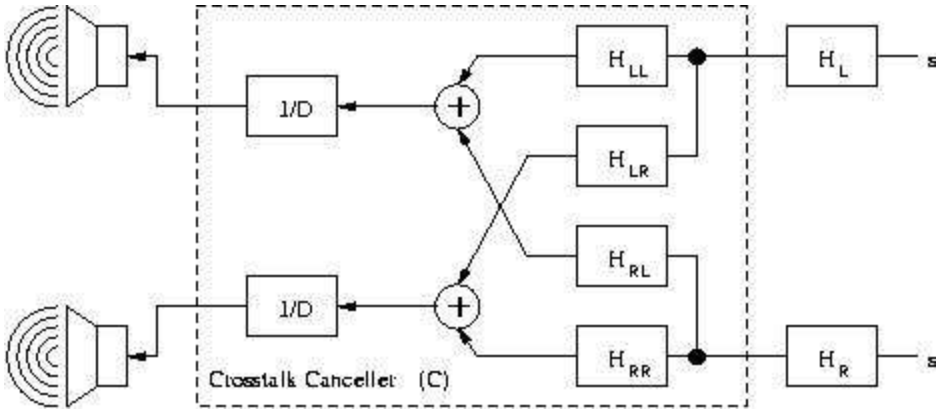


Figure 4.2: Atal and Schroeder crosstalk canceller, where $D = H_{LL}H_{RR} - H_{LR}H_{RL}$.

By further manipulating Equation 4.9, and making the assumption that the contralateral signals contain less power than the ipsilateral signals due to the shadowing effect of the head (a valid assumption as verified by Mouchtaris et. al. [105]), the left and right ear filters F_L and F_R respectively can be obtained as follows [105]:

$$F_L = \frac{H_L}{H_i} - \frac{H_c}{H_i}\frac{H_R}{H_i} \qquad F_R = \frac{H_R}{H_i} - \frac{H_c}{H_i}\frac{H_L}{H_i} \tag{4.10}$$

These two filters contain the desired HRTF response corresponding to the synthesis location as well as the necessary crosstalk cancellation components. By processing the monaural signal with these filters and delivering the resulting signals to the appropriate loudspeaker, the desired binaural audio effect described in Equation 4.1 will be obtained. A graphical illustration to the Atal and Schroeder crosstalk canceller is shown in Figure 4.2.

Finally, the crosstalk canceller solution as presented assumes a single listener. Crosstalk cancellation however can also be applied in a multi-listener scenario. Garas [53], provides a set of equations for a multi-listener crosstalk canceller in which N "program signals" are used to create M loudspeaker signals which in turn result in L ear signals. They also present solutions to these equations using algebraic methods.

Problems Associated with Crosstalk Cancellation

In theory, crosstalk cancellation completely removes the unwanted signals thereby allowing the desired binaural signals to be delivered to the corresponding ears. In practise, however, this is not the case. Given the use of HRTFs in the crosstalk canceller, its effectiveness

is limited by the variability in head size and shape of the human head and pinnae [58]. In addition, it has a small sweet spot and in order to function properly, the listener must remain stationary (e.g. no head movements) in the sweet spot [140] as movements as small as 74 - 100mm completely destroys the desired effect [105]. When the listener moves more than this allowable amount, the HRTFs used by the crosstalk canceller may be incorrect and the time required for the crosstalk signals to reach the contralateral ears and the attenuation factor may also change. As with headphone based systems, this problem can be overcome (or greatly reduced) by tracking the listener's head. Gardner [55] developed a system utilizing a magnetic tracker to dynamically obtain the position of the listeners head in order to produce a much more realistic and greater range 3D auditory display using loudspeakers. Given the dynamic updates of head movements, this system offers improved localization over existing, non-tracked loudspeaker displays as it allows for dynamic localization cues. Kyriakakis and Holman [105] also describe a loudspeaker based 3D audio display which allows for dynamic crosstalk cancellation. However, rather than using a magnetic tracker, they utilize a camera-based tracking system to track the listeners head and thereby eliminate the need of any tether required by the magnetic tracker. Furthermore, despite the theoretical generalization to N listeners as presented in the previous section, crosstalk cancellation for multiple listeners is an extremely complex and computationally expensive task to be of any practical use. As a result, crosstalk cancellation is typically restricted to a single user.

Finally, the topic of crosstalk cancellation is far more complex than presented here. This section simply provided the motivation and basic theory behind crosstalk cancellation. Greater details can be found in [33, 55, 53, 9].

4.2.2 Amplitude Panning

The difference in intensity between the sound reaching both ears forms the basis of the interaural level difference (ILD) cue and can be used by humans to localize a sound source. In the *amplitude panning* technique, the amplitude (intensity or output level) of the signal being delivered to each loudspeaker[1] is adjusted in some manner to simulate the directional properties of the ILD. In other words, by adjusting the amplitude of the signal applied to each loudspeaker through the use of a gain factor, the listener can perceive a phantom image (virtual source) emanating from some direction dependent on the gain factors [116]. Mathematically, amplitude panning can be described as:

$$b_i(t) = g_i(t)s_m(t), \qquad i = 1, \ldots, N \tag{4.11}$$

[1]Headphones can also be used when the number of loudspeakers is two.

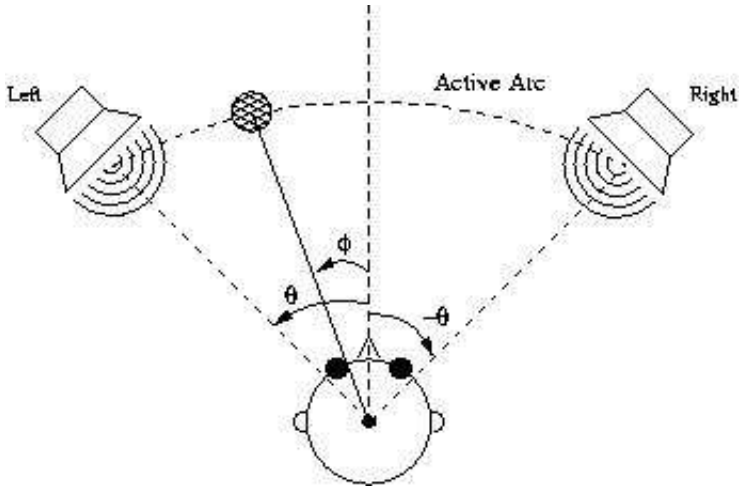


Figure 4.3: Stereo amplitude panning configuration.

where $b_i(t)$ is the signal output by loudspeaker i at time t, $s_m(t)$ is the "un-processed" sound applied to each of the loudspeakers at time t, g_i is the gain factor applied to the signal delivered to loudspeaker i, and finally, N is the total number of loudspeakers being used.

Various amplitude panning techniques exist which allow for a wide variety of loudspeaker configurations including both two and three dimensional configurations. Regardless of the technique used however, the general idea remains the same: compute the appropriate gain factors to create the impression of a virtual sound source at a specific position relative to the listener. Greater details regarding several of the panning techniques available for both two and three dimensional loudspeaker configurations are provided in the following sections, beginning with two-channel amplitude (stereo) panning, the most popular panning technique.

