

Shilling, R. and B. G. Shinn-Cunningham (2000). Virtual Auditory Displays. To appear in Handbook of Virtual Environment Technology. K. Stanney (ed), Lawrence Erlbaum, Associates, Inc., in press.

Keywords: sound, sound servers, sound cards, binaural, spatial audio, auditory displays, virtual audio, sonification, headphones, speakers, presence, immersion, cocktail party effect, supernormal auditory localization, spatial hearing, head related transfer functions, HRTF, acoustics, confusion, field of view, auditory masking, reverberation, intersensory integration, auditory scene analysis, diotic displays, dichotic displays, stereo, surround sound, transaural, auditory scene analysis, speech perception, cross-modal enhancement, pitch, timbre, intensity, distance, room modeling, radio communication

Virtual Environments Handbook

Chapter 4 Virtual Auditory Displays

Russell D. Shilling, Ph.D.
Naval Postgraduate School

Barbara Shinn-Cunningham, Ph.D.
Boston University

1. Introduction

Auditory processing is often given minimal attention when designing virtual environments or simulations. This lack of attention is unfortunate since auditory cues play a crucial role in everyday life. Auditory cues increase awareness of surroundings, cue visual attention, and convey a variety of complex information without taxing the visual system. The entertainment industry has long recognized the importance of sound to create ambience and emotion, aspects that are often lacking in virtual environments. In short, placing someone in a virtual world with an improperly designed auditory interface is equivalent to creating a “virtual” hearing impairment for the user.

Auditory perception, especially localization, is a complex phenomenon affected by physiology, expectation, and even the visual interface. Different methods for creating auditory interfaces will be considered. As will be discussed later in this chapter, spatialized audio using headphones is the only audio technique that is truly “virtual” since it reproduces azimuth, elevation, and distance and offers the sound engineer the greatest amount of control over the auditory experience of the listener. For many applications, especially using projections screens, speaker systems may be simpler to implement and provide benefits not available to headphone systems. Properly designed

speaker systems incorporating subwoofers may contribute to emotional context. The positives and negatives associated with each option will be discussed. It is impossible to include everything that needs to be known about designing auditory interfaces in a single chapter. Instead of trying to review all perceptual and technical issues relevant to creating virtual auditory displays, this chapter unapologetically focuses on issues of spatial auditory perception and the generation of spatial auditory cues, since this area has undergone rapid development with the advent of virtual environments. Unlike the visual channel, very little effort has been put into synthesizing sources for the auditory channel in virtual environments; consequently, the question of how to generate realistic sounds (rather than using sources from some pre-computed, stored library of source sounds) is not discussed in this chapter. Similarly, many important aspects of auditory perception are ignored or given relatively little consideration in this review.

In the chapter, some of the most important physical properties of sound are introduced. Then, basic perceptual abilities of the auditory system are described, with an emphasis on spatial hearing. Finally, general techniques for producing auditory stimuli (with and without spatial cues) will be discussed. A complete lexicon for understanding and developing auditory displays can be found in Letowski, Vause, Shilling, Ballas, Brungert & McKinley (2000). The technology involved in producing spatialized audio is rapidly changing and new products are being continually introduced to the market while others are removed. Any specific recommendations would quickly be dated. Thus, this chapter will not discuss specific devices in great detail. However, a brief overview of current technology and solutions is discussed at the conclusion of the chapter.

1.1 Why are virtual auditory interfaces important?

1.1.1 Environmental Realism and Ambience

If it does nothing else, an auditory interface should convey basic information about the virtual environment to the user. For instance, in the real world, dismounted soldiers are aware of the sound of their own footsteps, the sounds of other soldiers, sounds of the natural environment, and mechanical sounds from various types of equipment such as jeeps, artillery, rifles, etc. In “control room” situations such as nuclear power plants, Air Traffic Control Centers, or the bridge of a ship, sounds such as alarms, switches being toggled, and verbal communications with other people in the room (including sounds of stress or uncertainty) provide vital information for the user. The location of these voices, switches, and alarms also provides information concerning their function and importance. In the absence of these basic auditory cues, situational awareness is severely degraded, and the same is true in virtual environments.

The entertainment industry has recognized that sound is a vital aspect of creating the ambience and emotion for films. George Lucas, best recognized by the public for the stunning visual effects in his movies, has stated that sound is 50% of the movie experience (THX, 2000). In virtual environments, the argument is often erroneously made that sound is secondary. Again, we can create a compelling visual representation of soldiers slogging through the mud, traversing a jungle, or going house-to-house. However, unless we supply the appropriate background sounds (machinery, artillery, animals, footsteps, etc), the participant will probably feel detached from the action. The sound of footsteps is different depending on whether you're in the grass, on pavement, or in a hallway. Likewise, the action of an M-16 sounds different depending on whether you're inside a room or outside. These are the types of things that create ambience and emotion in film; the same should hold true in virtual environments.

1.1.2 Presence/Immersion

Presence can be defined as the "sense of being immersed in a simulation or virtual environment." Such a nebulous concept is difficult to quantify. Although definitive evidence is lacking, it is generally believed that the sense of presence is dependent upon auditory, visual, and tactile fidelity (Sheriden, 1996). Referring back to the previous section, it can be inferred that as environmental realism increases, the sense of presence increases. However, although realism probably contributes to the sense of presence, the inverse is not true. It has been demonstrated that, although virtual or spatial audio does not necessarily increase the perceived realism of the virtual environment, it does increase the sense of presence (Hendrix, 1996). Thus, if implemented properly, appropriately designed audio increases the overall sense of presence in a virtual environment or simulation.

1.1.3 Collision/Tactile Cueing

Auditory cues in a system using head-mounted displays (HMDs) are particularly important and should be given close attention during the design phase. In limited field-of-view (FOV) HMDs, sounds associated with collision detection may play a major role in the user being able to successfully move through the simulation. These auditory collision cues can be used to substitute for tactile collision cues.

1.1.4 Cross-Modal Enhancement

Response times to visual targets associated with localized auditory cues have been shown to decrease (Perrott, et al., 1990; Perrott, et al., 1991). It has also been shown that the latency of saccadic eye movements towards a target is reduced when the visual target is aligned with an auditory cue (Frens, Opstal, & Willigen, 1995). In this manner, a properly designed auditory interface may be used to "enhance" the FOV for an HMD by cueing the user to locations outside the limited FOV of the HMD. Appropriate care must be taken to properly design the auditory interface, as auditory location has been shown to affect perceived visual target location when the visual target is presented on a non-textured background (Radeau & Bertelson, 1976).

1.1.5 Cocktail Party Effect

In a multi-source sound environment, it is easier to discriminate and comprehend sounds if they are separated in space. This enhancement in comprehension with spatially disparate sound sources is often called the cocktail party effect (Yost, 1994). The ability to understand multiple speakers simultaneously can be useful when applied to multi-user environments such as teleconferencing (Begault, 1999) or multi-channel radio communications (Begault & Wenzel, 1992; Begault, 1993; Haas, Gainer, Wightman, Couch, & Shilling, 1997). Even when the perception of location is not optimal, communication is improved in multichannel situations when spatialized auditory displays are employed (Drullman & Bronkhorst, 2000).

1.1.6 Sonification of Data

Auditory cues can be used to represent information that is not normally available in the real world. For instance, if a user makes a response error, or is slow in responding, "auditory icons" can be created to indicate the deficiency to the user (Gaver, 1986). On the other hand, using a technique called "sonification" complex information presented visually can be simplified by supplementing it with an auditory representation correlated with some dimension or dimensions of the data set (Rabenhorst, et al., 1990). For example, in simulations designed to teach operators how to interpret complex controls and displays (radar, sonar, etc), multidimensional data can be "sonified" to make it easier to interpret. Likewise, sonification could be used as a learning aid to illustrate principles using multidimensional data in Physics and Chemistry.

1.1.7 Supernormal Auditory Localization

When designing auditory interfaces, it is possible to exaggerate normal auditory cues so that the listener is able to localize sound in the virtual world with better resolution than in the real world. One way of achieving supernormal localization is to exaggerate the normal head and ear size cues available to the listener (e.g., see Shinn-Cunningham, Durlach, and Held, 1998a, b). Supernormal auditory localization may be of best use in teleoperator applications, and virtual environments where there is a need to compensate for a limited field of view head mounted display. For example, spatialized auditory displays have been designed which exaggerate normal auditory cues and “zoom” the auditory display as the magnification on the visual display is increased (Shilling, et al., 1998; Shilling & Letowski, 2000). Of course, experience with such “supernormal” systems affects both the accuracy and resolution of perception, and both effects must be evaluated in order to determine the overall cost/benefit of such techniques for a given application (e.g., see Shinn-Cunningham, 2000a).

1.1.8 Virtual Auditory Displays (VAD)

While graphical displays are an obvious choice for displaying spatial information to a human operator (particularly after considering the spatial acuity of the visual channel), the visual channel is often overloaded, with operators monitoring a myriad of dials, gauges, and graphic displays. In these cases, spatial auditory information can provide invaluable information to an operator, particularly when the visual channel is saturated (Begault, 1993; Bronkhorst, et al., 1996; Shilling & Letowski, 2000). Spatial auditory displays are also being developed for use in applications for which visual information provides no benefit; for instance, in limited FOV applications or when presenting information to the blind. In these command/control applications, the primary goal is to convey unambiguous information to the human operator. Realism, *per se*, is not useful, except to the extent that it makes the operator’s task easier (i.e., reduces the workload). Conversely, spatial resolution is critical. In these applications, signal-processing schemes that could enhance the amount of information transferred to the human operator may be useful, even if the result is “unnatural,” as long as the user is able to extract this information (e.g., see Durlach, Shinn-Cunningham, and Held, 1993). It should be noted that when designing spatialized auditory displays for noisy environments such as cockpits, electronic noise cancellation technology should be employed and user’s hearing loss taken into account to make certain the displays are localizable (Begault, 1996). Also, for high-g environments, more work needs to be conducted to discover the contribution of g-forces to displacements in sound localization, the so-

called “audiogyral illusion” (Clark & Graybiel, 1949; DiZio, Held, Lackner, Shinn-Cunningham, and Durlach, 2000).

1.1.9 Enhancement of Perceived Quality of the Simulation

The importance of multi-modal interactions involving the auditory system cannot be ignored. It has been shown that using medium and high quality auditory displays can enhance the perception of quality in visual displays. Inversely, using low quality auditory displays reduces the perceived quality of visual displays (Storms, 1998).

2. Physical Acoustics

2.1 Properties of Sound

Simply put, sound is a pressure wave produced when an object vibrates rapidly back and forth. The diaphragm of a speaker produces sound by pushing against molecules of air, thus creating an area of high pressure (condensation). As the speaker diaphragm returns to its resting position, it creates an area of low pressure (rarefaction). This localized disturbance travels through the air as a wave of alternating low pressure and high pressure at approximately 344 m/sec or 1128 ft/sec (at 70° F) depending on temperature and humidity.

2.1.1 Frequency

If we play the musical note “A” as a pure sinusoid, there will be 440 condensations and rarefactions per second. The distance between two adjacent condensations or rarefactions equals the wavelength of the sound wave and is typically represented by the symbol λ . The velocity at which the sound wave is traveling is denoted as c . The time one full oscillatory cycle (condensation through rarefaction) takes is called the frequency (f) and is expressed in Hertz or “Cycles Per Second”. The relationship between frequency, velocity, and wavelength is given by $f = c/\lambda$.

From a modeling standpoint, this relationship is important when considering Doppler Shift. As a sound source is moving toward us, the perceived frequency increases because the wavelength is compressed as a function of the velocity (v) of the moving source. This compression can be explained by the equation $\lambda = (c - v) / f$. For negative velocities (i.e., for sources moving away), this expression describes a relative increase in the wavelength (and a concomitant decrease in frequency).

