A 3-D Sound Primer: Directional Hearing and Stereo Reproduction

Author(s): Gary S. Kendall

Source: *Computer Music Journal*, Winter, 1995, Vol. 19, No. 4 (Winter, 1995), pp. 23-46

Published by: The MIT Press

Stable URL: https://www.jstor.org/stable/3680989

Gary S. Kendall
Center for Music Technology
School of Music
Northwestern University
Evanston, Illinois 60208, USA
g-kendall@nwu.edu

A 3-D Sound Primer: Directional Hearing and Stereo Reproduction

Background

Imagine yourself standing outside of your home. You close your eyes. You hear the sound of cars moving by, people talking next door, birds singing overhead, construction work going on in the distance. What is it that gives this experience its particular vibrancy? It might be the individual qualities of the sound sources themselves, but, more likely, it is the clear sense that you are within a dynamic 3-D sound space. A crucial aspect of your awareness and appreciation of sound in everyday life is that sound comes from all directions. You can identify the location of auditory events in 3-D space instantly and effortlessly. Of course, you are familiar with your environment, and you know the kinds of sounds you are likely to hear and their probable locations. Also crucial to this experience is that you are constantly interacting with the environment as you move your head or change your location. You receive a continuous flow of information from all of your senses, which changes in response to your movements. This information helps you to update your cognitive model of the environment, which in turn establishes the spatial context in which you judge the location and other spatial properties of auditory events.

Contrast this everyday experience with that of listening to recorded music. Traditional stereo reproduction provides you with some spatial information, but not enough to recreate the full dimensionality of being in a room with a "live" musical performance. Rather than giving the feeling of being within a 3-D space, loudspeaker reproduction creates the impression that you are in front of the sound space, while headphone reproduction makes you feel as though the sound space is inside your head. Consider, too, that when you listen to a sound recording, you receive sensory information about the recorded "events," but you cannot interact with the events to update, test, and refresh your cognitive model of the environment. With few exceptions (such as music videos), you do not integrate information from your other senses while listening. You are relegated to the role of an immobile observer with impoverished sensory information.

Certain recordings have architectural associations that help to establish a spatial context. Classical orchestras perform in concert halls, and jazz combos perform in small clubs. Rock music, though, does not have such a clear archetypical, environmental context. It is amplified to begin with, and as such, is a creature of electronic reproduction. Rock music, in particular, reveals that listening to recorded music is a special idiom of everyday experience, with a complex set of conventions governing the presentation and apprehension of auditory events. Recorded music is a learned "cultural form" that we usually take to be a mediated approximation to the direct experience. Especially in the case of rock music, the recorded form of the music has become the primary archetype.

The primary difference in the two situations described above is essentially that of direct versus mediated experience. Today's enthusiasm for multimedia and virtual reality seems to be part of a cultural desire to create artistic/cultural forms without these intervening conventions of presentation, forms that stand in a much more immediate relationship to direct experience than do conventional audio or video recordings. Although the spatial properties of rock music are products of studio art, a need is still felt to invent a form of more-direct experience within the existing musical genre. An intrinsic part of that direct experience is 3-D sound.

Computer Music Journal, 19:4, pp. 23–46, Winter 1995

The Scope of 3-D Sound Technology

While it can be argued that some traditional stereo recordings produce effects that could be called 3-D, when we refer to "3-D sound," we generally mean that the listener hears sounds in directions that are not experienced with conventional stereo. A 3-D sound system should exceed conventional stereo's range of directions in either azimuth or elevation, or both. For example, a 3-D system for loudspeaker reproduction should be able to position sounds outside the boundaries of the loudspeakers, and a 3-D headphone system should be able to place sounds outside the head, to the listener's front and rear.

The key technical innovation that enables these advances over traditional stereo is the use of signal processing to superimpose directionally dependent transfer functions (DTFs) on the stereo output signals. These transfer functions must recreate the complex acoustic cues used by listeners in everyday life to determine the direction of a sound in 3-D space. Theoretically, the entire breadth of 3-D auditory phenomena that is experienced every day can be recreated with such a system. The construction and manipulation of the DTFs is much more easily accomplished with computer technology than with analog circuits, and this fact helps to explain why 3-D sound has matured so recently. Another important reason for this late start is that the phase information essential to 3-D listening is not well preserved on vinyl records. Therefore, the emergence of 3-D sound as a commercial phenomenon has been greatly facilitated by the development of compact discs and high-quality digital-to-analog converters.

While the key feature of a 3-D sound system is its ability to "directionalize" sound, there are a number of other perceptual attributes associated with listening to sounds in space that can, potentially, be designed into a 3-D system. For example, the opposite effect to directionalization is non-directionalization. Non-directionalized sounds are usually described as diffuse sound fields occupying a region of 3-D space (Kendall 1995). Another perceptual attribute is the perceived distance of the sound image. The manipulation of distance is usually accomplished by direct control of simulated reverberation (Chowning 1971), although sound images close to the head can be produced with directional transfer functions alone (as discussed later in this article, and, somewhat differently, in Kendall 1995). If there is reverberation, the listener may interpret it as presenting information about the environment, such as the size of the room or the reflective properties of the walls and furnishings. The sound images in the reverberant environment may be characterized by the listener according to their degree of "definition" and "spaciousness," critical qualities in the design of concert halls. (See Rasch and Plomp 1982 for an excellent review of subjective room acoustics.) It is possible to control these perceptual attributes by simulating the physical environments (Borish 1984; Kendall and Martens 1984; Kendall et al. 1986; Kleiner, Dalenback, and Svensson 1993). The listener's perception of these environmental attributes does not always require 3-D directionalization. (For example, traditional stereo recordings of classical music have excelled in this area.) But the marriage of directionalization and environmental simulation can produce a sense of "being there," in direct sensory contact with physical reality, that is never achieved with traditional stereo reproduction. Yet, one expects that the evolution of 3-D sound technology will be driven not so much by the modeling of physical reality as by the demands of creative artists who will invent new artistic idioms for 3-D sound.

The Goal of This Article

As the technology for 3-D sound advances, there is a need to summarize and re-explain the field. The goal of this article is to provide a primer on 3-D sound technology for people in the professional community who find this topic an increasingly important part of their fundamental knowledge, as well as for the upcoming young professionals who need a starting point in their own learning process. (The reader seeking more in-depth coverage is directed to the books by Begault 1994 and Blauert 1974.) This paper focuses on the core technical issue of 3-D sound: the scientific and engineering

means by which a person listening to a stereo reproduction system can perceive the direction of a sound in 3-D space. The discussion is organized in two sections. The first section discusses the scientific basis of directional hearing, while the second section discusses practical techniques for 3-D stereo reproduction.

The Scientific Basis of 3-D Sound

The scientific basis of 3-D sound is captured in the literature of three separate disciplines: physical acoustics, psychoacoustics, and auditory neurophysiology. Physical acoustics focuses on the sound waves that reach the listener's eardrums, and the acoustic phenomena that determine their specific properties. Psychoacoustics studies the relationship between the acoustic waves at the eardrums and the perception of spatial imagery reported by listeners. Auditory neurophysiology is concerned with understanding the neurological structures that give rise to the experience of sound.

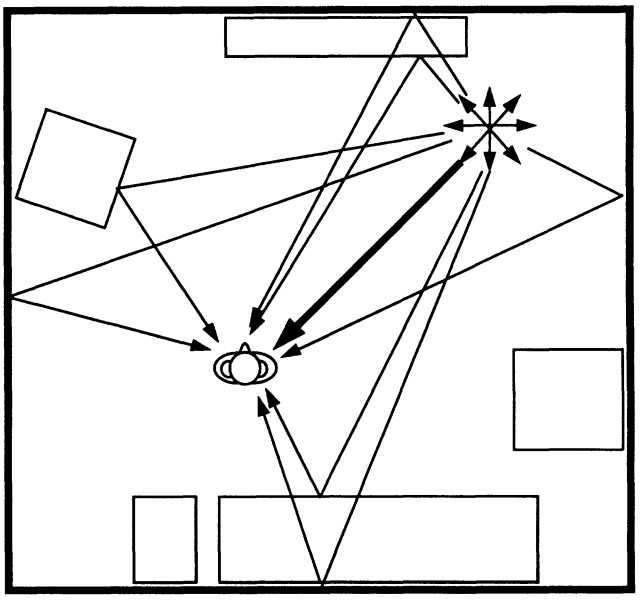
The discussion below will consider 3-D sound from the perspective of each one of these disciplines in turn. Knowledge gained within any one discipline is insufficient to understand many of the phenomena that we take for granted in everyday life, and as the technology for 3-D sound continues to develop, professionals increasingly need to draw upon multi-disciplinary sources of information.

Physical Acoustic Perspective

When an acoustic event occurs in the natural environment, sound waves from that event propagate in all directions. The waves encounter objects in the environment with which they interact by reflection and diffraction. The constructive and destructive interference of all the resulting waves creates a rich acoustic admixture that is further enriched when there are multiple sound sources.

One of the potential objects encountered in the environment is a listener. At the listener's position, sound waves are arriving at different times and from various directions. As shown in Figure 1, there is typically one straight-line path along

Figure 1. Depiction of sound events in an environment. There is one direct sound path (thick line) between the event and the listener, and many indirect sound paths (thin lines).



which the initial waves of each event first reach the listener. This initial *direct sound* provides the least-compromised information about the direction of the sound event. Later, sound waves are reflected back from objects in the environment, and arrive from many directions with different time delays. This *indirect sound* provides information about the environment and the relative position of the sound event within the environment, especially its distance from the listener. For as long as the sound event persists, direct sound and indirect sound are simultaneously present and virtually indistinguishable.

When a sound wave encounters a listener, there are two acoustic results depending on the frequency: (1) high-frequency energy is specularly reflected away, and (2) low-frequency energy diffracts and bends around the listener. In between, there is a transition band that is centered around 1,500 Hz, the frequency whose wavelength is approximately equal to the diameter of the head. This acoustic phenomenon can be thought of as analogous to ocean waves hitting the piling of a pier: small waves bounce off, while large waves bend around and go past it.

Figure 2. Relationship of the median, horizontal, and frontal (lateral) planes to the listener's head.

Figure 3. Specifying the position of a sound event relative to the head in terms of azimuth, elevation, and distance.

The sound waves that reach the listener's two eardrums are affected by the interaction of the original sound wave with the listener's torso, head, *pinnae* (outer ears), and ear canals. The composite of these properties can be measured and captured as a head-related transfer function (HRTF). The complexity of the interaction of the sound wave with the acoustics of the listener's body makes the HRTF at each ear strongly dependent on the direction of the sound.

When a sound event is equidistant from the two ears, the sound arrives at each ear from the same direction and the HRTFs are very similar (but not identical due to slight asymmetries of the head). The region in which sound sources are equidistant from the two ears is called the *median plane*. (The similarity of acoustic information is often given as the reason why localization accuracy is poor on the median plane.) There are two other names by which researchers refer to planes in 3-D space. One is the *horizontal plane* which is level with the listener's ears. The other is the *frontal plane* (or lateral plane), which divides the listener's head vertically between the front and the back. These planes are illustrated in Figure 2.

When the source is not equidistant from the ears, the signal arrives at each ear from a different direction and the HRTFs are far from identical. The ear nearest the sound source is called the *ipsilateral ear* and the ear farthest from the sound source is called the *contralateral ear.* The position of a sound source relative to the center of the listener's head is most conveniently captured as a vector expressed in terms of two angles, *azimuth* and *elevation*, and one scalar, *distance* (see Figure 3). Azimuth is measured as the angle between a projection of the vector onto the horizontal plane and a vector extending directly in front of the listener. A progressive movement from 0 to 360 degrees would take the source completely around the listener's head. (There is no general agreement as to whether 90 degrees azimuth represents the listener's left or right.) Elevation is measured as the angle formed between the vector and the horizontal plane rising to 90 degrees overhead or descending to −90 degrees below.