Two Dimensional Amplitude Panning

The typical two-channel (stereo) configuration is illustrated in Figure 4.3. The listener is placed symmetrically (in the horizontal plane) equidistant between the left and right loudspeakers, displaced by an angle of θ between each one (usually $\theta = 30^o$). A monaural sound s is applied to each of the loudspeakers. By scaling the amplitude of the signal applied to the left and right loudspeakers by the appropriate gain factors (g_l and g_r respectively), the virtual sound source can be positioned anywhere on the "active arc" (a semi circle between the two loudspeakers with radius equal to the distance between the listener and each of the loudspeakers [112]).

Several methods can be used to actually calculate the gain factors g_l and g_r. The most common technique is the *stereophonic law of sines*, demonstrated first by Blumlein and given by Bauer [10] as follows:

$$\frac{\sin\phi}{\sin\theta_o} = \frac{g_l - g_r}{g_l + g_r} \tag{4.12}$$

where, referring to Figure 4.3, ϕ is the azimuth (horizontal) angle between the listener and virtual sound source and θ is the angle between the listener and each of the loudspeakers.

Although the "stereophonic law of sines" can be used to place a sound source between the two loudspeakers, it has its limitations. It is valid for low frequency signals (e.g. below 500Hz) only, and when the listener is facing directly forward [116]. To account for head movements which may arise as the listener is tracking the virtual source, the *tangent law* introduced by Bennett [112], and given in Equation 4.13, may be used instead:

$$\frac{\tan\phi}{\tan\theta_o} = \frac{g_l - g_r}{g_l + g_r} \tag{4.13}$$

Equations 4.12 and 4.13 can be manipulated in order to determine the left and right gain values by assuming a constant virtual sound power level $C > 0$. This can be accomplished by ensuring the following:

$$g_l^2 + g_r^2 = C \tag{4.14}$$

Since the loudness level of the source can be a potential cue to source distance, keeping the virtual source power level constant ensures the perceived distance to the virtual source remains constant while it is being panned from one loudspeaker to the other along the active arc.

When the direction of the virtual sound source coincides with one of the loudspeakers (e.g. $\phi = \theta$), the sound will emanate from that particular loudspeaker only, producing accurate and correct results. Finally, although this law does produce accurate results when the virtual source is positioned at either of the loudspeakers (e.g. $\phi = \theta$) or in the center, directly in front of the listener (e.g. $\phi = 0^o$), it is not as accurate as when placing the virtual source between the center and either of the loudspeakers [63].

The two-channel stereo configuration can be extended to allow for the placement of one or more additional loudspeakers on the horizontal plane (the plane in which the two-channel stereo loudspeakers are placed) as done with the Dolby Stereo and Quadraphonics systems (see Sections 2.6.3 and 2.6.1 respectively). Amplitude panning can then be extended to account for the additional N loudspeakers, as done in the popular *pair-wise amplitude*

panning technique introduced by Chowning [28] which can produce sound sources in all azimuth directions given the use of a sufficient number of loudspeakers. In this technique, despite the availability of N channels (loudspeakers), sound is applied to two loudspeakers only, in a manner similar to the conventional two-channel stereo panning technique. Dolby Surround, Quadraphonics and two dimensional Ambisonics are examples of two dimensional panning techniques using greater than two channels (loudspeakers).

Three Dimensional Panning

The three dimensional panning technique is an extension of the two-channel, two dimensional technique. However, rather than having all loudspeakers at the same height (e.g. on the same plane as the listener's head), the height of some (or all) additional loudspeaker(s) will differ. In this configuration, all loudspeakers are positioned equidistant from the listener. In a manner similar to pairwise amplitude panning, sound is applied to a subset (three) of the loudspeakers only. A virtual sound source can be positioned anywhere on the triangle formed by the three loudspeakers [112, 115].

Currently, no general trigonometric method of three-dimensional amplitude panning for an arbitrary three-dimensional loudspeaker setup exists [116] and the calculation of the gains applied to the loudspeakers is very configuration dependent. However, as with the two-channel stereo configuration, the intensity (loudness) of the output sound heard by the listener can be kept at a constant level C by ensuring the following:

$$g_1^2 + g_2^2 + g_3^2 = C \tag{4.15}$$

where g_i is the gain applied to loudspeaker i.

Vector Base Amplitude Panning (VBAP)

The vector base amplitude panning technique (VBAP), introduced by Pulkki in 1996 [117] is an amplitude panning technique that can be used with an arbitrary number of loudspeakers. It supports two and three-dimensional loudspeaker configurations and allows the loudspeakers to be placed in any position provided they are "nearly" equidistant around the listener and that the listening room is not very reverberant [112].

VBAP can be applied to both two and three-dimensional loudspeaker configurations, including the traditional two-channel stereo setup and three channel, three dimensional setup. A formulation of the two-channel VBAP method as described by Pulkki in [112, 114, 115] and using the same notation, are presented in the following section.

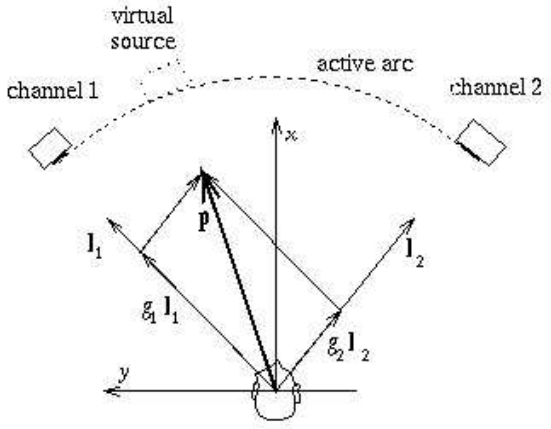


Figure 4.4: Vector base amplitude panning with a two-channel stereo configuration. Reprinted from [112].

Two-Dimensional Stereo Vector Base Amplitude Panning

In the stereo VBAP configuration, the two-channel stereo setup defines a two-dimensional vector base, with two unit length vectors $l_1 = [l_{11}\ l_{12}]^T$ and $l_2 = [l_{21}\ l_{22}]^T$, pointing to the left and right loudspeakers respectively as shown in Figure 4.4.

The unit length vector $p = [p_1\ p_2]^T$ which points in the direction of the virtual source can then be given as follows:

$$p = g_1 l_1 + g_2 l_2 \tag{4.16}$$

where g_1 and g_2 are the gain factors applied to the left and right loudspeaker respectively. In matrix form, it can be formulated as $p^T = gL_{12}$, or in other words,

$$p^T = gL_{12} \tag{4.17}$$

$$\begin{bmatrix} p_1 \\ p_2 \end{bmatrix} = \begin{bmatrix} l_{11} & l_{12} \\ l_{21} & l_{22} \end{bmatrix} \times \begin{bmatrix} g_1 \\ g_2 \end{bmatrix} \tag{4.18}$$

Assuming the inverse to matrix L_{12} exists, the gain factors can be solved for:

$$\begin{bmatrix} g_1 \\ g_2 \end{bmatrix} = \begin{bmatrix} l_{11} & l_{12} \\ l_{21} & l_{22} \end{bmatrix}^{-1} \times \begin{bmatrix} p_1 \\ p_2 \end{bmatrix} \tag{4.19}$$

Finally, the gain factors can be applied to the corresponding loudspeakers after they have been scaled with the following scaling factor g_{scaled}

$$g_{scaled} = \frac{\sqrt{C}g}{\sqrt{g_1^2 + g_2^2}} \tag{4.20}$$

VBAP can also be extended to allow an arbitrary number of loudspeakers in a two-dimensional configuration, where once again, the loudspeakers and the listener are on the same plane. As with the pairwise panning technique previously described, in the VBAP technique, two of the N loudspeakers are chosen and sound is applied to these two loudspeakers only. A complete discussion of how the two loudspeakers are actually chosen is given by Pulkki [112, 114].

