Perceptual Factors in Virtual Acoustic Displays

Elizabeth M. Wenzel
Aerospace Human Factors Research Division
NASA-Ames Research Center, Mail Stop 262-2
Moffett Field, CA 94035
U.S.A.
beth@eos.arc.nasa.gov

Abstract

Virtual acoustics, also known as 3-D sound and auralization, is the simulation of the complex acoustic field experienced by a listener within an environment. Going beyond the simple intensity panning of normal stereo techniques, the goal is to process sounds so that they appear to come from particular positions in three-dimensional space. Current techniques use digital signal processing to synthesize the acoustical properties that people use to localize a sound source. Thus, they provide the flexibility of a kind of digital binaural head, allowing a more active experience in which a listener can both design and move around or interact with a simulated acoustic environment in real time. Such simulations are being developed for a variety of application areas including architectural acoustics, advanced human-computer interfaces, telepresence and virtual reality, navigation aids for the visually-impaired, and as a test bed for psychoacoustical investigations of complex spatial cues.

The paper will discuss some recent technology developments in this area, the results of psychophysical experiments examining the perceptual validity of the synthesis technique, and factors which can enhance perceptual accuracy and realism.

Introduction

As with most research in information displays, virtual displays have generally emphasized visual information. Many investigators, however, have pointed out the importance of the auditory system as an alternative or supplementary information channel. [16] A three-dimensional auditory display can potentially enhance information transfer by combining directional and iconic information in a quite naturalistic representation of dynamic objects in the interface.

A three-dimensional auditory display can be realized with an array of real sound sources or loudspeakers. [17] An alternative approach is to generate externalized, three-dimensional sound cues over headphones in realtime using digital means. [60] Headphone presentation is desirable because it enables complete control over the acoustic waveforms delivered to the two ears, a feature which is critical to the control of apparent spatial position. The realtime capability, coupled with a head-tracking device, allows the user to experience virtual sounds interactively; virtual sources can thus be either moving or static and they can respond appropriately to the listener's head movements.

Performance Advantages of Virtual Acoustic Displays

Useful features of acoustic signals in general include the fact that they tend to produce an alerting or orienting response and that they can be detected more quickly than visual signals. [41, 42] Such characteristics are probably responsible for the most prevalent use of nonspeech audio in simple warning systems, such as the malfunction alarms used in aircraft cockpits or the siren of an ambulance. Another advantage of audition is that it is primarily a temporal sense and we are extremely sensitive to changes in an acoustic signal over time. [41] This feature tends to bring a new acoustical event to our attention and conversely, allows us to relegate sustained or uninformative sounds to the background. Consequently, audio is particularly suited to monitoring state changes over time, for example when the hard drive of your computer suddenly begins to malfunction (see Cohen, in [31]).

Acoustic displays can be further enhanced by taking advantage of the auditory system's ability to segregate, monitor, and switch attention among simultaneous streams of sound, particularly when such acoustic objects are distinguished by different locations in space. [9, 41] Another advantage of the binaural system, related to the so-called "cocktail party effect", is that the spatial separation of sounds improves the intelligibility of signals in a background of noise or other voices. [10, 14] Thus, the two primary performance advantages that can be expected from using spatial sound are enhanced situational awareness, or the direct comprehension of object relationships in a three-dimensional task space, and enhanced comprehension of multiple, simultaneous sound streams or voices (Table 1).

Some of the kinds of tasks that can benefit from spatial presentation of information are given in Table 2. In each case, the various roles that spatial cues could play -- direct representation of spatial information, spatial metaphors for displaying nonspatial information, and enhanced stream segregation -- are outlined.

For example, the direct representation of spatial information could be used in architectural design, where the ability to interactively simulate, or "auralize", room acoustics would be of great use in exploring and avoiding possible acoustically-undesirable effects that may not be obvious from a visual design. [21, 48] Similarly, when coupled with some type of range-finding device, artificial acoustical environments could be used as aids for the visually impaired by audibly representing the surfaces and obstacles through which a blind person must navigate. [34]

Spatial metaphors for displaying the organization of information could also be quite useful in navigating through large-scale databases, as in spatial mnemonic devices in which related topics are located in adjacent rooms of a metaphorical building. Similarly, complex, multidimensional variables from models in computational fluid dynamics could be represented spatially (e.g., representing the dispersion of errors for modeled vs. actual turbulence data via localized intensity cues; modeling aerodynamic flow patterns using a "virtual wind tunnel" [12]).

Segregation enhancement can be critical in applications involving both simultaneous speech channels, as in aviation communication systems,[4, 5, 6] and the kind of encoded nonspeech cues proposed for scientific "visualization" or sonification. Examples include aeronautical displays for air traffic control in which the controller hears communications from incoming traffic in positions which correspond to their actual location in the terminal area, [7] sonification displays for the acoustic representation of multi-dimensional statistical data, [30, 31] and alternative computer interfaces for the visually-impaired. [18]

Another aspect of auditory spatial cues is that, in conjunction with the other senses, they can reinforce the information content of a display and provide a greater sense of presence or realism in a manner not readily acheived by a single (usually visual) modality. [15, 47, 58] Similarly, in direct-manipulation tasks, auditory cues can provide an alternative medium for the representation of tactile or force-feedback cues. [64] This is a quite difficult interface problem for multimodal displays that is only beginning to be solved. [39] Intersensory synergism will be particularly important in applications involving telepresence, including advanced teleconferencing, [35] shared electronic workspaces, [19, 23] and monitoring telerobotic

Table 1. Performance advantages of the spatial presentation of sound.