As shown in Figure 4, the signals arriving at the eardrums can be examined from two perspectives:

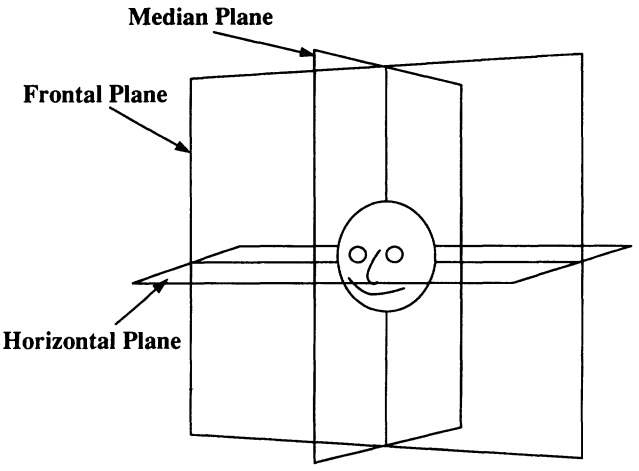


Figure 2

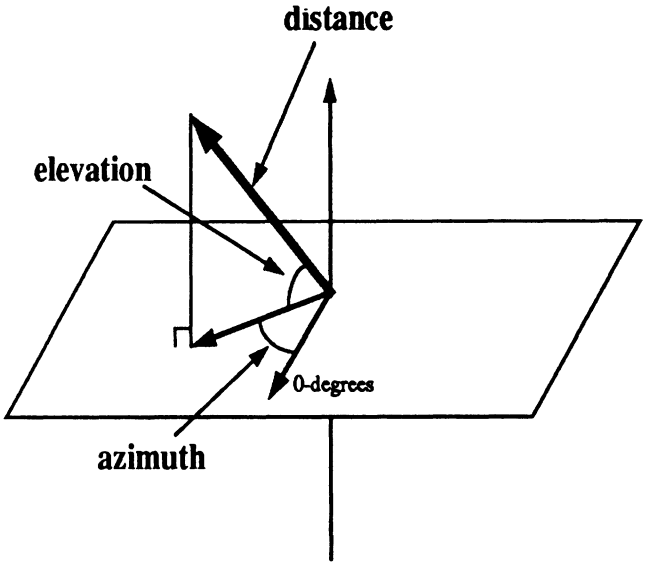


Figure 3

the time domain and the frequency domain. If we imagine that the sound event is a simple impulse, we can easily identify the features that are dependent just on the acoustics of the listener. From the standpoint of the time domain, the signals that reach the two ears are no longer impulsive. The energy has been spread over 1–3 msec by the acoustic interaction with the listener's body. Comparing the two ears, the sound arriving at the ipsilateral ear is generally more intense and arrives earlier than that

Figure 4. Time-domain and frequency-domain representations of HRTFs for the ipsilateral and contralateral ears. Adapted from Kendall et al. 1990. Used by permission of the Audio Engineering Society.

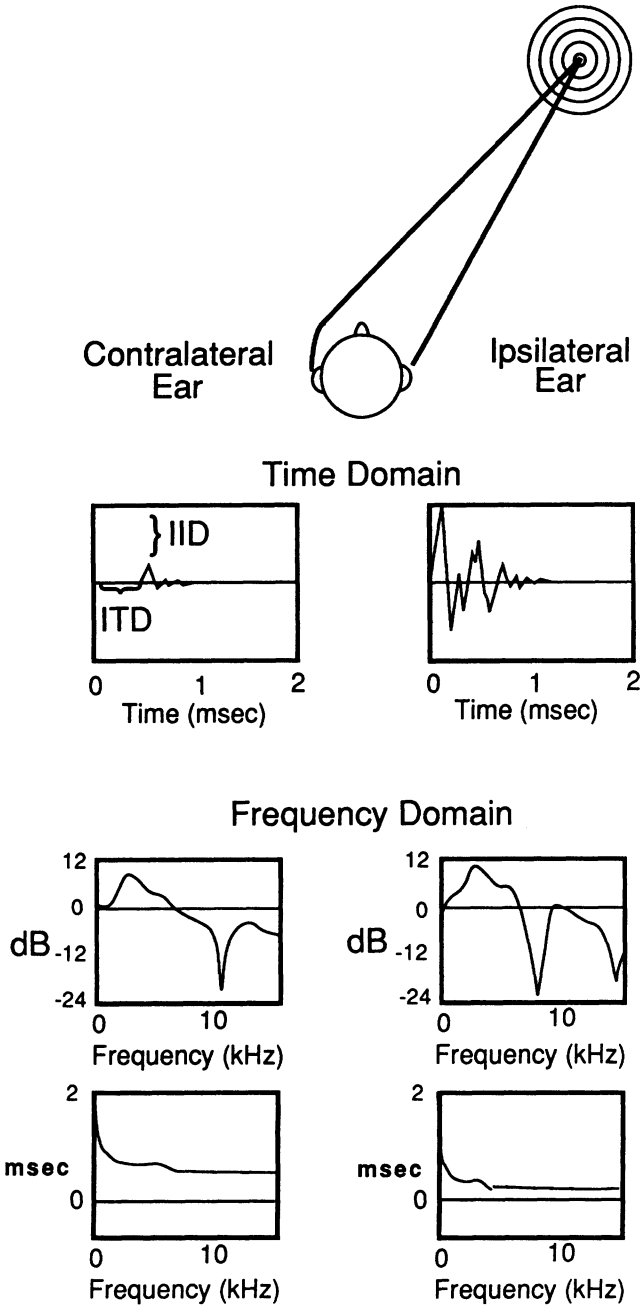
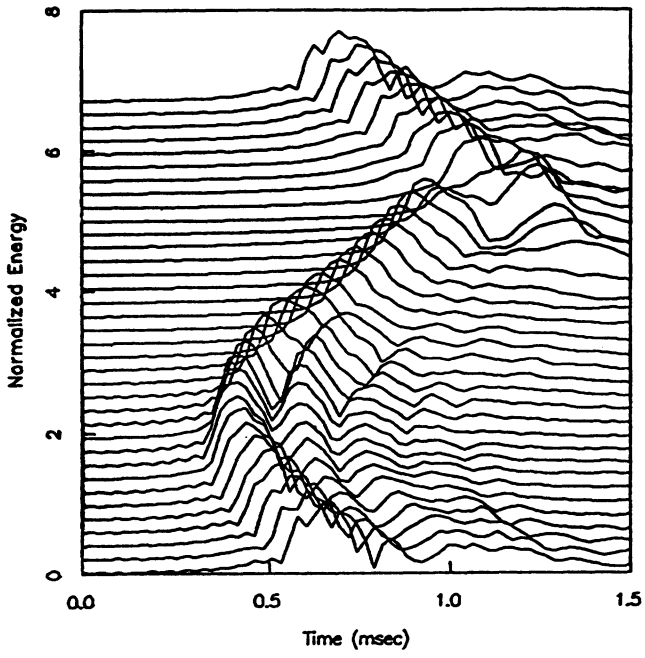


Figure 5. Energy-time curves measured at the eardrum position of the Kemar mannequin for 36 azimuth angles on the horizontal plane. The curve at the bottom of the graph was measured at 0 degrees azimuth (front), and subsequent curves proceed by 10-degree increments completely around the head to 350 degrees. From Kendall et al. 1990. Used by permission of the Audio Engineering Society.



at the contralateral ear. These differences between the two ears are called the interaural intensity difference (IID) and the interaural time difference (ITD), respectively. When a sound source is completely to the side, near 90 degrees azimuth on the horizontal plane, the ITD reaches a maximum near .7 to .8 msec.

A comparison of impulse responses measured for different locations will reveal few significant patterns. But, if those impulse responses are converted to energy-time curves (similar to those of Hiranaka and Yamasaki 1983), more significant trends emerge. These energy-time curves, also called envelope functions, capture the dispersion of the impulse's energy across time (while omitting the waveform's positive and negative excursions). Figure 5 shows energy-time curves measured at the eardrum position of the Kemar mannequin for 36 azimuth angles on the horizontal plane. Most significantly, one can see the variation in the delay of the initial sound that accompanies a change of azimuth. Around 270 degrees (the far contralateral side), the symmetry of sound circling the head in both directions disrupts the pattern of the peaks. There are also clear patterns in the delayed energy after the initial peak. (The delayed sound reduces gain between 150 and 270 degrees, probably reflecting a reduction in sound from the pinna.)

In the frequency domain, Figure 4 reveals that HRTF magnitude profiles vary tremendously. Com-

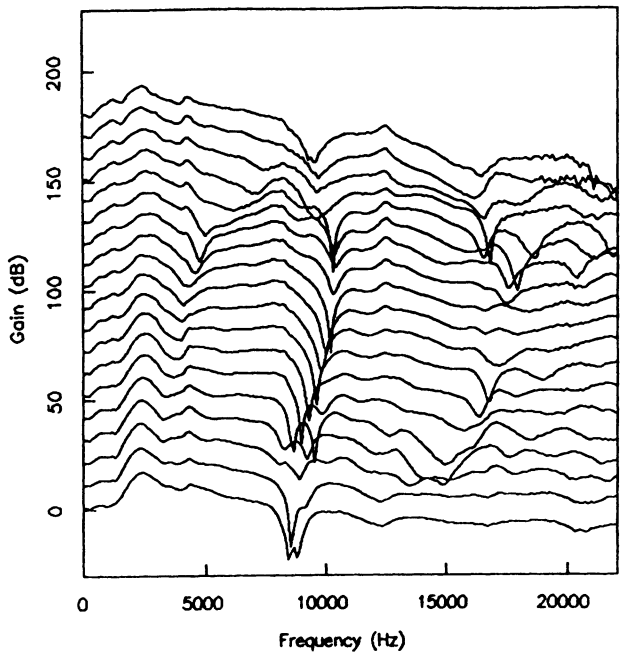
paring the two ears, we see that the magnitude profiles are more similar for low frequencies than for high frequencies. The differences become increasingly noticeable above the 1,500 Hz transition zone, because the head is increasingly effective at blocking waves at these higher frequencies.

Plots of the HRTF phase are typically difficult to interpret. The phase function "wraps" repeatedly from $-\pi$ to $+\pi$, because the time delays exceed the wavelengths of most frequencies. More-significant information is revealed when the phase is reinterpreted in terms of time delay, expressed either as *phase delay* or *group delay*. Phase delay reveals the time delay of each frequency, and group delay describes the time delay of the amplitude envelope of each frequency (see Smith 1985 for a more complete description). Figure 4 represents HRTF phase as phase delay. The delays are greatest for the lowest frequencies, because the diffraction of waves around the head causes the low-frequency waves to move more slowly than the high-frequency waves. Between 500 and 2,500 Hz there is a region in which delay makes a transition from a low-frequency region to a high-frequency plateau. The approximate center of this region lies at 1,500 Hz, clearly an important region for both magnitude and phase.

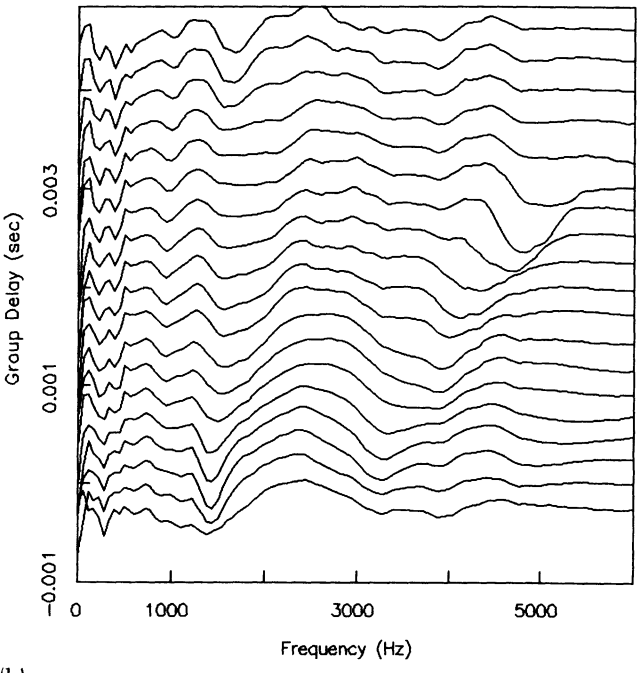
Numerous acoustic factors add complexity and richness to HRTFs. For example, there is a clear magnitude peak in the region around 3,000 Hz that is caused by the resonance of the ear canal. There are also notches and other fine details in the magnitude response, caused by constructive and destructive interference of the direct wave with sound reflected off the body. Reflected sound below 2,000 Hz is mainly from the torso, and above 4,000 Hz it is mainly from the pinnae; in between, there is a region of overlapping influence (Kuhn 1987).

A comparison of HRTFs measured for adjacent directions will reveal many significant patterns. Figure 6a illustrates the patterns that can be observed in the magnitude response of the ipsilateral ear on the horizontal plane between 0 and 180 degrees azimuth. For example, the bandwidth of the spectral peak near 3,000 Hz widens as the sound source moves from front to back. A deep notch in the 8,000 Hz region migrates upward in frequency as

Figure 6. Ipsilateral HRTFs measured at the eardrum position of the Kemar mannequin for 19 azimuth angles on the horizontal plane: magnitude response (a) (from Kendall et al. 1990), and phase response (b), expressed in group delay. The curve at the bottom of each graph was measured at 0 degrees azimuth (front) and the curve at top of each graph was measured at 180 degrees azimuth (rear). Figure 6a used by permission of the Audio Engineering Society.



