Virtual Audio Systems

B. Kapralos[1,2], *M. R. Jenkin*[3], *and E. Milios*[4]

[1]Faculty of Business and Information Technology,

[2]Health Education Technology Research Unit.

University of Ontario Institute of Technology. Oshawa, Ontario, Canada. L1H 7K4

[3]Department of Computer Science and Engineering,

Centre for Vision Research.

York University. Toronto, Ontario, Canada. M3J 1P3

[4]Faculty of Computer Science.

Dalhousie University. Halifax, Nova Scotia, Canada. B3H 1W5

Corresponding Author:

Bill Kapralos
Faculty of Business and Information Technology,
University of Ontario Institute of Technology.
2000 Simcoe St. North
Oshawa, Ontario, Canada. L1H 7K4.

Phone: 905-721-8668 x2882
Email: bill.kapralos@uoit.ca

Abstract

To be immersed in a virtual environment, the user must be presented with plausible sensory input including auditory cues. A virtual (three-dimensional) audio display aims to allow the user to perceive the position of a sound source at an arbitrary position in three-dimensional space despite the fact that the generated sound may be emanating from a fixed number of loudspeakers at fixed positions in space or a pair of headphones. The foundation of virtual audio rests on the development of technology to present auditory signals to the listener's ears so that these signals are perceptually equivalent to those the listener would receive in the environment being simulated. This paper reviews the human perceptual and technical literature relevant to the modeling and generation of accurate audio displays for virtual environments. Approaches to acoustical environment simulation are summarized and the advantages and disadvantages of the various approaches are presented.

1 Introduction

A virtual (three-dimensional) audio display allows a listener to perceive the position of a sound source, emanating from a fixed number of stationary loudspeakers or a pair of headphones, as coming from an arbitrary location 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, front-back, and up-down (Cohen & Wenzel, 1995). The simulation of realistic "spatial" sound cues in a virtual environment can contribute to a greater sense of "presence" or "immersion" than visual cues alone and at a minimum, adds a "pleasing quality" to the simulation (Shilling & Shinn-Cunningham, 2002). Furthermore, in certain situations a virtual sound source can be indistinguishable from the real source it is simulating (Kulkarni & Colburn, 1998; Zahorik, Wightman & Kistler, 1995). Despite these benefits, spatial sound is often overlooked in immersive virtual environments which often emphasize the generation of believable visual cues over other perceptual cues (Carlile, 1996; Cohen & Wenzel, 1995). Just as the generation of compelling visual displays requires an understanding of visual perception, the generation of effective audio displays requires an understanding of human auditory perception and the interaction between audition and other perceptual processes. In 1992 Wenzel provided a thorough and extensive review on the development of virtual audio displays. Although a thorough review of the state of the art at the time, Wenzel's review was published over 15 years ago and there have been significant advances in our understanding of human auditory processing and in the design of virtual audio displays since then. In this paper we focus on advances that have occurred in the field of spatial audio since Wenzel's 1992 review. This includes head-tracking and system latency (issues critical in the deployment of many realistic audio systems), modeling the room impulse response (wave-based and geometric-based room impulse response modeling, and diffraction modeling), spherical microphone arrays, and loudspeaker-based techniques (transaural audio, amplitude panning, and wave-field synthesis).

2 Human sound localization

The development of an effective virtual audio display requires an understanding of human auditory perception. Sound results from the rapid variations in air pressure caused from the vibrations of an object (or an object in motion) in the range of approximately 20 Hz to 20 kHz (Moore, 1989). We perceive these rapid variations in air pressure through the sense of hearing. Since sounds propagate omni-directionally (at least in an open environment), one of the most interesting properties of human hearing is our ability to localize sound in three-dimensions. The *duplex theory* is arguably the earliest theory of human sound localization (Strutt, 1907). Under the assumption of a perfectly spherical head without any external ears (pinnae) this theory explains many properties of human sound localization. Unless the sound source lies on the median plane (the plane equidistant from the left and right ears) the distance traveled by sound waves emanating from a sound source to the listener's left and right ears differs. This causes 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 interaural time delay (ITD) is the difference between the onset of sounds at the two ears (see Figure). When the wavelength of the sound wave is small relative to the size of the head, the head acts as an occluder and creates an acoustical shadow which attenuates the sound pressure level of the sound waves reaching the contralateral ear (Wightman & Kistler, 1993). The difference in sound level at the ipsilateral and contralateral ears is commonly referred to as the interaural level difference (ILD) although it is also referred to as the interaural intensity difference (IID) as well (see Figure).

Figure 1 about here.

ITDs provide localization cues primarily for low frequency sounds (< 1500 Hz) where the wavelength of the arriving sound is large relative to the diameter of the head thus allowing the phase difference between the sounds reaching the two ears to be unambiguous (Blauert, 1996). However, recent studies indicate that listeners can detect interaural delays in the envelopes of high frequency carriers (Middlebrooks & Green, 1990). Low frequency sounds corresponding to wavelengths greater than the diameter of the head experience *diffraction*, essentially the sound waves “bending around” the head to reach the contralateral ear. Hence, ILD cues for low frequency sounds are typically minuscule although in some cases, they may be as large as 5 dB (Wightman & Kistler, 1993). For frequencies in excess of 1500 Hz, where the head is larger than the wavelength, the sound waves are too small to bend around the head but are rather shadowed by the head. This results in detectable ILDs for lateral sources.

Studies by Mills (1958) indicate that the *minimum audible angle* (MAA), the minimum amount of sound source displacement that can be reliably detected, is dependent on both frequency and azimuth. Precision is best directly in front of the listener (0° azimuth) and decreases as azimuth increases to 75°. At an azimuth of 0°, the MAA is less than 4° for all frequencies between 200 and 4000 Hz and is as precise as 1° for a 500 Hz tone. More recent work has examined differences in MAAs in the azimuthal and vertical planes (Perrott & Saberi, 1990), and the interaction of MAAs with the precedence effect i.e. the ability of the auditory system to "combine" both the direct and reflected sounds such that they are heard as a single "entity" and localized in the direction corresponding to the direct sound (Saberi & Perrott, 1990).

Although the duplex theory explains sound localization on the horizontal plane with ILD and ITD cues, there are aspects of human sound localization for which it cannot account. For example, even listeners suffering form unilateral hearing loss are capable of localizing sound sources (Slattery & Middlebrooks, 1984). The duplex theory cannot differentiate the placement of a sound source on the median plane since both ITD and ILD cues are zero in either case. A further illustration of the ambiguity of the duplex theory is the so-called *cone of confusion* (see Figure). This is a cone centered on the interaural axis with the centre of the head as its apex. A sound source positioned on any point on the surface of the cone of confusion will have the same ITD values (Blauert, 1996; Mills, 1972).

Figure 2 about here.

In normal listening environments humans are mobile rather than stationary. Head movements are a crucial and natural component of human sound source localization, reducing front-back confusions and increasing sound source localization accuracy, (Thurlow, Mangels & Runge, 1967; Wallach, 1940; Wightman & Kistler, 1997). Head movements lead to changes in the ITD and ILD cues and in the sound spectrum reaching the ears (see Figure). We are capable of integrating these changes temporally in order to resolve ambiguous situations (Begault, 1999). Lateral head motions can also be used to distinguish frontal low frequency sound sources as being either above or below the horizon (Perrett & Noble, 1995, 1997).

Figure 3 about here.

It has been well established that sound source localization accuracy is dependent on the source spectral content. Various studies have demonstrated that sound source localization accuracy decreases as sound source bandwidth decreases (Hebrank & Wright, 1974; King & Oldfield, 1997; Roffler & Butler, 1968a). Studies have also demonstrated that, for optimal sound source localization, the sound source spectrum must extend from about 1 to 16 kHz (Hebrank & Wright, 1974; King & Oldfield, 1997).

2.1 Head-related transfer function

Batteau's work in the 1960's on the filtering effects introduced by the pinna of the ear was the next major advance in the study of human sound localization (Batteau, 1967). He observed that sounds reaching the ears interact with the physical makeup of the listener (in particular, the listener's head, shoulders, upper-torso, and most notable, the pinna of each ear) in a direction- and distance-dependent manner, and that this information can be used to estimate the distance and direction to the sound source. Collectively, these interactions are characterized by a complex response function known as the *head-related transfer function* (HRTF) or the *anatomical transfer function* (ATF) and encompass various sound localization cues including ITDs, ILDs, and changes in the spectral shape (frequency distribution) of the sound reaching a listener (Hartmann, 1999). With the use of HRTFs, many of the localization limitations inherent within models based on the use of ITD and ILD alone are overcome.

The left $H_L(\omega, \theta, \phi, d)$, and right $H_R(\omega, \theta, \phi, d)$ ear HRTFs are functions of four variables: ω, the angular frequency of the sound source, θ and ϕ, the sound source azimuth and elevation angles respectively, and d, the distance from the listener to the sound source (measured from the center of the listener's head) (Zotkin, Duraiswami & Davis, 2004). The HRTF itself can be decomposed into two separate components: the *directional transfer function* (DTF), which is specific to the particular sound source direction; and the common transfer function (CTF), which is common to all sound source locations (Middlebrooks & Green, 1990). When considering a sound source in the near-field (i.e. at a distance of less than approximately one meter) displaced from the median plane, HRTFs (and in particular the ILD component of the HRTF) are both direction- and distance-dependent across all frequencies (Brungart & Rabinowitz, 1999). Beyond approximately one meter, HRTFs are generally assumed to be independent of distance.

The pinna of individuals varies widely in size, shape, and general make-up. This leads to variations in the filtering of the sound source spectrum particularly when the sound source is to the rear of the listener and when the sound is within the 5-10 kHz frequency range.

2.2 Other factors affecting human auditory perception

In addition to sound source localization cues based on one's physical make-up, other "external" factors can alter the sound reaching a listener providing additional cues to the location of a sound source. Reverberation, the reflection of sound from objects or encountered surfaces, is a useful cue to sound localization. Reverberation is capable of providing information with respect to the physical "make-up" of the environment (e.g., size, type of material on the walls, floor, ceiling, etc.). Reverberation can also provide absolute sound source distance estimation independent of the overall sound source level due to variation in the direct-to-reverberant sound energy level as a function of sound source distance (Begault, 1994; Bekesy, 1960; Bronkhorst & Houtgast, 1999; Brungart, 1998; Carlile, 1996; Chowning, 2000; Coleman, 1963; Nielsen, 1993; Shinn-Cunningham, 2000a). Despite the importance of reverberation with respect to sound source localization, its presence can lead to a decrease in directional localization accuracy in both real and virtual environments and although this effect is of small magnitude, it is nevertheless measurable (Rakerd & Hartmann, 1985; Shinn-Cunningham, 2000b).

The frequency spectrum of a sound source varies with distance due to absorption effects caused by the medium (Naguib & Wiley, 2001). This high frequency attenuation is particularly important for distance judgements for larger distances (greater than approximately 15 m) but is largely uninformative for smaller distances.

Finally, a listener's prior experience with a particular sound source and environment (e.g., the source transmission path) can provide either a more accurate localization estimate or may help overcome ambiguous situations. For example, from infancy humans engage in conversations with each other. For normal listeners, speech is an integral aspect of communication. Consequently, one becomes familiar with the acoustic characteristics of speech (e.g., how loud a whisper or a yell may be, and who is speaking) and under normal listening conditions is capable of accurately judging the distance to a live talker (Brungart & Scott, 2001; Gardner, 1968).