Enhanced Situational Awareness

* Direct representation of spatial information.
* Omnidirectional Monitoring: "The function of the ears is to point the eyes."
* Reinforces (or replaces) information in other modalities.
* Enhances the sense of presence or realism.

Enhanced Multiple Channel Presentation

* The "Cocktail Party Effect:" Improves intelligibility, discrimination, and selective attention for sound sources in a background of noise or other sources.
* Enhanced Stream Segregation: Allows separation of multiple sounds into distinct "objects."

Table 2. Performance advantages associated with the spatial presentation of sound in various application areas.

Architectural Acoustics: Acoustical CAD/CAM Systems	
Direct representation of spatial information:	"auralization" of models of room acoustics
Enhanced source intelligibility / separation:	simultaneous sources
Data Spaces: Large-Scale Databases / Information Systems	
Symbolic representation via spatial location:	database navigation: an architecural / spatial metaphor for database organization
Enhanced intelligibility / separation:	simultaneous icons / symbologies
Data Visualization: Computational Fluid Dynamics, Virtual Wind Tunnel	
Direct representation:	airflow / noise patterns produced by an aircraft engine
Symbolic representation:	localized intensities => size of error measures dispersed over the sample grid of a flow model
Enhanced intelligibility / separation:	simultaneous icons / symbologies
Aeronautics: ATC Displays, Cockpit Warning Systems	
Direct representation:	incoming aircraft locations; left vs. right engine malfunctions
Symbolic representation:	different aircraft systems mapped to different locations (a cockpit data space)
Enhanced intelligibility / separation:	simultaneous radio communications
Telerobotic Control: Space Station Construction and Repair	
Direct representation:	contact cues; range-finding
Enhanced intelligibility / separation:	simultaneous icons / symbologies

activities in remote or hazardous situations.[64] Similarly, the interaction of the senses will be critical in purely virtual environments for visualization, large-scale data management and systems control,[11, 19] and entertainment.[29]

Techniques for Creating Spatial Displays

The success of spatial synthesis relies on understanding the acoustical cues that are used by human listeners to locate sound sources in space. In the original duplex theory of sound localization based on experiments with pure tones (sinewaves), interaural intensity differences (IIDs) were thought to determine localization at high frequencies because wavelengths smaller than the human head create an intensity loss or head shadow at the ear farthest from the sound source.[51] Conversely, interaural time differences (ITDs) were thought to be important for low frequencies since interaural phase (delay) relationships are nonambiguous only for frequencies below about 1500 Hz (wavelengths larger than the head.) The duplex theory, however, cannot account for the ability to localize sounds on the vertical median plane where interaural cues are minimal. Also, when subjects listen to sounds over headphones, they usually appear to be inside the head even though ITDs and IIDs appropriate to an external source location are present. Many studies now indicate that these deficiencies of the duplex theory reflect the important contribution to localization of the direction-dependent filtering that occurs when incoming sound waves interact with the outer ears. Experiments have shown that spectral shaping by the pinnae is highly direction-dependent, [55] that the absence of pinna cues degrades localization accuracy, [22] and that pinna cues are at least partially responsible for externalization or the "outside-the-head" sensation. [49]

The synthesis technique typically used in creating a spatial auditory display involves the digital generation of stimuli using location-dependent filters constructed from acoustical measurements made with small probe microphones placed in the ear canals of individual subjects or artificial heads for a large number of different source (loudspeaker) locations. [66] The filters constructed from these ear-dependent characteristics are examples of Finite Impulse Response (FIR) filters and are often referred to as Head-Related Transfer Functions (HRTFs). Acting something like a pair of graphic equalizers, the HRTFs capture the essential cues needed for localizing a sound source; the ITDs, IIDs, and the spectral coloration produced by a sound's interaction with the outer ears. Figure 1 illustrates these complex spectral effects. The panels, which plot representations in the frequency domain, show what happens to a broadband sound source (e. g., a train of noisebursts) delivered from two different locations after interaction with the outer ears. The differences between the left and right intensity curves are the IIDs at each frequency. Spectral phase effects (frequency-dependent phase, or time, delays) are also present in the measurements, but are not shown here for clarity. While the interaural phase differences are somewhat frequency-dependent, in practice they are sometimes approximated by inserting a single delay between the two ears that increases in size with increasing displacement from the median plane (e. g., a delay estimated from the peak of the cross-correlation function of the left and right HRTFs [28]).

Using these HRTF-based filters, it is possible to impose spatial characteristics on a signal such that it apparently emanates from the originally-measured location. Spatial synthesis can be achieved either by filtering in the frequency domain, a point-by-point multiplication of an input signal with the left and right HRTFs, or by filtering in the time domain, using the FIR representation and a somewhat more computationally-intensive, multiply-and-add operation known as convolution. Of course, the localization of the sound will also depend on other