2.1.2 Intensity

The intensity of the sound stimulus is determined by the amplitude of the waveform. Intensity is measured in decibels (dB). Decibels give the level of sound (on a logarithmic scale) relative to some reference level. One common reference level is $2 \times 10^{-5} \text{ N/m}^2$. Decibels referenced to this value are commonly used to describe sound intensity expressed in units of dB SPL. The sound level in dB SPL can be computed by the following equation:

$$\text{dB SPL} = 20 \log_{10} \left(\frac{\text{RMS sound pressure}}{2 \times 10^{-5} \text{ N/m}^2} \right)$$

The threshold of hearing is in the range of 0-10 dB SPL for most sounds, although the actual threshold depends on the spectral content of the sound. When measuring sound intensity in the “real world”, intensity is measured using a sound pressure meter. Most sound pressure meters allow one to collect sound level information using different scales which weight energy in different frequencies differently in order to approximate the sensitivity of the human auditory system to sound at low, moderate, or high intensity levels. These scales are known as the A, B, and C weighted scales, respectively. The B scale is rarely used; however, the C scale (dBC) is useful for evaluating noise levels in high intensity environments such as traffic noise and ambient cockpit noise. The frequency response of the dBC measurement is closer to a flat-response than dBA. In fact, when conducting “sound surveys” in a complex noise environment, it is prudent to measure sound level in both dBA and flat-response (or dBC) to make an accurate assessment of the audio environment.

Frequency, intensity, and complexity are physical properties of an acoustic waveform. The perceptual analogues for frequency, intensity, and complexity are pitch, loudness, and timbre respectively. This distinction between physical and perceptual measures of sound properties is an important one. Thus, it is critical to consider both physical and perceptual descriptions when designing auditory displays.

3. Psychophysics

The basic sensitivity of the auditory system is reviewed in detail in a number of textbooks (e.g., see Yost, 1994; Moore, 1997; Gelfand, 1998). This section provides a brief overview of some aspects of human auditory sensitivity that are important to consider when designing auditory virtual environments.

3.1 Frequency Analysis in the Auditory System

In the cochlea, acoustic signals are broken down into constituent frequency components by a mechanical Fourier-like analysis. Along the length of the cochlea, the frequency to which that section of the cochlea responds varies systematically from high to low frequencies. The strength of neural signals carried by the auditory nerve fibers arrayed along the length of the cochlea varies with the mechanical displacement of the corresponding section of the cochlea. As a result, each nerve fiber can be thought of as a frequency channel that conveys information about the energy and timing of the input signal within a restricted frequency region. At all stages of the auditory system, these multiple frequency channels are in evidence.

Although the bandwidth changes with the level of the input signal and with input frequency, to a crude first order approximation, one can think of the frequency selectivity of the auditory system as constant on a log-frequency basis (approximately 1/3-octave wide). Thus, a particular auditory nerve responds to acoustic energy at and near a particular frequency.

Humans are sensitive to acoustic energy at frequencies between about 20 Hz and 22,000 Hz. Absolute sensitivity varies with frequency. Humans are most sensitive to energy at frequencies around 2000 Hz, and are less sensitive for frequencies below and above this range.

The fact that input waveforms are deconstructed into constituent frequencies affects all aspects of auditory perception. Many behavioral results are best understood by considering the activity of the auditory nerve fibers, each of which responds to energy within about a third of an octave of its particular “best frequency.” For instance, the ability to detect a sinusoidal signal in a noise background degrades dramatically when the noise spectrum is within a third octave of the sinusoid frequency. When a noise is spectrally remote from a sinusoidal target, it causes much less interference with the detection of the sinusoid. These factors are important when one considers the spectral content of different sounds that are to be used in an auditory virtual environment. For instance, if one must monitor

multiple kinds of alerting sounds, choosing signals that are spectrally remote from one another will improve the users ability to detect and respond to the different signals.

3.2 Intensity Perception

Listeners are sensitive to sound intensity on a logarithmic scale. For instance, doubling the level of a sound source causes roughly the same perceived change in the loudness independent of the reference level. This logarithmic sensitivity to sound intensity gives the auditory system a large dynamic range. For instance, the range between just detectable sound levels and sounds that are so loud that they cause pain is roughly 110 – 120 dB (i.e., an increase in sound pressure by a factor of a million). The majority of the sounds encountered in everyday experience span a dynamic intensity range of 80 – 90 dB. Typical sound reproduction systems use 16 bits to represent the pressure of the acoustic signal (providing a useful dynamic range of about 90 dB), which is sufficient for most simulations.

While sound intensity (a physical measure) affects the loudness of a sound (a perceptual measure), loudness does not grow linearly with intensity. In addition, the same decibel increase in sound intensity can result in different increments in loudness, depending on the frequency content of the sound. Thus, intensity and loudness, while closely related, are not equivalent descriptions of sound.

3.3 Masking Effects

As mentioned above, when multiple sources are presented to a listener simultaneously or in rapid succession, the sources interfere with one another in various ways. For instance, a tone that is audible when played in isolation may be inaudible when a loud noise is presented simultaneously. Such effects (known as “masking” effects) arise from a variety of mechanisms, from physical interactions of the separate acoustic waves impinging on the ear to high-level, cognitive factors. For a more complete description of these effects than is given below, see Yost, 1994, pp. 153-167 or Moore, 1997, pp. 111-120.

“Simultaneous masking” occurs when two sources are played concurrently. However, signals do not have to be played simultaneously for them to interfere with one another perceptually. For instance, both “forward” masking (in which a leading sound interferes with perception of a trailing sound) and “backward” masking (in which a lagging sound interferes with perception of a leading sound) also occurs.

Generally speaking, many simultaneous and forward masking effects are thought to arise from peripheral interactions that occur at or before the level of the auditory nerve. For instance, the mechanical vibrations of the basilar membrane are nonlinear, so that the response of the membrane to two separate sounds may be less than the sum of the response to the individual sounds. These nonlinear interactions can suppress the response to what would (in isolation) be an audible event. In other words, because of nonlinearities in the transduction of acoustic signals, the response of the auditory nerve to a given signal may be less robust when a second signal is present.

Other, more central factors influence masking as well. For instance, backward masking may reflect higher-order processing that limits the amount of information extracted from an initial sound in the presence of a second sound. The term “informational masking” refers to all masking that cannot be explained by peripheral interactions in the transduction of sound by the auditory periphery. There is ample evidence that informational masking is an important factor in auditory perception. For instance, perceptual sensitivity in discrimination and detection tasks is often degraded when there is uncertainty about the characteristics of a target source (e.g., see Yost, 1994, pp. 219-220).

3.4 Pitch and Timbre

Just as sound intensity is the physical correlate of the percept of loudness, source frequency is most closely related to the percept of pitch. For sound waves that are periodic (including pure sinusoids, for instance), the perceived “pitch” of the sound is directly related to the inverse of the period of the sound signal. Thus, sounds with low pitch have relative long periods and sounds with high pitch generally have short periods. Many real-world sounds are not strictly periodic in that they have a temporal pattern that repeats over time, but has fluctuations from one cycle to the next. Examples of such pseudo-periodic signals include the sound produced by a flute or a vowel sound spoken aloud. The perceived pitch of such sounds is well predicted by the average period of the cyclical variations in the stimulus.

The percept of pitch is not uniformly strong for all sound sources. In fact, non-periodic sources, such as noise waveforms, often do not have a salient pitch associated with them. For relatively narrow sources that are aperiodic, a weak percept of pitch can arise that depends on the center frequency of the spectral energy of the signal. In fact, perceived pitch is affected by a wide variety of stimulus attributes, including temporal structure, frequency content, harmonicity, and even loudness. Although the pitch of a pure sinusoid is directly related to its frequency, there is no single physical parameter that can predict perceived pitch for more complex sounds. Nonetheless, for many sounds,

pitch is a very salient and robust perceptual feature that can be used to convey information to a listener. For instance, in music, pitch conveys melody. In speech, pitch conveys a variety of information (ranging from the gender of the speaker to the paralinguistic, emotional content of the speech).

The percept of timbre describes other attributes of sound sources, and is the sound property that enables a listener to distinguish an oboe from a trumpet. Like pitch, the percept of timbre depends on a number of physical parameters of sound, including spectral content and temporal envelope (such as the abruptness of the onset and offset of sound).

3.5 Temporal Resolution

The auditory channel is much more sensitive to temporal fluctuations in the sensory inputs than the visual or proprioceptive channel. For instance, the system can detect amplitude fluctuations in input signals up to 50 Hz (i.e., a duty cycle of 20 ms) very easily (e.g., see Yost, 1994, pp. 146-149). Sensitivity degrades slowly with increasing modulation rate, so that some sensitivity remains even as the rate approaches 1000 Hz (i.e., temporal fluctuations at a rate of 1 per ms). The system is also sensitive to small fluctuations in the spectral content of the input signal for roughly the same modulation speeds. Listeners not only can detect rapid fluctuations in an input stimulus, they can react quickly to auditory stimuli. For instance, reaction times to auditory stimuli are faster than for vision by 30-40 ms (an improvement of roughly 20%; e.g., see Welch and Warren, 1986, p. 25-3).

3.6 Spatial Hearing

The spatial acuity of the auditory system is far worse than that of the visual or proprioceptive systems. For a listener to detect an angular displacement of a source from the median plane, the source must be displaced laterally by about a degree. For a source directly to the side, the listener does not always detect a lateral displacement of 10 degrees. Auditory spatial acuity is even worse in other spatial dimensions. A source in the median plane must be displaced by as much as 15 degrees for the subject to perceive the directional change accurately. While subjects can judge relative changes in source distance, absolute distance judgements are often surprisingly inaccurate, even under the best of conditions.

Functionally, spatial auditory perception is distinctly different from that of the other “spatial” senses of vision and proprioception. For the other spatial senses, position is neurally encoded at the most peripheral part of the sensory system. For instance, the photoreceptors of the retina are organized topographically so that a source at a

particular position (relative to the direction of gaze) excites a distinct set of receptors. In contrast, spatial information in the auditory signals reaching the left and right ears of a listener must be computed from the peripheral neural representations. The way in which spatial information is carried by the acoustic signals reaching the eardrums of a listener has been the subject of much research. This section provides a brief review of how acoustic attributes convey spatial information to a listener and how the perceived position of a sound source is computed in the brain (for more complete reviews, see Mills, 1972; Middlebrooks and Green, 1991; Wightman and Kistler, 1993; Blauert, 1997).

3.6.1 Binaural Cues

The most robust cues for source position depend on differences between the signals reaching the left and right ears. Such *interaural* or *binaural* cues are robust specifically because they can be computed by comparing the signals reaching each ear. As a result, binaural cues allow the listener to factor out those acoustic attributes that arise from source *content* from those attributes that arise from source *position*.

Depending on the angle between the interaural axis and a sound source, one ear may receive the sound earlier than the other. The resulting interaural time differences (ITDs) are the main cue indicating the laterality (left/right location) of the direct sound. The ITD grows with the angle of the source from the median plane; for instance, a source directly to the right of the listener results in an ITD of 600-800 μs favoring the right ear. ITDs are most salient for sound frequencies below about 2 kHz, but occur at all frequencies in a sound. At higher frequencies, listeners use ITDs in signal *envelopes* to help determine source laterality, but are insensitive to differences in the interaural phase of the signal.