Three-Dimensional Stereo Vector Base Amplitude Panning

Consider three non-coplanar loudspeakers equidistant to the listener as illustrated in Figure 4.5. Once again, using Pulkki's notation, let vector $l_i = [l_{i1}\ l_{i2}\ l_{i3}]^T$ be the unit vector from the origin (the center of the imaginary sphere on which the loudspeakers are placed) to the i^{th} loudspeaker and let $p = [p_1\ p_2\ p_3]^T$ be the unit vector pointing from the origin to the direction of the virtual sound source. The vector p can be given as a linear combination of the three unit vectors l_i $(i = 1,\ 2,\ 3)$ in matrix notation as $p = g_1 l_1 + g_2 l_2 + g_3 l_3$ where g_i is the gain applied to loudspeaker i. In other words, $p^T = gL_{123}$ where $L_{123} = [l_1\ l_2\ l_3]^T$ or

$$\begin{bmatrix} p_1 \\ p_2 \\ p_3 \end{bmatrix} = \begin{bmatrix} l_{11} & l_{12} & l_{13} \\ l_{21} & l_{22} & l_{23} \\ l_{31} & l_{32} & l_{33} \end{bmatrix} \times \begin{bmatrix} g_1 \\ g_2 \\ g_3 \end{bmatrix} \tag{4.21}$$

By rearranging the above equation, vector g can be solved for assuming L_{123}^{-1} exists (it does exist if the vector base defined by L_{123} defines a three-dimensional space), as $g = p^T L_{123}^{-1}$ or,

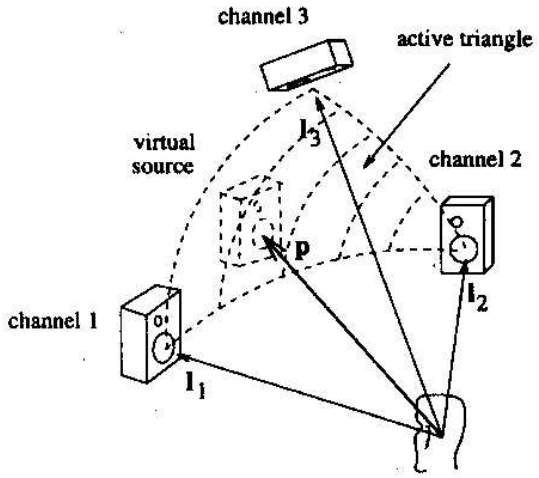


Figure 4.5: Vector base amplitude panning for a three-dimensional (three channel) configuration. Reprinted from [112].

$$\begin{bmatrix} g_1 \\ g_2 \\ g_3 \end{bmatrix} = \begin{bmatrix} l_{11} & l_{12} & l_{13} \\ l_{21} & l_{22} & l_{23} \\ l_{31} & l_{32} & l_{33} \end{bmatrix}^{-1} \times \begin{bmatrix} p_1 \\ p_2 \\ p_3 \end{bmatrix} \tag{4.22}$$

The gain factors may be applied to the corresponding loudspeakers after they have been scaled using the following scaling factor g_{scaled}:

$$g_{scaled} = \frac{\sqrt{C}g}{\sqrt{g_1^2 + g_2^2 + g_3^2}} \tag{4.23}$$

Finally, as in the two-dimensional case, the three loudspeaker VBAP technique can be extended to handle N loudspeakers. In this case, the sound will only be presented over three loudspeakers.

Chapter 5

Discussion

The previous chapters provide an overview on the field of 3D (spatial) sound as well as the underlying foundation on which it depends, the human auditory system. Various technologies available for the generation of 3D sound were presented, beginning with a historical description of some of the early techniques which did not necessarily produce "true" 3D sound. Such technologies include recording techniques such as monaural, stereo, binaural, Quadraphonic, Ambisonic and Surround Sound. Recording techniques typically involve the playback of a previously recorded sound field. The sound field is captured with a number of microphones (one for monaural, two for stereo and binaural and four with Quadraphonics and Ambisonics systems), in order to capture any inherent spatial cues. The recorded sounds are typically conveyed to the listener with an equal number of loudspeakers (e.g. each microphone has a corresponding loudspeaker). Rather than recording the actual sound field with a set of microphones at once, surround recordings (such as Dolby Stereo recordings), can also be produced by either recording or creating synthetic versions of each of the desired sounds (e.g. dialogue, special effects etc.) independently, possibly at different locations and then *mixing* each of the sounds in a *mixing studio* (e.g. assigning the sounds to the channels).

Recording techniques do not necessarily convey "true" 3D sound and are typically used for entertainment applications, such as in a cinema and home theaters. Monaural systems are incapable of providing any directional cues but can be used to convey non-directional information such as speech, even in a virtual environment. With stereo, when the listener is placed symmetrically between the two loudspeakers, the virtual sound source can be placed anywhere on a "line" between the two loudspeakers. Quadraphonic techniques allow sounds

to be placed both in front of and behind the listener but also require the listener to be placed in the center of the square loudspeaker arrangement they form. Ambisonics permit for greater flexibility with respect to loudspeaker and listener placement, sweet spot positioning and can potentially convey elevation information. As with all other recording techniques however, it does have a sweet spot and as the listener moves their head inside the sweet spot, certain timbral artifacts are generated. The surround sound systems introduced by Dolby have become standards in today's entertainment industry. Although these systems are capable of producing convincing audio effects they do place restrictions on the loudspeaker configuration and listener placement and do have a limited sweet spot.

Finally, recording techniques are not necessarily suited for real-time interactive virtual environments where a sound field is updated to account for movement of the listener or changes in the environment. Instead, these techniques are rather suited for playback of recorded soundtracks. Regardless, they have, nevertheless, paved the way for the modern, more perceptually correct systems which aim at simulating the human sound localization cues present in our everyday environment.