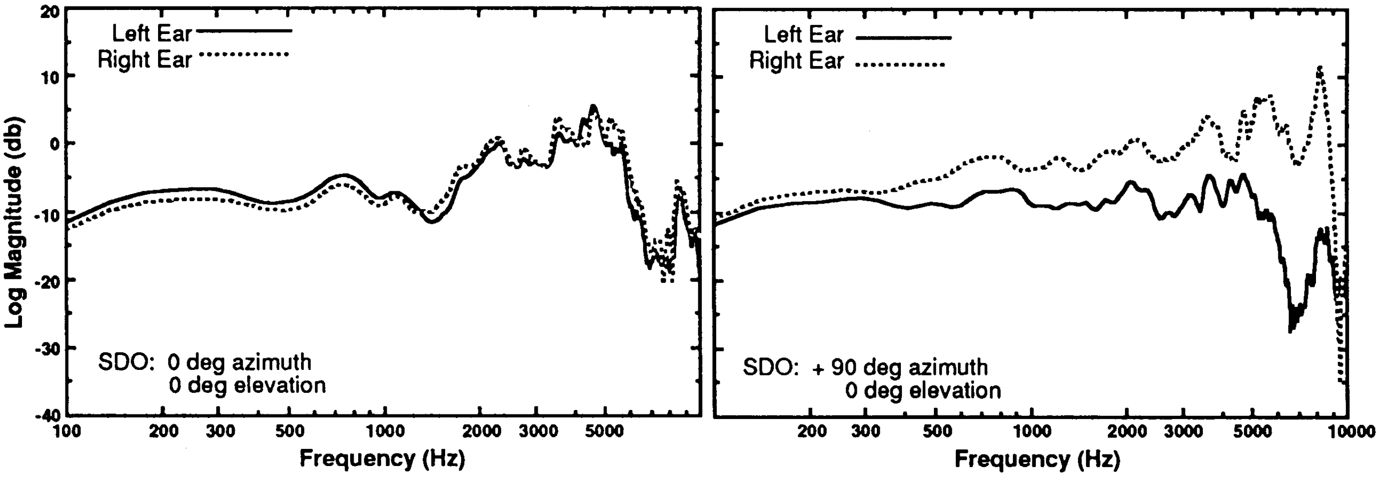

Figure 1. Illustration of the effects of spectral shaping by the pinnae. Magnitudes are plotted the left (solid line) and right (dashed line) ear canals of a single individual for stimuli delivered from loudspeakers at two different locations: directly in front (0° azimuth, 0° elevation), and directly to the right (90° azimuth).

Table 3. Examples of current 3D sound systems and their performance characteristics.

AKG Creative Audio Processor (Persterer [48])
headphone presentation, 32 sources/reflections, room modeling, large standalone hardware system.
HRTFs: individualized and artificial head.

Beachtron, Convolvotron, & Acoustetron, Crystal River Eng. (Foster [20])
headphone presentation, 2 anechoic sources to 16 anechoic sources or 4 sources plus 6 early reflections each, Doppler effects, variable source radiation, interactive head-tracking supported, PC host.
HRTFs: several individuals based on published behavioral data.

Focal Point, Gehring Research (Gehring [24])
headphone presentation, 1 to 2 anechoic sources per card, interactive head-tracking supported, MacIntosh and PC hosts.
HRTFs: unknown source.

HEAD Acoustics (Sonic Perceptions; Gierlich [25])
headphone presentation, 4 anechoic sources or 1 source plus 3 early reflections, standalone hardware system.
HRTFs: based on a structural model.

McKinley and Ericson [36], Wright-Patterson AFB
headphone presentation, 2 to 4 anechoic sources, interactive head-tracking supported, standalone hardware (lab and flightworthy systems).
HRTFs: individualized artificial head.

Roland Spatial Sound Processor (Chan [13])
headphone and loudspeaker presentation, 4 anechoic sources, standalone hardware system.
HRTFs: single individual's.

factors such as its original spectral content; narrowband sounds (sinewaves) are generally very difficult to localize while broadband, impulsive sounds are the easiest to locate. Filtering with HRTF-based filters cannot increase the bandwidth of the original signal, it merely transforms the energy and phase of the frequency components that are initially present.

In most current systems (Table 3), from one to four moving or static sources can be simulated (with varying degrees of fidelity) in an anechoic (free-field or echoless) environment by time-domain convolution of incoming signals with HRTF-based filters chosen according to the output of a head-tracking device. The head-tracking device allows the display to update the directional filters in real time to compensate for a listener's head motion so that virtual sources remain stable within the simulated environment. Motion trajectories and static locations at finer resolutions than the empirical data are generally simulated either by switching, or more preferably, by interpolating between the measured HRTFs. [63] Also, in some systems, a simple distance cue can be provided via real-time scaling of amplitude.

Psychophysical Validation

The literature on human sound localization indicates that several kinds of perceptual errors occur even for sounds in the real-world; in particular, there may be apparent front-back or up-down reversals, poor elevation accuracy, and failures of externalization (Figure 2). Such errors can be exacerbated when spatial cues are