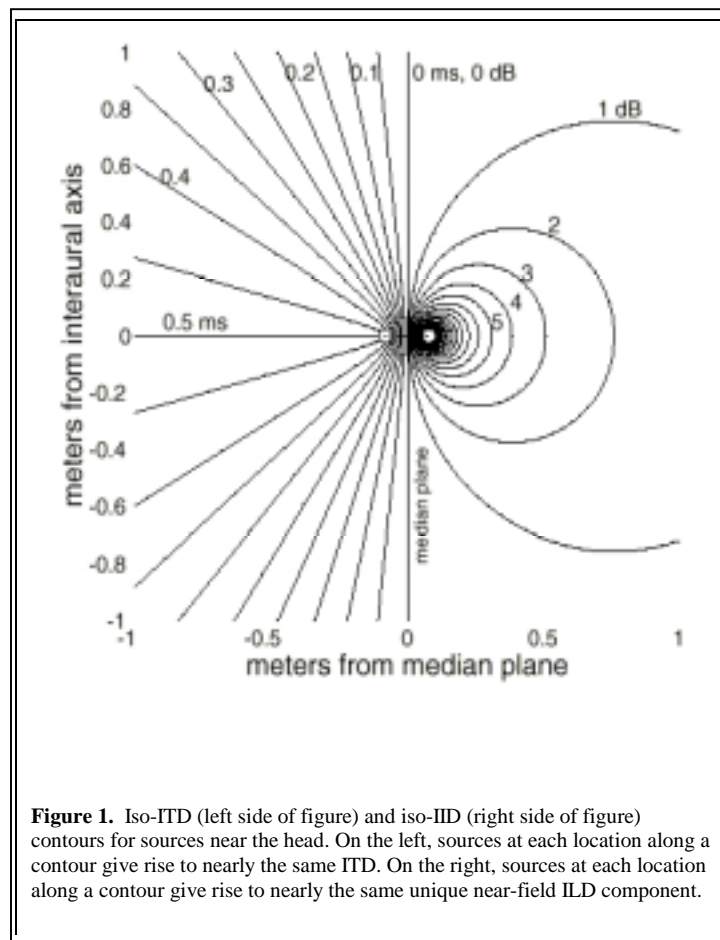
Listeners can reliably detect ITDs of 10-50 μs (depending on the individual listener), which grossly correspond to the ITDs that would result from a source positioned 1-5 degrees from the median plane. Sensitivity to changes in the ITD deteriorates as the reference ITD gets larger. For instance, the smallest detectable change in ITD around a reference source with an ITD of 600-800 μs (corresponding to the ITD of a source far to the side of the head) can be more than a factor of 2 larger than for a reference with zero ITD.

At the high end of the audible frequency range, the head of the listener reflects and diffracts signals so that less acoustic energy reaches the far side of the head (causing an “acoustic head shadow”). Due to the acoustic head

shadow, the relative intensity of the sound at the two ears varies with the lateral location of the source. The resulting interaural intensity differences (IIDs) generally increase with source frequency and the angle between the source and the median plane. IIDs are perceptually important for determining source laterality for frequencies above about 2 kHz.

When a sound source is within reach of the listener, extra large IIDs (at all frequencies) arise due to the differences in the relative distances from source to left and right ears (e.g., see Duda and Martens, 1998; Brungart and Rabinowitz, 1999). These additional IIDs are due to the differences in the relative distances from source to left and right ears and help to convey information about the relative distance and direction of the source from the listener (Shinn-Cunningham, Santarelli, and Kopco, 2000). Other low-frequency IIDs that may arise from the torso appear to help determine the elevation of a source (Algazi, Avendano, and Duda, 1999). Most listeners are able to detect IIDs of 0.5-1 dB, independent of source frequency.

The perceived location of a sound source usually is consistent with the ITD and IID information available. However, there are multiple source locations that cause roughly the same ITD and IID cues. For sources more than a meter from the head, the locus of such points is approximately a hyperbolic surface of rotation symmetric about the interaural axis that is known as the “cone of confusion” (see left side of Figure 1). When a sound is within reach of the listener, extra large IIDs provide additional robust, binaural information about the source location. For a simple spherical head model, low-frequency IIDs are constant on spheres



centered on the interaural axis (see right side of Figure 1;). The rate at which the extra-large IID changes with spatial

location decreases as sources move far from the head or near the median plane. In fact, once a source is more than a meter or so from the head, the contribution of this “near field” IID is perceptually insignificant (Shinn-Cunningham, Santarelli, and Kopco, 2000). In general, positions that give rise to the same binaural cues (i.e., the intersection of constant ITD and IID contours) form a circle centered on the interaural axis (Shinn-Cunningham, Santarelli, and Kopco, 2000). Since ITD and IID sensitivity is imperfect, the locus of positions that cannot be resolved from binaural cues may be more accurately described as a “torus of confusion” centered on the interaural axis. Such “tori of confusion” degenerate to the more familiar cones of confusion for sources more than about a meter from the listener.

3.6.2 Spectral Cues

The main cue to resolve source location on the torus of confusion is the spectral content of the signals reaching the ears. These spectral cues arise due to interactions of the outer ear (pinna) with the impinging sound wave that depend on the relative position of sound source and listener’s head (Batteau, 1967). Spectral cues only occur at relatively high frequencies, generally above 6 kHz. Unlike interaural cues for source location, spectral cues can be confused with changes in the spectrum of the source itself. Perhaps because of this ambiguity, subjects are more likely to make localization errors in which responses fall near the correct torus of confusion, but are not be in the right direction. Individual differences in the spectral filtering of the pinnae are large and are important when judging source direction (e.g., see Wenzel, Arruda Wenzel, Arruda, Kistler, and Wightman, 1993).

3.6.3 Anechoic Distance Cues

In general, the intensity of the direct sound reaching a listener (i.e., sound that does not come off of reflective surfaces like walls) decreases with the distance of the source. In addition, the atmosphere absorbs energy in high, audible frequencies as a sound propagates, causing small changes in the spectrum of the received signal with changes in source distance. If a source is unfamiliar, the intensity and spectrum of the direct sound are not robust cues for distance since they can be confounded with changes in the intensity or spectral content (respectively) of the signal emitted from the source. However, even for unfamiliar sources, overall level and spectral content provide *relative* distance information (Mershon, 1997).

3.6.4 Reverberation

Reverberation (acoustic energy reaching the listener from indirect paths, via walls, floors, etc.) generally has little effect on or degrades the perception of source direction (e.g., see Hartmann, 1983; Shinn-Cunningham, Zurek, and Durlach 1993; Shinn-Cunningham, 2000b; Begault, 1992). However, it actually aids in the perception of source distance (e.g., see Mershon, Ballenger, Little, McMurtry, and Buchanan, 1989; Shinn-Cunningham, Kopco, and Santarelli, 1999). At least grossly, the intensity of reflected energy received at the ears is independent of the position of the source relative to the listener (although it can vary dramatically from one room to another). As a result, the ratio of direct to reverberant energy provides an absolute measure of source distance for a given listening environment.

Reverberation not only provides a robust cue for source distance, it provides information about the size and configuration of the listening environment. For instance, information about the size and “spaciousness” of a room can be extracted from the pattern of reverberation in the signals reaching the ears. While many psychophysical studies of sound localization are performed in anechoic (or simulated anechoic) environments, reverberation is present (in varying degrees) in virtually all normal listening conditions. Anechoic environments (such as those used in many simulations and experiments) seem subjectively “unnatural” and “strange” to naïve listeners. Conversely, adding reverberation to a simulation causes all sources to seem more realistic and provides robust information about relative source distance. While reverberation may improve distance perception and improve the realism of the display, it can decrease accuracy in directional perception, albeit slightly and may interfere with the ability to extract information in the source signal (e.g., degrade speech reception) and to attend to multiple sources (e.g., see section 3.6.8).

3.6.5 Dynamic Cues

In addition to “static” acoustic cues like ITD and IID, changes in spatial cues with source or listener movement also influence perception of source position and help to resolve torus-of-confusion ambiguities (e.g., see Wightman and Kistler, 1999; Wallach, 1940). For instance, either a source directly in front or directly behind the listener would cause near zero ITDs and IIDs; however, a leftward rotation of the head results in either ITDs and IIDs favoring the right ear (for a source in front) or the left ear (for a source behind).

While the auditory system generally has good temporal resolution, the temporal resolution of the *binaural* hearing system is much poorer. For instance, investigations into the perception of moving sound sources implies that binaural information averaged over a time window lasting 100-200 ms, resulting in what has been termed “binaural sluggishness” (e.g., see Kollmeier and Gilkey, 1990; Grantham, 1997).

3.6.6 Effects of Stimulus Characteristics on Spatial Perception

Characteristics of the source itself affect auditory spatial perception in a number of ways. For instance, the bandwidth of a stimulus can have a large impact on the perceived location of a source. As a result, one must consider how nonspatial attributes of a source in a virtual environment will impact the spatial perception of the signal. In cases where one can design the acoustic signal (i.e., if the signal is a warning signal or some other arbitrary waveform), these factors should be taken into consideration when one selects the source signal.

For instance, the spectral filtering of the pinnae cannot be determined if the sound source does not have sufficient bandwidth. This makes it difficult to unambiguously determine the location of a source on the torus of confusion for a narrowband signal. Similarly, if the source signal does not have energy above about 5 kHz, spectral cues will not be represented in the signals reaching the ears and errors along the torus of confusion are more common (e.g., Gilkey and Anderson, 1995).

Ambiguity in narrowband source locations arises in other situations as well. For instance, narrowband, low-frequency signals in which ITD is the main cue can have ambiguity in their heard location because the auditory system is only sensitive to interaural phase. Thus, a low-frequency sinusoid with an ITD of 1/2 cycle favoring the right ear may also be heard far to the left side of the head. However, binaural information is integrated across frequency so that ambiguity in lateral location is resolved when interaural information is available across a range of frequencies (Trahiotis and Stern, 1989; Brainard, Knudsen, and Esterly, 1992; Stern and Trahiotis, 1997). When narrowband sources are presented, the heard location is strongly influenced by the center frequency of the source (Middlebrooks, 1997).

While spectral bandwidth is important, temporal structure of the source signal is also important. In particular, onsets and offsets in a signal make source localization more accurate, particularly when reverberation and echoes are present. A gated or modulated broadband noise will generally be more accurately localized in a reverberant room (or

simulation) than a slowly gated broadband noise (e.g., see Rakerd and Hartmann, 1985; Rakerd and Hartmann, 1986).

3.6.7 Top-Down Processes in Spatial Perception

Experience with or knowledge of the acoustics of a particular environment also affects auditory localization (e.g., see Clifton, Freyman, and Litovsky, 1993; Shinn-Cunningham, 2000). Results show that “top-down” processing of auditory information due to implicit learning and experience affects performance. In other words, spatial auditory perception is not wholly determined by stimulus parameters, but also by the state of the subject. Although such effects are not due to conscious decision, they can measurably alter auditory localization and spatial perception.

3.6.8 Benefits of Binaural Hearing

Listeners benefit from receiving different signals at the two ears in a number of ways. As discussed above, ITD and IID cues allow listeners to determine the location of sound sources. However, in addition to allowing listeners to locate sound sources in the environment, binaural cues allow the listener to selectively attend to sources coming from a particular direction. This ability is extremely important when there are multiple competing sources in the environment (e.g., see Bronkhorst, 2000).

Imagine a situation in which there is both a speaker (whom the listener is trying to attend) and a competing source (that is interfering with the speaker). If the speaker and competitor are both directly in front of the listener, the competitor degrades speech reception much more than if the competitor is off to one side, spatially separated from the speaker. This “binaural advantage” arises in part because when the competitor is to one side of the head, the energy from the competitor is attenuated at the far ear. As a result, the signal-to-noise ratio at the far ear is larger than when the competitor is in front. In other words, the listener has access to a cleaner signal in which the speaker is more prominent when the speaker and noise are spatially separated. However, the advantage of the spatial separation is even larger than can be predicted on the basis of energy: spatial information can be used to “squellch” signals from directions other than the direction of interest.

A homologous benefit can be seen under headphones. In particular, if one varies the level of a signal until it is just detectable in the presence of a masker, the necessary signal level is much lower when the ITD of the signal and masker are different than when they are the same. The difference between these thresholds, referred to as the

masking level difference (MLD), can be as large as 10-15 dB for some signals (e.g., see Durlach and Colburn, 1978; Zurek, 1993).