The primary human sound localization cues are interaural time and intensity differences (ITD and ILD respectively), the head related transfer function (HRTF) and reverberation. ITD and ILD cues involve the differences in time and intensity between the sound arriving at the left and right ears (e.g. a sound closer to the left ear will arrive at the left ear first and will also be of greater intensity due to the shadowing effect of the head). The HRTF describes the position dependent, complex interaction of a sound wave with the torso, shoulders, head and particularly the pinna (outer ear) of a listener. Reverberation refers to the collection of the reflections of a sound wave arising when a sound wave encounters objects in, or the boundaries (e.g. walls, ceiling and floor) of, some environment. Although ITD and ILD cues are fairly simple to model and implement, systems which rely on these cues solely are of limited use as they are incapable of providing 3D sound spatialization. In other words, there are locations which cannot be simulated using these cues alone or locations which may be ambiguous to the listener. For example, the listener may not be able to distinguish between a sound directly in front or directly in back of them (front-back-confusion) using ITD and/or ILD cues alone. As with the human auditory system, the ability for a user of a 3D sound system to spatialize a sound source and eliminate ambiguous situations can be achieved with the incorporation of HRTF cues into the system. Although the inclusion of HRTF cues greatly improves the spatialization abilities of a 3D sound system, as will be described there are various shortcomings associated with their use. Most notable, the HRTF is position dependent (e.g. the HRTF differs for each position in three-dimensional space) and due to differences between the physical make-up amongst individuals, the HRTF differs widely amongst individuals

HRTF responses can be obtained by either solving the wave equation, taking into account the complex interaction between the sound waves and the head, torso and par-

ticularly the pinnae or by physically measuring them from humans or anthropomorphic manakins. Solving the wave equation, taking into account these complex interactions, is currently beyond our computational and analytical scope and numerical approximations are used instead. These approximations however typically place many restrictions and make many simplifying assumptions (e.g. assume a spherical head), making them of limited use as well. Therefore, HRTFs are measured instead. However, the use of measured HRTFs has its share of problems. In particular, the process of collecting a set of HRTFs is very tedious, time consuming and requires specialized equipment and environments. Given these considerations, it is impractical to employ individualized HRTFs for each user of s 3D sound system and despite the errors which may result (e.g. greater rate of front-back confusions), non-individualized HRTFs, measured from anthropomorphic dummies, very good localizers or by averaging the HRTFs of many people, are used instead. Furthermore, since it is clearly impractical to measure the HRTFs for every position, a subset of all possible positions is sampled instead. Since the space is sampled, there will be positions in which there is no corresponding HRTF and in order to synthesize a sound to such a location without a corresponding response, some method of interpolation must be employed, thereby leading to potential performance decreases.

The addition of reverberation cues can greatly improve the performance of a 3D sound system, even when the system employs HRTF cues. Reverberation is a strong cue to sound source distance estimation, it can provide environmental information (e.g. size of a room, whether the room is "open" or contains many objects, composition of objects in the room etc.) and at the very least, can provide a certain ambiance and "warmth" to the simulation. Artificial reverberation algorithms can be used to easily add reverberation effects. Although these methods do not necessarily recreate the acoustics of some particular environment they are capable of providing good results with respect to late reverberation. The inclusion accurate reverberation information is of course not a trivial task, and may be very complicated depending on the environment being simulated. Various methods are available to model the reflection patterns of a particular environment. Typically, these methods involve measuring or simulating the response of a particular room and then using this response to filter a sound source before presenting the sound to the listener. This response can be measured directly (in a manner similar to HRTFs) or it can be artificially computed using for example, wave based and geometric techniques. Similarly to HRTF wave-based methods, the goal is to solve the wave equation in order to determine the sound pressure patterns of some particular environment. Solving the wave equation is of course a difficult task and is therefore, approximated using numerical approximation techniques, such as finite element methods, instead. These approximations however are extremely computationally demanding making them impractical for all but the simplest, static environments and when the sounds of interest are of low frequency.

Geometric methods, such as ray-tracing and image sources, model the reflection paths

followed by a sound while it is propagating from the source to the receiver (listener), taking into account such phenomenon as reflection, diffraction, atmospheric absorption and attenuation of the sound as it travels through the medium. These methods are however valid when the objects encountered by the propagating sound waves are larger than the waves themselves, assume specular reflection and require a substantial amount of time to compute. These restrictions typically limit their use to low order reflections (e.g. early reverberation) and simple, static environments with limited user movement. In other words, they are not well suited for interactive "real-time" virtual environments.

Conveying of sound to the users of a 3D audio system (or any other type of audio system for that matter), is accomplished using either headphones or loudspeakers. Headphones ensure the listener is isolated from any non-desired external sounds (un-wanted noise) and that the signal intended for the right (left) ear is delivered to the right (left) ear exclusively. In other words, with a headphone based system, crosstalk, where a portion of the signal intended for the left (right) is heard by the right (left) ear, is not an issue. Despite the benefits they may offer, there are several serious drawbacks associated with the use of headphones. Most importantly, sounds conveyed through headphones appear to be emanating from inside the listeners head (inside-the-head localization), leading to an increased number of front-back confusions and decreased sound localization performance. In addition, comfortability may become an issue, especially after wearing the headphones for an extended period of time. Finally, when used in any virtual environment, the physical presence of headphones is a constant reminder to the listeners that they are actually in a synthetic world.

Two loudspeaker based techniques were introduced, transaural audio and amplitude panning. Transaural audio involves the presentation of left and right binaural audio signals to the corresponding left and right ear. It typically involves the use of two loudspeakers however, generalized theoretical solutions for N loudspeakers and M listeners have been proposed. Transaural audio can overcome many of the shortcomings associated with headphone based systems (e.g. inside-the-head localization, front-back confusions etc.), however, before being of practical use, some method of crosstalk cancellation must be employed to (ideally) eliminate the inherent crosstalk. Various crosstalk cancellation schemes have been proposed and in theory, they actually do eliminate crosstalk. Such schemes however make many assumptions which are not necessarily valid (e.g. spherical head) and result in a very small sweet spot, making them of limited practical use. Furthermore, although theoretically crosstalk cancellation can be applied to multiple listeners, in practise, multiple listener transaural audio systems are beyond our reach, thereby typically restricting transaural audio to "desk-top" applications.

Amplitude panning techniques involve the manipulation of the amplitude (intensity or output level), of the signal applied to each loudspeaker in order to simulate the directional properties of the interaural level difference (ILD) cue. Amplitude panning can be applied to

both two and three dimensional loudspeaker configurations however, typically, a subset of the available loudspeakers are used at any time depending on the position and orientation of the listener. Amplitude panning techniques are simple to implement and since they do not employ any HRTF cues, are rather computationally efficient making them practical for use in real-time applications such as interactive virtual environments. Despite these benefits however, they also have their share of limitations. In particular, the position of virtual sounds is generally restricted to outside the loudspeaker "enclosure" (e.g. virtual sounds to be positioned inside the loudspeaker enclosure), therefore not allowing for sounds to be placed near the listener. In addition, with certain panning techniques, such as pair wise and triplet panning, the quality of the perceived virtual sound source position depends on the direction of the virtual source since the number of active loudspeakers (e.g. number of loudspeakers outputting the sound), varies depending on the direction of the virtual sound source [113]. Furthermore, panning techniques typically require the loudspeakers to be equidistant from the listener and that the listening environment is not too reverberant.

In conclusion, the generation of 3D sound for an interactive, immersive virtual environment is actually a difficult and computationally extensive task. The field of spatial sound has progressed extremely quickly over the last 60 to 70 years and various promising technologies have emerged. In addition, various 3D sound systems are currently available and can be quite good, however they are typically only capable of providing accurate spatial sound under restricted conditions. Plenty of work remains to be done in order to allow the generation of convincing spatial audio for use in an interactive real time virtual environment, regardless of the listener and environmental context.

Bibliography

[1] V. R. Algazi, R. O. Duda, D. M. Thompson, and C. Avendano. The CIPIC HRTF database. In *2001 IEEE ASSP Workshop on Applications of Signal Processing to Acoustics*, pages 111–123, New Paltz, NY. USA, October 2001.

[2] J. B. Allen and D. A. Berkley. Image method for efficiently simulating small-room acoustics. *Journal of the Acoustical Society of America*, 65(4):943–950, April 1979.