3 Auralization

Kleiner, Dalenbäck, & Svensson (1993) define auralization 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 taking into account the environmental acoustics (e.g., the environmental context of a listening room or the "room acoustics") and the listener's characteristics. Auralization is typically defined in terms of the *binaural room impulse response* (BRIR). The BRIR represents the response of a particular acoustical environment and human listener to sound energy and captures the room acoustics for a particular sound source and listener configuration. The direct sound, reflection (reverberation), diffraction, refraction, sound attenuation, and absorption properties of a particular room configuration (e.g., the "room acoustics") are captured by the *room impulse response* (RIR). The listener-specific portion of the BRIR is defined in terms of the HRTF (Kleiner, Dalenback & Svensson, 1993).

Within a real environment, the BRIR can be measured by generating an impulsive sound with known characteristics through a loudspeaker positioned within the room and measuring the response of the arriving sound (with probe microphones) at the ears of the observer (either an actual human listener or an anthropomorphic "dummy head") positioned in the room. The recorded response then forms the basis of a filter that is used to process source sound material (anechoic or synthesized sound before presenting it to the listener). When the listener is presented with this filtered sound, the direct and reflected sounds of the environment are reproduced in addition to directional filtering effects introduced by the original listener (Väänänen, 2003). However, physically measuring the BRIR in this manner is highly restrictive; the measured response is dependent upon the room configuration with the original sound source and listener positions. Only that particular room and sound source/receiver configuration can be "re-created" exactly. Movement of the sound source, the receiver, or changes to the room itself (e.g., introduction of new objects or movement of existing objects in the room) necessitates BRIR re-measurement. A sample BRIR measured in a "moderate sized, reverberant classroom" at the right ear of a listener with the sound source at an azimuth and elevation of 45° and 0° respectively, and at a distance of 1m is provided in Figure .

Figure 4 about here.

Although not necessarily separable, for reasons of simplicity and practicality the BRIR is commonly approximated by considering the RIR and HRTF separately and then combining them to approximate the BRIR (Kleiner, Dalenback & Svensson, 1993). The RIR is used to model the effects of the room while sound reaching the “head” is modeled with an HRTF pair corresponding to the geometry of the listener in order to recreate “binaural listening” (Begault, 1994). This approach is taken by a variety of auralization systems including NASA’s SLAB (Wenzel, Miller & Abel, 2000a,b). Under this approach to auralization, the HRTF filtering accounts for most of the computational complexity and can be impractical for interactive (real-time) systems (Hacihabiboğlu & Murtagh, 2006). In order to limit the computational complexity, often only the early portion of the room impulse response (the first 80–100 ms) is modeled and only reflections within this portion are filtered with the corresponding HRTFs. The latter portion is then modeled as exponentially decaying noise using statistical methods and techniques (Garas, 2000), and artificial reverberation methods such as feedback delay networks (Jot, 1992; Jot, Cerveau & Warusfel, 1997; Kuttruff, 2000). Hacihabiboğlu and Murtagh (2006) describe a perception-based method for selecting a small number of early reflections in a geometric room acoustics model without affecting the spatialization capabilities of the system.

3.1 Receiver modeling: determining the HRTF

In theory, the HRTF can be determined by solving the wave equation, taking into consideration the interaction of the wave with the head, upper torso, and pinna. However, such an approach is impractical given the computational and analytical complexity associated with it. As a result, various approximations have been developed. One approach involves ignoring the pinna and torso altogether and assuming a spherical head. This ignores the filtering effects introduced by the pinna despite the fact that the interaction of a sound wave with the pinna is the major contributor to the HRTF. Consequently, such approximations lead to decreased performance when employed in a three-dimensional audio display. More sophisticated mathematical models must deal with difficult issues associated with modeling the HRTFs, including (Duda, 1993):

1. Approximation of the effect of wave propagation and diffraction using simple low-order filters;
2. The complicated relationship between azimuth, elevation, and distance in the HRTF;
3. The quantitative evaluation criteria; and
4. The large variation among the HRTFs of different individuals.

In light of these problems, most practical systems are based on measured HRTFs whereby an individual's left and right ear HRTFs for a sound source at a position $\vec{p}$ relative to the listener are measured. This is accomplished by outputting an excitation signal $s(n)$ with known spectral characteristics from a loudspeaker placed at position $\vec{p}$ and measuring the resulting impulse response at the left (h_L) and right (h_R) ears using small microphones inserted into the individual's left and right ear canals (Begault, 1994). 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). Applying the discrete Fourier transform (DFT) to the time domain impulse responses h_L and h_R results in the left $H_L(\omega, \theta, \phi, d)$ and right $H_R(\omega, \theta, \phi, d)$ ear HRTFs respectively. When measuring HRTFs it is common to assume a far-field sound source model and to model attenuation loss with distance separately (Martens (2000) describes an audio display that does account for sound source distance in simulated HRTFs at close range). This reduces the time needed to estimate the HRTF and simplifies the mathematical representation of the HRTF at the cost of reduced accuracy. Even with this simplification, it is not practical to measure HRTFs at every possible direction. Instead, as described below, the set of discrete-measured HRTFs are interpolated to form a complete HRTF space.

In order to minimize the influence of reverberation, HRTF measurements are typically made in an anechoic chamber. Alternatively, if collected within a reverberant environment, the resulting time-domain measurements can be "windowed" to reduce reverberation effects. For example, Gardner (1998) employed a Hanning window to attenuate the reflections of HRTFs collected in a reverberant environment.

3.1.1 Non-individualized ("generic") HRTFs

Optimal results are achieved when an individual's own HRTFs are measured and used (Wenzel, Arruda & Kistler, 1993). However, the process of collecting a set of *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. It is therefore impractical to use individualized HRTFs and as a result, generalized (or generic) *non-individualized* HRTFs are often used instead. Non-individualized 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"). Several non-individualized HRTF datasets are freely available to the research community (Algazi, Duda, Thompson & Avendano, 2001; Gardner & Martin, 1995; Grassi, Tulsi & Shamma, 2003; Ircam & AKG Acoustics, 2002). Although practical, the use of non-individualized HRTFs can be problematic. A large variation between the measured HRTFs across individuals is due to a number of factors, including (Carlile, 1996):

Variation of each person's pinna: The pinna of each individual differs with respect to size, shape, and general make-up, leading to differences in the filtering of the sound source spectrum, particularly at higher frequencies. Higher frequencies are attenuated by a greater amount when the sound source is to the rear of the listener as opposed to the front of the listener. In the 5 kHz to 10 kHz frequency range, the HRTFs of individuals can differ by as much as 28 dB (Wightman & Kistler, 1989). This high frequency filtering is an important cue to sound source elevation perception and in resolving front-back ambiguities (Begault, 1994; Middlebrooks, 1992; Roffler & Butler, 1968a,b; Wenzel, Arruda & Kistler, 1993). The left and right ear HRTF measurements of three individuals for a sound source located at an azimuth and elevation of 90° and 0° respectively provided in Figure illustrate the individual differences. Studies have demonstrated that non-individualized HRTFs reduce localization accuracy, especially with respect to elevation. Wenzel, Wightman, & Kistler (1988) examined the effect of non-individualized HRTFs measured from average listeners when presented to listeners who were “good localizers”. They found that the use of non-individualized HRTFs resulted in a degradation of the subjects’ ability to determine the elevation of a sound source. A similar study performed by Begault and Wenzel (1993) in which subjects localized a speech stimuli as opposed to broadband noise resulted in a decrease in elevation judgments as well. In addition to the filtering effects introduced by the pinna, 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 audio display will be 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) (Kendall, 1995).

Differences in the measurement procedures: Currently no universally accepted approach for measuring HRTFs exists (Begault, 1994). The *non-blocked ear canal* approach uses measurements in one of three main positions of the ear canal: i) deep in the ear canal, ii) in the middle of the ear canal, and iii) at the ear canal entrance (Carlile, 1996). Particularly when taken near the ear drum, such measurements account for the individual localization characteristics of the listener, including the ear canal response (Algazi, Avendano & Thompson, 1999). The non-blocked ear canal approach is often impractical as it requires both measuring the response within the small ear canal and the use of probe microphones with low sensitivity and a non-flat frequency response (Møller, 1992). With the *blocked ear canal* approach the response of the ear canal is suppressed by physically blocking the ear canal (Møller, Hammershoi, Jensen & Sorensen, 1995). Blocked ear canal measurements are simpler, more comfortable, and less obtrusive than placing probe microphones within the ear canal or close to the ear drum. Furthermore, the HRTF measurement position within the ear canal is not critical since the HRTF at the eardrum can be determined by incorporating a simple position-independent transfer function compensation factor that is measured away from the ear canal (Algazi, Avendano & Thompson, 1999).

Perturbation of the sound field by the microphone: The microphones used to measure the response, due to their size, perturb the sound field over the wavelengths of interest (Carlile, 1996).

Variations in the relative position of the head: When measuring human subject HRTFs, measurements may be quite sensitive to variations in the subject's head position; even small head movements during the measurement procedure can result in a large variation in the measured HRTF within one subject.

Figure 5 about here.

In recent years a number of approaches have been developed to increase the efficiency of the HRTF process. For example, Zotkin, Duraiswami, Grassi & Gumerov (2006) present an efficient method for HRTF collection that relies on the acoustical principle of reciprocity

(Morse & Ingard, 1968). In contrast to traditional HRTF measurement procedures, they swap the speaker and microphone positions. A “microspeaker” is inserted into the individual’s ear while a number of microphones are positioned around the individual. Upon emitting an impulsive sound from the microspeaker, the resulting HRTF at each microphone location is measured simultaneously. There are small observable differences between reciprocally measured HRTFs and directly measured HRTFs. However, results of preliminary perceptual experiments indicate that reciprocally measured HRTFs can be reasonably interchanged with directly measured HRTFs in virtual audio applications as the errors introduced by such an exchange are within the errors inherent with measured HRTFs (Zotkin, Duraiswami, Grassi & Gumerov, 2006).

3.1.2 Interpolation of HRTFs

One of the simplest interpolation methods for HRTFs is based on linear interpolation. The desired HRTF is obtained by taking a weighted average of measured HRTFs surrounding the direction of interest (Freeland, Wagner, Biscainho & Dinz, 2002). Although simple, such an approach does not preserve a number of features including interaural time delays (Zotkin, Duraiswami & Davis, 2004). Interaural time delays must therefore be removed from the HRTFs before they are interpolated and re-introduced in a later post-processing operation. Furthermore, linear interpolation results in HRTFs that are acoustically different from the actual measured HRTFs of the desired target location (Kulkarni & Colburn, 1993). However, Wenzel and Foster (1993) found that localization errors associated with linearly interpolated (normal or minimum phase) non-individualized

HRTFs are relatively small when compared to the localization errors associated with the use of non-individualized HRTFs. More complex interpolation schemes have also been used (Algazi, Duda & Thompson, 2004; Carlile, Jin & Raad, 2000; Freeland, Biscainho & Diniz, 2004).