(a)



(b)

the source moves toward the back, and then virtually disappears. The 4,000 Hz region shows a deep notch between 100 and 130 degrees in azimuth. Figure 6b reveals related trends in group delay.

These frequency-domain profiles can also be viewed from the perspective of the differences between the two ears. In complex ways, IIDs and ITDs vary across frequency. Figure 7 shows the frequency-dependent IID and the ITD (expressed as group delay) for a single direction.

When the distance of the originating sound event changes, HRTFs change very little if the event is more than 2 m from the head. Beyond 2 m, the sound wave from the acoustic event is approximately planar. (This means that HRTFs recorded at least 2 m from the head can be used to simulate sound sources farther away, provided that environmental cues to distance are also present.) Less than 2 m from the head, the sound waves from the acoustic event are more spherical, the effective angle between the sound event and the individual ears changes, and the HRTFs diverge significantly from those recorded farther away. Figure 8 shows a series of HRTFs recorded at varying distances directly in front of the head. The perception of distance close to the head appears to depend on these alternative HRTFs.

A comparison of HRTFs from different individuals will reveal that spectral features do not entirely match. The magnitude of individual HRTFs will vary in gross shape, as well as in details. Figure 9 compares the ipsilateral HRTFs of two individuals on the frontal plane. Although there are considerable differences in shape and detail, it can be seen that the overall trends are quite similar. For example, both individuals show the same trend in the upward migration of notch frequencies as elevation rises. This suggests that while individuals possess heads of different sizes and pinnae of different shapes, the acoustic processes that forge the individual HRTFs are the same. Nonetheless, interaural phase differences will be especially affected by head size because of the difference in the separation of the ears. The magnitude of interaural phase cues for children must vary considerably from those for adults.

Figure 7. Frequency-dependent interaural magnitude difference and the interaural group-delay difference for a sound source at 90 degrees in the horizontal plane. (Original data was measured with the Kemar mannequin and then smoothed.)

Figure 8. Magnitude response of sources located at 0 degrees azimuth, 0 degrees elevation, at distances of 90 inches (solid line), 24 inches (dashed line), and 4.5 inches (dotted line).

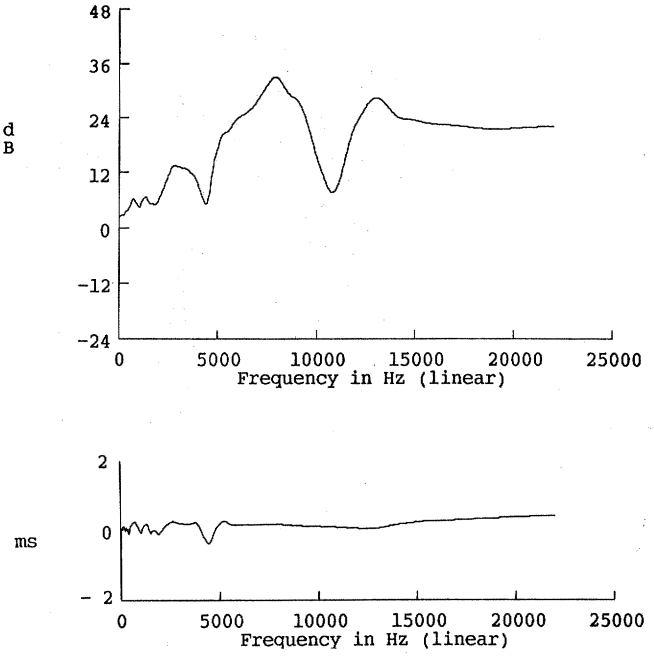


Figure 7

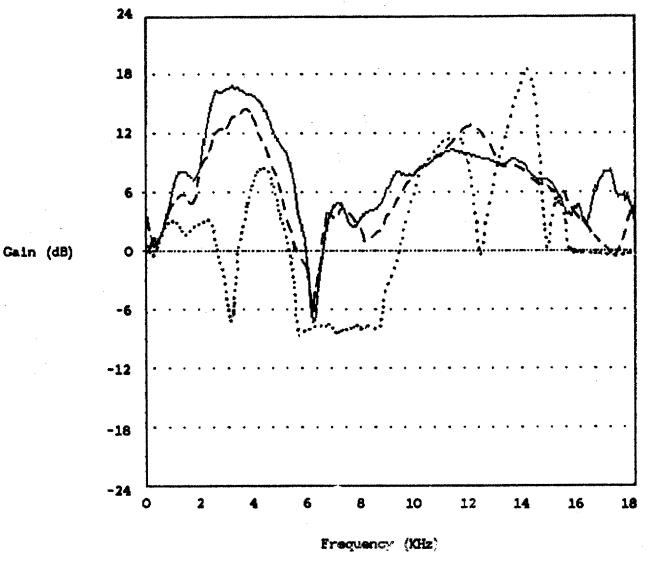


Figure 8

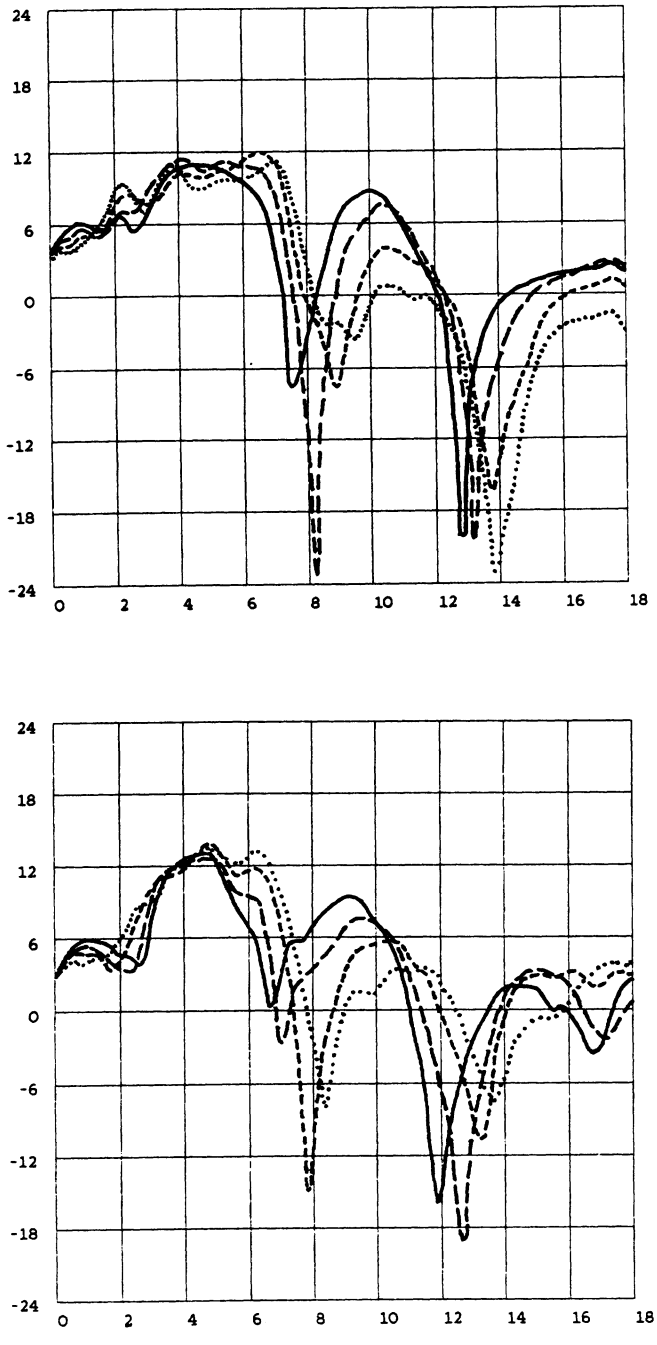


Figure 9. The HRTFs on the frontal plane for two subjects. The sound increases in elevation (solid line, 0 degrees; long dashes, 10 degrees; short dashes, 20 degrees; and dotted line, 30 degrees) (from Kendall and Martens 1984).

HRTF Measurement Techniques

HRTFs are generally measured by recording test signals in one of three positions: (1) at the blocked entrance of the ear canal, with a miniature microphone capsule; (2) within the ear canal, using a probe tube; or (3) at the eardrum position, employing a dummy head. In all three cases, the head must be kept perfectly still during measurement, and environmental sound must be eliminated. Measurements made at each position have a stable, fixed relationship to measurements made at other positions (Moller 1992). For example, measurements made with a probe tube placed at least 15 mm into the ear canal are closely related to those at the eardrum position. There is a fixed ratio between the magnitude spectra of the two, up to around 7,000 Hz. Above 7,000 Hz (and sometimes below), notches in the two measurements are offset from each other and create push-pull spectral differences. (There is typically a poor signal-to-noise ratio in the notches, which may cause inaccuracies when one transforms one type of measurement into another.)

Measurements made at the ears must be processed to isolate the part that represents the actual HRTFs. The acoustic signals measured at the ears can be represented as the products of the transfer functions of the source, $S(\omega)$, and the recording equipment, $T(\omega)$, with the ipsilateral ear, $H_i(\omega)$, or the contralateral ear, $H_c(\omega)$:

$$S(\omega)\, T(\omega)\, H_i(\omega) \text{ or } S(\omega)\, T(\omega)\, H_c(\omega).$$

A reference measurement without a human subject is the product of the source and recording equipment alone, $S(\omega)$ and $T(\omega)$. Therefore, the HRTFs can be isolated by dividing the reference from the measurements in the ears:

$$[S(\omega)\, T(\omega)\, H_i(\omega)] \,/\, [S(\omega)\, T(\omega)] = H_i(\omega)$$

and

$$[S(\omega)\, T(\omega)\, H_c(\omega)] \,/\, [S(\omega)\, T(\omega)] = H_c(\omega)$$

This computation is typically performed by first transforming the time-domain measurements to the frequency domain via the Fast Fourier Transform (FFT), where the complex-valued division can be performed directly. Alternatively, the complex-valued frequency data can be converted to magnitude and phase, after which, the complex division is achieved by subtracting the gain in dB and the phase of the reference measurement from the ear-

measurement data. The impulse response for HRTF is then computed by transforming the frequency-domain HRTF to the time domain via the inverse FFT.

Psychoacoustic Perspective

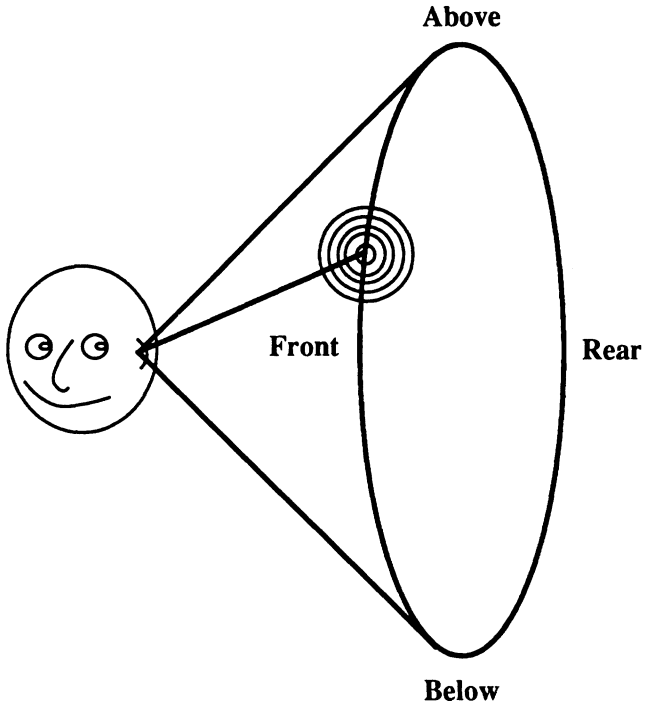
A listener's judgment of the direction of an acoustic event is dominated by the sound that reaches the listener along the shortest, most direct path (otherwise the judgment of the direction of the event would be ambiguated by the indirect sound). This preference given to the initial sound is called the *precedence effect* (Wallach et al. 1949) or the *law of the first wavefront* (Blauert 1971). Even these initial sound waves are radically transformed in comparison to those of the original event. The sound arriving at each ear is spectrally modified by the HRTF, each ear has a different transformation, and the transformation changes as the head and/or the source moves. The auditory system performs the phenomenal task of integrating the information arriving at the two ears into a single, fused perceptual image of the acoustic event in space, extracting the directional information, and reconstructing an estimate of the original source spectrum. This is accomplished even though there is no direct, structural representation of spatial information in the peripheral auditory system, as there is in the peripheral visual system when light is focused onto the retina. (No wonder that research into three-dimensional sound has lagged behind research into three-dimensional vision!)