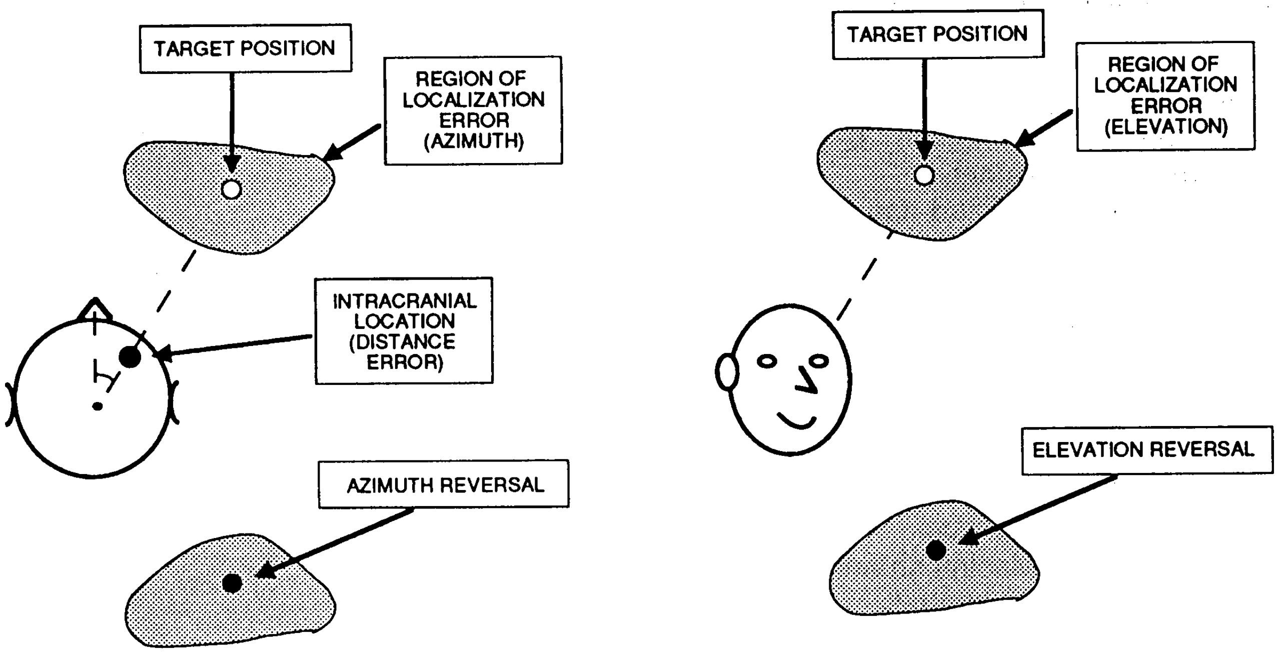
TARGET POSITION
REGION OF LOCALIZATION ERROR (AZIMUTH)
INTRACRANIAL LOCATION (DISTANCE ERROR)
AZIMUTH REVERSAL
TARGET POSITION
REGION OF LOCALIZATION ERROR (ELEVATION)
ELEVATION REVERSAL

Figure 2. Illustration of the types of perceptual errors observed for human sound localization..

synthesized. The only conclusive test of the adequacy of the spatial simulation in a virtual acoustic display is an operational one in which the localization of real and synthesized stimuli are directly compared in psychophysical studies. Wightman and Kistler [67] compared subjects' localization judgments of static sources in the free-field with their judgments of virtual (headphone-presented) sources synthesized from the subjects' own HRTFs. Presumably, synthesis using individualized HRTFs would be the most likely to replicate the free-field experience for a given listener. In general, Wightman and Kistler [67] reported that localization accuracy for the free-field and headphone stimuli was comparable for the experienced listeners of their study. However, some minor degradation of localization accuracy was observed; in particular, source elevation was less well-defined for virtual sources.

If virtual sources are to be used in a general-purpose 3-D auditory display, it may not be feasible to measure the HRTFs from each potential listener. It may also be the case that the user of such a display may not have the opportunity for extensive training. Thus, a critical issue for the design of virtual acoustic displays is the degree to which the general population of listeners can obtain adequate localization cues from stimuli based on non-individualized HRTFs. A recent study by Wenzel, Arruda, Wightman, and Kistler [62] suggests that it may be feasible to use non-individualized transfer functions to synthesize 3-D auditory display cues, so long as the HRTFs that are used come from a subject whose measurements have been "behaviorally-calibrated" and are thus correlated with accurate localization performance in both free-field and headphone conditions. Sixteen inexperienced listeners judged the apparent spatial location of 24 target locations presented over loudspeakers in the free-field or over headphones. The headphone stimuli were generated digitally using HRTFs measured in the ear canals of a representative subject, a "good localizer" (SDO), from the experiment by Wightman and Kistler. [67] Localization performance was quite good for 12 of the subjects, with judgments for the non-individualized stimuli nearly identical to those in the free-field. Two of the subjects showed poor elevation performance in both free-field and headphone conditions while another two subjects showed inconsistent behavior, with poor elevation accuracy in only the synthesized conditions.

In general, these data suggest that most listeners can obtain useful directional information from an auditory display without requiring the use of individually-tailored HRTFs, particularly for the dimension of azimuth. However, a caveat is important regarding the existence of localization reversals. The data described above were based partially on analyses in which reversals had been resolved (i.e., coded as if subjects had indicated the correct hemisphere) with reversal rates computed as a separate statistic. For free-field versus simulated free-field stimuli, the 8 experienced listeners in the Wightman and Kistler study exhibited front-back reversal rates of about 6 vs. 11% while the 16 inexperienced listeners using non-individualized HRTFs showed average rates of about 19 vs. 31%. [62] In both studies, the majority of the reversals were the result of front sources apparently reversing to the rear. Note that the existence of real-source reversals indicates that these confusions are not strictly the result of the simulation.

In a recent study at NASA-Ames, Begault [3] investigated the effects of synthetic reverberation (based on an extension of a ray-tracing model) on inexperienced listeners' localization and externalization of static, virtual sound sources using the same HRTFs as in Wenzel, et al. [62] He found that, compared to anechoic stimuli, adding reverberation to speech stimuli nearly eliminated non-externalized judgments but tended to decrease localization accuracy. Reversals, on the other hand, were relatively unaffected. Average rates were 33% for both anechoic and reverberant conditions, although some individual differences were observed in the relative bias toward front-to-back versus back-to-front reversals in the two conditions.

Engineering Compromises vs. Perceptual Validity

In implementing a real-time spatial auditory display, many engineering compromises must be made to achieve a practical system. Table 4 is an attempt to examine such compromises by comparing manufacturers' engineering specifications for the systems listed in Table 3 with traditional psychoacoustical parameters that are likely to be most relevant to adequate perceptual performance during spatial synthesis. This comparison is often difficult to make because the most perceptually-relevant specifications are not given, or the measurement technique is not stated clearly enough for one to assess the relevance of the value that is given. For example, the total latency through a system is critical for assessing the adequacy of the real-time simulation of moving sources and/or moving listeners. Measurements of the minimum audible movement angle (MAMA) [43, 46] suggest that this total latency should not be more than about 50 msec for moderate source/listener velocities (less than 360°/sec). Frequently, however, a system latency is simply not given, or the value that is quoted is probably based not on the overall system latency, but on the time it takes the particular DSP chip being used to load in a new filter. While the latter specification is certainly of interest, it may or may not be well correlated with the interactive, real time performance of the system.