3.1.3 HRTF personalization

Several current research efforts are examining the development of HRTF *personalization* for individual users of a virtual audio display. These studies take advantage of the similarities observed in the HRTFs amongst individuals with similar pinna structure. Zotkin, Hwang, Duraiswami & Davis (2003) describe a system where seven anatomical features in an image of the outer ear are located using image processing techniques. Greater details regarding these features are provided by Algazi, Duda, Thompson & Avendano (2001). A set of "similar" HRTFs is chosen from the CIPIC HRTF dataset based on a comparison between the measured features and corresponding features associated with HRTFs in the dataset (Algazi, Duda, Thompson & Avendano, 2001). Middlebrooks (1999a,b) describes a procedure for scaling the non-individualized DTF component of the HRTF. The procedure involves multiplying the frequency domain representation of the direct transfer function (DTF) by a scaling factor and is based on two observations: i) the directional sensitivity at one frequency at the ear of an individual is similar to the directional sensitivity at some other frequency for another individual, and ii) frequencies in which subjects demonstrated directional sensitivity showed an inverse relationship with the subject's physical anatomy (e.g., head size, and pinna structures). The scaling factors for an individual user are estimated based on a comparison between certain anthropomorphic measures including pinna cavity height, head width of the user, and the individual used to obtain the non-individualized HRTFs. Instead of relying on these anthropomorphic measures, Middlebrooks, Macpherson & Onsan (2000) later developed a psychophysical procedure for determining the scaling factors.

3.1.4 HRTF simplification

Although HRTFs differ amongst individuals, not all features of the HRTF are necessarily perceptually significant. This has lead to various data reduction models of the HRTF such as principal components analysis (PCA) (Kapralos & Mekuz, 2007; Martens, 1987; Kistler & Wightman, 1992), and genetic algorithms (Cheung, Trautmann & Horner, 1998), whose goal is to represent the HRTF with a reduced number of basis spectra. Using the DTFs of 36 individuals, Jin, Leong, Leung, Corderoy & Carlile (2003) constructed a two-pass PCA-based statistical model of the DTF to provide a compressed representation of the DTF. With their model, seven PCA coefficients accounted for 60% of the variation across individual DTFs. Experiments conducted to test the validity of the reduced model found that accurate virtual sound source localization could be achieved even when accounting for only 30% of the individual DTF variation. Kulkarni, Isabelle & Colburn (1995, 1999) modeled the HRTF as a minimum-phase function together with a position-dependent and frequency independent interaural time delay. Theoretical and psychophysical results indicate the adequacy of the approach when considering brief, anechoically measured HRTFs (Kulkarni, Isabelle & Colburn, 1999).

3.1.5 Equalization of the measured HRTF

In addition to containing the actual impulse response due to the head, pinna, and upper torso (shoulders), measured HRTFs are corrupted by the transfer functions of the loudspeaker, headphones, and electronic measurement system (Gardner, 1998). Various *equalization* methods have been developed in order to compensate for the response of the measurement and playback systems. These methods typically involve "filtering" the measured HRTF with a filter that is essentially an approximation to the inverse of the unwanted response. Details regarding a number of HRTF equalization techniques including *free-field equalization*, *diffuse-field equalization*, and *measurement equalization* are provided by Gardner (1998).

3.1.6 Head tracking and system latency

HRTFs are defined in a head-centered coordinate system. This implies that the position of the listener's head must be tracked in terms of both position and orientation if the HRTF is to be combined with the RIR to establish the BRIR. Current head tracking technology introduces position and orientation inaccuracies and latency leading to position and orientation estimation errors (Allison, Harris, Jenkin, Jasiobedzka & Zacher, 2001). A survey of tracking technologies is available from Foxlin (2002) and Rolland, Davis & Baillot (2001). For a spatial auditory system, Wenzel (1999) defines *total system latency* or *end-to-end latency* as the time between the transduction of an event or action and the time at which the consequences of that particular action causes an equivalent change in the virtual sound source. System latency involves each component comprising the virtual environment including head trackers, audio hardware, and filters (Vorländer, 2008). Several studies have examined the perceptual effects of system latency with respect to virtual environments but the consequences associated with position and orientation tracking error and latency during dynamic sound localization remain largely unknown. Available studies examining the effect of latency on sound localization are inconsistent (Brungart, Simpson, McKinley, Kordik, Dallman & Ovenshire, 2004). However, according to Wenzel (2001), localization remains accurate even with system latencies of up to 500 ms, although accuracy decreases slightly for shorter duration sounds particularly at higher latencies. Recent studies have found that head tracker latencies of 70 ms or less do not have a substantial impact on sound localization ability even with short duration sounds (Brungart, Kordik & Simpson, 2006; Brungart, Simpson, McKinley, Kordik, Dallman & Ovenshire, 2004). This of course does not imply latency can be completely ignored since there are other tasks, such as tracking a virtual sound source, where latency is critical. In an immersive virtual environment where visual imagery and auditory cues are both present, differences in the latency requirements of the two systems exist. The reason is that the perception of an audio/visual event as asynchronous is more easily detected when the audio precedes the video (Dixon & Spitz, 1980).

3.2 Modeling the room impulse response (RIR)

There are two major approaches to computationally modeling the RIR i) *wave-based modeling* where numerical solutions to the wave equation are used to compute the RIR, and ii) *geometric modeling* where sound is approximated as a ray phenomenon and traced through the scene to construct the RIR. Although the focus here is on recreating the acoustics of a particular environment by estimating the RIR, reverberation effects can be added synthetically through the use of artificial reverberation models. In their simplest form, synthetic techniques present the listener with delayed and attenuated versions of a sound source. These delays and attenuation factors do not necessarily represent the simulated physical properties of the environment. Rather, they are adjusted until a desirable effect is achieved. The approach is capable providing convincing late reverberation effects (Dattorro, 1997; Funkhouser, Tsingos, Carlbom, Elko, Sondhi, West, Pingali, Min & Ngan, 2004). Such techniques are widely used by the recording industry to add a pleasing "lively" aspect to voice and music and can convey a particular environmental setting (Warren, 1983). A discussion of artificial reverberation models is beyond the scope of this review. Further details can be found in (Ahnert & Feistel, 1993; Dattorro, 1997; Funkhouser, Tsingos, Carlbom, Elko, Sondhi, West, Pingali, Min & Ngan, 2004; Jot, 1992, 1997; Moorer, 1978; Schroeder, 1962).

3.2.1 Wave-based RIR modeling

The objective of wave-based methods is to solve the wave equation which is also known as the *Helmholtz-Kirchoff* equation (Tsingos, Carlbom, Elko, Funkhouser & Kubli, 2002), to recreate the RIR that models a particular sound field. An analytical solution to the wave equation is rarely feasible hence wave-based methods use numerical approximations such as finite element methods, boundary element methods, and finite difference time domain methods instead (Savioja, 1999). Numerical approximations sub-divide the boundaries of a room into smaller elements. 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 (Funkhouser, Tsingos, Carlbom, Elko, Sondhi, West, Pingali, Min & Ngan, 2004). The acoustical radiosity method, a modified version of the image synthesis radiosity technique, is an example of such an approach (Nosal, Hodgson & Ashdown, 2004; Shi, Zhang, Encarnacão & Göbel, 1993).

The numerical approximations associated with wave-based methods are computationally prohibitive making them impractical except for the simplest static environments. Furthermore, their computational complexity increases linearly with the volume of the room and the number of volume elements. Aside from basic or simple environments, such techniques are currently beyond our computational ability for interactive virtual environment applications.

3.2.2 Geometric (ray-based) acoustical modeling

Many acoustical modeling approaches adopt the hypothesis of “geometric acoustics” that assumes that sound and rays behave in a similar manner. The acoustics of an environment is then modeled by tracing (following) these “sound rays” as they propagate through the environment while accounting for any interactions between the sound rays and any objects/surfaces they may encounter. Mathematical models are used to account for sound source emission patterns, atmospheric scattering, the medium’s absorption of sound ray energy as a function of humidity, temperature, frequency, and distance (Bass, Bauer & Evans, 1972). At the receiver, the RIR is obtained by constructing an *echogram* which describes the distribution of incident sound energy (rays) at the receiver over time. The equivalent room impulse response can be obtained by post-processing the echogram (Kuttruff, 1993). Examples of geometric acoustic-based methods include *image sources* (Allen & Berkley, 1979), *ray tracing* (Krokstad, Strom & Sorsdal, 1968), *beam tracing* (Funkhouser, Tsingos, Carlbom, Elko, Sondhi, West, Pingali, Min & Ngan, 2004), *phonon tracing* (Bertram, Deines, Mohring, Jegorovs & Hagen, 2005) and *sonel mapping* (Kapralos, Jenkin & Milios, 2006).

Many ray-based methods assume that all interactions between a sound ray (wave) and objects/surfaces in the environment are specular in nature despite the fact that in natural settings other phenomena (e.g., diffuse reflections, diffraction, and refraction) influence a sound wave while it propagates through the environment. As a result, these methods are only valid for higher frequency sounds where reflections are primarily specular (Calamia &

Svensson, 2007). The wavelength of the sound waves and any phenomena associated with it, including diffraction, are typically ignored (Calamia, Svensson & Funkhouser, 2005; Kuttruff, 2000; Torres, Svensson & Kleiner, 2001; Tsingos, Funkhouser, Ngan & Carlbom, 2001).

One computational problem associated with ray-based approaches involves dealing with the large number of potential interactions between a propagating sound ray and the surfaces it may encounter. A sound incident on a surface may be simultaneously reflected specularly, reflected diffusely, be refracted, and be diffracted. Typical solutions to modeling such effects include the generation and emission of multiple "new" rays at each interaction point. Such approaches lead to exponential running times making them computationally intractable except for the most basic environments and only for very short time periods. An alternative to deterministic approaches to estimate the type of interaction between an acoustical ray and an incident surface are probabilistic approaches such as Russian roulette (Hammersley & Handscomb, 1964). Russian roulette was initially introduced to the field of particle physics simulation to terminate random paths whose contributions were estimated to be small. With a Russian roulette approach at each sound ray/surface interaction point only one interaction occurs probabilistically (e.g., the sound ray may be either absorbed, reflected specularly, reflected diffusely, etc.), based on the characteristics of the surface and the sound ray, and the value of a randomly generated number. In contrast to deterministic approaches whereby a sound ray is terminated when its energy has decreased beyond some threshold value or after it has been reflected a pre-set number of times, with Russian roulette the sound ray is terminated only when the interaction is determined to be absorption. This ensures that the path length of each sound ray is maintained at a manageable size yet due to its probabilistic nature, arbitrary size paths may be explored. Sonel mapping employs a Russian roulette solution in order to provide a computationally tractable solution to room acoustical modeling (Kapralos, Jenkin & Milios, 2005, 2006). Finally, with ray-based methods only a subset of the actual paths from the sound source to the listener are actually followed; certain paths may be missed altogether. To overcome this limitation, rather than emitting and tracing a single ray from the sound source, multiple rays bundled into a beam can be emitted and traced instead. Such an approach was first introduced by Whitted (1980) in the field of computer graphics and this technique has inspired various other approaches including cone tracing whereby a single ray is replaced by a cone (Amanatides, 1984), and beam tracing, which replaces a ray with a beam (Funkhouser, Tsingos, Carlbom, Elko, Sondhi, West, Pingali, Min & Ngan, 2004).