Classical psychoacoustics focused on the separation of the two ears, and proposed the *duplex theory of sound localization* (Rayleigh 1907). Experimenters attempted to construct a theory of localization by compositing results from many experiments conducted with the ultimate acoustic building blocks—sine waves. These experiments demonstrated that *interaural differences*, that is, differences in the acoustic signals simultaneously presented to the left and right ears, strongly affect spatial perception; IID and ITD each make a significant impact on perceptual judgments in a separate frequency range. Above 1,500 Hz there is acoustic shadowing by the head, and localization judgments are dominated by the intensity difference between the ears (IID). Below 1,500 Hz, the head is not a significant acoustic obstacle, there is a less-significant intensity difference, and localization judgments are dominated by the time difference between the ears (ITD). (Consider too that above 1,500 Hz, ongoing phase differences would often exceed 360 degrees, making it impossible to judge time delay on the basis of these phase differences.) The differentiation in perceptual processing appears to be coupled to the acoustic properties of the head.

These observations do not, however, provide sufficient explanation for human localization. In fact, IID and ITD only affect the extent of the *lateralization* of the sound source, that is, its perceived position along the interaural axis, a left/right axis between the ears. With only IID and ITD, a listener cannot determine whether an acoustic event is in front, above, behind, or below. This ambiguity of location at a given degree of lateralization has been called the *cone of confusion* (Woodworth 1954) (see Figure 10). It is now commonly accepted that the seeming uncertainty of spatial location on the cone of confusion is disambiguated by the complex acoustic profiles of the HRTFs. The classic psychoacoustic experiments supporting the duplex theory of localization did not utilize the frequency-dependent interaural magnitude difference and interaural phase difference typical of HRTFs. Then too, the duplex theory ignored the influence of alternative temporal cues above 1,500 Hz, such as interaural onset differences (see Blauert 1974 for a comprehensive review). Acoustic events in natural environments also exhibit ongoing perturbations that provide additional opportunities for grasping interaural temporal cues. The classical psychoacoustic stimuli were impoverished, and the results are only partially useful in understanding localization in everyday listening situations.

Modern psychoacoustic research has turned its attention to binaural hearing and the role of HRTFs in localization. In the broadest context, *binaural* means combining information from the two ears (as opposed to *monaural* , which means using information from one ear or from each ear indepen-

Figure 10. The cone of confusion (based on Woodworth 1954; adapted from Kendall et al. 1990). Used by permission of the Audio Engineering Society.



dently). Use of the word "binaural" also implies the kind of frequency-dependent interaural cues typical of HRTFs. This change in the focus of research is also accompanied by a shift toward the use of broadband stimuli, rather than sine waves.

Even though HRTFs are rich in acoustic detail, perceptual research suggests that the auditory system is selective in the acoustic information that it utilizes in making judgments of sound direction. Evidence reveals that monaural phase information is irrelevant to spatial perception, and that interaural phase information is extremely important. Wightman and Kistler (1992) have demonstrated that low-frequency ITD is the dominant localization cue for sounds that contain energy below 2.5 kHz. For sounds that lack this low-frequency energy, IID provides the most likely basis for localization. It is still unclear, though, how much influence high-frequency time differences might have, since experiments have shown that the time differences between the temporal envelopes of high-frequency sounds are easily detectable (Henning 1974). Although the majority of research focuses on binaural cues, there is research into monaural spectral cues that suggests they are important for sound sources at the sides (Musicant and Butler 1985). There is also evidence that elevation in particular is influenced by the spectral content of the sound source itself (which is received at both ears), such that high-pitched/bright sounds are typically localized higher than low-pitched/dark sounds (Butler 1973).

There are important differences between the vertical and horizontal dimensions in the resolution with which people can judge the spatial location of a sound source, an effect that Blauert terms *localization blur* (Blauert 1974). The highest resolution is evident in the horizontal dimension, especially in front of the listener where the minimum audible angle is 2 degrees or less, depending on the exact nature of the experimental task. That angle increases to near 10 degrees at the sides, and narrows to near 6 degrees in the rear. By comparison, the resolution in the vertical dimension is low. The vertical minimum-audible angle in front is near 9 degrees, and it steadily increases overhead until it reaches 22 degrees. (See Blauert 1974 for a summary of research in this area.) Spatial acuity is apparently not as important for auditory perception as it is for visual perception.

While front/back discrimination is possible on the basis of the full acoustic information in HRTFs, it is also clear that *head movement* plays a dominant role in resolving front/back confusions (Wallach 1940). This is particularly important for sound sources located near the median plane, where other acoustic information provides few interaural differences. Figure 11 illustrates how the location of sound sources in front and in back of the listener is disambiguated by turning the head toward the right. For a sound source in front of the listener, turning the head toward the right causes the left ear to receive sound earlier and with greater intensity. For a sound source behind the listener, it is the right ear that receives the earlier and more intense sound. Wallach's classic experiments also clearly demonstrated that dynamic interaural cues would override HRTFs when the two were placed into conflict.

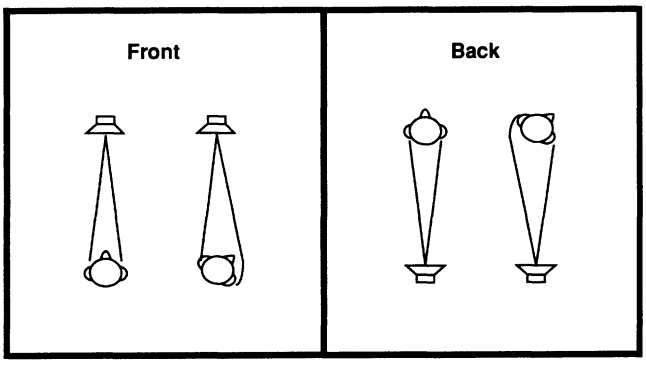


Figure 11. A dynamic head turn to the right disambiguates whether a sound source is in front or in back of the listener (adapted from Kendall et al. 1990). Used by permission of the Audio Engineering Society.

Individual Differences

There is debate at present concerning the impact of individual differences and the extent to which people can localize with HRTFs other than their own. Individual HRTFs vary tremendously, and interaural differences are strongly affected by differences in head size and pinnae size and orientation. It appears that some individuals' HRTFs improve other individuals' localization accuracy (Butler and Belendiuk 1977; Wightman and Kistler 1989), but that large differences in head size can undermine localization (Morimoto and Ando 1983). Wenzel, Wightman, and Kistler (1993) report that elevation judgments and front-back differentiation are more likely to degrade with non-individualized HRTFs. At the same time, it appears that effective localization can occur in many cases in which the ears receive directional transfer functions (DTFs) whose details differ significantly from measured HRTFs. Kendall and Rodgers (1982) used low-order filters to create cartoon-like approximations of natural HRTFs, while Martens (1987) and Kendall, Martens, and Wilde (1990) describe using principal-components analysis to create artificial DTFs. Comparison of results suggests the following:

1. Individuals generally localize better with their own HRTFs than with those of others.
2. Some individuals have HRTFs that are superior, and these HRTFs can sometimes improve the others' localization.
3. In order for one individual's HRTFs to work for another, the head sizes must be approximately the same.
4. Localization can be achieved with synthetic DTFs whose details differ from measured HRTFs.

Neurophysiological Perspective

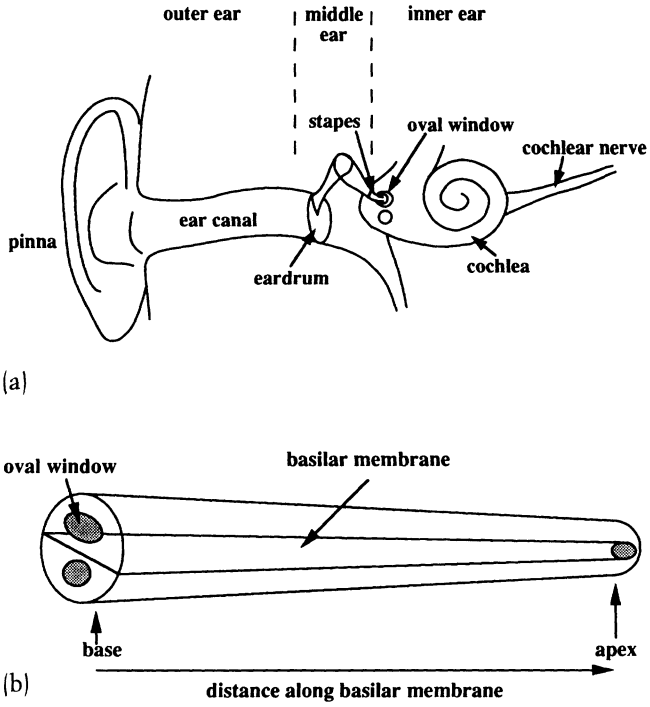
Although neurophysiology is not part of the educational background of many computer music and audio professionals, it is an area from which many of the most important new ideas and discoveries about hearing continue to come. This is especially true for directional hearing. (For comprehensive reviews, see Phillips and Brugge 1985; Casseday and Covey 1987; and Kuwada and Yin 1987.) The terminology and perspective of neurophysiology are quite distinct from those of physical acoustics and psychoacoustics. The purpose of this section is to familiarize the reader with this important context for understanding directional hearing and, in particular, to point out the special adaptations in the auditory system for sound localization. Although an attempt is made to introduce terminology somewhat gently, it is undoubtedly helpful if the reader has some basic familiarity with the field, especially the physiology of the auditory system.

Peripheral System

While the pinna is clearly adapted to auditory localization, the peripheral neurological system has little or no specialization for directional hearing. The peripheral neurological system transforms the acoustic ear signals into neural activity and seems most clearly designed to capture the spectral/temporal decomposition of incoming acoustic waves. The primary function of the signal decomposition appears to be identifying the sound source, namely, the sounding object and its excitation. This strongly conditions the structure of the neural mechanisms that underlie human localization, since, at the level of the peripheral neurological system, source information commingles with spatial information.

The acoustic signal at the outer ear is converted to mechanical energy by the linkage of the eardrum to the middle ear (see Figure 12). This mechanical

Figure 12. Peripheral auditory system: physical structure showing pinna, eardrum, middle ear, oval window, and cochlea (a); conceptual representation of the uncoiled cochlea, which is divided down the middle by the basilar membrane (b).



energy is converted to fluid pressure by the linkage of the stapes (stirrup) to the flexible oval window at the base of the cochlea. Stapes motion at the oval window initiates a traveling wave of displacement down the basilar membrane. As this wave travels, it is increasingly damped by the changing mass and shape of the basilar membrane. From base to apex, the wavelength lengthens and its velocity decreases. The extent of membrane displacement is related to the spectral content of the wave, such that maximal displacement occurs near the base for high-frequency components and near the apex for low-frequency components. Sensory receptor cells located along the basilar membrane, called inner and outer hair cells, respond maximally at a characteristic frequency. These frequencies run from high to low along the membrane from base to apex, and are arranged nearly logarithmically. Thus, distance along the basilar membrane is approximately proportional to the log of the characteristic frequency. In this way spectral information is spatially mapped onto a neurological representation. There is no spatial representation of location as there is in the peripheral visual system.

The motion of the basilar membrane causes displacement of the cilia of the hair cells, and changes the cell potential. The resulting potential can be viewed as containing an AC part and a DC part. The AC part captures the temporal changes of the waveform itself, while the DC part can be viewed as the average value of the potential over a period. At high frequencies, the DC part is the only response. For example, above 5 kHz, the temporal structure of a sine waveform is not individually resolved (it has no AC part) and the inner hair cells respond only to the temporal envelope (captured by the DC part). The neurological representation of temporal information therefore shifts gradually from the waveform itself at low frequencies to the signal envelope at high frequencies. (Thus, it appears that the most appropriate time-delay representation for low frequencies is phase delay, and for high frequencies is group delay.)

Neural Pathways

The basilar membrane creates a neural representation of the acoustic activity taking place in the physical world, and this information is initially transformed and retained in the action potential firing patterns of fibers innervating (or, furnishing neural connections to) the basilar membrane from the cochlear nucleus (CN). These auditory nerve fibers bifurcate up to the anteroventral cochlear nucleus (AVCN) and down to the dorsal cochlear nucleus (DCN). (Follow Figures 13a and 13b for a diagrammatic representation.) The goal of the central neurological system and subsequent neurological processing is to construct a representation of information about the physical world that is useful for survival, including the identity of sound sources and their locations.