The binaural advantage affects both signal detection (e.g., see Gilkey and Good, 1995) and speech reception (e.g., see Bronkhorst and Plomp, 1988). It is one of the main factors contributing to the ability of listeners to monitor and attend multiple sources in complex listening environments (i.e., the “cocktail party effect;” see, for example, Yost, 1997). Thus, the binaural advantage is important for almost any auditory signal of interest. In order to get these benefits of binaural hearing, the signals reaching the listener must have appropriate ITDs and/or IIDs.

3.6.9 Adaptation to Distorted Spatial Cues

While a naïve listener responds to ITD, IID, and spectral cues based on their everyday experience, listeners can learn to interpret cues that are not exactly like those that occur naturally. For instance, listeners can learn to adapt to unnatural spectral cues when given sufficient long-term exposure (Hofman, Van Riswick, and Van Opstal, 1998). Short-term training allows listeners to learn how to map responses to spatial cues to different spatial locations than normal (Shinn-Cunningham, Durlach, and Held, 1998). These studies imply that for applications in which subjects can be trained, “perfect” simulations of spatial cues may not be necessary. However, there are limits to the kinds of distortions of spatial cues to which a listener can adapt (Shinn-Cunningham, Durlach and Held, 1998b; Shinn-Cunningham, 2000a).

3.6.10 Intersensory Integration of Spatial Information

Acoustic spatial information is integrated with spatial information from other sensory channels (particularly vision) to form spatial percepts (e.g., see Welch and Warren, 1986). In particular, auditory spatial information is combined with visual (and/or proprioceptive) spatial information to form the percept of a single, multisensory event, especially when the inputs to the different modalities are correlated in time (e.g., see Warren, Welch et al., 1981). When this occurs, visual spatial information is much more potent than that of auditory information so that the perceived location of the event is dominated by the visual spatial information (although auditory information does affect the percept to a lesser degree; e.g., see Pick, Warren et al., 1969; Welch and Warren, 1980). “Visual capture” refers to the perceptual dominance of visual spatial information, describing how the perceived location of an auditory source is captured by visual cues.

Summarizing these results, it appears that the spatial auditory system computes source location by combining all available acoustic spatial information. Perhaps even more importantly, a priori knowledge and information from other sensory channels can have a pronounced effect on spatial perception of auditory and multisensory events.

3.7 Auditory Scene Analysis

Listeners in real-world environments are faced with the difficult problem of listening to many competing sound sources that overlap in both time and/or frequency. The process of separating out the contributions of different sources to the total acoustic signals reaching the ears is known as “auditory scene analysis” (e.g., see Bregman, 1990).

In general, the problem of grouping sound energy across time and frequency to reconstruct each sound source is governed by a number of basic (often intuitive) principles. For instance, naturally-occurring sources are often broadband, but changes in the amplitude or frequency of the various frequency components of a single source are generally correlated over time. Thus, co-modulation of sound energy in different frequency bands tends to “group” these signal elements together and cause them to fuse into a single perceived source. Similarly, temporal and spectral proximity both tend to promote grouping so that signals close in time or frequency are grouped into a single perceptual source (sometimes referred to as a “stream”). Spatial location also can influence auditory scene analysis such that signals from the same or similar locations are grouped into a single stream. Other factors affecting streaming include (but are not limited to) harmonicity, timbre, and frequency or amplitude modulation.

For development of auditory displays, these grouping and streaming phenomena are very important, because they can directly impact the ability to detect, process, and react to a sound. For instance, if a masker sound (comprised of a number of constituent modulated sinusoids) is played simultaneously with a target sinusoid that is also modulated, the ability to detect the target improves if the masker sinusoids are modulated with the same envelope (different from the target). This process cannot be explained by peripheral mechanisms, since the peripheral masking produced is independent of whether or not the masker components are temporally correlated. Thus, perceptual segregation of two perceptual “streams” reduces their interaction and interference, improving performance on signal detection, speech intelligibility, temporal discrimination and other tasks.

3.8 Speech Perception

Arguably the most important acoustic signal is that of speech. The amount of information transmitted via speech is larger than for any other acoustic signal. For many applications, accurate transmission of speech information is the most critical component of an auditory display.

Speech perception is affected by many of the low-level perceptual issues discussed in previous sections. For instance, speech can be masked by other signals, reducing a listener's ability to determine the content of the speech signal. Speech reception in noisy environments improves if the speaker is located at a different position than the noise source(s), particularly if the speaker and masker are at locations giving rise to different interaural level differences. Speech reception is also affected by factors that affect the formation of auditory streams, such as comodulation, harmonic structure, and related features. However, speech perception is governed by many high-level, cognitive factors that do not apply to other acoustic signals. For instance, the ability to perceive a spoken word improves dramatically if it is heard within a meaningful sentence rather than in isolation. Speech information is primarily conveyed by sound energy between 200 and 5000 Hz. For systems in which speech communication is critical, it is important to reduce the amount of interference in this range of frequencies or it will impede speech reception.

4. Spatial Simulation

Spatial auditory cues can be simulated using headphone displays or loudspeakers. Headphone displays generally allow more precise control of the spatial cues presented to the listener, both because the signals reaching the two ears can be controlled independently and because there is no indirect sound reaching the listeners (i.e., no echoes or reverberation). However, headphone displays are generally more expensive than loudspeaker displays and may be impractical for applications in which the user does not want to wear a device on the head. While it is more difficult to control the spatial information reaching the listener in a loudspeaker simulation, loudspeaker-based simulations are relatively simple and inexpensive to implement and do not physically interfere with the user.

Simulations using either headphones or speakers can vary in complexity from providing *no* spatial information to providing nearly all naturally-occurring spatial cues. This section reviews both headphones and speaker approaches to creating spatial auditory cues.

4.1 Headphone Simulation

4.1.1 Diotic Displays

The simplest headphone displays present identical signals to both ears (“diotic” signals). With a diotic display, all sources are perceived as *inside* the head (not “externalized”), at midline. This internal sense of location is known as “lateralization” not “localization” (Plenge, 1974). While a diotic display requires no spatial auditory processing, it also provides no spatial information to the listener. Such displays may be useful if the location of the auditory object is not known or if spatial auditory information is unimportant. However, diotic displays are the least realistic headphone display. In addition, as discussed in Section 3.4.8, benefits of spatial hearing can be extremely useful for detection and recognition of auditory information. For instance, when users are required to monitor multiple sounds sources, spatialized auditory displays are clearly superior to diotic displays (Hass, Gainer, Wightman, Couch, & Shilling, 1997).

4.1.2 Dichotic Displays

While normal interaural cues vary with frequency in complex ways, simple frequency-independent ITDs and IIDs affect the perceived lateral position of a sound source (e.g., see Durlach and Colburn, 1978). Stereo signals that only contain a frequency dependent ITD and/or IID are herein referred to as “dichotic” signals (although the term is sometimes used to refer to any stereo signal in which left and right ears are different).

Generation of a constant ITD or IID is very simple over headphones since it only requires that the source signal be delayed or scaled (respectively) at one ear. Just as with diotic signals, dichotic signals result in sources that appear to be located on an imaginary line inside the head, connecting the two ears. Varying the ITD or IID causes the lateral position of the perceived source to move toward the ear receiving the louder and/or earlier-arriving signal. For this reason, such sources are usually referred to as “lateralized” rather than “localized.”

Dichotic headphone displays are simple to implement, but are only useful for indicating whether a sound source is located to the left or right of the listener. On the other hand, when multiple sources are lateralized at different locations (using different ITD and/or IID values), some binaural unmasking can be obtained (see Section 3.4.8).

The left and right signals of commercial stereo recordings generally contain simple ITD and IID cues in the direct sounds, but also contain reverberation and echoes. When these signals are played over headphones, sources are usually lateralized, but actually may seem more “realistic” than other signals due to the reverberation.

4.1.3 Spatialized Audio

Using signal-processing techniques, it is possible to generate stereo signals that contain most of the normal spatial cues available in the real world. In fact, if properly rendered, spatialized audio can be practically indistinguishable from free-field presentation (Langendijk & Bronkhorst, 2000). When coupled with a headtracking device, spatialized audio provides a true virtual auditory interface. Using a spatialized auditory display, a variety of sound sources can be presented simultaneously at different directions and distances. One of the early criticisms of spatialized audio was that it was expensive to implement; however, as hardware and software solutions have proliferated, it has become feasible to include spatialized audio in most systems. Spatialized audio solutions can be fit into any budget, depending on the desired resolution and number of sound sources required. Most head-mounted displays are currently outfitted with headphones of sufficient quality to reproduce spatialized audio, making it relatively easy to incorporate spatialized audio in an immersive VR system.

4.1.3.1 Head-Related Transfer Functions

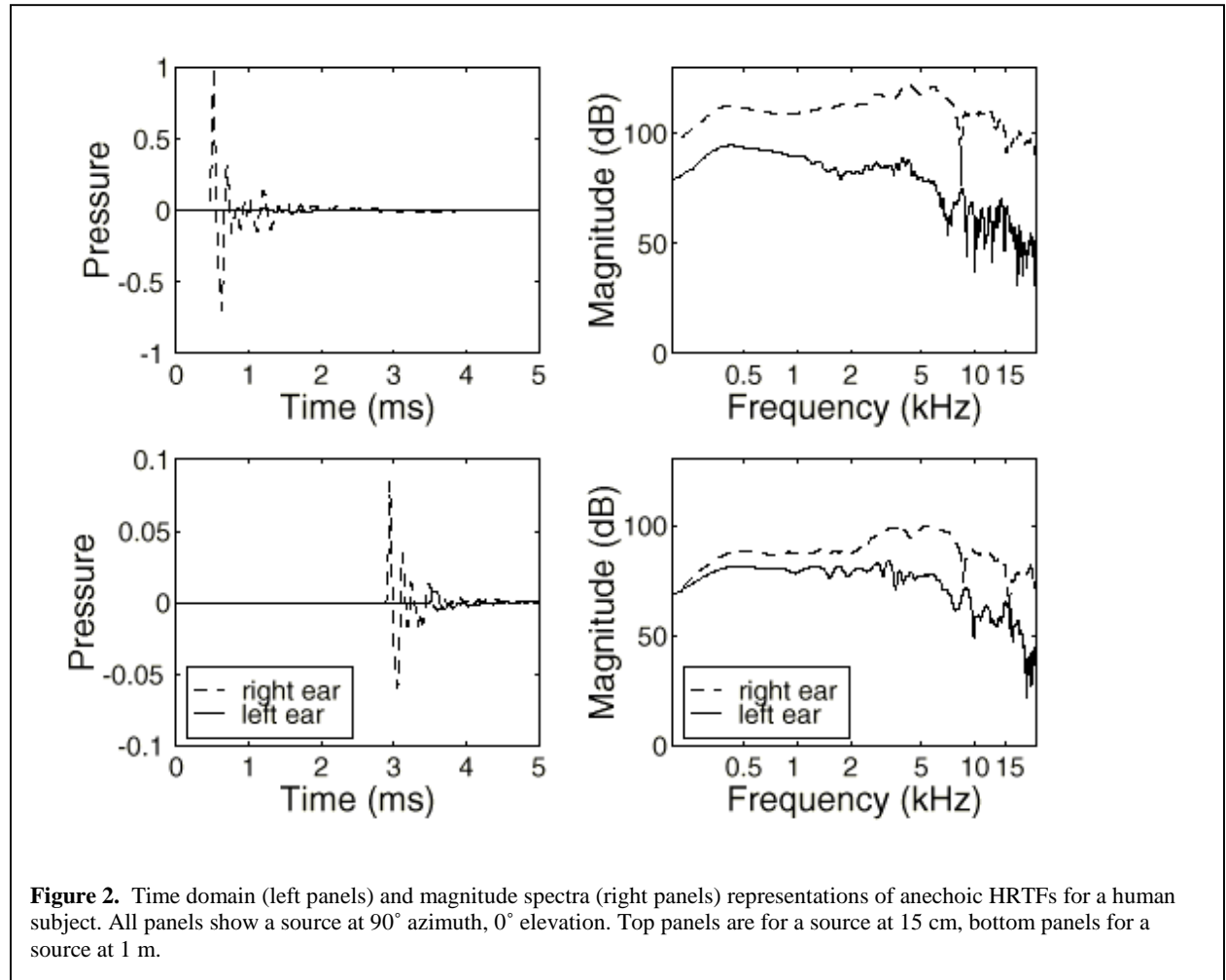
In order to simulate any source somewhere in space over headphones, one must simply play a stereo headphone signal that recreates at the eardrums the exact acoustic waveforms that would actually arise from a source at the desired location. This is generally accomplished by empirically measuring the transfer functions that describe how an acoustic signal at a particular location in space is transformed as it travels to and impinges on the head and ears of the listener. Then, in order to simulate an arbitrary sound source at a particular location, the appropriate transfer functions are used to filter the desired (known) source signal. The resulting stereo signal is then corrected to compensate for the transfer characteristics of the display system (for instance, to remove any spectral shaping of the headphones) and presented to the listener.