Similarly, the perceptual adequacy of a simulation can be affected by the spatial resolution of the set of HRTFs being used to synthesize spatial cues during motional simulation. Ideally, one would like the minimum spatial resolution of the measurements to be approximately equivalent to the spatial resolution of the human auditory system as measured by such methods as the minimum audible angle (MAA; approximately 1-5°) [38, 44, 45] or absolute judgment paradigms (20-30°). [67] Since it may not always be practical to measure HRTFs at all possible angles of incidence, motion trajectories and static locations at finer resolutions than the empirical data can be simulated by interpolating, or computing a weighted average of the measured impulse responses closest to the desired target location. Fast-memory limitations in synthesis devices may also make it advantageous to reduce the number of HRTFs that must be stored as much as possible.

While simple linear interpolation seems a reasonable approach, comparisons of the acoustical features of interpolated and non-interpolated impulse responses reveal obvious spectral discrepancies. [68] A simple example illustrates the nature of the errors introduced. If one were to average two impulse responses, one with an undelayed impulse (at τ_0) and one with an impulse at time delay τ (e.g., a different azimuth), the correct result would be a phase response which reflects an impulse at time $\tau/2$ (the in-between azimuth) and a flat magnitude response, if the original magnitude responses were also flat. However, temporal interpolation results in a cosinusoidal magnitude response and a phase response reflecting two half-height impulses at time τ_0 and time τ. This simple case is clearly incorrect. The net effect of such errors for synthesis with the much more complex stimuli represented by HRTFs is less clear. It may be that despite errors occurring for interpolation at each ear, the interaural relationships remain relatively intact while the complex (monaural) spectral cues largely responsible for elevation perception and

Table 4. Perceptual vs. Engineering Specifications for current 3D sound systems

	Simultaneous Sources	Frequency Bandwidth	Dynamic Range	Spatial Resolution	Motional Resolution
Auditory System	7 ± 2? (Short-term memory limits)	20 Hz - 20 kHz HRTFS: nominal 20 Hz - 15 kHz	100+dB	Relative (MAA): az: 1°-5° el: 4°-5° Absolute: 20°-30°	velocity-dependent: v: 8°-360° /sec MAMA: 4°-21° latency: 500-58 ms
AKG	1-32 sources/ reflections HRTFs: ~100 pts	22-25 kHz (sample rates: 44.1-50 kHz)	85-90 dB practical (16-bit+3 A/D,D/A, 32-bit arith. FP)	az: ?° el: ?°	latency: ?
Crystal River	1-28 sources/ reflections HRTFs: 75-512 pts	22-25 kHz (sample rates: 44.1-50 kHz)	85-90 dB practical (16-bit A/D,D/A, 24-bit arith.)	az: 30° el: 18° interpolated	system latency: 30-40 msec; head-tracked
Focal Point	1-2 sources HRTFs: ~100 pts	22 kHz (sample rate: 44.1 kHz)	85-90 dB practical (16-bit A/D,D/A, 24-bit arith.)	az: ?° el: ?° interpolated	3-6 msec filter update (not a true latency?) head-tracked
HEAD Acoustics	1-8 sources/ reflections HRTFs: ?	22-24 kHz (sample rates: 44.1-48 kHz)	85-90 dB practical (16-bit conv. 24-bit arith.)	az: 5° el: 2° switches	latency: ?
ALCS (Wright-Patt) lab & flightworthy	1-4 sources HRTFs: ~128 pts	20 kHz (sample rate: 40 kHz)	85-90 dB practical (16-bit A/D,D/A, 16-bit arith.)	if az only: 1° az: 15° el: 15° interpolated	system latency: ~10 msec? head-tracked
Roland RSS	1-4 sources HRTFs: ? pts	22-24 kHz (sample rates: 44.1-48 kHz)	95-100 dB practical (18, 20-bit A/D,D/A 24-bit arith.)	az: ?° el: ?°	latency: ?

disambiguation of the cones-of-confusion may be more likely to be disrupted. Further, the impact of such errors may be mitigated by using relatively small interpolation intervals on the order of the MAA. Thus, it is important to determine whether interpolation is a viable technique as well as the minimal resolution required to achieve a perceptually-adequate simulation.

During informal listening, the locations of virtual images and smooth motion achieved by systems such as the Convolvotron are not obviously distorted when using stimuli such as music. However, signals with approximately stationary spectra, like white noise, result in an unnatural dynamic, comb-filtering effect when the listener and/or source is in motion. Such an effect is probably due to the dynamic roving of the spuriously-located impulses as the stimuli move from empirical to interpolated locations. A method for minimizing such effects has been suggested by Wightman, Kistler, and Arruda.[68] They use minimum-phase approximations of measured HRTFs which avoid the differences in onset (and therefore locations of the peaks in the impulse responses) that occur across source locations. Briefly, the magnitudes of the minimum-phase filters are the same as the original filters, the phase is derived from the magnitude spectra, and the interaural delay is represented by a single delay estimated from the peaks of the cross-correlations of the left and right-ear HRTFs. [28] Recently, this method has been implemented in the Convolvotron and appears to reduce the comb-filtering effect described above. Four-way linear interpolations of the impulse responses derived from the minimum-phase HRTFs are computed as before, while the interaural delay estimates are interpolated separately and inserted at the end of the filtering process.