3.2.3 Diffraction modeling

Auralization methods based on geometric (ray) acoustics typically ignore wavelength and any associated phenomena including diffraction. A limited number of research efforts have investigated acoustical diffraction modeling. The beam tracing approach of Tsingos, Funkhouser, Ngan & Carlbom (2001) includes an extension capable of approximating diffraction. Their frequency domain method is based on the *uniform theory of diffraction* (UTD) (Keller, 1962). Tsingos and Gascuel (1997) developed an occlusion and diffraction auralization method that utilizes computer graphics hardware to perform fast sound visibility calculations accounting for specular reflections, absorption, and diffraction caused by partial occluders. In later work Tsingos and Gascuel (1998) introduced another occlusion and diffraction method based on the Fresnel-Kirchoff optics-based approximation to diffraction (Hecht, 2002). Similarly, sonel mapping also accounts for diffraction effects using a modified version of the Huygens-Fresnel principle (Kapralos, Jenkin & Milios, 2007). Calamia and Svensson (2007) describe an edge-subdivision strategy for interactive acoustical simulations that allows for fast time-domain edge diffraction calculations with relatively low error when compared with more numerically accurate solutions. Their approach allows for a trade-off between computation time and accuracy enabling the user to choose the necessary speed and the error tolerable for a specific modeling scenario. In contrast to the highly detailed physical approaches, Martens and Herder (1999) describe a perceptually-based solution to modeling the diffraction of sound.

3.3 Spherical microphone arrays

A viable alternative to the methods discussed above for generating three-dimensional sound is a technique that involves recording the sound field using an array of microphones and subsequently reproducing it with the ultimate goal of reconstructing the original sound field (Abhayapala & Ward, 2002; Meyer & Elko, 2002). Various microphone array configurations including linear, circular, and planar have well developed theoretical models. Microphone arrays have also been applied to various applications such as speech enhancement in conference rooms and auralization of sound fields measured in concert halls

(Rafaely, 2004). *Equiangle sampling* (Driscoll & Healy, 1994), *Gaussian sampling*, and *nearly uniform sampling* (Rafaely, 2005) represent available sampling approaches. Irrespective of the sampling technique utilized, in order to avoid aliasing, the sampling must be band-limited and the number of microphones required to sample up to the Nth-order harmonic of a signal must be $(N+1)^2$ (Rafaely, 2005). In theory, one can sample up to any order harmonic. However, due to the complexity associated with sampling second- and higher-order harmonics, sampling is typically restricted to measuring the zeroth- and first-order of a sound field. A system capable of recording second-order sound fields has only recently been introduced (Poletti, 2000). Abhayapala and Ward presented the theory (using spherical harmonics analysis) and guidelines for a higher-order system and provided an example of a third-order system for operation in the frequency range of 340 Hz to 3.4 kHz (Abhayapala & Ward, 2002). Rafaely (2005) presents a spherical-harmonics-based design and analysis for a spherical microphone array framework covering various factors including array order, input noise, microphone positioning, and spatial aliasing. Recording the sound field and reproducing it a later time is not a novel idea. In the early 1970's Ambisonics introduced a microphone technique that can be used to perform a synthesis of spatial audio (Furness, 1990).

4 Conveying sound to the user

Independent of the technology used to generate spatial sound, the generated sounds must be conveyed to the listener with some appropriate technology. The most common approaches are the use of either loudspeakers or headphones worn by the listener. Headphones and loudspeakers each have their respective advantages and disadvantages; either may produce more favorable results depending on the application. This section examines the delivery of spatial sound using both headphones and loudspeakers.

4.1 Headphone-based systems

Headphones provide a high level of channel separation thereby minimizing any *crosstalk* that arises when the signal intended for the left (or right) ear is also heard by the right (or left) ear. Headphones can also isolate the listener from external sounds and reverberation that may be present in the environment ensuring that the acoustics of the listening environment or the listener's position in the room does not affect the listener's perception (Gardner, 1998). Headphones typically deliver the auditory stimuli to the listener's ears through the air. The human auditory system is also sensitive to pressure wave propagation through the bones of the skull (Bekesy, 1960; Tonndorf, 1972). Bone conduction headsets which allow sound to be delivered to the user via direct application of vibrators to the skull, are small, comfortable, and provide the privacy and portability offered by traditional headphones. Moreover, they ensure that the pinna and ear canal remain unobstructed (Walker & Stanley, 2005). Generally, their use has been restricted to monaural applications, although investigations for their application in audio display designs is ongoing (Tonndorf, 1972; Walker & Stanley, 2005). While headphone-based systems offer potential benefits, there are shortcomings to their use as well. Headphones may be uncomfortable and cumbersome to wear, especially when worn for long periods. Additionally, unless the relevant spatial information is accounted for (e.g., inclusion of reverberation, and HRTFs), sounds conveyed through headphones will not be properly "externalized" but will rather be perceived as originating inside the head. This is referred to as *inside-the-head localization* (IHL).

Inside-the-head localization is the false impression that the sound originates from inside the listener's head. The sound is perceived as moving left and right inside the head along the interaural axis, with a bias towards the rear of the head (Kendall, 1995). 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 (Cohen & Wenzel, 1995).

IHL results from various factors, including the lack of a correct environmental context (e.g., lack of reverberation and HRTFs). IHL can be greatly reduced by ensuring the sounds delivered to the listener's ears reproduce the sound as it would be heard naturally. In other words, the listener should be provided with a "realistic spectral profile of the sound at each ear" (Semple, 1998). Although the externalization of a sound source is difficult to accurately predict it does increase the more "natural" the sound becomes (Begault, 1992). This of course implies some means of tracking the position and orientation of the listener's head and dynamically updating the HRTFs.

4.1.1 Headphone equalization

No headphone is perfect and its effects must be accounted for in the generation of an accurate three-dimensional audio display. This process is known as *headphone equalization*. The headphone transfer function represents both the characteristics of the headphone transducer itself as well as the transfer function between the headphone transducer and the ear drum (or at the point in the ear canal or outer ear where it was measured) (Kulkarni & Colburn, 2000). It is measured in a manner similar to measuring HRTFs but unlike the HRTF, the headphone transfer function does not vary as a function of sound source location. Once the transfer function has been obtained, equalization filters can be used to remove the effects of the headphone transfer function from headphone-conveyed sound. Møller (1992) provides a detailed description of headphone equalization.

The spectral features of the headphone transfer function can be significant and may contain peaks and notches with magnitude and bandwidth similar to the magnitude and bandwidth of the peaks and notches of HRTFs (Kulkarni & Colburn, 2000). However, there is an ongoing debate as to the influence of the headphone transfer function on localization. Studies have shown that the headphone transfer function can influence the resulting ITD due to group delays that vary between the ears and the placement location of the headphones. Studies have also shown that the headphone transfer function varies across individuals with substantially larger differences than those between individual HRTFs (Hammershøi & Møller, 2002; Møller, Hammershoi, Jensen & Sorensen, 1995). That being said, studies by Martin, Mcanally & Senova (2001) and Mcanally & Martin

(2002) that have investigated the use of a cochlear filter model to filter both HRTFs and headphone transfer functions, reveal that degradation of localization abilities is unlikely to result from differences in the transfer function arising from alternative headphone placements given that i) the variability of the magnitude of filtered HRTFs is considerably greater than the magnitude of filtered headphone transfer functions, and ii) group delays are considerably less than the minimum discriminable interaural time difference.

4.2 Loudspeaker-based systems

There are various loudspeaker-based systems that do not incorporate "true" three-dimensional sound technologies but have nevertheless found widespread use for entertainment purposes. Such systems include Quadraphonics$^{\text{TM}}$, Dolby Stereo$^{\text{TM}}$, Dolby Digital$^{\text{TM}}$, and Dolby 5.1$^{\text{TM}}$. Further details regarding such systems and recording/playback techniques in general are provided by Rumsey (2001). Since the emphasis here is on systems that aim to recreate a particular sound field as though in the natural setting, entertainment-based systems are not discussed further. Irrespective of the loudspeaker techniques employed, the intended effect is typically restricted to a small region of space known as the *listener sweet spot*. Deviation from this region causes serious degradations in system performance.

4.2.1 Transaural audio

In contrast to headphone-based systems, when loudspeakers are used there is no isolation between the signals intended for the left and right ears. In a typical two loudspeaker (stereo) scenario the signal received at the left and right ears 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 (Gardner, 1998). 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 will reach the right ear H_{LR}, while a delayed and attenuated portion of the right loudspeaker signal will reach the left ear H_{RL} (see Figure). *Transaural audio* is an example of a loudspeaker technique that employs crosstalk cancelation to ideally remove the unwanted crosstalk signals (Casey, Gardner & Basu, 1996).

Figure 6 about here.

Crosstalk cancelation Bauer (1961) first proposed crosstalk cancelation in order to allow for the delivery of HRTF-based (binaural) audio using a pair of loudspeakers. Two years later, Atal and Schroeder (1963) actually implemented the first crosstalk canceler. The Atal and Schroeder crosstalk canceler involves adding a delayed and inverted version of the crosstalk signal to the opposite loudspeaker output. In the process of combating crosstalk, distortion is introduced to the signal intended for each ear by adding the inverted and delayed signal emitted at the opposite loudspeaker. An additional round of crosstalk cancelation is thus necessary to eliminate the crosstalk associated with these signals. Fortunately, the corrections become smaller for each round of crosstalk cancelation and it is therefore possible to write a closed-form equation to compensate for the crosstalk (Kyriakakis, Tsakalides & Holman, 1999).

In theory, crosstalk cancelation completely removes the unwanted signals, thereby allowing the desired binaural signals to be delivered to the corresponding ears. However, in practice this is not the case. The signal introduced in the crosstalk cancelation process is a function of the listener's HRTF. As a result, its effectiveness is limited by the variability in size and shape of the human head and pinna (Gardner, 1998). The technique also has a small listener sweet spot; to function properly, the listener must remain stationary in the sweet spot as head movements as small as 74 - 100 mm completely destroy the desired effect (Mouchtaris, Reveliotis & Kyriakakis, 2000). As with headphone-based systems, this problem can be significantly reduced by tracking the listener's head. Gardner (1998) developed a system that utilized a magnetic head tracker in order to produce a realistic and larger range three-dimensional audio display using loudspeakers. Gardner's system offers improved localization over non-tracked loudspeaker displays as it allows for dynamic localization cues. Mouchtaris, Reveliotis & Kyriakakis (2000) describe a loudspeaker-based three-dimensional audio display which produces dynamic crosstalk cancelation using a camera-based head tracking system thus eliminating the tether associated with magnetic trackers.

4.2.2 Amplitude panning

In the *amplitude panning* technique, the amplitude (intensity or output level) of the signal being delivered to each loudspeaker (or headphone) is adjusted in some manner to simulate the directional properties of the ILD. By adjusting the amplitude of the signal applied to each loudspeaker through the use of a gain factor, the listener perceives a virtual sound source emanating from a direction that is dependent on the gain factors (Pulkki, 2001). Amplitude panning techniques allow for a wide variety of loudspeaker set-ups including both two- and three-dimensional configurations. The general idea is to compute the appropriate gain factors for each loudspeaker to create the impression of a virtual sound source at a specific position relative to the listener. The *stereophonic law-of-sines* (Bauer, 1961), and the *tangent law* (Bennett, Barker & Edeko, 1985) can be used to compute the gain for each channel in the typical two-channel (stereo) configuration. Using *pair-wise amplitude panning* techniques (Chowning, 1971), the two-channel methods can be extended to N loudspeakers by choosing and outputting the sound signal to pairs of loudspeakers simultaneously in a manner similar to the conventional two-channel panning technique. Three-dimensional panning is an extension of the two-channel, two-dimensional technique. Sound is applied to a subset of three loudspeakers only and a virtual sound source is positioned anywhere on the triangle formed by the three loudspeakers. Currently, no general trigonometric method of three-dimensional amplitude panning for an arbitrary three-dimensional loudspeaker setup exists and the calculation of the gains applied to the loudspeakers is configuration dependent (Pulkki, 2001).