At the beginning of the neural processing, the source information and the directional information are confounded. The most direct strategy for segregating the two is to extract directional information from the differences between the ears, i.e., binaural information. The auditory neurological system forms symmetric left and right neural pathways for this binaural information. To simplify the discussion of these binaural pathways, we will trace the evolution of one path; same-side connections will

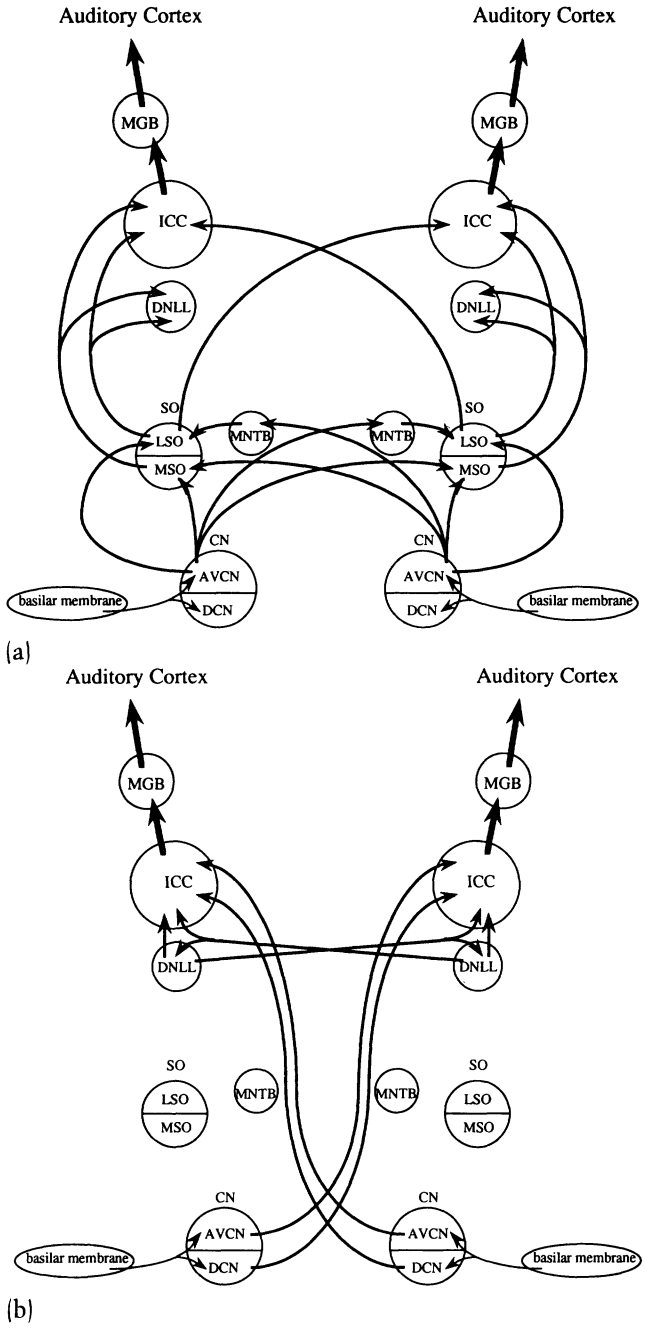


Figure 13. Representation of the primary auditory neural pathways important for directional hearing: projections to and from the superior olives (SO) constitute the heart of the binaural system (a); monaural pathways and the integration of binaural information in the DNLL (b). Abbreviations are explained in the text.

be referred to as ipsilateral, and opposite-side connections as contralateral.

The origin of the binaural pathways is the AVCN, which is the source of projections to both the ipsilateral and contralateral superior olive (SO) (Stotler 1953). Projections in and out of the SO are represented in Figure 13a. The medial superior olive (MSO) is innervated by both the ipsilateral and contralateral cochlear nuclei. Its input is dominated by low-frequency fibers that retain the fine temporal structure from the basilar membrane. There is strong evidence suggesting that the MSO is a coincidence detector for interaural time differences (Goldberg and Brown 1968). The lateral superior olive (LSO) is directly innervated only by the ipsilateral cochlear nucleus. It is connected to the contralateral cochlear nuclei through an intermediate connection in the contralateral medial nucleus of the trapezoid body (MNTB). The MNTB appears to provide an inhibitory input to the LSO. Both inputs are dominated by high-frequency fibers. Evidence suggests that the LSO detects IIDs (Boudreau and Tsuchitani 1968).

The LSO and MSO project to and converge on two targets, the inferior colliculus central nucleus (ICC) and the dorsal nuclei of the lateral lemniscus (DNLL). This gives rise to the possibility that IID and ITD information is conjoined. Moreover, both ipsilateral and contralateral LSO project to the ICC, suggesting that information from both binaural pathways are combined, though only the ipsilateral projection includes LSO low frequencies. The ICC is also the target of projections from the contralateral AVCN and the DCN (see Figure 13b). These projections contain monaural, rather than binaural information. In the ICC, the targets of the MSO and LSO lie within the target of the AVCN and overlap with each other, giving rise to the possibility that monaural source information is recombined with binaural information. The ipsilateral DNLL projects to the contralateral DNLL (Figure 13b), providing a clear opportunity for integrating information from both binaural pathways, which can then be passed on through projections to the ipsilateral and contralateral ICC. The DNLL is also connected to the greater superior colliculus (not shown in Figure 13), providing binaural auditory information with a path to motor centers.

The inferior colliculus has been the site of much work on IID and ITD. Research with low-frequency tones reveals neurons that respond to a "characteristic delay" (Rose et al. 1966). Similar results have been found with amplitude-modulated high-frequency tones (Yin, Kuwada, and Sujaku 1984).

The "phase locking" that occurs with the envelope of the high-frequency tones is just like that of the low-frequency tones. Thus, there appears to be a single system of ITD detection that extends from the phase of low-frequency tones to the envelope of high-frequency tones.

Although less clear in mammals, research with barn owls has shown that a spatial referent map of auditory space exists in the equivalent to the inferior colliculus (Knudsen and Konishi 1978). Individual neurons respond to acoustic stimulation from a narrow spatial region, and neighboring cells respond to sources in adjacent spatial regions. Not only that, but azimuth is associated with ITDs and elevation with IIDs (Moiseff and Konishi 1981).

After the convergence of binaural and monaural information in the ICC, pathways ascend to the medial geniculate body (MGB) and then the auditory cortex (shown in Figures 13a and 13b). One might expect that a spatial referent map would be found in the auditory cortex of mammals. Instead, spatial information appears to be coded in the temporal firing pattern of a group of neurons (Middlebrooks et al. 1994). This allows spatial information to be projected on top of other neural maps.

The Stereo Reproduction of 3-D Sound

Many 3-D sound advocates share a vision of an ideal home audio system that would include a computational engine with sufficient power to synthesize the full 3-D acoustics of a simulated environment. In fact, simultaneous simulated environments would be needed to place each sound into its most appropriate environment. (For example, upper strings need lots of reverberation, while electric basses are best left dry; dialog might be in a small room, while the orchestra in the background is in a large hall.) Each simulated sound source and each of its simulated reflections would be processed by a pair of directional filters that capture the directional properties of the listener's head (Kendall and Martens 1984). These filters would change instantaneously in response to the listener's head movements or to changes in the simulated environment. If there were more than one listener, the changes would have to occur independently for each person. The directional filters would be based on each listener's HRTFs, or on an idealized set matched to each listener. Any influence of the reproduction equipment or environment would be eliminated.

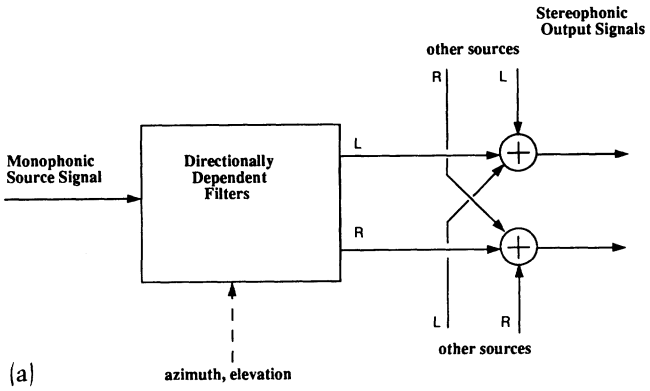
Many factors keep us from realizing this vision today. One is that the computational burden placed on this system has no apparent bound. Many engineering shortcuts must be incorporated before a practical system would begin to approach the functionality described above. Crafting a system that effectively communicates to the listener is probably more important than matching the acoustics of physical reality, since we already know that the auditory system is selective in the information it utilizes. More important, however, is that today's implementations of directional filters are far from perfect. We are still improving our understanding of how to reproduce 3-D sound.

Cohen (1989) and Begault (1991) have raised warning flags about the lack of thorough discussion of problem areas, and about the overly optimistic predictions and claims for 3-D sound, especially by commercial companies. A few of the key problems are: front/back reversals, timbral discoloration, differences in listener performance, and differences due to the acoustics of the sound source. In itself, 3-D sound reproduction can seem a complicated topic. For example, while headphone and loudspeaker reproduction share many technical issues and goals, they also present different problems. (See Moller 1992 for an excellent technical summary.)

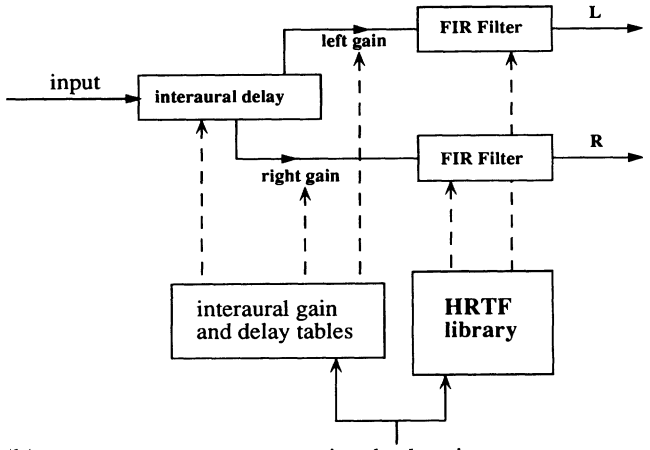
Directional Filtering

Whether the reproduction occurs through headphones or loudspeakers, some essential aspects of the computational simulation remain the same. For example, as shown in Figure 14a, each potential sound source and each simulated reflection starts off as a single, monophonic signal that eventually must be split to form a left/right stereo pair. Each channel of the stereo pair must be processed by directionally dependent filters which change in response to the intended source location. All of the resulting left/right stereo pairs are summed together to form a composite stereophonic output signal that is eventually reproduced through head-

Figure 14. Directional filtering: a single, monophonic source is split to form a left/right stereo pair, which is processed by directionally dependent digital filters and summed to form a composite stereophonic output signal (a); implementation details with FIR filters and independent interaural delay and gain controls (b).



(a)



(b)

phones or loudspeakers. These directional filters can be implemented in a variety of ways. Figure 14b shows the details of a typical implementation. There are left and right finite-impulse-response (FIR) filters, whose coefficients are the HRTF impulse responses themselves retrieved from an HRTF library. The HRTFs would usually incorporate the interaural time and intensity differences, but these can also be implemented separately by independent gain and delay controls. (Separate interaural delay control can reduce the number of coefficients needed to implement the FIR filters.)

Equalization

Before headphone or loudspeaker reproduction occurs, the output signals must be equalized to eliminate two potential "errors" in the reproduction process. The first is that components of reproduction equipment, especially the transducers themselves, superimpose their own characteristics on the output signals. The second is that the path of the sound from the transducers to the listener's eardrum will superimpose information itself. In both cases, we must attempt to compensate so that the sound reaching the listener's eardrums is as close as possible to the intended signals. The actual signal at the eardrum is a product of the transfer functions of the program material, $P(\omega)$, the directional filter, $D(\omega)$, the reproduction equipment, $E(\omega)$, and the sound transmission path to the eardrums, $T(\omega)$:

$$P(\omega)\, D(\omega)\, E(\omega)\, T(\omega).$$

The influence of the reproduction equipment and the sound-transmission path could be eliminated by dividing out the transfer functions, $E(\omega)$ and $T(\omega)$. One can obtain these transfer functions by direct measurement, but in most practical situations the measurements are only near-approximations, $E'(\omega)$ and $T'(\omega)$, of the actual transfer functions present during reproduction. So the equalization of the reproduced material is:

$$[P(\omega)\, D(\omega)\, E(\omega)\, T(\omega)] \,/\, [E'(\omega)\, T'(\omega)] \approx P(\omega)\, D(\omega)$$

The division by $E'(\omega)$ and $T'(\omega)$ must occur before reproduction, either as the final step in the signal processing, or it could be rolled into the directional filters:

$$\{D(\omega) \,/\, [E'(\omega)\, T'(\omega)]\}\, P(\omega)\, E(\omega)\, T(\omega).$$

In this case, the coefficients for the directional filters that are stored in the processor's memory are already equalized. Headphone and loudspeaker reproduction share the need for equalization, but at a more-detailed level they present some very different problems that require specific solutions.