The pairs of spatial filters that describe how sound is transformed as it impinges on the listener are known as Head-Related Transfer Functions (HRTFs). HRTFs describe how to simulate the *direct* sound reaching the listener from a particular position, but do not generally include any reverberant energy. Empirically-measured HRTFs vary mainly with the direction from head to source, but also vary with source distance (particularly for sources within reach of the listener). For sources beyond about a meter away, the main effect of distance is to change the overall gain of the HRTFs. In the time domain, the HRTF pair for a particular source location give the pressure waveforms

that would arise at the ears if a perfect impulse were presented from the spatial location in question. Often, HRTFs are represented in the frequency domain by taking the Fourier Transform of the time domain impulse responses.

HRTFs contain most of the spatial information present in real-world listening situations. In particular, ITD and IID are embodied in the relative phase and magnitude (respectively) of the linear filters for the left and right ears. Spectral cues and source intensity are present in the absolute frequency-dependent magnitudes of the two filters.

Figure 2 shows two HRTF pairs from a human subject in the time domain (left side of figure) and in the frequency domain (magnitude only, right side of figure). All panels are for a source at azimuth 90° and elevation 0° . The top two panels show the HRTF for a source very close to the head (15 cm from the center of the head). The bottom two panels show the HRTF for a source 1 m from the head. In the time domain, it is easy to see the interaural differences in time and intensity, while the frequency domain representation shows the spectral notches that occur in HRTFs as well as the frequency-dependent nature of the interaural intensity difference. The interaural intensity differences are larger in all frequencies for the nearer source (top panels), as expected. In the time domain, the 1 m source must traverse a greater distance to reach the ears than the near source, resulting in additional time delay before the energy reaches the ears (note time onset differences in the impulse responses in the left top and left bottom panels).



4.1.3.2 Room Modeling

HRTFs generally do *not* include reverberation or echoes, although it is possible to measure transfer functions (known as room transfer functions) that incorporate the acoustic effects of the room. While possible, such approaches are generally not practical since such filters vary with listener and source position in the room as well as the relative position of listener and source to produce a combinatorially large number of transfer functions. In addition, such filters can be an order of magnitude longer than traditional HRTFs, increasing both computational and storage requirements of the system.

There has been substantial effort devoted to developing computational models for room reverberation (e.g., see the discussion in Shinn-Cunningham, Lehnert, Kramer, Wenzel, and Durlach (1997)). The required computations are quite intensive; in order to simulate each individual echo, one must calculate the distance the sound wave has

traveled, how the waveform was transformed by every surface upon which it impinged, and the direction from which it is arriving at the head. The resulting waveform must then be filtered by the appropriate anechoic HRTF based on the direction of incidence with the head.

If one looks at the resulting echoes as a function of time from the initial sound, the number of echoes in any given time slice increases exponentially with time (since the number of echoes grows exponentially with time as each echo impinges on multiple surfaces to effectively create new sources). At the same time the level of each individual echo decreases rapidly, both due to energy absorption at each reflecting surface and due to the increased pathlength from source to ear. Many simulations only “spatialize” a small number of the loudest, earliest-arriving echoes, and then add random noise that dies off exponentially in time (uncorrelated at the two ears) to simulate later arriving echoes that are dense in time and arriving from essentially random directions. Even with such simplifications, the computations necessary to generate “realistic” reverberation (particularly in a system that tries to account for movement of the listener) can be overwhelming (e.g., see Shinn-Cunningham, Lehnert et al., 1997).

Figure 3 shows the room impulse response at the right ear for a source located at 45° azimuth, 0° elevation, and distance 1 m. This impulse response was measured in a moderate-sized classroom in which significant reverberant energy persists for as long as 450 ms. The initial few ms of the response are shown in the inset. In the inset, the

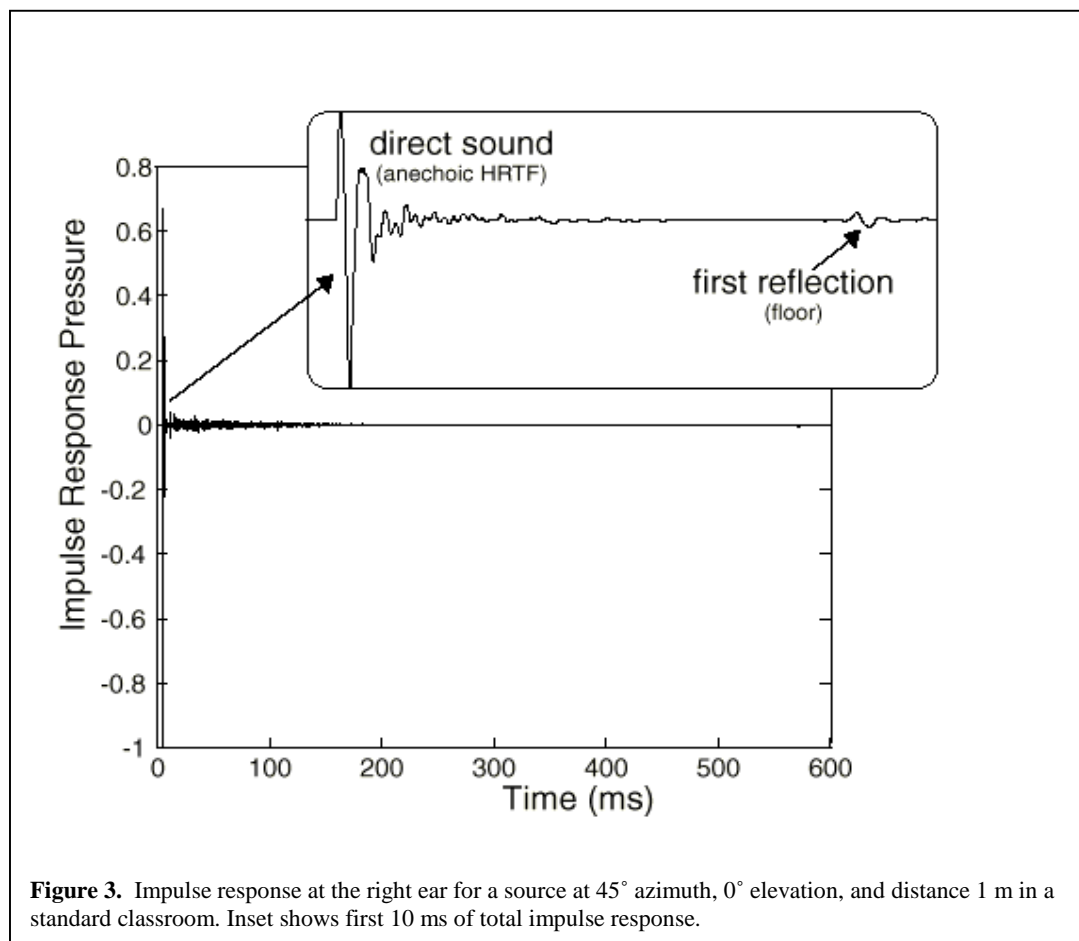


Figure 3. Impulse response at the right ear for a source at 45° azimuth, 0° elevation, and distance 1 m in a standard classroom. Inset shows first 10 ms of total impulse response.

initial response is that caused by the sound wave that travels directly from the source to the ear. The first reflection is also evident at the end of the inset, at a much reduced amplitude. In the main figure, the decay of the reverberant energy can be seen with time.

Development of tractable reverberation algorithms for real time systems is an ongoing area of research. The extent to which listeners are sensitive to the interaural and spectral details in reverberant energy is also not well understood and requires additional research. Nonetheless, there is clear evidence that inclusion of reverberation can have a dramatic impact on the subjective realism of a virtual auditory display and can aid in perception of source distance.

4.1.3.3 Practical Limitations on Spatialized Audio

While in theory, HRTF simulation should yield stimuli that are perceptually indistinguishable from natural experience, a number of practical considerations limit the realism of stimuli simulated using HRTFs. Measurement of HRTFs is a difficult, time-consuming process. In addition, storage requirements for HRTFs can be prohibitive. As a result, HRTFs are typically measured only at a single distance, relatively far from the listener, and at a relatively sparse spatial sampling. Changes in source distance are simulated simply by scaling the overall signal intensity. Since the HRTFs are only measured for a finite number of source directions at this single source distance, HRTFs are interpolated to simulate locations for which HRTFs are not measured. While this approach is probably adequate for sources relatively far from the listener and when some inaccuracy can be tolerated, the resulting simulation cannot perfectly recreate spatial cues for all possible source locations (Wenzel & Foster, 1993). Individual differences in HRTFs are very important for some aspects of sound source localization (particularly for distinguishing front/back and up/down). However, most systems employ a standard set of HRTFs that are not matched to the individual listener. Unfortunately, using these “nonindividualized” HRTFs reduces the accuracy and externalization of auditory images, but still results in useful performance increases (Begault & Wenzel, 1993). A great deal of ongoing research is focused on developing ways to tailor HRTFs to an individual listener without explicitly measuring HRTFs for that individual. For instance, researchers are exploring a variety of HRTF compression schemes in which individual differences are encoded in a small number of parameters that can be quickly or automatically fit to the individual (e.g., see Kistler and Wightman, 1991; Middlebrooks and Green, 1992).

Nonetheless, many “typical” systems cannot simulate source position along a donut of confusion since they do not use individualized HRTFs.

The most sophisticated spatialized audio systems use trackers to measure movement of the listener and update the HRTFs in real time to produce appropriate dynamic spatial cues. The use of head tracking dramatically increases the accuracy of azimuthal localization (Sorkin, Kistler, & Elvers, 1989). However, the time lag in such systems (from measuring the listener movement, choosing the new HRTF, and filtering the ongoing source signal) can be greater than 100 ms. While the binaural system is sluggish, the resulting delay can nonetheless be perceptible. Real-time systems are also too complex and costly for some applications. Instead, systems may compute signals off-line and either ignore or limit the movement of the user; however, observers may hear sources at locations inside or tethered to the head (i.e., moving with the head) with such systems.

Many simulations do not include any echoes or reverberation in the generated signals. Although reverberation has little impact (or degrades) perception of source direction, it is important for distance perception. In addition, anechoic simulations sound subjectively artificial and less realistic than do simulations with reverberation.

4.2 Simulation Using Speakers

The total acoustic signal reaching each ear is simply the sum of the signals reaching that ear from each source in an environment. Using this property, it is possible to vary spatial auditory cues (like ITD, IID, and spectrum) by controlling the signals played from multiple speakers arrayed around the listener. In contrast with headphone simulations, the signals at the two ears cannot be independently manipulated; i.e., changing the signal from any of the speakers changes the signals reaching both ears. As a result, it is difficult to precisely control the interaural differences and spectral cues of the binaural signal reaching the listener to mimic the signals that would occur for a real world source. However, various methods for specifying the signals played from each loudspeaker exist to simulate spatial auditory cues using loudspeakers.