Recently, some formal perceptual studies have been conducted to determine the computational trade-offs that can be tolerated during interpolation in a virtual acoustic display. Wightman, et al. [68] compared localization accuracy (static target locations) for non-interpolated (normal HRTFs) vs. interpolated stimuli synthesized from the five subjects' own HRTFs (none were SDO). Interpolations of both normal HRTFs and minimum-phase HRTFs were tested and the size of the interpolation intervals was varied simultaneously in both azimuth (15° and 30° intervals) and elevation (12° and 24° intervals). In general, their data indicated that localization accuracy was comparable for the non-interpolated and minimum-phase interpolations using the smallest interpolation intervals. Increasing the azimuth and elevation intervals to 30° and 24°, respectively, only slightly reduced localization accuracy. Interpolations of normal HRTFs, on the other hand, resulted in decreased accuracy (primarily increased front-back confusions) for even the smallest interpolation intervals.

Wenzel and Foster [63] investigated stimulus conditions based on non-individualized transforms, a common circumstance for most virtual acoustic displays. Three subjects' localization judgments were compared for four synthesis conditions: stimuli synthesized from normal, non-interpolated HRTFs, simple linear interpolations of normal HRTFs, stimuli synthesized from non-interpolated minimum-phase approximations of the HRTFs, and linear interpolations of the minimum-phase HRTFs. The empirical HRTFs were those of a "good localizer" (SDO) from previous studies [62, 67] whose measurements are one of the data sets provided with the Convolvotron. Interpolation intervals were varied both independently and simultaneously in azimuth (15°-60° intervals) and elevation (9°-36° intervals) in an attempt to assess the relative contribution of each dimension to potential localization errors during the interpolation process. In contrast to Wightman et al., [68] localization accuracy was largely unaffected by interpolation of either normal or minimum-phase HRTFs, even for the largest interpolation intervals (60°), when compared to non-interpolated stimuli. However, Wightman, et al., used the subjects' own HRTFs while Wenzel and Foster used non-individualized transforms. The data were actually quite similar to the previous study [62] in which inexperienced subjects' free-field localization was compared to localization of virtual sources synthesized from SDO's non-interpolated transforms, particularly in terms of the high confusion rates observed for the synthesized stimuli. It appears that, for static sound sources, the effect of interpolation is relatively small compared to the impact on localization accuracy of using non-individualized HRTFs. Thus, it may be that several sets of HRTFs from subjects with different physical (and therefore acoustical) characteristics should be made available with a virtual acoustic display so that the best match possible may be made between listeners and the data sets. The fact that the subject who was physically most similar to SDO in the study by Wenzel and Foster had the lowest overall confusion rates tends to support this view.

Enhancing Perceptual Accuracy

So far, the psychophysical data using static sources suggest that the primary difficulties for synthesizing spatial information in virtual acoustic displays will be ensuring reliable elevation discrimination and the elimination or, at least, minimization of reversals. Reversals are probably due in large part to the static nature of the stimulus and the ambiguity resulting from the so-called cones of confusion (see Blauert [8]). Cone-of-confusion effects alone, however, cannot explain a front-to-back response bias, and it is probable that higher-level cognitive factors like visual dominance play a substantial role in auditory localization. [59] It is also possible, as Asano, Suzuki, and Sone [2] have claimed, that reversals tend to diminish as subjects gain experience with the impoverished stimuli provided by static anechoic sources, whether real or simulated. Subjects may eventually learn to discriminate the subtle and unfamiliar, location-dependent spectral cues that allow them to reliably resolve the cones of confusion. The higher overall reversal rates for the inexperienced listeners of Wenzel, et al. [62] compared to the more experienced subjects of Wightman and Kistler [67] tend to support this view. A similar process of adaptation may be required to learn unfamiliar spectral cues for elevation. Thus, it may be that some form of adaptation or training will usually be needed to take full advantage of a virtual acoustic display.

Again, another problem in synthesizing veridical acoustic images over headphones is the fact that such stimuli sometimes fail to externalize. Environmental cues such as the ratio of direct to reflected energy and reverberation time appear to enhance the externalization of images. [3, 49] Further, just as we come to learn the characteristics of a particular room or concert hall, the localization of virtual sounds may improve if the listener is allowed to become familiar with sources as they interact in a particular artificial acoustic world. For example, perhaps dynamic simulation of an asymmetric room would tend to aid the listener in distinguishing front from rear locations by strengthening spectral or timbral differences. By taking advantage of the head-tracker in the real-time system, we can close the loop between the auditory, visual, vestibular, and kinesthetic systems and study the effects of dynamic interaction with relatively complex, but known, acoustic environments. However, the specific parameters used in such a model must be investigated carefully if localization accuracy is to remain intact. [3] It may be possible to discover an optimal trade-off between environmental parameters which enhance externalization while minimizing the impact of the resulting expansion of the spatial image that can interfere with the ability to judge the direction of the source.

Whether distance can be reliably controlled beyond mere externalization also awaits further research. Humans appear to be quite poor at judging the absolute distance of sound sources. Distance judgments depend at least partially on the intensities of sound sources, but the relationship is not a simple one and interacts heavily with factors like stimulus familiarity and reverberation. [37] Enhancing the ability to make relative, rather than absolute, distance judgments may be a more fruitful approach and at least crude manipulations of relative distance should be possible in a virtual acoustic display. Further understanding of the role of environmental cues, and the ability to synthesize such cues interactively, may eventually improve the reliable identification of source distances.