Headphone Reproduction

It might seem intuitively obvious that headphone reproduction provides the most controlled method for reproducing directional cues, but the task is far more difficult than one might expect. Headphone reproduction of traditional stereo recordings creates the impression that sound events are originating

inside the head, with a bias toward the rear. Even with the addition of cues for IID and ITD, auditory images move only left and right inside the head along the interaural axis. True 3-D sound should mean that images are perceived outside the head (with *externalization*) and that frontal images are not easily confused with rear images (few *front/back confusions*). This has proven difficult to achieve through the use of directional filters with standard stereo headphones alone. Such systems tend to be successful in some spatial regions (such as the left and right sides) and much less successful in others (such as in front). Externalization is aided by the presentation of ambient sound with interaural incoherence that mimics the acoustical properties of a late reverberant field (Kendall 1995). Through informal experimentation, the author discovered that front/back discrimination can be improved through modifications to HRTFs that exaggerate front/back spectral differences.

Head Tracking

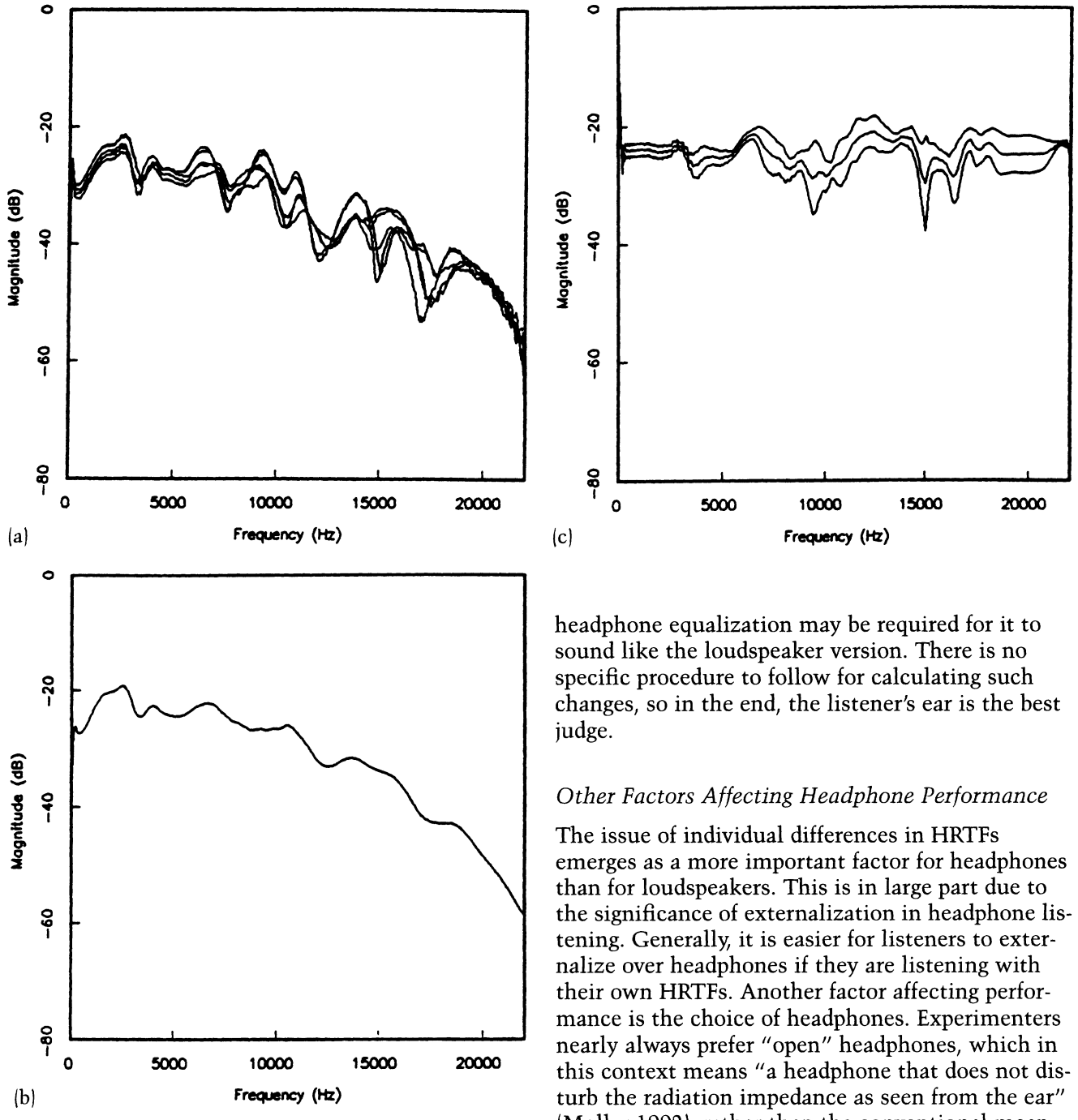
A truly categorical improvement can be achieved by combining the headphones' directional filters with head tracking. A head-tracking system combines a sensor for the direction and orientation of the listener's head with computer control of the directional filters. The computer receiving this spatial information continuously updates the direction of the filters to maintain the absolute position of the sound source within the environment, even as the listener's head moves. This simulates the kind of interaction the listener experiences in the natural environment, where a sound position remains invariant, fixed in its position within the environment, as the head turns. Head tracking is therefore an essential ingredient in any virtual reality system. Even changing ITD and IID in response to head movement without directional filters produces front/back discrimination due to the dominance of dynamic interaural cues over HRTFs (Wallach 1940). That such interaction is missing in traditional headphone reproduction strongly suggests why sounds are internalized inside the head: if the head turns and nothing changes at the eardrums, there is only one place the sound could be coming from—the middle of the head. We experience this every day when we listen to ourselves talk. The auditory system is sensitive to time lags between the movement of the head and the change in the directional filters, but no data is available that describes the relationship between localization performance and head-tracking latency.

Equalization

Even with the best headphones, the headphone system must be equalized to compensate for the acoustic properties of the transducers and the coupling to the ears. The response of headphone transducers varies from one model to another, and tends to be deficient in very high and/or very low frequencies. These deficiencies cannot always be compensated for by equalization. Dramatically increasing the gain for a spectral region in which the transducer is deficient would overdrive the transducer and create nonlinearities. Another consideration is that the coupling to the ears changes with each reseating of the headphones, and therefore no one measurement provides a sufficient basis for an equalization function (see Figure 15). It is recommended that the equalization function be calculated by critical-band smoothing of the measured spectra, followed by averaging the representative measurements.

Combining an equalization function with DTFs will produce an overall spectral profile at the ears that mimics sound sources in an open space. This is called *free-field* equalization. An alternative approach is to mimic microphone equalization and to base the equalization function on the average response for all sound directions. This method attempts to provide an equalization similar to traditional stereophony, and is applicable to recordings with room reflections that arrive from all directions. This is called *diffuse-field* equalization. If one of the goals of equalization is to match the perceived coloration to a standard (like sound sources in free field or traditional stereophony), it raises the question, What is the most appropriate standard? For audio work in which the program material is heard over loudspeakers as well as headphones, the appropriate standard is usually the loudspeaker version. In that case, additional adjustments to the

Figure 15. Headphone equalization: magnitude response measured for five reseatings of STAX SR Lambda earphones (a); critical-band smoothed, mean magnitude function, which is inverted for equalization (b); magnitude response of reseatings measured with equalization showing mean and one standard deviation above and below (c) (Martens 1991).



headphone equalization may be required for it to sound like the loudspeaker version. There is no specific procedure to follow for calculating such changes, so in the end, the listener's ear is the best judge.

Other Factors Affecting Headphone Performance

The issue of individual differences in HRTFs emerges as a more important factor for headphones than for loudspeakers. This is in large part due to the significance of externalization in headphone listening. Generally, it is easier for listeners to externalize over headphones if they are listening with their own HRTFs. Another factor affecting performance is the choice of headphones. Experimenters nearly always prefer "open" headphones, which in this context means "a headphone that does not disturb the radiation impedance as seen from the ear" (Moller 1992), rather than the conventional meaning that the ears are "open" to environmental sound. Moller (1992) provides an analysis that explains the basis for this preference. Electrostatic headphones are among the best.

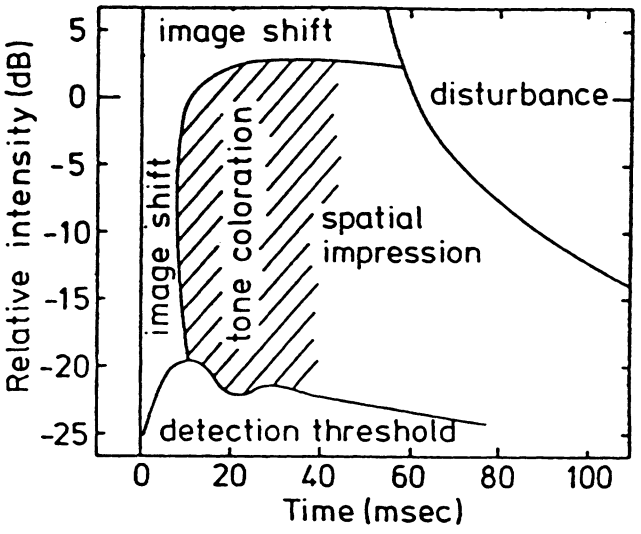


Figure 16. Barron's summary diagram of perceptual effects (Barron 1971). The axes represent the level and arrival time of the second loudspeaker relative to the first. See text for an explanation of terms. Adapted by Rasch and Plomp (1982) and used by permission of Academic Press.

Loudspeaker Reproduction

From the auditory system's point of view, loudspeaker reproduction is a special category of environmental listening—the sound waves from the two loudspeakers arrive from two directions, and are usually offset by some time and intensity differences just like the direct sound of an acoustic event followed by a strong reflection. The auditory system appears to "view" the second loudspeaker just as it would a room reflection: it must make the best sense it can out of the signals and construct a mental image of acoustic events in space.

It should come as no surprise that one of the methods that hearing scientists use to study the perception of environmental acoustics is simulating direct and reflected sound with loudspeakers. Barron performed a particularly important study in 1971, using a pair of loudspeakers to identify how direct sound from one loudspeaker in the front interacted with sound from a second loudspeaker 40 degrees to the side. The intensity and time delay of the second loudspeaker could be varied through a continuous range, simulating the effect of reflections coming from a wall at various distances. Although the study's focus was investigating concert hall acoustics, the results are interesting from the standpoint of stereo loudspeaker reproduction, because in most loudspeaker reproduction settings the listener is closer to one loudspeaker than another. Barron summarized the perceptual results of the two loudspeaker interactions with the diagram shown in Figure 16.

The most important observation to be made from this diagram is that there are many different subjective effects that result from the interaction of intensity and time delay. These percepts are described by the terms listed below. Although these particular results would undoubtedly change with alterations in the experimental setup (such as increasing or decreasing the angle between the loudspeakers, or rotating the listener's head position), the perceptual categories would likely stay the same. Here are some quotes from Barron on each term:

detection threshold — "Reflections below threshold produce no audible effects." (The second loudspeaker is not perceived to be making any difference in the sound.)

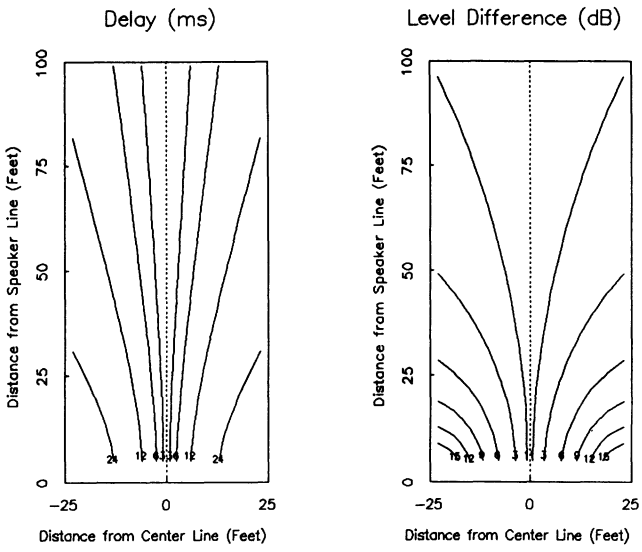
disturbance — "Echo disturbances." (The delayed sound is perceived as a separate source interfering with the intelligibility of the leading source.)

image shift — "The apparent source moved from the direct sound loudspeaker toward the reflection loudspeaker. . . . The effect is very similar to that observed when the balance control of a stereo system is adjusted." (A single image is perceived, emanating from a location between the loudspeakers.)

spatial impression — ". . . the source appeared to broaden, the music beginning to gain body and fullness. One had the impression of being in a three-dimensional space." (This is an effect one would wish to have in concert halls.)

tone coloration — "For certain delay reflections, . . . the tone of the music appeared to sharpen. . . . One explanation of this colouration effect is the interference effect between a signal and a delayed version of itself, producing a comb filter." (The sound source is perceived as emanating from the leading loudspeaker, but the timbre of the sound source appears to be modified.)