To reduce the variability of audio signals reaching the ears, careful attention should be given to speaker placement and room acoustics. If speaker systems are not properly placed and installed in a room, even the best sound systems will sound inferior. Improperly placed speakers can reduce speech comprehension, destroy the sense of immersion, and dramatically reduce bass response (Holman, 2000). This is especially true when dealing with small rooms. Unfortunately, speaker placement will vary depending on the dimensions and shape of the room, as

well as the number of speakers employed. If the system is mobile, the sound system will have to be readjusted for every new location, unless the simulation incorporates its own enclosure. If the simulation will be housed in different sized rooms, the audio system (amplifiers and speakers) must have enough headroom to accommodate both large enclosures as well as small. An equalizer should be used to frequency balance the system. When possible, acoustical tile and diffusers should be employed where appropriate to reduce reverberation.

4.2.1 Nonspatial Display

Many systems use free field speakers in which each speaker presents an identical signal. Such systems are analogous to diotic headphone systems; although simple to develop, these displays (like diotic headphone displays) provide no spatial information to the user. Such systems can be used when spatial auditory information is unimportant and when segregation of simultaneous auditory signals is not critical. For instance, if the only objects of interest are within the visual field and interference between objects is not a concern, this kind of simplistic display may be adequate.

4.2.1 Stereo Display

The analog of the dichotic headphone display presents signals from two speakers simultaneously in order to control the perceived laterality of a “phantom” source. For instance, simply by varying the level of otherwise identical signals played from a pair of speakers can alter the perceived laterality of a phantom source. Most commercial stereo recordings are based on variations of this approach.

Imagine a listener sitting equidistant from two loudspeakers positioned symmetrically in front of the listener. When the left speaker is played alone, the listener hears a source in the direction of the left speaker (and ITD and IID cues are consistent with a source in that leftward direction). When the right speaker is played alone, the listener hears a source in the direction of the right speaker. When identical signals at identical levels are played from both speakers, each ear receives two direct signals, one from each of the symmetrically placed speakers. To the extent that the listener’s head is left-right symmetric, the total direct sound in each ear is identical, and the resulting percept will be of a single source at a location that gives rise to zero ITD and zero IID (e.g., in the listener’s median plane). Varying the relative intensity of otherwise identical signals played from the two speakers causes the gross ITD and IID cues

to vary systematically, producing a phantom source whose location between the two speakers varies systematically with the relative speaker levels (e.g., see Bauer, 1961; Zurek and Shinn-Cunningham, 1997).

This simple “panning” technique produces a robust perception of a source at different lateral locations; however, it is nearly impossible to precisely control the exact location of the phantom image. In particular, the way in which the perceived direction changes with relative speaker level depends upon the location of the listener with respect to the two loudspeakers. As the listener moves outside a restricted area (the “sweet spot”), the simulation degrades rather dramatically. In addition, reverberation can distort the interaural cues, causing biases in the resulting simulation. Nonetheless, such systems provide some information about source laterality.

4.2.2 Lessons from the Entertainment Industry

The ability to generate an accurate spatial simulation using loudspeakers increases dramatically as the number of speakers used in the display increases. With an infinite number of speakers around the listener, one would simply play the desired signal from the speaker at the desired location of the source to achieve a “perfect” reproduction. Panning between multiple pairs of speakers (for instance, speakers arrayed in front of and behind the listener) is often used to improve spatial simulations using loudspeakers.

These “Surround Sound” technologies are primarily seen in the entertainment industry and are implemented via a variety of formats. Surround sound systems find their genesis in a three-channel system created for the movie “Fantasia” in 1939. “Fantasound” speakers were located in front of the listener at left, middle, and right. The “surround” speakers consisted of approximately 54 speakers surrounding the audience and carried a mixture of sound from the left and right front speakers (i.e., it was not true stereo) (Garity & Hawkins, 1941). At \$85,000 per theater, few theaters were ever equipped to play “Fantasound”.

Currently, the most common Surround Sound format is the 5.1 speaker system in which speakers are located at the left, middle, and right in front of the listener, and left and right behind the listener. All sound below approximately 80 Hz is funneled through the so-called 0.1 speaker, a subwoofer. The middle speaker, located in front of the listener, reproduces most of the speech information to the listener. Typical 5.1 Surround formats are Dolby Digital Surround and Digital Theater Systems (DTS). Newer surround sound formats include THX Surround EX, which is a 7.1 speaker system (adding a center speaker behind the listener). Another format is the Sony Dynamic Digital Sound® (SDDS®) system, a 7.1 speaker system adding a left center and right center speaker in

front of the listener. A 10.2 speaker system is now on the horizon (Holman, 2000). In the future, even greater numbers of speakers and more complex processing are likely to become standard. However, it is important to note that adding additional speakers may be detrimental to producing a sense of immersion, especially in a small room. As the number of speakers increases, explicit care must be taken to assure that the sound field in the room is diffuse enough that the speakers themselves are not obtrusive. If the user notices the speakers in the room, the illusion of reality will be destroyed. To negate this possibility, extreme care must be taken to account for room acoustics, speaker design and placement, and the location of the listener in the room (Holman, 2000).

4.2.3 Cross-Talk Cancellation and Transaural Simulations

More complex signal-processing schemes can also be used to control spatial cues using loudspeakers. In such approaches, the total signal reaching each ear is computed as the sum of the signals reaching that ear from each of the speakers employed. By considering the timing and content of each of these signals, one can try to reproduce the exact signal desired at each ear.

The earliest such approach attempted to recreate the sound field that a listener would have received in a particular setting from stereo recordings taken from spatially separated microphones. In the playback system, two speakers were positioned at the same relative locations as the original microphones. The goal of the playback system was to play signals from the two speakers such that the total signal at each ear was equal to the recorded signal from the nearer microphone. To the extent that the signal from the far speaker was acoustically cancelled, the reproduction would be accurate. Relatively simple schemes involving approximations of the acoustic alterations of the signals as they impinged on the head were used to try to accomplish this “cross-talk cancellation.”

As signal processing approaches have been refined and knowledge of the acoustic properties of HRTFs improves, more sophisticated algorithms have been developed. In particular, it is possible to calculate the contribution of each speaker to the total signal at each ear by considering the HRTF corresponding to the location of the speaker. The total signal at each ear is then the sum of the HRTF-filtered signals coming from each speaker. If one also knows the location and source of the signal that is to be simulated, one can write equations for the desired signals reaching each ear as HRTF-filtered versions of the desired source. Combining these equations yields two frequency-dependent, complex-valued equations that relate the signals played from each speaker to the desired signals at the ears. To the extent that one can find and implement solutions to these equations, it is possible (at least in theory) to recreate the desired binaural signal by appropriate choice of the signals played from each speaker. The problem with such

approaches is that the simulation depends critically on the relative location of the speakers and the listener. In particular, if the listener moves his head outside of the sweet spot, the simulation degrades rapidly. It can be difficult to compute the required loudspeaker signals and the computations are not particularly stable, numerically. To the extent that the HRTFs used in the equations are not matched to the listener and reverberation distorts the signals reaching the ears, the derived solutions are even less robust.

Head trackers can be used in conjunction with multispeaker simulations in order to improve the simulation. However, this requires that computations be performed in real time and significantly increases the cost and complexity of the resulting system.

In general, it is possible to generate relatively realistic “phantom” sources using multiple loudspeakers whose lateral position changes with changes in the speaker waveforms. More complex simulations in which spectral cues are simulated are often too difficult to include. Loudspeaker simulations often achieve reasonable results for their cost; however, the accuracy of the simulation is much poorer than with headphone simulations, primarily because the signals at each ear cannot be independently controlled.

5. Design Considerations

5.1 Hardware

Despite the importance of audio cues in human existence, there are relatively few platforms available for easily implementing spatial audio into virtual environments. The most common platforms used today for implementing spatial audio in research and VE are the AuSim Convolvotron, Lake Huron, and Tucker Davis Technologies PD-1 Power SDAC and expander.

The AuSim GoldServer is based on the original Acoustetron II design pioneered by Crystal River Engineering (CRE; Wenzel, Wightman, & Foster, 1988). Although the original Acoustetron II is no longer commercially available, AuSIM acquired the rights to the device and is now producing boxes and replacement parts for the original CRE devices. AuSIM also claims that their new product line will not only fill the niche for Virtual Environments applications, but will also provide a more flexible research tool than its predecessor (personal communication). A typical system will cost between \$10k and \$16k and claims to be relatively easy to integrate into a variety of VE applications.

Hailing from Australia, Lake Technology offers a variety of hardware and software systems using either headphones or speakers. On the hardware side, the Huron System is a multiple DSP system that will present multiple sound sources to multiple listeners using positional tracking. The Huron can be configured using multiple speakers or headphones. Combined with an extensive set of software development tools, the Huron system is another good option for integrating sound in virtual environments. For instance, the software package “Multiscape” is advertised to be able to create a multiple user interactive auditory virtual environment complete with geometric rooms with interconnecting doors. Finally, in cooperation with Dolby, Lake has recently helped develop, and now supports, headphone systems that simulate Dolby 5.1 Surround Sound.

The Tucker-Davis Technologies (TDT) PD-1 Power SDAC system is primarily a research tool. However, with some creative programming, it too can be used as a sound server for simulations (Shilling et al, 2000). The PD-1 combines up to 27 DSPs, allowing researchers to create complex spatial audio scenarios including numerous reflections and incorporating headtracking. It also provides real-time convolution, allowing the research to convolve “live” signals.

A software solution currently under development for research in spatial sound is Sound Lab (Wenzel, Miller, and Abel, 2000). Sound Lab (SLAB) is designed to run on a regular PC platform and provides low-level control of a variety of signal processing functions, including the number of reflections, the number of filter taps, and the positions of reflections. No special purpose hardware is required for the software to run. In addition to allowing normal HRTF processing and the inclusion of reflections, the software allows for the manipulation of acoustic radiation patterns, spherical spreading loss, and atmospheric absorption. Preliminary analysis of the dynamic performance of SLAB shows it achieves a very low 24 ms latency on a 450 MHz Pentium II, although the latency would be larger for a more complex acoustic model. At this time, SLAB continues to be developed.

5.2 Defining the Auditory Environment

When creating the auditory portion of a virtual environment, careful attention should be placed on what is absolutely essential for the task. The adage of motion picture sound designers, “see it, hear it” (Holman 1997; Yewdall, 1999), is also valid for designing audio for virtual environments. The sense of immersion experienced in a movie theater is a carefully orchestrated combination of expertly designed sound effects and skillfully applied auditory ambiences. It is also interesting to note that “realistically” rendered sound is often perceived as

emotionally flat in motion pictures. Sound effects are often designed as exaggerated versions of reality to convey emotion or to satisfy the viewers' expectations of reality (Holman, 1997). Sound design in VE needs to balance the need for accurate reproduction with the need to make the user emotionally involved in a synthetic environment. When dealing with issues of emotionality in VE, sound should be considered as synonymous with emotion.