While non-interactive room modeling or auralization, has been implemented for some time, [1, 25, 26, 29, 32, 48, 50, 52, 54] recently some progress has been made toward interactively synthesizing reverberant cues. For example in one system (the Convolvotron), the walls, floor, and ceiling in an environment are simulated by using HRTF-based filters to place the "mirror image" of a sound source behind each surface to account for the specular reflection of the source signal. [21] The filtering effect of surfaces such as wood or drapery can also be modeled with a separate filter whose output is delayed by the time required for the sound to propagate from each reflection being represented. Such dynamic modeling requires enormous computational resources for real-time implementation in a truly interactive (head-tracked) display. Currently it is not practical to render more than the first one or two reflections from a small number of surfaces. For example, the Convolvotron requires approximately 200 million operations per second to render the direct path plus six first-order reflections for a single sound source, or a total of seven acoustic images that can be interactively-updated in real time (up to four sources and twenty-eight images in an Acoustetron). Future work in this area will examine the perceptual consequences of using dynamic reflection models and eventually extend the approach to more realistic models of acoustic environments.

Creating an Integrated Virtual Acoustic Display

Although simulation of the spatial cues used by the listener has received the most attention in recent work, it represents only one component of a virtual acoustic display. The development and validation of HRTF-based processing and the environmental modeling techniques summarized in Figure 3 could clearly encompass years of research, yet in many ways the problems that need to be solved in spatial synthesis are at least fairly well understood. The same cannot be said for the real-time generation of other qualities of acoustic sources that are also critical for simulating acoustic objects in a virtual display. In addition to spatial location, various acoustic features such as temporal onsets and offsets, timbre, pitch, intensity and rhythm, can specify the identities of individual objects and convey meaning about discrete events or ongoing actions in the world and their relationships to one another. One can systematically manipulate these features, effectively creating an auditory symbology for computer interfaces that operates on a continuum from literal everyday sounds, such as the clunk of a file being thrown into the trash can, to a completely abstract mapping of statistical data into multidimensional sound parameters. Unfortunately, both current synthesis technology and understanding of the perceptual consequences of simulating such complex, multidimensional stimuli are not yet well developed.

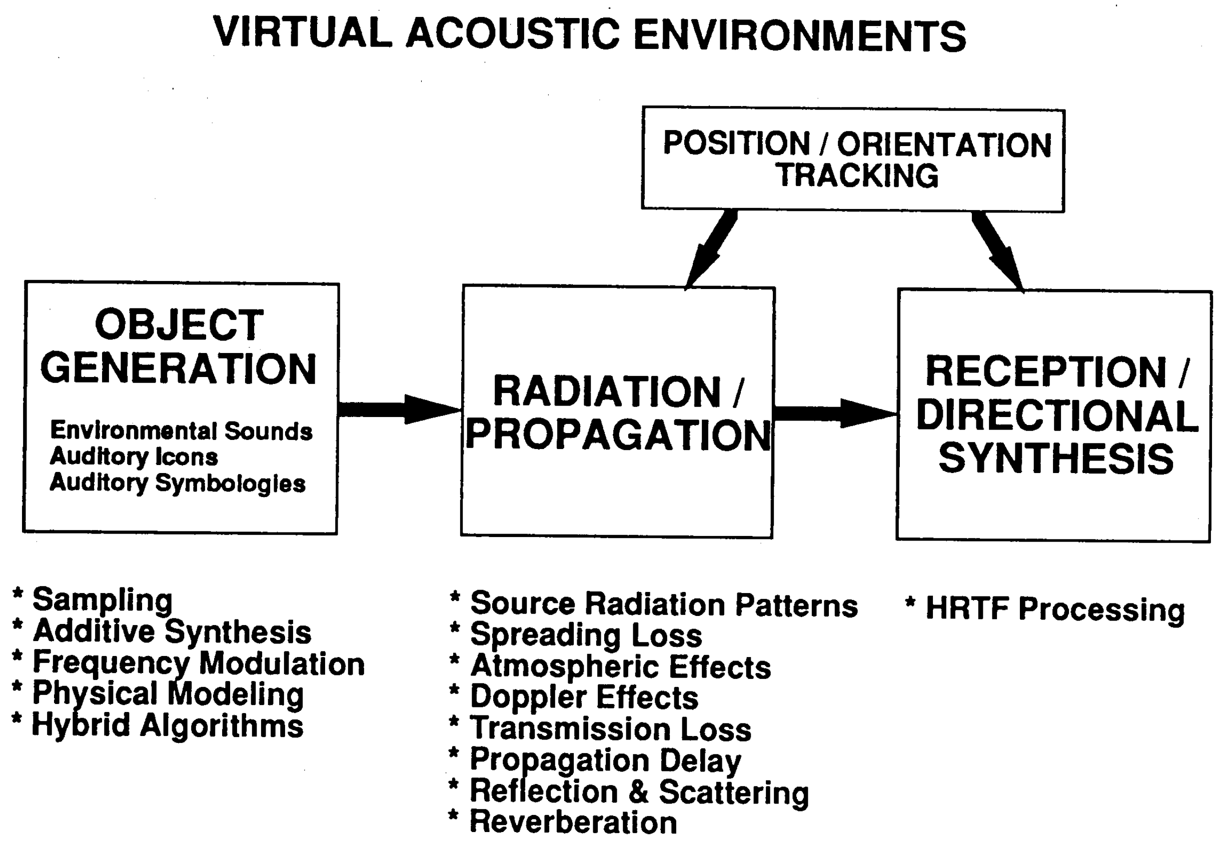

Figure 3. The primary components required for simulation of acoustic objects in a virtual acoustic environment.