Figure 17. Distribution of time and intensity differences across seating locations between stereo loudspeakers (Kendall, Wilde, and Martens 1989).



Although not depicted in Mr. Barron's diagram, the precedence effect is assumed to be active in a region overlapping those of the other terms.

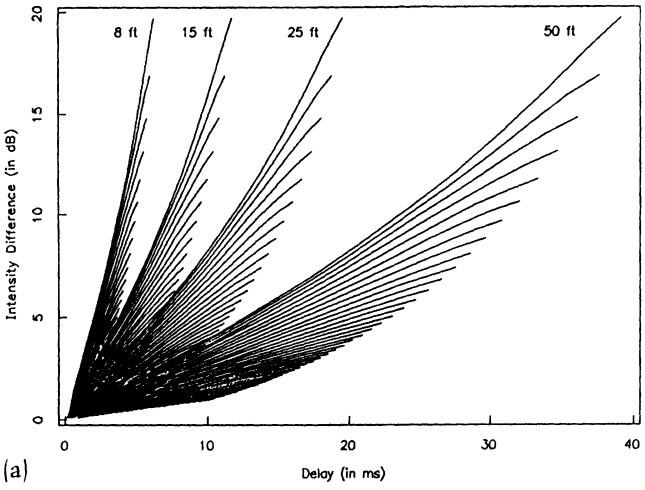
Consider the following mental experiment—a listener is sitting between two loudspeakers, one of which is moved progressively farther and farther away, creating greater and greater time and intensity differences between the sound coming from the two loudspeakers. Mr. Barron's diagram and terms provide us with a guide to what the listener experiences. At the beginning, the two loudspeakers are equally distant and there is no time or intensity difference; the listener hears a sound source that is located between the loudspeakers. As one loudspeaker is moved far enough away that the time delay is less than approximately 1.0 msec, the listener hears a single sound image that is shifted away from its original position and toward the closer loudspeaker, or "image shift." As the loudspeaker gets farther away and the time delay increases beyond 1.0 msec, the listener will hear a single sound image that is located in the closer loudspeaker, which is the "precedence effect." As the distance increases, the listener would perceive "tone coloration" and then "spatial impression." There is eventually an upper limit to the time delay at which the precedence effect is released and the delayed sound from the second loudspeaker begins to be heard. The exact delay at which precedence is released depends upon qualities of the sound source, and is reported to vary from 8 to 70 msec, with a typical limit of about 35 msec. This is further complicated because precedence is more pronounced for transient sound sources, such as struck or plucked musical instruments, than it is for continuous sound sources, such as blown or bowed musical instruments. When the precedence effect releases, the listener will report hearing sound images in each loudspeaker. When the loudspeakers are separated by a sufficiently great distance, the listener may report hearing an "echo disturbance."

Large-Space Reproduction

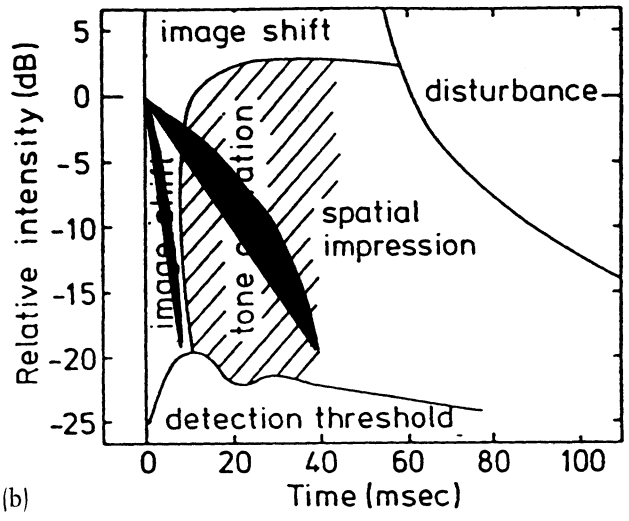
This range of percepts is typical of what is experienced by listeners to stereo reproduction in large halls where the listener's seat location is the primary factor determining the potential spatial imagery. Figure 17 illustrates the time and intensity differences distributed over the listener seating locations for a pair of stereo loudspeakers (which are, in this example, 50 feet apart). The time differences are dependent on the absolute difference in the time of arrival between signals from the two loudspeakers. This means that the farther apart the loudspeakers are, the greater the range of time differences experienced by listeners in the coverage area. The intensity difference is dependent on the *ratio* of the distances to the individual loudspeakers. This means that when the loudspeakers are moved farther apart, the pattern of intensity differences stays the same; it is just spread over a larger area. It is important to observe that the distribution patterns for time and intensity are similar but not identical, and listeners in every seating position experience a unique combination of time and intensity differences. The single exception is the middle line of seating positions, which is equidistant from each loudspeaker, where the time and intensity differences are zero.

The distribution of time and intensity combinations across the entire coverage area can be summarized as shown in Figure 18a. The shape of the essential distribution pattern stays the same while

Figure 18. Time delay and intensity differences in loudspeaker reproduction: distribution of time and intensity differences form wing-like patterns for loudspeakers separated by 8 ft, 15 ft, 25 ft, and 50 ft (each line traces time and intensity differences for a cross section of the listening area) (a); time and intensity difference information for the 8-ft and 50-ft separation, superimposed over Barron's summary diagram (b). Figure 18b adapted by Rasch and Plomp (1982) and used by permission of Academic Press.



(a)



(b)

the time differences are compressed when the loudspeakers are closer together. These distribution patterns can also be superimposed on Mr. Barron's original summary diagram as shown in Figure 18b to reveal the range of spatial percepts associated with large and small reproduction settings.

With directional filters, 3-D sound is relatively robust in the range of time differences associated with the size of the head, but it is completely overwhelmed at time delays commonly experienced with loudspeakers in large rooms. This virtually rules out using DTFs except for those listeners who are located along the center line. (This is why the best 3-D solution for large listening spaces is to use an array of loudspeakers.) The only alternative strategy left for stereo reproduction is to target the directional cues on one selective region of the audience at a time. For example, one could add a time delay and intensity difference to the loudspeaker signals that compensate for the naturally occurring differences, shifting the line of listeners who experience no time or intensity difference away from the center toward another part of the audience. It is not a perfect strategy, since time and intensity patterns do not quite match, but in this way, some subset of the listeners could always be experiencing a 3-D cue.

Near-Field Reproduction

When the listener's position relative to the loudspeakers is fixed and known in advance, as can occur most easily in near-field reproduction settings such as living rooms and audio control rooms, 3-D sound will be most successful. Figure 19 shows an idealized loudspeaker reproduction setting and illustrates the transmission paths by which sound reaches the listener's eardrums.

The acoustic signals arriving at the eardrums have superimposed on them the HRTF for the loudspeaker's direction relative to the ipsilateral ear, typically 30 degrees in the horizontal plane (represented as H_{30}). Equalization should divide out the responses of the reproduction system and the H_{30} HRTF for the transmission path.

There are also acoustic signals that reach the ears from the loudspeakers on the other side of the head. For example, the signal from the left loudspeaker arrives at the right ear. These signals have superimposed on them the HRTF for the loudspeaker direction relative to the contralateral ear, typically 330 degrees in the horizontal plane (represented as H_{330}). These signals reaching the ears on the opposite side from each loudspeaker are typically referred to as acoustic *cross talk*. Cross talk creates constructive and destructive acoustic interference with the signals arriving directly from the closest loudspeakers. Figure 20 shows the change in magnitude response at the ears that results from

Figure 19. Paths by which signals arrive at the listener's eardrums in near-field loudspeaker reproduction.

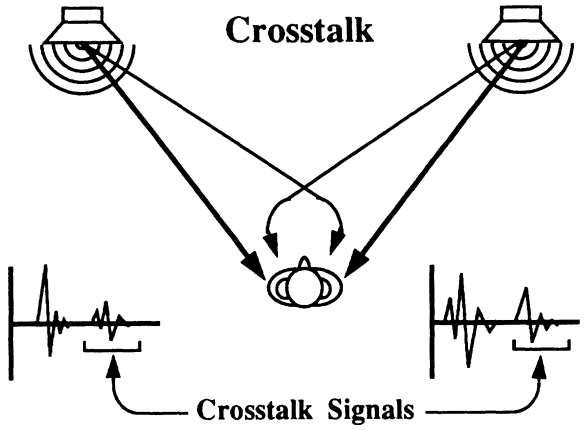


Figure 19

Figure 20. Magnitude response measured at listener's right ear in stereo reproduction: one loudspeaker on ipsilateral side (dotted line); two loudspeakers with cross talk (solid line).

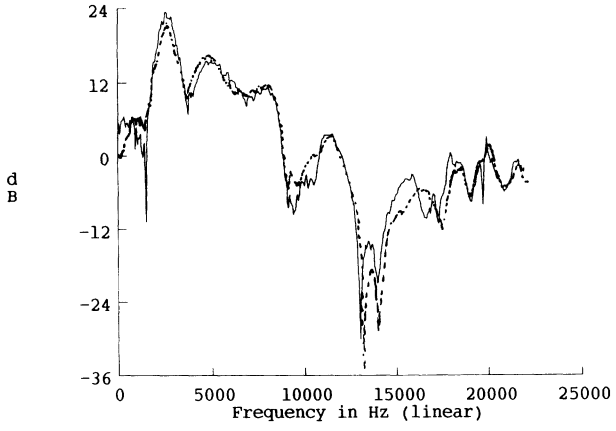
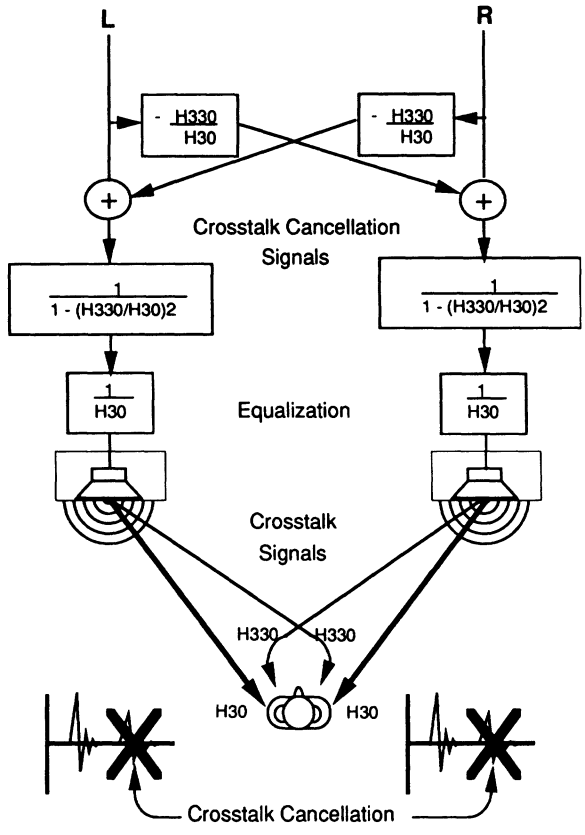


Figure 20

Figure 21. Schroeder-Atal method for cross-talk cancellation; H_{30} represents the HRTF of an ipsilateral loudspeaker placed at 30 degrees azimuth, and H_{330} represents the HRTF of a contralateral loudspeaker placed at 330 degrees azimuth.



cross talk, and the deep notch created around 2 kHz. Even though we are accustomed to the presence of cross talk and typically ignore it, one can learn to hear it in a reproduction environment that is free of room reflections. Even in the best of reproduction settings, cross talk is taken to be a natural part of the color of reproduced sound.