[3] D. B. Anderson and M. A. Casey. The sound dimension. *IEEE Spectrum*, pages 46–50, March 1997.

[4] B. S. Atal and M. R. Schroeder. Apparent sound source translator. U.S. Patent 3, 236, 949, February 23 1963.

[5] B. Bartlett and J. bartlett. *Stereo Microphone techniques*. Focal Press, 1 edition, 1991.

[6] H. E. Bass, H. J. Bauer, and L. B. Evans. Atmospheric absorption of sound: analytical expressions. *Journal of the Acoustical Society of America*, 52(3):821–825, 1972.

[7] D. W. Bateau. Listening with the naked ear. In S. J. Freedman, editor, *Neuropsychology of Spatially Oriented Behavior*. Dorsey Press, Homewood, IL. USA, 1968.

[8] D. W. Batteau. The role of the pinna in human sound localization. *Proceedings of the Royal Society*, 168:158–180, 1967.

[9] J. Bauck and D. H. Cooper. Generalized transaural stereo. In *Proceedings of the* 93*rd Audio Engineering Society Conference*, San Francisco, CA. USA, 1992. preprint 3401.

[10] B. Bauer. Phasor analysis of some stereophonic phenomena. *Journal of the Acoustical Society of America*, 33:1536–1539, November 1961.

[11] B. Bauer. Stereophonic earphones and binaural loudspeakers. *Journal of the Audio Engineering Society*, 9:148–151, 1961.

[12] E. B. Becker, G. F. Carey, and J. T. Oden. *Finite Elements, An Introduction, Volume 1*. Prentice Hall Publishers, Englewood Cliffs, NJ. USA, 1981.

[13] D. R. Begault. Perceptual effects of synthetic reverberation on three-dimensional audio systems. *Journal of the Audio Engineering Society*, 40(11):895–904, 1992.

[14] D. R. Begault. Auditory and non-auditory factors that potentially influence virtual acoustic imagery. In *Proceedings of the Audio Engineering Society* 16*th International Conference on Spatial Sound Reproduction*, pages 1–14, Rovanciemi, Finland, 1999. Audio Engineering Society.

[15] D. R. Begault and E. M. Wenzel. Headphone localization of speech. *Human Factors*, 35:361–376, 1993.

[16] R. Begault. *3-D Sound for Virtual Reality and Multimedia*. Academic Press Professional, Cambridge, MA. USA, 1994.

[17] J. Blauert. *Spatial Hearing: The Psychophysics of Human Sound Localization*. MIT Press, Cambridge, MA. USA, 1983.

[18] P. Brown and R. O Duda. A structural model for binaural synthesis. *IEEE Transactions on Speech and Audio processing*, 6(5):476–488, September 1998.

[19] D. S. Brungart. Control of perceived distance in virtual audio displays. In *Proceedings of the 20th Annual International Conference of the IEEE in Medicine and Biology Society.*, volume 20, pages 1101–1104, Hong Kong., 1998.

[20] D. S. Brungart and W. M. Rabinowitz. Auditory localization of nearby sources. head related transfer functions. *Journal of the Acoustical Society of America*, 106(3):1465–1479, 1999.

[21] D. S. Brungart and W. R. Rabinowitz. Auditory localization in the near-field. In *Proceedings of the International Conference on Auditory Displays*, Palo Alto, CA. USA, November 4-6 1996.

[22] 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*, 8:425–440, 2001.

[23] M. D. Burkhard and R. M. Sachs. Anthropometric manikin for acoustic research. *Journal of the Acoustical Society of America*, 58(1):214–222, 1975.

[24] D. M. Burrston, M. P. Hollier, and M. O. Hawksford. Limitations of dynamically controlling the listening position in a 3D ambisonic environment. In 102 *Audio Engineering Society*, Munich, Germany, March 22-25 1997. Audio Engineering Society. preprint 4460.

[25] S. Carlile. *Virtual Auditory Space: Generation and Application*. R. G. Landes Company, Austin, TX. USA, 1996.

[26] M. Casey, W. Gardner, and S. Basu. Vision steered beamforming and transaural rendering for the artificial life interactive video environment (ALIVE). In *Audio Engineering Society* 99*th Conference*, New York, NY. USA, 1996. Audio Engineering Society.

[27] C. I. Cheng and G. H. Wakefield. Introduction to head related transfer functions (HRTFs): Representation of hrtfs in time, frequency and space. *Journal of the Audio Engineering Society*, 49(4):231–249, 2001.

[28] J. Chowning. The simulation of moving sound sources. *Journal of the Audio Engineering Society*, 19(1):2–6, 1971.

[29] J. M. Chowning. Digital sound synthesis, acoustics and perception: A rich intersection. In *Proceedings of the COST G-6 Conference on Digital Audio Effects (DAFX-00).*, Verona, Italy, December 2000.

[30] M. Cohen and E. Wenzel. The design of multidimensional sound interfaces. In W. Barfield and T. A. Furness III, editors, *Virtual Environments and Advanced Interface Design*, chapter 8, pages 291–346. Oxford University Press Inc., New York, NY. USA, 1995.

[31] M. F. Cohen and J. R. Wallace. *Radiosity and Realistic Image Synthesis*. Morgan Kaufmann Publishers, Inc., San Francisco, CA. USA, 1993.

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

[33] D. H. Cooper and J.L. Bauck. Prospects of transaural recording. *Journal of the Audio Engineering Society*, 37:3–19, 1989.

[34] Aureal Corporation. 3D audio primer. Technical report, Aureal Corporation, 1998.

[35] J. Dattorro. Effect design: Part 1: Reverberator and other filters. *Journal of the Audio Engineering Society*, 45(9):660–684, 1997.

[36] C. S. Desai and J. F. Abel. *Introduction to the Finite Element Method*. Von Nostrand Reinhold, New York, NY. USA, 1972.

[37] Wells Deutsche. Technical report, Deutsche Wells Radio Training Centre, Cologne Germany, 1998.

[38] A. Digenis. The implementation of ambisonics for restoring quadraphonic recordings. BA (Hons) Recording Arts Undergraduate Thesis, SAE Technology College. Sydney, Australia, 2002.

[39] R. Dressler. Pro Logic surround decoder principles of operation. Technical report, Dolby Laboratories, San Francisco, CA. USA, 1998.

[40] R. O. Duda. Modeling head related transfer functions. In *Proceedings of the Twenty-Seventh Conference on Signals, Systems and Computers*, Alisomar, CA. USA, October 31 - November 3 1993.

[41] R. O. Duda, C. Avendano, and V. R. Algazi. An adaptable ellipsoidal head model for the interaural time difference. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing ICASSP'99*, pages 965 –968, Phoenix, AZ. USA, 1999.

[42] J. M. Eargle. *Handbook of Recording Engineering*. Van Nostrand Reinhold, New York, NY. USA, 1996.

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

[44] G. Eckel. The LISTEN vision. In *Preconference Proceedings of ACM SIGGRAPH and Eurographics Campfire on Acoustic Rendering for Virtual Environments*, pages 55–58, SnowBird, UT. USA, May 26-29 2001.

[45] R. Elen. Whatever happened to Ambisonics? *AudioMedia Magazine*, November 1991.