Vector base amplitude panning A more recent method of calculating the gain factors is the *vector base amplitude panning* (VBAP) technique (Pulkki, 1997, 2001). This technique can be used with an arbitrary number of loudspeakers and supports both two- and three-dimensional loudspeaker configurations. It allows the loudspeakers to be placed in any position provided that they are nearly equidistant from the listener and that the listening room is not overly reverberant.

In the stereo VBAP configuration the two-channel stereo setup is treated as a two-dimensional vector base defined by two unit length vectors, each vector with its origin at the listener and pointing to one of the two loudspeakers. A third unit vector points to the direction of the virtual sound source and is formulated as a linear combination of the two loudspeaker vectors. The two loudspeaker scaling factors (gains) are calculated using simple linear algebra techniques. The formulation of two-dimensional VBAP can be generalized to handle a three-dimensional loudspeaker configuration where three equidistant loudspeakers are conceptualized as positioned on an imaginary unit radius sphere. Three loudspeaker unit vectors point from the listener's position to each of the three loudspeakers, and a fourth unit vector points to the position of the virtual sound source. The virtual sound source can then be mapped to a location within the "active triangle" formed by the three loudspeakers (see Figure). As with the two-dimensional stereo configuration, the vector pointing to the virtual sound source is expressed as a linear combination of the three loudspeaker vectors and the appropriate gain is calculated (using simple linear algebra techniques) and used to scale the signal output to each loudspeaker.

Figure 7 about here.

The VBAP technique is a relatively simple and computationally efficient method allowing for the maximum virtual sound source localization accuracy possible with amplitude panning (Pulkki, 2001). In the three-dimensional configuration, maximum localization accuracy is proportional to the physical dimensions of the active triangle (Pulkki, 2001).

Although the dimension of the active triangle can be decreased by increasing the number of loudspeakers, increasing the number of loudspeakers is sometimes impossible. As with all pair-wise and triplet-wise amplitude panning techniques the virtual sound source spreads when it is panned between loudspeakers. Finally, although VBAP allows for accurate virtual sound source localization on the azimuthal plane (particularly near the median plane), the localization of virtual sound sources which do not lie on the azimuthal plane (e.g., non-zero elevation) is unpredictable since it is listener dependent. However, with a large number of loudspeakers elevation localization becomes acceptable (Pulkki & Karjalainen, 2001).

4.2.3 Wave field synthesis

The *wave field synthesis* (WFS) method involves audio signals fed to a large number of closely-spaced loudspeakers so that a highly natural sound field is produced, including the reproduction of the wavefront curvature that would result from real sound sources (Berkhout, de Vries & Vogel, 1993; Boone, de Vries & van Tol, 1995). Thus the WFS method allows for the simultaneous reproduction of an arbitrary number of virtual sound sources (Berkhout, de Vries & Vogel, 1993). WFS is based on Huygens' principle which states that at every time instant every point on a primary wavefront can be thought of as a continuous emitter of secondary wavelets combining to produce a new wavefront in the direction of propagation. Given a wave field (that is specified regarding pressure and normal particle velocity) on a boundary surface S of a closed volume V free of any sources, the sound pressure at any point within V can be determined. Loudspeakers that surround the listening area are driven to produce a volume flux proportional to the normal component of the particle velocity of the original wave field at each corresponding position (Boone & de Bruijn, 2000). For practical purposes (e.g., hardware and computational power requirements) rather than using multiple planes of loudspeakers to enclose the listener, linear loudspeaker arrays are used. This leads to several problems, most notable of which is that sound reproduction is correct for wave field components in the horizontal plane only (Boone, 2001).

Unlike other loudspeaker-based systems whose intended effect is restricted to the listener sweet spot, WFS systems generate a wave field with natural time and space properties that envelops the listening area (de Vries & Boone, 2004). Multiple listeners are free to move about within this area without fear of losing the correct acoustical impression. This has made WFS an attractive approach for applications such as sound enhancement in theaters, multi-purpose auditoria, and the reproduction of multi-channel recordings (de Vries & Boone, 2004). However, WFS is impractical in many virtual reality settings due to several inherent limitations, most notable, the requirement that the distance between loudspeakers be as small as possible in order to avoid spatial aliasing; the highest frequency that can be represented is inversely proportional to the spacing between loudspeakers (Pulkki, 2001; Verheijen, 1998). This results in the requirement for a large number of loudspeakers and extensive computation.

5 Concluding remarks

Human sound localization is an extremely sophisticated process. The sound field itself is the result of the complex interactions between sound waves and objects in the environment. The listener transduces these sound waves in very sophisticated ways and our perception of these waves is finely tuned to subtle acoustical effects. Correctly simulating these complexities is an arduous task. Nevertheless, over the past few decades the field of virtual audio has progressed steadily and promising technologies have emerged. For example, Sound Lab (SLAB) is an object-oriented software-based virtual acoustical environment that allows for real-time virtual audio rendering on a standard Windows-based PC (Miller & Wenzel, 2002). Spat is a real-time modular spatial sound processing software system that allows for the reproduction and control of localized sound sources in three-dimensional space (Jot, 1999). Doerr, Rademacher, Huesgen & Kubbat (2007) describe a low cost software-based three-dimensional audio system capable of providing basic three-dimensional sound information to users wearing headphones and equipped with a head tracker. Jin, Tan, Kan, Lin, von Schaik et al. (2005) describe a three-dimensional audio playback system that employs head-tracking with an unlimited number of simultaneous sound sources. Their method relies on the use of a 500 - 900 MBytes/s sound buffer that contains pre-processed HRTF data for 385 (closely-spaced) head orientations which can be presented to the user at interactive rates. Despite the march of progress, considerable research and development remains to be done to facilitate the generation of convincing virtual sound for use in interactive real-time virtual environments.

Acknowledgments

The thorough review of the manuscript and corresponding comments and suggestions by Dr. Douglas S. Brungart and two anonymous reviewers is very much appreciated. The financial support of the Natural Science and Engineering Research Council of Canada (NSERC)is gratefully acknowledged. The authors also thank Antonin Pribetic for his editorial assistance.

References

Abhayapala, T. D. & Ward, D. B. (2002). Theory and design of high order sound field microphones using spherical microphone array. In *Proceedings of the 2002 International Conference on Acoustics Speech and Signal Processing* (pp. 1949–1952). Orlando, FL. USA.

Ahnert, W. & Feistel, R. (1993). EARS auralization software. *Journal of the Audio Engineering Society*, *41(11)*, 894–904.

Algazi, V. R., Avendano, C. & Thompson, D. (1999). Dependence of subject and measurement position in binaural signal acquisition. *Journal of the Audio Engineering Society*, *47(11)*, 937–947.

Algazi, V. R., Duda, R. O. & Thompson, D. M. (2004). Motion-tracked binaural sound. *Journal of the Audio Engineering Society*, *52(11)*, 1142–1156.

Algazi, V. R., Duda, R. O., Thompson, D. M. & Avendano, C. (2001). The CIPIC HRTF database. In *2001 IEEE Workshop on Applications of Signal Processing to Acoustics* (pp. 111–123). New Paltz, NY. USA.

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

Allison, R. S., Harris, L. R., Jenkin, M., Jasiobedzka, I. & Zacher, J. E. (2001). Tolerance of temporal delay in virtual environments. In *Proceedings of the IEEE Conference on Virtual Reality* (pp. 247 – 254). Yokohama, Japan.

Amanatides, J. (1984). Ray tracing with cones. In *Proceedings of the 11th Annual Conference on Computer Graphics and Interactive Techniques (SIGGRAPH 1984)* (pp. 129–135). Minneapolis, MN. USA.

Atal, B. S. & Schroeder, M. R. (1963). Apparent sound source translator. U.S. Patent $3,236,949$.

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

Batteau, D. W. (1967). The role of the pinna in human localization. *Proceedings of the Royal Society of London*, *168(11)*, 158–180.

Bauer, B. (1961). Phasor analysis of some stereophonic phenomena. *Journal of the Acoustical Society of America*, *33(11)*, 1536–1539.

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

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

Begault, D. R. & Wenzel, E. M. (1993). Headphone localization of speech. *Human Factors*, *35(2)*, 361–376.

Begault, R. (1994). *3-D Sound for Virtual Reality and Multimedia.* Cambridge, MA. USA: Academic Press Professional.

Bekesy, G. V. (1960). *Experiments in Hearing.* New York, NY. USA: McGraw Hill.

Bennett, J. C., Barker, K. & Edeko, F. O. (1985). A new approach to the assessment of stereophonic sound system performance. *Journal of the Audio Engineering Society*, *33(5)*, 314–321.

Berkhout, A. J., de Vries, D. & Vogel, P. (1993). Acoustic control by wave field synthesis. *Journal of the Acoustical Society of America*, *93(5)*, 2764–2778.

Bertram, M., Deines, E., Mohring, J., Jegorovs, J. & Hagen, H. (2005). Phonon tracing for auralization and visualization of sound. In *Proceedings of IEEE Visualization 2005* (pp. 151–158). Minneapolis, MN. USA.

Blauert, J. (1996). *The Psychophysics of Human Sound Localization* (Revised Ed.). Cambridge, MA. USA: MIT Press.

Boone, M. M. (2001). Acoustic rendering with wave field synthesis. In *Proceedings of the ACM SIGGRAPH and Eurographics Campfire on Acoustic Rendering for Virtual Environments* (pp. 37–45). Snowbird, UT. USA.

Boone, M. M. & de Bruijn, W. P. J. (2000). On the applicability of distributed mode loudspeaker panels for wave field synthesis based sound reproduction. In *Proceedings of the 108th Convention of the Audio Engineering Society.* Paris, France.

Boone, M. M., de Vries, D. & van Tol, P. F. (1995). Spatial sound field reproduction by wave field synthesis. *Journal of the Audio Engineering Society*, *43(12)*, 1003–1012.

Bronkhorst, A. W. & Houtgast, T. (1999). Auditory distance perception in rooms. *Nature*, *397(6719)*, 517–520.

Brungart, D. S. (1998). Control of perceived distance in virtual audio displays. In *Proceedings of the 20th Annual International Conference of the IEEE Engineering in Medicine and Biology Society.* (pp. 1101–1104). Hong Kong, China.

Brungart, D. S., Kordik, A. J. & Simpson, B. D. (2006). Effects of headtracker latency in virtual audio displays. *Journal of the Audio Engineering Society*, *54(1/2)*, 32–44.

Brungart, D. S. & Rabinowitz, W. M. (1999). Auditory localization of nearby sources. Head related transfer functions. *Journal of the Acoustical Society of America*, *106(3)*, 1465–1479.

Brungart, D. S. & Scott, K. R. (2001). The effects of production and presentation level on the auditory distance perception of speech. *Journal of the Acoustical Society of America*, *110(1)*, 425–440.

Brungart, D. S., Simpson, B. D., McKinley, R. L., Kordik, A. J., Dallman, R. C. & Ovenshire, D. A. (2004). The interaction between head-tracker latency, source duration, and response time in the localization of virtual sound sources. In *Proceedings of the 2004 International Conference on Auditory Display* (pp. 1–7). Sydney, Australia.

Calamia, P. T. & Svensson, U. P. (2007). Fast time-domain edge-diffraction calculations for interactive acoustic simulations. *EURASIP Journal on Applied Signal Processing, Special Issue on Spatial Sound and Virtual Acoustics*, *2007*, Article ID 63560, 10 pages. doi:10.1155/2007/63560.