Cross-talk Cancellation

The first significant 3-D loudspeaker reproduction system was achieved by Schroeder and Atal in 1963. Despite the early date, it has served as the foundation for most 3-D loudspeaker systems ever since. To deliver to the ears the HRTFs associated with an illusory source location, this system has both to equalize for the H_{30} HRTF of the loudspeaker location and to eliminate the cross-talk signals with the H_{330} HRTF. It eliminates the cross-talk signals by issuing from the near loudspeaker a signal that could acoustically cancel the cross-talk signal from the far loudspeaker. This is represented in Figure 21. (The system is actually a bit more complex than described here.) The Schroeder-Atal system has many descendants, among the best of which could be considered the system described by Cooper and Bauck (1988).

All of the variants of this system are constrained by a set of assumptions that produce practical limitations. Just as with headphones, because there are individual differences in HRTFs, equalization is seldom perfect. This becomes particularly problematic for the cancellation signals, which must match the listener's H_{330} HRTF. Most importantly, to can-

cel the high-frequency content of the HRTFs, there must be an exact match between the signals arriving at the head and the cancellation signal. This is undermined by individual differences in HRTFs. In fact, cross-talk cancellation systems seldom cancel high-frequency signals, which are typically localized toward the loudspeakers even when the low- to mid-range signals are localized toward the side or rear. Small variances in the head position relative to the loudspeakers can cause total phase reversals of the cancellation signal and dense combing. It is typical that a shift in head position of less than 20 cm will totally collapse the imagery.

Alternative Approaches

An alternative to this approach was reported by Kendall and Rodgers (1982), who achieved loudspeaker localization with low-order digital filters that provided simple approximations of HRTFs without the benefit of cross-talk cancellation. Another alternative was achieved by Lowe and Lees (1991), who took a purely empirical approach and constructed very effective DTFs by direct experimentation with gated sinusoids (thereby capturing interaural onset delays). Some of the same problems associated with cross-talk cancellation affect these alternative approaches as well. Variations in head position cause inaccuracies in the high-frequency information arriving at the ears. (Because cross talk is never eliminated, the left and right loudspeaker signals combine acoustically at the ears and cause phase shifts and cancellations.) The primary advantages are that these systems are less sensitive to the listener's seating location. Kendall and Martens (1984) reported that circular sound paths retain their general shape and deform in a graceful manner, even as the listener moves far off center. Lowe and Lees reported that listeners were able to rotate their heads and orient toward the sound sources.

Reproduction Environment

Even with these alternative approaches, the loudspeaker reproduction environments often inhibit the creation of images in one or more spatial regions, due to early reflected sound in the reproduction environment or asymmetries in the reproduction equipment. Environmental reflections of sound arriving within 1 msec will corrupt the HRTFs. Therefore sound reflections near the loudspeakers or listener must be eliminated. This is considerably easier to manage in control rooms than in living rooms. Most susceptible are rear images, which often shift to the front or cling close to the listener's head, and side images, which collapse toward the front due to shifts in the location of the listener's head.

Conclusion

Both headphone and loudspeaker reproduction of directional cues present tractable problems and can be very successful in controlled reproduction settings. Headphone reproduction with head-tracking provides the most resilient form of reproduction, but it is also the most complicated and expensive, due to the overhead of dynamic filtering and head-tracking. Loudspeaker reproduction, even when limited to near-field monitoring, is more convenient but less resilient than headphone reproduction.

As the technology for reproducing directional cues becomes increasingly refined (and less expensive), different technical issues begin rising to the surface. The progressive increase in the level of complexity, from reproducing directional cues for a single source to reproducing full spatial environments, necessitates a tremendous increase in computational bandwidth. Simulated natural environments must be able to contain many individual sound sources and to replicate the reflected sound arriving at the listener from all directions. Also, interactivity is an essential element in breaking down the autonomy of auditory experience. The engineering and computational requirements of interactive spatial sound are tremendous, but must be met if we are to fulfill the aesthetic visions being born today in the minds of artists and audiences.

Acknowledgments

Appreciation and thanks to Bill Martens and Marty Wilde for many years of support and friendship

while studying auditory localization. Thanks to Doug Keislar for his careful and thoughtful comments on the manuscript. Thanks to Mike Pisaro, Matt Moller, and Lonny Chu for their suggestions. Some parts of this article are based on Kendall (1992).

References

Barron, M. 1971. "The Subjective Effects of First Reflections in Concert Halls—The Need For Lateral Reflections." *Journal of Sound and Vibration* 15:475–494.

Begault, D. R. 1991. "Challenges to the Successful Implementation of 3-D Sound." *Journal of the Audio Engineering Society* 39(11):864–870.

Begault, D. R. 1994. *3-D Sound for Virtual Reality and Multimedia.* Boston: Academic Press.

Blauert, J. 1971. "Localization and the Law of the First Wavefront in the Median Plane." *Journal of the Acoustical Society of America* 50:466–470.

Blauert, J. 1974. *Räumliches Hören.* Stuttgart: S. Hirzel Verlag. Also available as: *Spatial Hearing,* 1983, trans. John S. Allen. Cambridge, Massachusetts: MIT Press.

Borish, J. 1984. "Extension of the Image Model to Arbitrary Polyhedra." *Journal of the Acoustical Society of America* 75:1827–1839.

Boudreau, J. C., and C. Tsuchitani. 1968. "Binaural Interaction in the Cat Superior Olive S Segment." *Journal of Neurophysiology* 31:442–454.

Butler, R. A. 1973. "The Relative Influence of Pitch and Timbre on the Apparent Location of Sound in the Median Sagittal Plane." *Perception and Psychophysics* 14:255–258.

Butler, R. A., and K. Belendiuk. 1977. "Spectral Cues Utilized in the Localization of Sound in the Median Sagittal Plane." *Journal of the Acoustical Society of America* 61:1264–1269.

Casseday, J. H., and E. Covey. 1987. "Central Auditory Pathways in Directional Hearing." In William A. Yost and George Gourevitch, eds. *Directional Hearing.* New York: Springer-Verlag.

Chowning, J. 1971. "The Simulation of Moving Sound Sources." *Journal of the Audio Engineering Society* 19:2–6.

Cohen, E. A. 1989. "3D Sound Fiction, Fantasy, and Fact." Paper presented at the 87th Audio Engineering Society Convention, New York.

Cooper, D. H., and J. L. Bauck. 1988. "Prospects for Transaural Recording." Preprint 2734, 85th Audio Engineering Society Convention, November. New York: Audio Engineering Society.

Goldberg, J. M., and P. B. Brown. 1968. "Functional Organization of the Dog Superior Olivary Complex: An Anatomical and Electrophysiological Study." *Journal of Neurophysiology* 31:639–656.

Henning, G. B. 1974. "Detectability of Interaural Delay in High-frequency Complex Waveforms." *Journal of the Acoustical Society of America* 55(1):84–90.

Hiranaka, Y., and H. Yamasaki. 1983. "Envelope Representations of Pinna Impulse Responses Relating to Three-dimensional Localization of Sound Sources." *Journal of the Acoustical Society of America* 73(1):291–296.

Kendall, G., et al. 1986. "Image Model Reverberation from Recirculating Delays." Preprint 2408, 81st Audio Engineering Society Convention. December 1986.

Kendall, G. S. 1992. "Directional Sound Processing in Stereo Reproduction." *Proceedings of the 1992 International Computer Music Conference,* pp. 261–264. San Francisco: International Computer Music Association.

Kendall, G. S. 1995. "The Decorrelation of Audio Signals and Its Impact on Spatial Imagery (with an Appendix on Image Distance)." *Computer Music Journal* 19(4):71–87.

Kendall, G. S., and W. L. Martens. 1984. "Simulating the Cues of Spatial Hearing in Natural Environments." *Proceedings of the 1984 International Computer Music Conference.* San Francisco: International Computer Music Association.

Kendall, G. S., W. L. Martens, and M. D. Wilde. 1990. "A Spatial Sound Processor for Loudspeaker and Headphone Reproduction." *The Sound of Audio. Proceedings of the AES 8th International Conference.* New York: Audio Engineering Society.

Kendall, G. S., and C. A. P. Rodgers. 1982. "The Simulation of Three-dimensional Headphones Cues for Headphone Listening." *Proceedings of the 1982 International Computer Music Conference.* San Francisco: International Computer Music Association.

Kendall, G. S., M. D. Wilde, and W. L. Martens. 1989. "Production and Reproduction of Three-Dimensional Sound." Paper presented at the 87th Audio Engineering Society Convention, New York.

Kleiner, M., B. Dalenback, and P. Svensson. 1993. "Auralization—An Overview." *Journal of the Audio Engineering Society* 41(11):861–875.

Knudsen, E. I., and M. Konishi. 1978. "Space and Frequency are Represented Separately in Auditory Midbrain of the Owl." *Journal of Neurophysiology* 41:870–884.

Kuhn, G. F. 1987. "Physical Acoustics and Measure-

ments Pertaining to Directional Hearing." In William A. Yost and George Gourevitch, eds. *Directional Hearing*. New York: Springer-Verlag.
Kuwada, S., and T. C. T. Yin. 1987. "Physiological Studies of Directional Hearing." In William A. Yost and George Gourevitch, eds. *Directional Hearing*. New York: Springer-Verlag.
Lowe, D. D., and J. W. Lees. 1991. "Sound Imaging Process." US Patent no. 5,046,097.
Martens, W. 1987. "Principal Components Analysis and Resynthesis of Spectral Cues to Perceived Direction." *Proceedings of the 1987 International Computer Music Conference*. San Francisco: International Computer Music Association.
Martens, W. L. 1991. Directional Hearing on the Frontal Plane: Necessary and Sufficient Spectral Cues. PhD dissertation, Northwestern University.
Middlebrooks, J. C., et al. 1994. "A Panoramic Code for Sound Localization by Cortical Neurons." *Science* 264:842–844.
Moiseff, A., and M. Konishi. 1981. "Neuronal and Behavioral Sensitivity to Binaural Time Differences in the Owl." *Journal of Neuroscience* 1:40–48.
Moller, H. 1992. "Fundamentals of Binaural Technology." *Applied Acoustics* 36:171–218.
Morimoto M., and Y. Ando. 1983. "On the Simulation of Sound Localization." *Journal of the Acoustical Society of Japan* 74:873–887.
Musicant, A. D., and R. A. Butler. 1985. "Influence of Monaural Spectral Cues on Binaural Localization." *Journal of the Acoustical Society of America* 77(1):202–208.
Phillips, D. P., and J. F. Brugge. 1985. "Progress in Neurophysiology of Sound Localization." *Annual Review in Psychology* 36:245–274.
Rasch, R. A., and R. Plomp. 1982. "The Listener and the Acoustic Environment." In D. Deutsch, ed., *The Psychology of Music*. New York: Academic Press.
Rayleigh, Lord [Strutt, J. W.] 1907. "On Our Perception of Sound Direction." *Philosophical Magazine*, Sixth Series. 13:214–323.
Rose, J. E., et al. 1966. "Some Neural Mechanisms in the Inferior Colliculus of the Cat which May Be Relevant to Localization of a Sound Source." *Journal of Neurophysiology* 29:288–314.
Schroeder, M. R., and B. S. Atal. 1963. "Computer Simulation of Sound Transmission in Rooms." *IEEE Conv. Record*, 7:150–155.
Smith, J. O. 1985. "An Introduction to Digital Filter Theory: Part II. Theoretical Foundations of Digital Filters." In J. Strawn, ed., *Digital Audio Signal Processing*. Los Altos, California: William Kaufmann.
Stotler, W. A. 1953. "An Experimental Study of the Cells and Connections of the Superior Olivary Complex of the Cat." *Journal of Comparative Neurology* 98:401–432.
Wallach, H. 1940. "The Role of Head Movements and Vestibular and Visual Cues in Sound Localization" *Journal of Experimental Psychology* 27(4):339–368.
Wallach, H., E. B. Newman, and M. R. Rosenzweig. 1949. "The Precedence Effect in Sound Localization." *The American Journal of Psychology* 62:315–336.
Wenzel, E. M., F. L. Wightman, and D. J. Kistler. 1993. "Localization Using Nonindividualized HRTFs." *Journal of the Acoustical Society of America* 94(1):112–123.
Wightman, F. L., and D. J. Kistler. 1989. "Headphone Simulation of Free-field Listening. II. Psychophysical Validation." *Journal of the Acoustical Society of America* 85:858–867.
Wightman, F. L., and D. J. Kistler. 1992. "The Dominant Role of Low-frequency Interaural Time Differences in Sound Localization." *Journal of the Acoustical Society of America* 91(3):1648–1661.
Woodworth, R. S. 1954. *Experimental Psychology*, rev. ed. New York: Rinehart and Winston.
Yin, T. C. T., S. Kuwada, and Y. Sujaku. 1984. "Interaural Time Sensitivity of High-frequency Neurons in the Inferior Colliculus." *Journal of the Acoustical Society of America* 76:1401–1410.