[46] R. Elen. Ambisonics: The surround alternative. In *Proceedings Surround* 2001, Los Angeles, CA. USA, December 2001.

[47] K. Farrar. Soundfield microphone. *Wireless World*, 85:99–102, 1979.

[48] H. G. Fisher and S. J. Freedman. The role of the pinna in auditory localization. *Journal of Auditory research*, 8:15–26, 1968.

[49] H. Fletcher and W. A. Munson. Loudness, its definition, measurement and calculation. *Journal of the Acoustical Society of America*, 5:82–108, 1933.

[50] T. Funkhouser, I. Carlbom, G. Elko, G. Pingali, M. Sondhi, and J. West. A beam tracing approach to acoustic modeling for interactive virtual environments. In *Siggraph '98*, pages 21–32, Orlando, FL. USA, 1998.

[51] T. Funkhouser, N. Tsingos, and J.M. Jot. Survey of methods for modeling sound propagation in interactive virtual environment systems. *Presence*, 2003. To appear.

[52] E. A. Gamble. Minor studies from the psychological laboratory of Wellesley College. Intensity as criterion in estimating the distance in sounds. *Psychological Review*, 16:416, 1909.

[53] J. Garas. *Adaptive 3D Sound Systems*. Kluwer Academic Publishers, Norwell, MA. USA, 2000.

[54] M. B. Gardner. Distance estimation of 0^o or apparent 0^o oriented speech signals in anechoic space. *Journal of the Acoustical Society of America*, 45:47–53, 1969.

[55] W. Gardner. *3-D Audio Using Loudspeakers*. Kluwer Academic Publishers, Norwell, MA. USA, 1998.

[56] W. G. Gardner. Efficient convolution without input-output delay. In 97*th Convention of the Audio Engineering Society*, pages 127–135, San Francisco, CA. USA, November 1994.

[57] W. G. Gardner. Reverberation algorithms. In M. Kahrs and K. Brandenburg, editors, *Applications of Digital Signal Processing to Audio and Acoustics*, chapter 2. Kluwer Academic Publishing, Norwell, MA. USA, 1998.

[58] W. G. Gardner. 3D audio and acoustic environment modeling. Technical report, Wave Arts Inc., Arlington, MA. USA, March 1999.

[59] W. G. Gardner and K. D. Martin. HRTF measurements of a KEMAR. *Journal of the Acoustical Society of America*, 97(6):3907–3908, 1995.

[60] M. A. Gerzon. Compatibility of and conversion between multispeaker systems. In 93*rd Convention of the Audio Engineering Society*, October 1992.

[61] P. Giddings. Multi-channel sound reproduction moving into the nineties. Engineering Harmonics, February 15 1991.

[62] W. Grantham. Spatial hearing and related phenomena. In B. C. J. Moore, editor, *Hearing*, Handbook of Perception and Cognition, chapter 9, pages 297–345. Academic Press Inc., San Diego, CA. USA, 1995.

[63] D. Griesinger. Stereo and surround panning in practice. In *Audio Engineering Society 112th Convention*, pages 1–6, Munich, Germany, May 10-13 2002. Audio Engineering Society.

[64] D. Griffin. *Echoes of Bats and Men*. Anchor Books Doubleday and Co., Garden City, NY. USA, 1938.

[65] Y. Haneda, S. Makino, Y. Kaneda, and N. Kitawaki. Common-acoustical-pole and zero modeling of head related transfer functions. *IEEE Transactions on Speech and Audio Processing*, 7(2):188–196, March 1999.

[66] C. M. Harris. Absorption of sound in air versus humidity and temperature. *Journal of the Acoustical Society of America*, 40:148–159, 1966.

[67] W. M. Hartmann. The physical description of signals. In B. C. J. Moore, editor, *Hearing*, Handbook of Perception and Cognition, chapter 1, pages 1–40. Academic Press Inc., San Diego, CA. USA, 1995.

[68] W. M. Hartmann. Listening in a room and the precedence effect. In R. H. Gilkey and T. R. Anderson, editors, *Binaural and Spatial Hearing in Real and Virtual Environments*, pages 191–810. Lawrence Erlbaum Associates, Mahwah, NJ. USA, 1997.

[69] D. D. Hearn and M. P. Baker. *Computer Graphics C Version*. Prentice-Hall, Upper Saddle River, NJ. USA, 2 edition, 1997.

[70] P. S. Heckbert and J. M. Winget. Finite element methods for global illumination. Technical Report UCP/CSD 91/643, Computer Science Division (EECS), University of California, Berkeley, CA, USA., July 1991.

[71] M. Hibbing. XY and MS microphone techniques in comparison. *Journal of the Audio Engineering Society*, 37:823–830, October 1989.

[72] C. Hugonnet and P. Walder. *Stereophonic Sound Recording*. John Wiley and Sons Ltd., West Sussex, England, 1995.

[73] J. Hull. Surround sound past, present and future. Technical report, Dolby Laboratories Inc., San Francisco, CA. USA, 1999.

[74] J. Huopaniemi. *Virtual Acoustics and 3D Sound in Multimedia Signal Processing*. PhD thesis, Department of Electrical and Communications Engineering, Helsinki University of Technology, Espoo, Finland, November 1999.

[75] U. Ingard. A review of the influence of meteorological conditions on sound propagation. *Journal of the Acoustical Society of America*, 25:405–411, 1953.

[76] Ircam and AKG Acoustics. LISTEN HRTF Databse, 2002. http://www.ircam.fr/equipes/salles/listen/index.html.

[77] D. Foley J, A. van Dam, S. K. Feiner, J. F. Hughes, and R. L. Phillips. *Introduction to Computer Graphics*. Addison-Wesley Publishing Co., Reading, MA. USA, 1994.

[78] Lord Raleigh J. W. Strutt. Our perception of sound direction. *Philosophical Magazine*, 13:214–232, 1907.

[79] J. M. Jot. An analysis/synthesis approach to real-time artificial reverberation. In *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing*, pages II.221–II.224, San Francisco, CA. USA, 1992. IEEE Press.

[80] J-M. Jot. Real-time spatial processing of sounds for music, multimedia and interactive human-computer interfaces. *Multimedia Systems*, 7:55–69, 1999.

[81] Y. Kahana, P. A. Nelson, M. Petyt, and S. Choi. Numerical modelling of the transfer functions of a dummy-head of the external ear. In *Proceedings of the 16th International Conference of the Audio Engineering Society*, Rovaniemi, Finland, April 10-12 1999.

[82] G. Kendall. A 3D sound primer: Directional hearing and stereo reproduction. *Computer music Journal*, 19(4):23–46, 1995.

[83] S. M. Khanna and M. R. Stinson. Specification of the acoustical input to the ear at high frequencies. *Journal of the Acoustical Society of America*, 77:511–589.

[84] M. Kleiner, D. I. Dalenback, and P. Svensson. Auralization - an overview. *Journal of the Audio Engineering Society*, 41(11):861–875, 1993.

[85] A. Kraemer. Two speakers are better than 5.1. *IEEE Spectrum*, pages 71–74, May 2001.

[86] G. F. Kuhn. Model for the interaural time differences in the azimuthal plane. *Journal of the Acoustical Society of America*, 62(1):157–167, July 1977.