Calamia, P. T., Svensson, U. P. & Funkhouser, T. A. (2005). Integration of edge-diffraction calculations and geometrical-acoustics modeling. In *Proceedings of Forum Acusticum 2005*. Budapest, Hungary.

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

Carlile, S., Jin, C. & Raad, V. V. (2000). Continuous virtual auditory space using HRTF interpolation: Acoustic and psychophysical errors. In *Proceedings of the First IEEE Pacific-Rim Conference on Multimedia* (pp. 220–223). Sydney, Australia.

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

Cheung, N. M., Trautmann, S. & Horner, A. (1998). Head-related transfer function modeling in 3-D sound systems with genetic algorithms. *Journal of the Audio Engineering Society*, *46(6)*, 531–539.

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

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

Cohen, M. & Wenzel, E. (1995). The design of multidimensional sound interfaces. In W. Barfield & T. Furness (Eds.), *Virtual Environments and Advanced Interface Design* chapter 8, (pp. 291–346). New York, NY. USA: Oxford University Press Inc.

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

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

Dixon, N. F. & Spitz, L. (1980). The detection of auditory visual desynchrony. *Perception*, *9(6)*, 719–721.

Doerr, K., Rademacher, H., Huesgen, S. & Kubbat, W. (2007). Evaluation of a low-cost 3D sound system for immersive virtual reality training systems. *IEEE Transactions on Visualization and Computer Graphics*, *13(2)*, 204–212.

Driscoll, J. R. & Healy, D. M. (1994). Computing Fourier transforms and convolutions on the 2-sphere. *Advances in Applied Mathematics*, *15(2)*, 202–250.

Duda, R. O. (1993). Modeling head related transfer functions. In *Proceedings of the 27th Conference on Signals, Systems and Computers*. Alisomar, CA. USA.

Foxlin, E. (2002). Motion tracking requirements and technologies. In K. Stanney (Ed.), *Handbook of Virtual Environment Technology* chapter 8, (pp. 163–210). Mahwah, NJ. USA: Lawrence Erlbaum Associates Inc.

Freeland, F. P., Biscainho, L. P. & Diniz, P. R. (2004). Interpositional transfer function for 3D-sound generation. *Journal of the Audio Engineering Society*, *52(9)*, 915–930.

Freeland, F. P., Wagner, L., Biscainho, P. & Dinz, P. R. (2002). Efficient HRTF interpolation in 3D moving sound. In *Proceedings of the Audio Engineering Society 22nd International Conference on Virtual, Synthetic and Entertainment Audio* (pp. 106–114). Espoo, Finland.

Funkhouser, T., Tsingos, N., Carlbom, I., Elko, G., Sondhi, M., West, J. E., Pingali, G., Min, P. & Ngan, A. (2004). A beam tracing method for interactive architectural acoustics. *Journal of the Acoustical Society of America*, *115(2)*, 739–756.

Furness, R. K. (1990). Ambisonics – an overview. In *Proceedings of the 8th International Conference of the Audio Engineering Society* (pp. 181–189). Washington, DC. USA.

Garas, J. (2000). *Adaptive 3D Sound Systems*. Norwell, MA. USA: Kluwer Academic Publishers.

Gardner, M. B. (1968). Distance estimation of 0° or apparent 0° oriented speech signals in anechoic space. *Journal of the Acoustical Society of America*, *45(1)*, 47–53.

Gardner, W. (1998). *3-D Audio Using Loudspeakers*. Norwell, MA. USA: Kluwer Academic Publishers.

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

Grassi, E., Tulsi, J. & Shamma, S. (2003). Measurement of head-related transfer functions based on the empirical transfer function estimate. In *Proceedings of the 2003 International Conference on Auditory Display* (pp. 119–122). Boston, MA. USA.

Hacihabiboğlu, H. & Murtagh, F. (2006). Perception-based simplification for binaural room auralization. In *Proceedings of the 12th International Conference on Auditory Display* (pp. 268–271). London, UK.

Hammershøi, D. & Møller, H. (2002). Methods for binaural recording and reproduction. *Acta Acustica*, *88(3)*, 302–311.

Hammersley, J. M. & Handscomb, D. C. (1964). *Monte Carlo Methods*. New York, NY. USA: Chapman and Hall.

Hartmann, W. (1999). How we localize sound. *Physics Today* (pp. 24–29). http://www.aip.org/pt/nov99/locsound.html.

Hebrank, J. & Wright, D. (1974). Spectral cues used in the localization of sound sources in the median plane. *Journal of the Acoustical Society of America*, *56(6)*, 1829–1834.

Hecht, E. (2002). *Optics* (4 Ed.). San Francisco, CA. USA: Pearson Education Inc.

Ircam & AKG Acoustics (2002). LISTEN HRTF database. http://www.ircam.fr/equipes/salles/listen/index.html.

Jin, C., Leong, P., Leung, J., Corderoy, A. & Carlile, S. (2003). Enabling individualized virtual auditory space using morphological measurements. In *Proceedings of the First IEEE Pacific-Rim Conference on Multimedia* (pp. 235–238). Sydney, Australia.

Jin, C., Tan, T., Kan, A., Lin, D., von Schaik, A., Smith, K. & McGinity, M. (2005). Real-time, head-tracked 3D audio with unlimited simultaneous sounds. In *Proceedings of the 2005 International Conference on Auditory Display (ICAD)* (pp. 1–4). Limerick, Ireland.

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

Jot, J. M. (1997). Efficient models for reverberation and distance rendering in computer music and virtual audio reality. In *Proceedings of the 1997 International Computer Music Conference*. Thessaloniki, Greece.

Jot, J. M. (1999). Real-time spatial processing of sounds for music, multimedia and interactive human-computer interfaces. *Multimedia Systems*, *7(1)*, 55–69.

Jot, J. M., Cerveau, L. & Warusfel, O. (1997). Analysis and synthesis of room reverberation based on a statistical time-frequency model. In *Proceedings of 103rd Convention of the Audio Engineering Society*. New York, NY. USA.

Kapralos, B., Jenkin, M. & Milios, E. (2005). Acoustical modeling using a Russian roulette strategy. In *Proceedings of the 118th Convention of the Audio Engineering Society*. Barcelona, Spain.

Kapralos, B., Jenkin, M. & Milios, E. (2006). Sonel mapping: A stochastic acoustical modeling system. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*. Toulouse, France.

Kapralos, B., Jenkin, M. & Milios, E. (2007). Diffraction modeling for interactive virtual acoustical environment. In *Proceedings of the 2nd International Conference on Computer Graphics Theory and Applications (GRAPP) 2007*. Barcelona, Spain.

Kapralos, B. & Mekuz, N. (2007). Application of dimensionality reduction techniques to HRTFs for interactive virtual environments. In *Proceedings of the ACM Advances in Computer Entertainment (ACE) 2007* (pp. 268–271). Salzburg, Austria.

Keller, J. B. (1962). Geometrical theory of diffraction. *Journal of the Optical Society of America*, *52(2)*, 116–130.

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

King, R. B. & Oldfield, S. R. (1997). The impact of signal bandwidth on auditory localization: Implications for the design of three-dimensional audio displays. *Human Factors*, *39(2)*, 287–295.

Kistler, D. J. & Wightman, F. L. (1992). A model of head-related transfer functions based on principle components analysis and minimum phase reconstruction. *Journal of the Acoustical Society of America, 91(3)*, 1637–1647.

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

Krokstad, A., Strom, S. & Sorsdal, S. (1968). Calculating the acoustical room response by the use of a ray tracing technique. *Journal of Sound and Vibration, 8(1)*, 118–125.

Kulkarni, A. & Colburn, H. S. (1993). Evaluation of a linear interpolation scheme for approximating HRTFs. *Journal of the Acoustical Society of America, 93(4)*, 2350.

Kulkarni, A. & Colburn, H. S. (1998). Role of spectral detail in sound-source localization. *Nature, 396(6713)*, 747–749.

Kulkarni, A. & Colburn, H. S. (2000). Variability in the characterization of the headphone transfer-function. *Journal of the Acoustical Society of America, 107(2)*, 1071–1074.

Kulkarni, A., Isabelle, S. K. & Colburn, H. S. (1995). On the minimum-phase approximation of head-related transfer functions. In *Proceedings of the IEEE Workshop on the Applications of Signal Processing to Audio and Acoustics* (pp. 84–87). New Paltz, NY. USA.

Kulkarni, A., Isabelle, S. K. & Colburn, H. S. (1999). Sensitivity of human subjects to head-related transfer-function phase spectra. *Journal of the Acoustical Society of America, 105(5)*, 2821–2840.

Kuttruff, H. (2000). *Room Acoustics* (Fourth Ed.). London, England: Spon Press.

Kuttruff, K. H. (1993). Auralization of impulse responses modeled on the basis of ray-tracing results. *Journal of the Audio Engineering Society*, *41(11)*, 876–880.

Kyriakakis, C., Tsakalides, P. & Holman, T. (1999). Surrounded by sound. *IEEE Signal Processing Magazine*, *16(1)*, 55–66.

Martens, W. L. (1987). Principal components analysis and resynthesis of spectral cues to perceived direction. In *Proceedings of the 1987 International Computer Music Conference* (pp. 274–281). Champaine-Urbana, IL. USA.

Martens, W. L. (2000). Efficient auralization of small, cluttered spaces: Simulating sonic obstructions at close range. In Kuwano, S. & T.Kato (Eds.), *Proceedings of the 7th Western Pacific Regional Acoustics Conference* (pp. 317–320). Kumamoto, Japan.

Martens, W. L. & Herder, J. (1999). Perceptual criteria for eliminating reflectors and occluders from the rendering of environmental sound. In *137th Meeting of the Acoustical Society of America and the 2nd Convention of the European Acoustics Association* (p. S53). Berlin.

Martin, R. L., Mcanally, K. I. & Senova, M. A. (2001). Free-field equivalent localization of virtual audio. *Journal of the Audio Engineering Society*, *49(1/2)*, 14–22.

Vorländer, M. (2008). *Auralization.* Berlin, Germany: Springer-Verlag.

Mcanally, K. I. & Martin, R. L. (2002). Variability in the headphone-to-ear-canal transfer function. *Journal of the Audio Engineering Society*, *50(4)*, 263–266.

Meyer, J. & Elko, G. W. (2002). A spherical microphone array for spatial sound recording. *Journal of the Acoustical Society of America*, *111(5)*, 2346.

Middlebrooks, J. C. (1992). Narrow-band sound localization related to external ear acoustics. *Journal of the Acoustical Society of America*, *92(5)*, 2607–2624.

Middlebrooks, J. C. (1999a). Individual differences in external-ear transfer functions reduced by scaling in frequency. *Journal of the Acoustical Society of America*, *106(2)*, 1480–1492.

Middlebrooks, J. C. (1999b). Virtual localization improved by scaling nonindividualized external-ear transfer functions in frequency. *Journal of the Acoustical Society of America*, *106(2)*, 1493–1510.

Middlebrooks, J. C. & Green, D. M. (1990). Directional dependence of interaural envelope delays. *Journal of the Acoustical Society of America*, *87(5)*, 2149–2162.

Middlebrooks, J. C., Macpherson, E. W. & Onsan, Z. A. (2000). Psychophysical customization of directional transfer functions. *Journal of the Acoustical Society of America*, *108(6)*, 3088–3091.