On the hardware side, a simple game card solution may be adequate if auditory spatialization and fidelity are not paramount. However, if there are multiple sound sources and/or a multi-user interface, an audio server such as the AuSIM Acoustetron or the Lake Huron System may be necessary. Just as we take photos and films of the visual environment during the development process, it is a good idea to make audio recordings, including sound level measurements, when developing the auditory interface. One of the current efforts in the Virtual Environment's in Training Technology program sponsored by the Office of Naval Research is to develop a systematic approach for obtaining baseline data concerning the content of an auditory environment. In addition to cataloguing the different sounds in a real environment, it is also important to systematically measure the intensity of sounds being experienced by the listener. In this manner, the VE developer has a highly detailed reference with which to compare the real world auditory environment with the virtual auditory environment. Two systems are currently being evaluated. The first system uses a portable Sony TCD-D8 DAT recorder coupled with Sennheiser microphone capsules to produce a crude spatialized recording (Figure 4, left). The microphone capsules will be inserted into an observer's auditory meatus (ear canal). In this manner, a complete spatialized recording can be made of the auditory environment, completely externalized with azimuth and elevation cues. The 2nd system (Figure 4, right) is more robust, using a larger set of microphones produced by Core Sound which can clip to a set of eyeglasses to produce a binaural recording. However, since the microphones are not inside the ear canals, this system will not include spectral cues.. Although pinnae cues cannot be utilized, the latter system has more robust construction and will be more tolerant of extreme conditions, especially if the recordings are made outdoors. Both systems can be clipped to

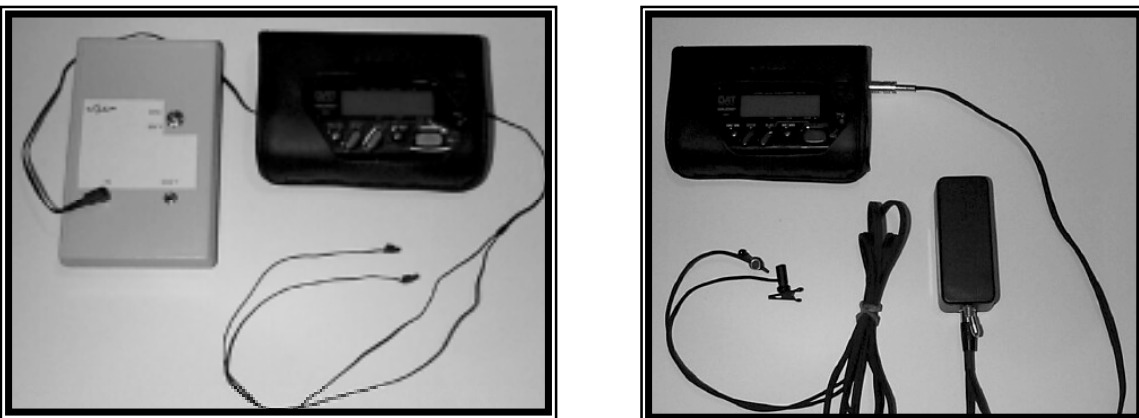


Figure 4. Two portable systems for making binaural recordings. On the left is a system using microphone capsules placed in the auditory meatus using Sennheiser microphone capsules (Tucker Davis Technologies). On the right is a binaural recording system from CORE Sound.

the belt and will be used in conjunction with a real time logging and event analyzer (CEL 593). The complete data set including sound recordings and sound measurements will be stored on CDROM for ease of use. The digital recordings also allow for spectral analyses to be conducted on specific auditory stimuli contained on the tape so that synthesized versions of those stimuli can be constructed. If concussive or low frequency sounds need to be recorded, microphones with wider frequency ranges should be employed for separate recordings. } Given the wide dynamic range involved with recording sounds ranging from concussive events to footsteps and the necessity that recordings be absolutely clean and accurate, the best solution may be to rely on professionals for making appropriate recordings. Different combinations of microphones and recording equipment produce vastly different sound qualities. Choosing the appropriate combination for a particular application is still more of an art than a science. In addition, there are many high-quality commercially available sound libraries for obtaining a wide variety of sound effects and ambiances (Holman, 1997; Yewdall, 1999).

5.3 How Much Realism Is Necessary?

Much of the research devoted to developing and verifying virtual display technology emphasizes the subjective “realism” of the display; however, this is not the most important consideration for all applications. In some cases, signal processing that improves “realism” actually interferes with the amount of information a listener can extract. For instance, the inclusion of echoes and reverberation can significantly increase the perceived realism of a display and improve distance perception. However, echoes can degrade perception of source direction. For applications in which information about the direction of the source is more important than the realism of the display or perception of source distance, including echoes and reverberation may be ill advised.

In headphone-based systems, realism is enhanced with the use of individualized HRTFs, particularly in the perception of up/down and front/back position. Ideally, HRTFs should also be sampled in both distance and direction at a spatial density dictated by human sensitivity. Thus, while the most “realistic” system would use individualized HRTFs that are sampled densely in both direction and distance, most systems use generic HRTFs sampled coarsely in direction and at only one distance.

If a particular application requires the user to extract three-dimensional spatial information from the auditory display, HRTFs may have to be tailored to the listener to preserve directional information and reverberation may

have to be included to encode source distance. On the other hand, if a particular application only makes use of one spatial dimension (for instance, to indicate the direction that a blind user must turn), coarse simulation of ITD and IID cues (even without detailed HRTF simulation) is probably adequate.

If information transfer is of primary importance, it may be useful to present acoustic spatial cues that are intentionally distorted so that they are perceptually more salient than are more “realistic” cues. For instance, it may be useful to exaggerate spatial auditory cues to improve auditory resolution; however, such an approach requires that listeners are appropriately trained with the distorted cues (Shinn-Cunningham, Durlach and Held, 1998).

The processing power needed to simulate the most realistic virtual auditory environment possible is not always cost effective. For instance, the amount of computation needed to create realistic reverberation in a virtual environment may not be justifiable when source distance perception and subjective realism are not important. Other acoustic effects are often ignored in order to reduce the computational complexity of the acoustic simulation, including the non-uniform radiation pattern of a realistic sound source, spectral changes in a sound due to atmospheric effects, and Doppler shift of the received spectrum of moving sources. The perceptual significance of many of these effects is not well understood; further work must be done to examine how these factors affect the realism of the display as well as what perceptual information such cues may convey.

In command and control applications, the goal is to maximize information transfer into the human operator; subjective impression (i.e., “realism”) is unimportant. In these applications, both technological and perceptual issues must be considered to achieve this goal. If nonverbal warnings or alerts are created, the stimuli must be wideband enough to be localizable. In addition, stimuli should be significantly intense to be at least 15 dB above background noise level. Stimulus onset should be fairly gradual so as not to be excessively startling to the user. In many instances, one may want the stimuli to be aesthetically pleasing to the user. As can be imagined, creating acceptable spatialized auditory displays is no trivial chore and should involve formal evaluations to ensure perceptual accuracy and system usability. For applications in which speech is the main signal of interest, basic interaural cues are important for preserving speech intelligibility, particularly in noisy, multi-source environments. On the other hand, there is probably little benefit gained from including the detailed frequency-dependence of normal HRTFs. In entertainment applications, cost is the most important factor; the precision of the display is unimportant as long as the simulation is subjectively satisfactory. For scientific research, high-end systems are necessary in order to allow careful examination of normal spatial auditory cues. In clinical applications, the auditory display must only be able

to deliver stimuli that can distinguish listeners with normal spatial hearing from those with impaired spatial hearing. Such systems must be inexpensive and easy to use, but there is no need for a “perfect” simulation.

References

Algazi, V.R., C. Avendano and R. O. Duda (2001). "Elevation localization and head-related transfer function analysis at low frequencies," *J. Acoust. Soc. Am.*, Vol. 109, No. 3, pp. 1110-1122.

Batteau, D.W. (1967). The Role of the Pinna in Human Localization. *Proceedings of the Royal Society Of London, B168*, 158-180.

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

Begault, D.R. (1993). Head-up Auditory Displays for Traffic Collision Avoidance System Advisories: A preliminary investigation. *Human Factors*, 35(4), 707-717.

Begault, D.R. (1993). Call Sign Intelligibility Improvement Using a Spatial Auditory Display. *NASA Technical Memorandum* 104014.

Begault, D.R. (1996). Virtual Acoustics, Aeronautics, and Communications. Presented at the 101st Convention of the Audio Engineering Society, Nov 8-11, Los Angeles, CA.

Begault, D.R. (1999). Virtual Acoustic Displays for Teleconferencing: Intelligibility Advantage for "Telephone-Grade" Audio. *Journal of the Audio Engineering Society*, 47(10), 824-828.

Begault, D.R., & Wenzel, E.M. (1992). Techniques and Applications for Binaural Sound Manipulation in Human-Machine Interfaces. *The International Journal of Aviation Psychology*, 2(1), 1-22.

Begault, D.R., & Wenzel, E.M. (1993). Headphone Localization of Speech. *Human Factors*, 35(2), 361-376.

Besing, J. M. and J. Koehnke (1995). A test of virtual auditory localization. *Ear and Hearing*, 16(2): 220-229.

Blauert, J. (1997). *Spatial Hearing (2e)*. Cambridge, MA: MIT Press.

Brainard, M. S., E. I. Knudsen, et al. (1992). Neural derivation of sound source location: Resolution of ambiguities in binaural cues. *Journal of the Acoustical Society of America*, 91: 1015-1027.

Bregman, A. S. (1990). Auditory Scene Analysis: The Perceptual Organization of Sound. Cambridge, MA, MIT Press.

Bronkhorst, A. W. (2000). "The cocktail party phenomenon: A review of research on speech intelligibility in multiple-talker conditions." *Acustica* 86: 117-128.

Bronkhorst, A. W. and R. Plomp (1988). The effect of head-induced interaural time and level differences on speech intelligibility in noise. *Journal of the Acoustical Society of America*, 83: 1508-1516.

Bronkhorst, A.W., Veltman, J.A., Veltman, J.A., & van Breda, L. (1996) Application of a Three-Dimensional Auditory Display in a Flight Task. *Human Factors*, 38(1), 23-33.

Brugge, J. F., R. A. Reale, and J. E. Hind (1997). Auditory cortex and spatial hearing. *Binaural and*

Spatial Hearing in Real and Virtual Environments. R. Gilkey and T. Anderson. New York, Erlbaum: 447-474.

Brungart, D. S. and W. M. Rabinowitz (1999). Auditory localization of nearby sources I: Near-field head-related transfer functions. Journal of the Acoustical Society of America **in press**.

Clark, B. and Graybiel, A. (1949). The Effect of Angular Acceleration on Sound Localization: The Audiogyral Illusion. The Journal of Psychology, 28, 235-244.

Clifton, R. K., R. L. Freyman, et al. (1993). Listener expectations about echoes can raise or lower echo threshold. Journal of the Acoustical Society of America.

Dizio, P., R. Held, J. R. Lackner, B. G. Shinn-Cunningham, and N. I. Durlach (2000). "The effect of changes in magnitude and direction of the resultant linear force on head-centric auditory localization." Experimental Brain Research **submitted**.

Drullman, R. and Bronkhorst, A.W. (2000). Multichannel Speech intelligibility and talker recognition using monaural, binaural, and three-dimensional auditory presentation. Journal of the Acoustical Society of America, 107(4): 2224-2235.

Duda, R. O. and W. L. Martens (1998). Range dependence of the response of a spherical head model. Journal of the Acoustical Society of America 104(5): 3048-3058.

Durlach, N. I. and H. S. Colburn (1978). Binaural phenomena. Handbook of Perception. E. C. Carterette and M. P. Friedman. New York, Academic Press. 4: 365-466.

Durlach, N. I., B. G. Shinn-Cunningham, and R. M. Held (1993). Supernormal auditory localization. I. General background. Presence 2(2): 89-103.

Frens, M. A., van Opstal, A. J. & van der Willigen, R. F. (1995). Spatial and temporal factors determine auditory-visual interactions in human saccadic eye movements. Perception and Psychophysics, 57, 802-816.

Garity, W.E., & Hawkins, J.A. (1941). Fantasound. Journal of the Society of Motion Picture Engineers, August.

Gaver, W. W. (1986). Auditory icons: Using sound in computer interfaces. Human-Computer Interaction, 2, 2, 167-177.

Gelfand, S. A. (1998). Hearing: An Introduction to Psychological and Physiological Acoustics. New York: Marcel Dekker, Inc.

Gilkey, R. H. and T. R. Anderson (1995). The accuracy of absolute localization judgments for speech stimuli. Journal of Vestibular Research 5(6): 487-497.