[87] A. Kulkarni and H. S. Colburn. Evaluation of a linear interpolation scheme for approximating HRTFs. *Journal of the Acoustical Society of America*, 93(4), 1993.

[88] H. Kuttruff. *Room Acoustics*. Elsevier Science Publishers, New York, NY. USA, 1991.

[89] C. Kyriakakis. Fundamental and technological limitations of immersive displays. *Proceedings of the IEEE*, 86(5):941–951, May 1998.

[90] Dolby Laboratories. Dolby digital - the sound of the future here today. Technical report, Dolby Laboratories, San Francisco, CA. USA, 1998.

[91] V. Larcher and J. M. Jot. Techniques d'interpolation de filtres audio-numeriques. application a la reproduction spatiale des sons sur ecouteurs. In *Proceedings of the 4th congress of the French Society of Acoustics*, Marseille, France, April 1997.

[92] T. Lund. Enhanced localization in 5.1 production. In *109th Audio Engineering Society Convention*, Los Angeles, CA. USA, September 2000. Audio Engineering Society.

[93] J. Mackenzie, J. Huopaniemi, V. Vlimki, and I. Kale. Low-order modeling of head-related transfer functions using balanced model truncation. *IEEE Signal Processing Letters*, 4(2):39–41, 1997.

[94] P. Massarani and S. Muller. Transfer-function measurement. *Journal of the Audio Engineering Society*, 49(6):443–471, 2001.

[95] MEDLINEplus. Medical encyclopedia, 2002. U.S. National Library of Medicine and the National Institute of Health.

[96] D. H. Mershon. Phenomenal geometry and the measurement of perceived auditory distance. In R. H. Gilkey and T. R. Anderson, editors, *Binaural and Spatial Hearing in Real and Virtual Environments*, chapter 13, pages 257–274. Laurence Erlbaum Associates Inc., Mahwah, NJ. USA, 1997.

[97] D. H. Mershon, W. L. Ballenger, W. L. Little, P.L. Mcmurtry, and J. L. Buchanan. Effects of room reflectance and background noise on perceived auditory distance. *Perception*, 18:403–416, 1989.

[98] D. H. Mershon and J. N. Bowers. Absolute and relative cues for the auditory perception of egocentric distance. *Perception and Psychophysics*, 8:311–322, 1979.

[99] D. H. Mershon and L. E. King. Intensity and reverberation as factors in the auditory perception of egocentric distance. *Perception and Psychophysics*, 18:409–415, 1975.

[100] J. C. Middlebrooks. Narrow-band sound localization related to external ear acoustics. *Journal of the Acoustical Society of America*, 92:2607–2624, 1992.

[101] E. Milios, B. Kapralos, A. Kopinska, and S. Stergiopoulos. Sonification of range information for 3D space perception. *IEEE Transactions on Rehabilitation Engineering*, 2001.

[102] B. C. J. Moore. *An Introduction to the Psychology of Hearing*. Academic Press Limited, San Diego, CA. USA, 3 edition, 1989.

[103] B. C. J. Moore, B. R. Glassberg, and Thomas Baer. A model for the prediction of thresholds, loudness and partial loudness. *Journal of the Audio Engineering Society*, 45(4):224–239, 1997.

[104] R. F. Moore. *Elements of Computer Music*. Prentice-Hall, Englewood Cliffs, NJ. USA, 1990.

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

[106] C. Moy. The elements of musical perception. Technical report, Headwize, 2000.

[107] M. Naguib and H. Wiley. Estimating the distance to a sound: Mechanisms and adaptations for long-range communications. *Animal Behavior*, 62:825–837, 2001.

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

[109] S. R. Oldfield and S. P. A. Parker. Acuity of sound localization: a topography of auditory space II: Pinna cues absent. *Perception*, 13:601–617, 1984.

[110] A. V. Oppenheim and R. W. Schafer. *Discrete Time Signal Processing*. Prentice Hall, Englewood Cliffs, NJ. USA, 1989.

[111] C. J. Plack and R. P. Carlyon. Loudness perception and intensity coding. In B. C. J. Moore, editor, *Hearing*, Handbook of Perception and Cognition, chapter 2, pages 123–160. Academic Press Inc., San Diego, CA. USA, 2 edition, 1995.

[112] V. Pulkki. Virtual sound source positioning using vector base amplitude panning. *Journal of the Audio Engineering Society*, 45(6):456–466, 1997.

[113] V. Pulkki. Uniform spreading of amplitude panned virtual sources. In *Proceedings of the 1999 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, pages W99–1 – W99–4, Mohonk Mountain House, New Paltz, NY. USA, 1999.

[114] V. Pulkki. Localization of amplitude-panned virtual sources I: Stereophonic panning. *Journal of the Audio Engineering Society*, 49(9):739–751, September 2001.

[115] V. Pulkki. Localization of amplitude-panned virtual sources II: Two- and three-dimensional panning. *Journal of the Audio Engineering Society*, 49(9):753–767, September 2001.

[116] V. Pulkki. *Spatial Sound Generation and Perception by Amplitude Panning Techniques*. PhD thesis, Department of Electrical and Communications Engineering, Helsinki University of Technology, Helsinki, Finland, August 2001.

[117] V. Pulkki, J. Huopaniemi, T. Huotilainen, and M. Karjalainen. DSP approach to multichannel audio. In *Proceedings of the International Computer Music Conference (ICMC'96)*, pages 93–96, 1996.

[118] R. Rabenstein, O. Schips, and A. Stenger. Acoustic rendering of buildings. In *5th International Conference on Building Simulation*, Prague, Czech Republic, September 8-10 1997. International Building Performance Simulation Association.

[119] A. V. Rabinovich. The effect of distance in the broadcasting studio. *Journal of the Acoustical Society of America*, 7:199–203, 1936.

[120] D. W. Robinson and R. S. Dadson. A re-determination of the equal-loudness relations for pure tones. *British Journal of Applied Physics*, 7:166–181, 1956.

[121] M. Robinson, J. Laurence, A. Hogue, J. E. Zacher, A. German, and M. Jenkin. IVY: Basic design and construction details. In *Proceedings of the 12th International Conference on Artifial Reality and Telexistence*, Tokyo, Japan, December $4 - 6$ 2002.

[122] H. Robjohns. Stereo microphone techniques explained - part one. *Sound on Sound*, February 1997.

[123] H. Robjohns. Stereo microphone techniques explained - part one. *Sound on Sound*, March 1997.

[124] H. Robjohns. You are surrounded: Surround sound explained - part 1. *Sound on Sound*, August 2001.

[125] H. Robjohns. You are surrounded: Surround sound explained - part 3. *Sound on Sound*, October 2001.

[126] S. K. Roffler and R. A. Butler. Factors that influence the localization of in the vertical plane. *Journal of the Acoustical Society of America*, 43:1255–1259, 1968.

[127] S. K. Roffler and R. A. Butler. Localization of tonal stimuli in the vertical plane. *Journal of the Acoustical Society of America*, 43:1260–1266, 1968.

[128] L. D. Rosenblum, M. S. Gordon, and L. Jarquin. Echolocation by moving and stationary listeners. *Ecological Psychology*, 12(3):181–206, 2000.

[129] T. D. Rossing, R. F. Moore, and P. A. Wheeler. *The Science of Sound*. Benjamin Cummings, San Francisco, CA. USA, 2002.