Miller, J. D. & Wenzel, W. E. (2002). Recent developments in SLAB: A software-based system for interactive spatial sound synthesis. In *Proceedings of the 2002 International Conference on Auditory Display* (pp. 403–408). Kyoto, Japan.

Mills, A. W. (1958). On the minimum audible angle. *Journal of the Acoustical Society of America*, *30(4)*, 237–246.

Mills, W. (1972). Auditory localization. In J. V. Tobias (Ed.), *Foundations of Modern Auditory Theory*, Volume 2 (pp. 303–348). New York, NY. USA: Academic Press.

Møller, H. (1992). Fundamentals of binaural technology. *Applied Acoustics*, *36(3/4)*, 171–218.

Møller, H., Hammershoi, D., Jensen, C. B. & Sorensen, M. F. (1995). Transfer characteristics of headphones measured on human ears. *Journal of the Audio Engineering Society*, *43(4)*, 203–217.

Moore, B. C. J. (1989). *An Introduction to the Psychology of Hearing.* San Diego, CA. USA: Academic Press Limited.

Moorer, J. A. (1978). About this reverberation business. *Computer Music Journal*, *3(2)*, 13–28.

Morse, P. M. & Ingard, K. U. (1968). *Theoretical Acoustics.* Princeton, NJ. USA: Princeton University Press.

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

Naguib, M. & Wiley, H. (2001). Estimating the distance to a sound: Mechanisms and adaptations for long-range communications. *Animal Behavior*, *62(5)*, 825–837.

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

Nosal, E., Hodgson, M. & Ashdown, I. (2004). Improved algorithms and methods for room sound-field prediction by acoustical radiosity in arbitrary polyhedral rooms. *Journal of the Acoustical Society of America*, *116(2)*, 970–980.

Perrett, S. & Noble, W. (1995). Available response choices affect localization of sound. *Perception and Psychophysics*, *57(2)*, 150–158.

Perrett, S. & Noble, W. (1997). The effect of head rotations on vertical plane sound localization. *Journal of the Acoustical Society of America*, *102(4)*, 2325–2332.

Perrott, D. R. & Saberi, K. (1990). Minimum audible angle thresholds for sources varying in both elevation and azimuth. *Journal of the Acoustical Society of America*, *87(4)*, 1728–1731.

Poletti, M. A. (2000). A unified theory of horizontal holographic sound systems. *Journal of the Audio Engineering Society*, *48(12)*, 1155–1182.

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

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

Pulkki, V. & Karjalainen, M. (2001). Directional quality of 3-D amplitude panned virtual sources. In *Proceedings of the 2001 International Conference on Auditory Display* (pp. 239–244). Espoo, Finland.

Rafaely, B. (2004). Plane-wave decomposition of the sound field on a sphere by spherical convolution. *Journal of the Acoustical Society of America*, *116(4)*, 2149–2157.

Rafaely, B. (2005). Analysis and design of spherical microphone arrays. *IEEE Transactions on Speech and Audio Processing*, *13(1)*, 135–143.

Rakerd, B. & Hartmann, W. M. (1985). Localization of sound in rooms, II: the effects of a single reflecting surface. *Journal of the Acoustical Society of America*, *78(2)*, 524–533.

Roffler, S. K. & Butler, R. A. (1968a). Factors that influence the localization of sound in the vertical plane. *Journal of the Acoustical Society of America*, *43(6)*, 1255–1259.

Roffler, S. K. & Butler, R. A. (1968b). Localization of tonal stimuli in the vertical plane. *Journal of the Acoustical Society of America*, *43(6)*, 1260–1266.

Rolland, J. P., Davis, L. & Baillot, Y. (2001). A survey of tracking technology for virtual environments. In W. Barfield & T. Caudell (Eds.), *Fundamentals of Wearable Computers and Augmented Reality* (pp. 67–112.). Mahwah, NJ. USA: Lawrence Erlbaum Associates Inc.

Rumsey, F. (2001). *Spatial Audio.* Woburn, MA. USA: Focal Press.

Saberi, K. & Perrott, D. R. (1990). Laterization thresholds in which the precedence effect is asusmed to operate. *Journal of the Acoustical Society of America*, *87(4)*, 1732–1737.

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

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

Semple, M. N. (1998). Sounds in a virtual world. *Nature*, *396(6713)*, 723–724.

Shi, J., Zhang, A., Encarnacão, J. & Göbel, M. (1993). A modified radiosity algorithm for integrated visual and auditory rendering. *Computers and Graphics*, *17(6)*, 633–642.

Shilling, R. D. & Shinn-Cunningham, B. (2002). Virtual auditory displays. In K. Stanney (Ed.), *Handbook of Virtual Environment Technology* (pp. 65–92). Mahwah, NJ. USA: Lawrence Erlbaum Associates.

Shinn-Cunningham, B. G. (2000a). Distance cues for virtual auditory space. In *Proceedings of the IEEE 2000 International Symposium on Multimedia Information Processing* (pp. 227–230). Sydney, Australia.

Shinn-Cunningham, B. G. (2000b). Learning reverberation: Considerations for spatial auditory displays. In *Proceedings of the 2000 International Conference on Auditory Display (ICAD)* (pp. 126–134). Atlanta, GA. USA.

Slattery, W. H. & Middlebrooks, J. C. (1984). Monaural sound localization: Acute versus chronic unilateral impairment. *Hearing Research*, *75(1-2)*, 38–46.

Strutt, J. W. (1907). On our perception of sound direction. *Philosophical Magazine*, *13*, 214–232.

Thurlow, W. R., Mangels, J. W. & Runge, P. S. (1967). Head movements during sound localization. *Journal of the Acoustical Society of America*, *42(2)*, 489–493.

Tonndorf, J. (1972). Bone conduction. In J. V. Tobias (Ed.), *Foundations of Modern Auditory Theory* (pp. 195–237). New York, NY. USA: Academic Press Inc.

Torres, R. R., Svensson, P. & Kleiner, M. (2001). Computation of edge diffraction for more accurate room acoustics auralization. *Journal of the Acoustical Society of America*, *109(2)*, 600–610.

Tsingos, N., Carlbom, I., Elko, G., Funkhouser, T. & Kubli, B. (2002). Validation of acoustical simulations in the "Bell Labs Box". *IEEE Computer Graphics and Applications*, *22(4)*, 28–37.

Tsingos, N., Funkhouser, T., Ngan, A. & Carlbom, I. (2001). Modeling acoustics in virtual environments using the uniform theory of diffraction. In *Proceedings of the 28th Annual Conference on Computer Graphics and Interactive Techniques (SIGGRAPH 2001)* (pp. 545–552).

Tsingos, N. & Gascuel, J. (1998). Fast rendering of sound occlusion and diffraction effects for virtual acoustic environments. In *104th Convention of the Audio Engineering Society* (pp. 1–14). Amsterdam, The Netherlands.

Tsingos, N. & Gascuel, J. D. (1997). A general model for the simulation of room acoustics. In *Proceedings of the 24th Annual Conference on Computer Graphics and Interactive Techniques (SIGGRAPH 2001).* Los Angeles, CA USA.

Verheijen, E. N. G. (1998). *Sound Reproduction by Wave Field Synthesis.* PhD thesis, Technical University Delft, The Netherlands.

Väänänen, R. (2003). *Parameterization, Auralization and Authoring of Room Acoustics for Virtual Reality Applications.* PhD thesis, Helsinki University of Technology, Helsinki, Finland.

de Vries, D. & Boone, M. M. (2004). Wave field synthesis and analysis using array technology. In *Proceedings of the 1999 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics* (pp. 15–18). New Paltz, NY. USA.

Walker, B. N. & Stanley, R. M. (2005). Evaluation of bone-conduction headsets for use in multitalker communication environments. In *Proceedings of the Human Factors and Ergonomics Society 49th Annual Meeting* (pp. 1615–1619). Orlando, FL. USA.

Wallach, H. (1940). The role of head movements and vestibular and visual cues in sound localization. *Experimental Psychology*, *27(4)*, 339–368.

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

Wenzel, E. M. (1999). Effects of increasing system latency on localization of virtual sounds. In *Proceedings of the Audio Engineering Society 16th International Conference on Spatial Sound Reproduction* (pp. 42–50). Rovaniemi, Finland.

Wenzel, E. M. (2001). Effect of increasing system latency on localization of virtual sounds with short and long duration. In *Proceedings of the 2001 International Conference on Auditory Display* (pp. 185–190). Espoo, Finland.

Wenzel, E. M., Arruda, M. & Kistler, D. J. (1993). Localization using nonindividualized head-related transfer functions. *Journal of the Acoustical Society of America*, *94(1)*, 111–123.

Wenzel, E. M. & Foster, S. H. (1993). Perceptual consequences of interpolating head-related transfer functions during spatial synthesis. In *Proceedings of the IEEE Workshop on Applications of Signal Processing to Audio and Acoustics* (pp. 102–105). New Paltz, NY. USA.

Wenzel, E. M., Wightman, F. L. & Kistler, D. J. (1988). Acoustic origins of individual differences in sound localization behavior. *Journal of the Acoustical Society of America*, *84(S1)*, S79.

Wenzel, E. W., Miller, J. D. & Abel, J. S. (2000a). A software-based system for interactive spatial sound synthesis. In *Proceedings of the 2000 International Conference on Auditory Display* (pp. 151–156). Atlanta, GA. USA.

Wenzel, E. W., Miller, J. D. & Abel, J. S. (2000b). Sound Lab: A real-time, software-based system for the study of spatial hearing. In *Proceedings of the 108th Convention of the Audio Engineering Society*, Preprint 5140. Paris, France.

Whitted, T. (1980). An improved illumination model for shaded display. *Communications of the ACM*, *23(6)*, 343–349.

Wightman, F. L. & Kistler, D. J. (1989). Headphone simulation of free-field listening. I: Stimulus synthesis. *Journal of the Acoustical Society of America*, *85(2)*, 858–867.

Wightman, F. L. & Kistler, D. J. (1993). Sound localization. In W. Yost, A. Popper & R. Fay (Eds.), *Springer Handbook of Auditory Research: Human Psychophysics*, Volume 3 (pp. 155–192). New York NY. USA: Springer-Verlag Inc.

Wightman, F. L. & Kistler, D. J. (1997). Factors affecting the relative salience of sound localization cues. In R. H. Gilkey & T. R. Anderson (Eds.), *Binaural and Spatial Hearing in Real and Virtual Environments* chapter 1, (pp. 1–23). Mahwah, NJ. USA: Lawrence Erlbaum Associates.

Zahorik, P., Wightman, F. & Kistler, D. (1995). On the discriminability of virtual and real sound sources. In *Proceedings of the IEEE Workshop on the Applications of Signal Processing to Audio and Acoustics* (pp. 76–79). New Paltz, NY. USA.

Zotkin, D. N., Duraiswami, R. & Davis, L. (2004). Rendering localized spatial audio in a virtual auditory space. *IEEE Transactions on Multimedia*, *6(4)*, 553–564.

Zotkin, D. N., Duraiswami, R., Grassi, E. & Gumerov, N. A. (2006). Fast head-related transfer function measurement via reciprocity. *Journal of the Acoustical Society of America*, *120(4)*, 2202–2215.

Zotkin, D. N., Hwang, J., Duraiswami, R. & Davis, L. S. (2003). HRTF personalization using anthropometric measurements. In *Proceedings of the 2003 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics* (pp. 157–160). New Paltz, NY. USA.