Gilkey, R. H. and M. D. Good (1995). Effects of frequency on free-field masking. Human Factors 37(4): 835-843.

Grantham, D. W. (1997). Auditory motion perception: Snapshots revisited. Binaural and Spatial Hearing in Real and Virtual Environments. R. Gilkey and T. Anderson. New York, Erlbaum: 295-314.

Haas, E. C., Gainer, C., Wightman, D., Couch M., and Shilling, R.D. (1997). Enhancing System Safety with 3-D Audio Displays. Proceedings of the Human Factors and Ergonomics Society 41st Annual Meeting. 868-872, Albuquerque, NM.

Hartmann, W. M. (1983). Localization of sound in rooms. Journal of the Acoustical Society of America 74: 1380-1391.

Hendrix, C. and Barfield, W. (1996). The sense of presence in auditory virtual environments. Presence **5**(3): 290-301.

Holman, T. (1997). *Sound for Film and Television*. Boston, MA. Focal Press.

Holman, T. (2000). *5.1 Surround Sound Up and Running*. Boston, MA. Focal Press.

Hofman, P. M., J. G. A. Van Riswick, et al. (1998). Relearning sound localization with new ears. Nature Neuroscience **1**(5): 417-421.

Kistler, D. J. and F. L. Wightman (1991). A model of head-related transfer functions based on principal components analysis and minimum-phase reconstruction. Journal of the Acoustical Society of America **91**: 1637-1647.

Koehnke, J., J. Besing, C. Goulet, M. Allard, and P. M. Zurek (1992). Speech intelligibility, localization, and binaural detection with monaural and binaural amplification. Journal of the Acoustical Society of America **92**: 2434.

Koehnke, J. and J. M. Besing (1996). A procedure for testing speech intelligibility in a virtual listening environment. Ear and Hearing **17**(3): 211-217.

Kollmeier, B. and R. H. Gilkey (1990). Binaural forward and backward masking: Evidence for sluggishness in binaural detection. Journal of the Acoustical Society of America **87**(4): 1709-1719.

Langendijk, E. H., and Bronkhorst, A.W. (2000). Fidelity of three-dimensional-sound reproduction using a virtual auditory display. Journal of the Acoustical Society of America **107**(1): 528-537.

Letowski, T., Vause, N., Shilling, R., Ballas, J., Brungert, D. & McKinley, R. (2000). *Human Factors Military Lexicon: Auditory Displays*. ARL Technical Report, ARL-TR-xxxx; APG (MD), in print

Litovsky, R. Y. (1998). Physiological studies of the precedence effect in the inferior colliculus of the kitten. Journal of the Acoustical Society of America **103**(6): 3139-3152.

Mershon, D. H. (1997). Phenomenal geometry and the measurement of perceived auditory distance. Binaural and Spatial Hearing in Real and Virtual Environments. R. Gilkey and T. Anderson. New York, Erlbaum: 257-274.

Mershon, D. H., W. L. Ballenger, A. D. Little, P. L. McMurty, and J. L. Buchanan (1989). Effects of room reflectance and background noise on perceived auditory distance. Perception **18**: 403-416.

Middlebrooks, J. (1994). A panoramic code for sound location by cortical neurons. Science **264**: 842-844.

Middlebrooks, J. C. (1997). Spectral shape cues for sound localization. Binaural and Spatial Hearing in Real and Virtual Environments. R. Gilkey and T. Anderson. New York, Erlbaum: 77-98.

Middlebrooks, J. C. and D. M. Green (1991). Sound localization by human listeners. Annual Review of Psychology **42**: 135-159.

Middlebrooks, J. C. and D. M. Green (1992). Observations on a principal components analysis of head-related transfer functions. Journal of the Acoustical Society of America **92**: 597-599.

Miller, J. D., Abel, J. S. and Wenzel, E. M (1999) Implementation issues in the development of a real-time, Windows-based system to study spatial hearing. Journal of the Acoustical Society of America, **105**, 1193.

Mills, A. W. (1972). Auditory localization. Foundations of Modern Auditory Theory. J. V. Tobias. New

York, Academic Press: 303-348.

Moore, B. C. J. (1997). An Introduction to the Psychology of Hearing (4e). San Diego, CA: Academic Press.

Perrott, D.R., Saberi, K., Brown, K. and Strybel T. (1990). Auditory psychomotor coordination and visual search behavior. Perception & Psychophysics, **48**, 214-226.

Perrott, D.R., Sadralodabai, T., Saberi, K., and Strybel T. (1991). Aurally Aided visual search in the central visual field: Effects of visual load and visual enhancement of the target. Human Factors, **33**, 389-400

Perrott, D.R., Constantino, B., and Cisneros, J. (1993). Auditory and visual localization performance in a sequential discrimination task. Journal of the Acoustical Society of America, **93(4)**, 2134-2138.

Pick, H. L., D. H. Warren, and J. C. Hay (1969). Sensory conflict in judgements of spatial direction. Perception and Psychophysics **6**: 203-205.

Plenge, G. (1974). On the differences between localization and lateralization. Journal of the Acoustical Society of America, **56(3)**, 944-951.

Poon, P. W. F. and J. F. Brugge (1993). Virtual space receptive fields of single auditory nerve fibers. Journal of Neurophysiology **70**: 667-676.

Rabenhorst, D.A., Farrell, E.J., & Jameson, D. (1990). Complementary Visualization and Sonification of Multi-Dimensional Data. IBM Technical Report RC 15467 (#68449).

Radeau, M. and Bertelson, P. (1976). The effect of a textured visual field on modality dominance in a ventriloquism situation. Perception & Psychophysics **20(4)**: 227-235.

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

Rakerd, B. and W. M. Hartmann (1986). Localization of sound in rooms. III. Onset and duration effects. Journal of the Acoustical Society of America **80**: 1695-1706.

Sheriden, T.B. (1996). Further Musings on the Psychophysics of Presence. Presence **5(2)**: 241-246.

Shilling, R.D. , and T. Letowski. (2000). Using Spatial Audio Displays to Enhance Virtual Environments and Cockpit Performance. NATO Research and Technology Agency Workshop entitled, "What is essential for Virtual Reality to Meet Military Human Performance Goals", The Hague., The Netherlands.

Shilling, R.D., Wightman, D., Couch M., Beutler, R. and Letowski, T. (1998). The Use of Spatialized Auditory Displays in an Aviation Simulation. Proceedings of the 16th Applied Behavioral Sciences Symposium, Colorado Springs, CO.

Shinn-Cunningham, B. G., N. I. Durlach, et al. (1998a). Adapting to supernormal auditory localization cues I: Bias and resolution. Journal of the Acoustical Society of America **103(6)**: 3656-3666.

Shinn-Cunningham, B. G., N. I. Durlach, et al. (1998b). Adapting to supernormal auditory localization cues II: Changes in mean response. Journal of the Acoustical Society of America **103(6)**: 3667-3676.

Shinn-Cunningham, B. G. (2000a). "Adapting to remapped auditory localization cues: A decision-theory model." Perception and Psychophysics **62(1)**: 33-47.

- Shinn-Cunningham, B. G. (2000b). Learning reverberation: Implications for spatial auditory displays. International Conference on Auditory Displays, Atlanta, GA.
- Shinn-Cunningham, B. G., S. Santarelli, et al. (2000). "Tori of confusion: Binaural localization cues for sources within reach of a listener." Journal of the Acoustical Society of America **107**(3): 1627-1636.
- Shinn-Cunningham, B. G., N. Kopco, and S. G. Santarelli. (1999). Computation of acoustic source position in near-field listening. 3rd International Conference on Cognitive and Neural Systems, Boston, MA.
- Shinn-Cunningham, B. G., H. Lehnert, G. Kramer, E. M. Wenzel, and N. I. Durlach (1997). Auditory Displays. Binaural and Spatial Hearing in Real and Virtual Environments. R. Gilkey and T. Anderson. New York, Erlbaum: 611-663.
- Shinn-Cunningham, B. G., P. M. Zurek, and N. I. Durlach (1993). Adjustment and discrimination measurements of the precedence effect. Journal of the Acoustical Society of America **93**: 2923-2932.
- Sorkin, R.D., Kistler, D.S. & Elvers, G.C. (1989). An Exploratory Study of the Use of Movement-Correlated Cues in an Auditory Head-Up Display, Human Factors, **31**(2), 161-166.
- Stern, R. M. and C. Trahiotis (1997). Binaural mechanisms that emphasize consistent interaural timing information over frequency. 11th International Symposium on Hearing: Auditory Physiology and Perception, Grantham, UK.
- Storms, Russell L. (1998). Auditory-visual cross-modal perception phenomena. Doctoral Dissertation. Naval Postgraduate School, Monterey, California.
- Trahiotis, C. and R. M. Stern (1989). Lateralization of bands of noise: Effects of bandwidth and differences of interaural time and phase. Journal of the Acoustical Society of America **86**(4): 1285-1293.
- Wallach, H. (1940). "The role of head movements and vestibular and visual cues in sound localization." Journal of Experimental Psychology **27**: 339-368.
- Warren, D. H., R. B. Welch, and T. J. McCarthy (1981). The role of visual-auditory 'compellingness' in the ventriloquism effect: implications for transitivity among the spatial senses. Perception and Psychophysics **30**: 557-564.
- Welch, R. and D. H. Warren (1986). Intersensory interactions. Handbook of Perception and Human Performance. K. R. Boff, L. Kaufman and J. P. Thomas. New York, John Wiley and Sons, Inc. II: Cognitive Processes and Performance: 25.1-25.36.
- Welch, R. B. and D. H. Warren (1980). Immediate perceptual response to intersensory discrepancy. Psychological Bulletin **88**: 638-667.
- Wenzel, E. M., Arruda, M., Kistler, D.J., and Wightman, F.L. (1993). Localization using nonindividualized head-related transfer functions. Journal of the Acoustical Society of America **94**: 111-123.
- Wenzel, E.M. and Foster S.H. (1993). Perceptual Consequences of Interpolating Head-Related Transfer Functions During Spatial Synthesis. Proceedings of the 1993 Workshop on the Applications of Signal Processing to Audio and Acoustics, Oct 17-20, New York, N.Y.
- Wenzel, E. M., Miller, J. D., and Abel, J. S. (2000) Sound Lab: A real-time, software-based system for the study of spatial hearing. Proceedings of the 108th Convention of the Audio Engineering Society, Paris, Feb. Preprint 5140.

Wenzel, E.M., Wightman, F.L., and Foster, S.H. (1988). Development of a Three-Dimensional Audiotry Display System. SIGCHI Bulletin, **20**, 52-57.

Wightman, F. L. and D. J. Kistler (1989). Headphone simulation of free-field listening. II. Psychophysical validation. Journal of the Acoustical Society of America **85**: 868-878.

Wightman, F. L. and D. J. Kistler (1993). Sound localization. Human Psychophysics. W. A. Yost, A. N. Popper and R. R. Fay. New York, Springer Verlag: 155-192.

Wightman, F. L. and D. J. Kistler (1999). Resolution of front-back ambiguity in spatial hearing by listener and source movement. Journal of the Acoustical Society of America **105**(5): 2841-2853.

Yewdall, D.L. (1999). Practical Art of Motion Picture Sound. Boston, MA. Focal Press.

Yost, W. A. (1994). Fundamentals of Hearing: An Introduction (3e). San Diego, CA: Academic Press.

Yost, W. A. (1997). The cocktail party problem: Forty years later. Binaural and Spatial Hearing in Real and Virtual Environments. R. Gilkey and T. Anderson. New York, Erlbaum: 329-348.

Zurek, P. M. (1993). Binaural advantages and directional effects in speech intelligibility. Acoustical Factors Affecting Hearing Aid Performance. G. Studebaker and I. Hochberg. Boston, MA, College-Hill Press.

Zurek, P. M. and B. G. Shinn-Cunningham (1997). Localization cues in intensity-difference stereophony. Journal of the Acoustical Society of America **under review**.