[130] L. Savioja. *Modeling Techniques for Virtual Acoustics*. PhD thesis, Helsinki University of Technology, Telecommunications Software and Multimedia Laboratory, Helsinki, Finland, 1999.

[131] P. Scheiber. Four channels and compatibility. *Journal of the Audio Engineering Society*, 19:267–279, 1971.

[132] P. Scheiber. Multidirectional sound system. U.S. Patent 3,746,792, July 1973.

[133] M. R. Schroeder. Natural sounding artificial reverberation. *Journal of the Audio Engineering Society*, 10(3):219–233, 1962.

[134] M. R. Schroeder and B. F. Logan. Colorless artificial reverberation. *Journal of the Audio Engineering Society*, 9(3):209–214, 1961.

[135] M. N. Semple. Sounds in a virtual world. *Nature*, 396:723–724, December 1998.

[136] C. W. Sheeline. *An Investigation of the Effects of Direct and reverberant Signal Interaction on Auditory Distance Perception.* PhD thesis, Department of Hearing and Speech, Stanford University, Stanford, CA. USA, 1982.

[137] R. D. Shilling and B Sinn-Cunningham. Virtual auditory displays. In *Handbook of Virtual Environments*, pages 65–92. Lawrence Erlbaum Associates, Mahwah, NJ. USA, 2002.

[138] B. G. Shin-Cunningham. Distance cues for virtual auditory space. In *Proceedings of the First IEEE Pacific-Rim Conference on Multimedia.*, Sydney, Australia, December 2000.

[139] B. G. Shin-Cunningham. Learning reverberation: Considerations for spatial auditory displays. In *Proceedingsof the International Conference on Auditory Displays*, pages 126–134, Atlanta, GA. USA, December 2000.

[140] A. Sibbald. Transaural acoustic crosstalk cancellation. Technical report, Sensaura Ltd., Middlesex, UK, 1999.

[141] A. Sibbald. Chaotic waves for 3D audio. Technical report, Sensaura Ltd., Middlesex, UK, 2001.

[142] W. H. Slattery and J. C. Middlebrooks. Monaural sound localization: acute versus chronic unilateral impairment. *Hearing Research*, 75:38–46, 1984.

[143] J. M. Speigle and J. M. Loomis. Auditory distance perception by translating observers. In *Proceedings of the IEEE Symposium on Research Frontiers in Virtual Reality*, pages 92–95, New York, NY. USA, 1993.

[144] J. C. Steinberg and W. B. Snow. Physical factors in auditory perspective. *Bell Systems Technical Journal*, 13:245–259, 1953.

[145] S.S. Stevens. On the physical law. *Psychology Review*, 64:153–181, 1957.

[146] S.S. Stevens. Perceived level of noise by mark VII and decibels (E). *Journal of the Acoustical Society of America*, 51:575–601, 1972.

[147] T. G. Stockholm. High speed convolution and correlation. In *Proceedings of the American Federation of Information Processing Societies*, pages 229–233, 1966.

[148] R. Streicher and A. Everest. *The New Stereo Soundbook*. Audio Engineering Associates, Pasadena, CA. USA, 2 edition, 1998.

[149] M. Supra, M. Cotzin, and K. M. Dallenbach. Facial vision: The perception of obstacles by the blind. *The American Journal of Psychology*, 57:133–183, 1944.

[150] C. J. Tan and W.S. Gan. User defined spectral manipulation of hrtf for improved localization in 3D sound systems. *Electronics Letters*, 34(25):2387–2389, December 1998.

[151] G. Theile. On the naturalness of two-channel stereo sound. *Journal of the Audio Engineering Society*, 39(10):761– 767, 1991.

[152] W. R. Thurlow, J. W. Mangels, and P. S. Runge. Head movements during sound localization. *Journal of the Acoustical Society of America*, 42:489–493, 1967.

[153] H. Tremaine. *Audio Cyclopedia*. Howard W. Sams and Co., Inc, Indianapolis, IN. USA, 2 edition, 1974.

[154] C. Uy. "Seeing" sounds: Echolocation by blind humans. *the Harvard Brain: Harvard's Undergraduate Neuroscience Magazine.*, 1, 1994.

[155] von Bekesy. *Experiments in Hearing*. McGraw Hill, New York, NY. USA, 1960.

[156] H. Wallach. The role of head movements and vestibular and visual cues in sound localization. *Experimental Psychology*, 27:339–368, 1940.

[157] H. Wallach, E. B. Newman, and M. R. Rosenzweig. The precedence effect in sound localization. *Journal of Psychology*, 52:315–336, 1949.

[158] D. B. Ward and G. W. Elko. A new robust system for 3D audio using loudspeakers. In *Proceedings 2000 IEEE International Conference on Acoustics, Speech and Signal Processing*, volume 2, pages II781 –II784. IEEE, 2000.

[159] R. M. Warren. *Auditory Perception: A New Analysis and Synthesis*. Cambridge University Press, New York, NY. USA, 1983.

[160] 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*, 94(1):111–123, 1993.

[161] E. M. Wenzel, F. L. Wightman, and D. J. Kistler. Acoustic origins of individual differences in sound localization behavior. *Journal of the Acoustical Society of America*, 84(S79), 1988.

[162] J. West. Five-channel panning laws: An analytical and experimental comparison. Master's thesis, Faculty of Music Engineering Technology, University of Miami, Coral Gables, FL. USA, May 1998.

[163] F. L. Wightman and D. J. Kistler. Headphone simulation of free-field listening. I: Stimulus synthesis. *Journal of the Acoustical Society of America*, 85(2), February 1989.

[164] F. L. Wightman and D. J. Kistler. Sound localization. In W. Yost, A. Popper, and R. Fay, editors, *Springer Handbook of Auditory Research: Human Psychophysics*, volume 3, pages 155–192. Springer-Verlag Inc., New York NY. USA, 1993.

[165] F. L. Wightman and D. J. Kistler. Factors affecting the relative salience of sound localization cues. In R. H. Gilkey and T. R. Anderson, editors, *Binaural and Spatial Hearing in Real and Virtual Environments*, chapter 1, pages 1–23. Lawrence Elbaum Associates, Mahwah, NJ. USA, 1997.

[166] F. L. Wightman, D. J. Kistler, and M. Arruda. Perceptual consequences of engineering compromises in synthesis of virtual auditory objects. *Journal of the Acoustical Society of America*, 92:2332, 1992.

[167] R. Wolfson and J. M. Pasachoff. *Physics with Modern Physics*. HarperCollins College Publishers, New York, NY. USA, 2 edition, 1995.

[168] R. S. Woodworth and G. Schlosberg. *Experimental Psychology*. Holt, Rinehard and Winston, New York, NY. USA, 1962.

[169] P. Zahorik. Auditory distance perception: A literature review. Phd Preliminary Examination, University of West Maddison, Deptartment of Psychology, August 1996.

[170] P. Zahorik. Assessing auditory distance perception using virtual acoustics. *Journal of the Acoustical Society of America*, 111(4):1832–1846, 2002.

[171] D. Zotkin, R.Duraiswami, and L.Davis. Creation of virtual auditory spaces. In *Proceedings of the International Conference on Acoustics, Speech and Signal Processing*, pages 2113–2116, Orlando, FL. USA, May 2002.