Figure Captions

Figure 1: Interaural time delay and level difference example. The sound source is closer to the left ear and will thus reach the left ear before reaching the right ear. Furthermore, the level of the sound reaching the left ear will be greater as the sound reaching the right ear will be attenuated given the acoustical shadow introduced by the head.

Figure 2: Cone of confusion. A sound source positioned on any point on the surface of the cone of confusion will have the same ITD values.

Figure 3: Head rotations to resolve front-back ambiguities (viewed from above). When the sound source is directly in front of the listener, the distance between the left and right ears (d_l and d_r respectively) is the same. Rotating the head in the counter-clockwise direction will increase the distance between the left ear and the sound source d_l, while rotating the head in the clockwise direction will increase the distance between the right ear and the sound source d_r. These changes provide sound source localization cues.

Figure 4: BRIR measured at the right ear of a listener in a "moderate sized reverberant classroom" at the right ear of a listener with the sound source at an azimuth and elevation of 45° and 0° respectively, and at a distance of 1m. Reprinted with permission from Shilling & Shinn-Cunningham (2002).

Figure 5: Left and right ear HRTF measurements of three individuals for a source at an azimuth and elevation 90° and 0° respectively. Reprinted with permission from Begault

(1994).

Figure 6: Crosstalk defined. 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 will reach the right ear H_{LR} while a delayed and attenuated portion of the right loudspeaker signal will reach the left ear H_{RL}.

Figure 7: Vector base amplitude panning for a three-dimensional (three channel) configuration. The virtual sound source can be mapped to a location within the "active triangle" formed by the three loudspeakers. Adapted with permission from Pulkki (1997).

Figures

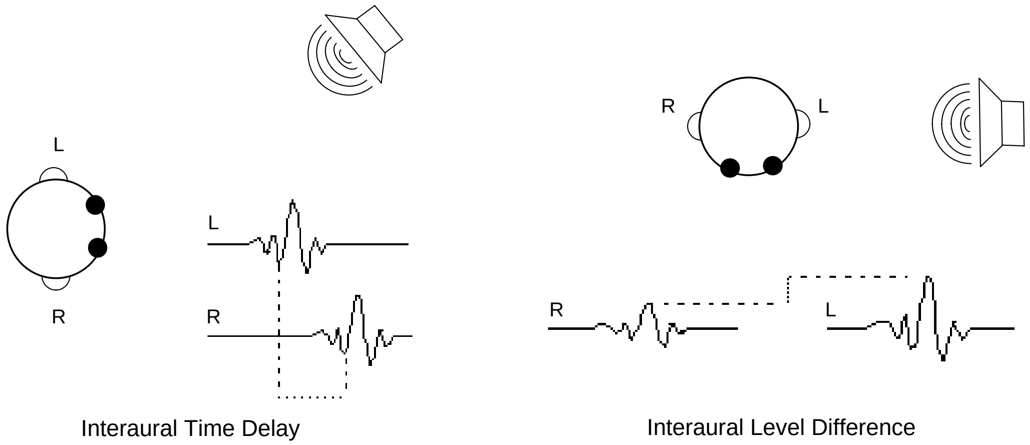

Figure 1: Interaural time delay and level difference example. The sound source is closer to the left ear and will thus reach the left ear before reaching the right ear. Furthermore, the level of the sound reaching the left ear will be greater as the sound reaching the right ear will be attenuated given the acoustical shadow introduced by the head.

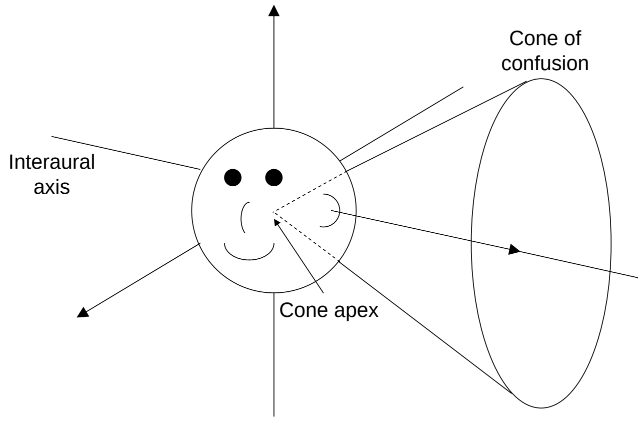

Figure 2: Cone of confusion. A sound source positioned on any point on the surface of the cone of confusion will have the same ITD values.

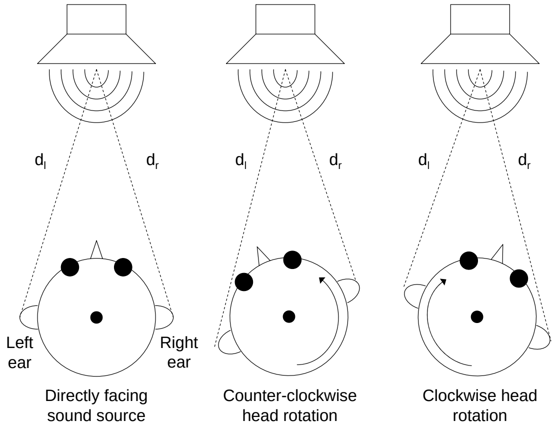

Figure 3: Head rotations to resolve front-back ambiguities (viewed from above). When the sound source is directly in front of the listener, the distance between the left and right ears (d_l and d_r respectively) is the same. Rotating the head in the counter-clockwise direction will increase the distance between the left ear and the sound source d_l, while rotating the head in the clockwise direction will increase the distance between the right ear and the sound source d_r. These changes provide sound source localization cues.

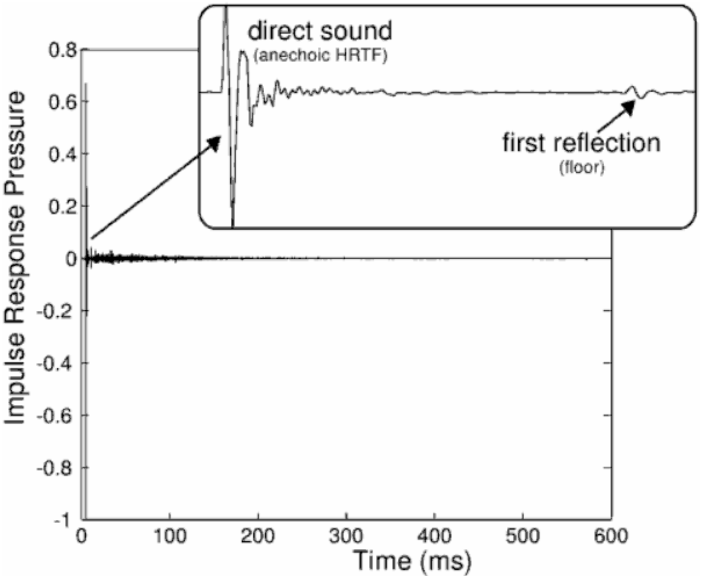

Figure 4: BRIR measured at the right ear of a listener in a “moderate sized reverberant classroom” at the right ear of a listener with the sound source at an azimuth and elevation of 45° and 0° respectively, and at a distance of 1m. Reprinted with permission from Shilling & Shinn-Cunningham (2002).

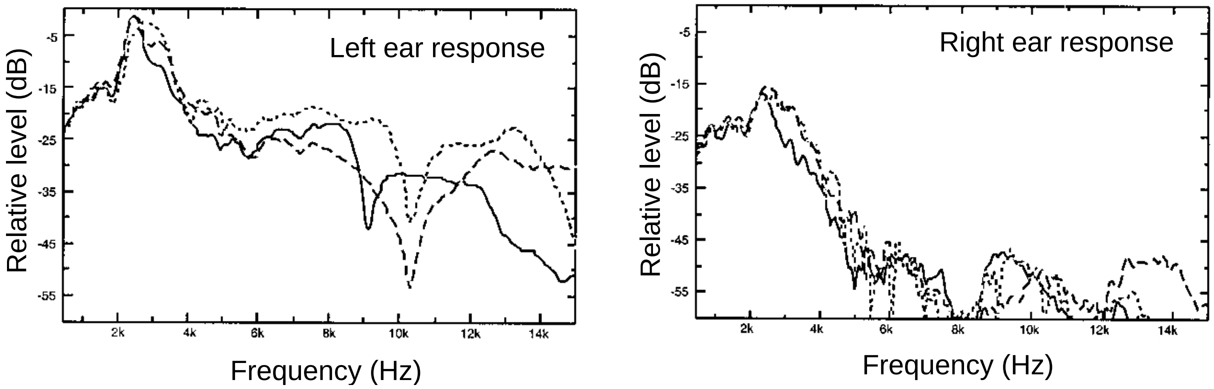

Figure 5: Left and right ear HRTF measurements of three individuals for a source at an azimuth and elevation 90° and 0° respectively. Reprinted with permission from Begault (1994).

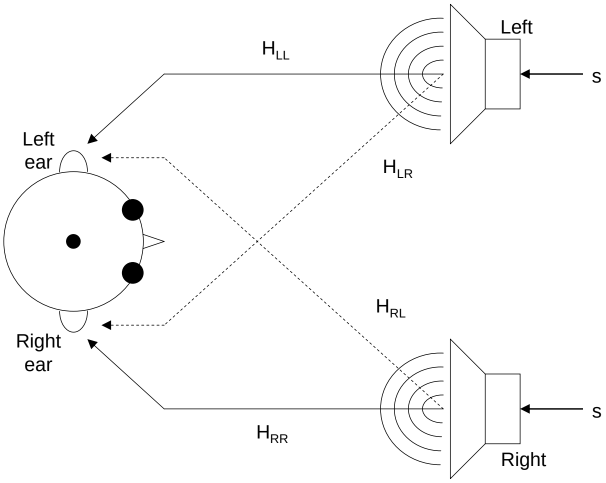

Figure 6: Crosstalk defined. In addition to the desired signal coming from the left and right loudspeakers H_{LL} and H_{RR} respectively, a delayed and attenuated portion of the left loudspeaker signal will reach the right ear H_{LR} while a delayed and attenuated portion of the right loudspeaker signal will reach the left ear H_{RL}.

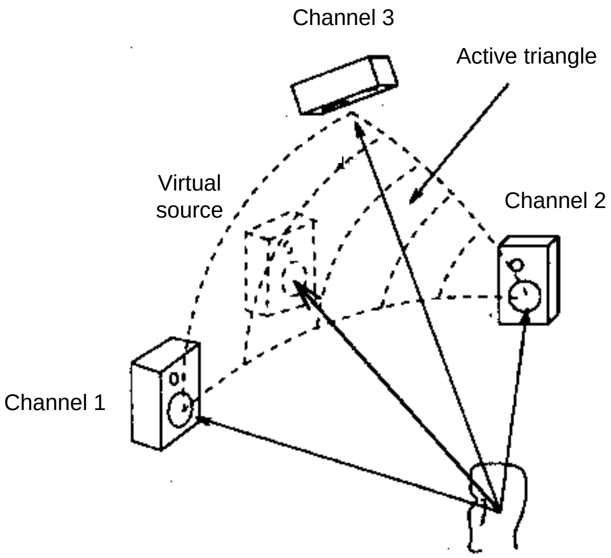

Figure 7: Vector base amplitude panning for a three-dimensional (three channel) configuration. The virtual sound source can be mapped to a location within the "active triangle" formed by the three loudspeakers. Adapted with permission from Pulkki (1997).