The ideal synthesis device would be able to flexibly generate the entire continuum of nonspeech sounds described above as well as be able to continuously modulate various acoustic parameters associated with these sounds in real time. As implied by Figure 3, such a device or devices would act as the generator of acoustic source characteristics which would then serve as the inputs to a sound spatialization system. Thus, initially at least, source generation and spatial synthesis would remain as functionally separate components of an integrated acoustic display system. While there would necessarily be some overhead cost in controlling separate devices, the advantage is that each component can be developed, upgraded and utilized as standalone components so that systems are not locked into an outmoded technology.

Current devices available for generating nonspeech sound sources tend to fall into two general categories; "samplers", which digitally store sounds for later real-time playback, and "synthesizers", which rely on analytical or algorithmically-based sound generation techniques originally developed for imitating musical instruments (see Scaletti in [31]). With samplers, many different sounds can be reproduced (nearly) exactly, but substantial effort and storage media are required for accurately pre-recording sounds and there is usually limited real-time control of acoustic parameters. Synthesizers, on the other hand, afford a fair degree of real-time, computer-driven control.

Most widely available synthesizers and samplers are based on MIDI (Musical Instrument Digital Interface) technology. The baud-rate of such devices (31.25 Kbs), especially when connected to standard serial computer lines (19.2 Kbs), is still low enough that continuous real-time control of multiple sources/voices will frequently "choke" the system. In general, synthesis-based MIDI devices such as those which use frequency modulation (FM), are more flexible than samplers in the type of real-time control available but less general in terms of the variety of sound qualities that can be reproduced. For example, it is difficult to generate environmental sounds such as breaking or bouncing objects from an FM synthesizer (see Gaver in [31]).

Large-scale systems designed for sound production and control in the entertainment industry or in music composition incorporate both sampling and digital synthesis techniques and are much more powerful. However, they are also quite expensive, require specialized knowledge for their use, and are primarily designed for off-line sound design and post-production . A potential disadvantage of both types of devices is that they are primarily designed with musical performance and/or sound effects in mind. This design emphasis is not necessarily well-suited to the generation and control of sounds for the display of information, and again, tends to require that the user/designer have specialized knowledge of musical/production techniques.

The most general systems would be based on synthesis via physical or acoustical models of sound source characteristics. A simpler but less versatile approach would use playback of sampled sounds and/or conventional MIDI devices as in most current systems. Since very general physical models are both difficult (perhaps impossible) to develop as well as computationally-intensive, a more practical and immediately achievable system might be a hybrid approach that uses techniques like real-time manipulation of simple parameters, such as the pitch, filter bandwidth, or intensity, of sampled sounds, and real-time interpolation between sampled sounds, analogous to "morphing" in computer graphics. The E-mu Morpheus synthesizer is an example of this kind of approach. Recently, several commercial synthesizer companies have announced new products based on physical modeling techniques. A soundcard being developed by MediaVision is based on digital waveguides; [56] the Yameha VL1 keyboard synthesizer is based on an unspecified physical modeling approach; and the Macintosh-based, Korg SynthKit allows "construction" of sounds via interconnection of a visual programming language composed of modular units representing hammer-strikes, bows, reeds, etc.

A few cue-generation systems have been integrated for virtual environments and data sonification using currently-available devices, [57, 64] while a few designers are developing special-purpose hardware and software systems for acoustic displays. [30, 31, 53, 65] However, far more effort needs to be devoted to the development of sound-generation technology specifically aimed at information display. Perhaps even more critical is the need for further research into lower-level sensory and higher-level cognitive determinants of acoustic perceptual organization, [9, 31] since these results should serve to guide technology development.

Conclusions

While much work remains to be done in the area of auditory spatial synthesis, the basic technology needed for adding at least minimal spatial cues to auditory displays is now available. Admittedly, this currently comes at a fairly high cost, although this is bound to become less of a factor as digital signal processing technology becomes cheaper in the near future. What is more critical for the future success of virtual acoustic displays is a systematic approach to understanding the perceptual and practical contraints involved in developing the real-time technology needed for generating source characteristics in as general a manner as possible. In the near term, this probably means continuing to use approaches based on standard synthesis and sampling techniques, or some hybrid version of both. In the long term, researchers must begin to think about instantiating more complex approaches to simulation; for example, using techniques based on physical modeling. [27, 31, 33, 40] In parallel with the development of a sonification infrastructure, it is also critical to conduct formal validation studies of the systems we develop in the context of real-world tasks.

Acknowledgements

Work supported by NASA and by the Naval Command, Control and Ocean Surveillance Center, San Diego. Some of the material presented here was based on the chapter, "Spatial Sound and Sonification," by E. M. Wenzel that recently appeared in the book edited by Kramer.[31] Thanks also to my colleague, Durand Begault, for his valuable input to the paper.

References

1. Allen, J. B., and Berkley, D. A. "Image Model for Efficiently Modeling Small-Room Acoustics." *Journal of the Acoustical Society of America*, **65** (1979): 943-950.

2. Asano, F., Suzuki, Y., & Sone, T. Role of spectral cues in median plane localization. *J. Acoust. Soc. Am.*, 88 (1990), 159-168.

3. Begault, D. R. "Perceptual Effects of Synthetic Reverberation on Three-Dimensional Audio Systems." *Journal of the Audio Engineering Society,* **40** (1992): 895-904.

4. Begault, D. R. "Head-Up Auditory Displays for Traffic Collision Avoidance System Advisories: A Preliminary Investigation." *Human Factors,* **35** (1993): 707-717..

5. Begault, D. R. "Intelligibility Improvement for Call Signs Using a Spatial Auditory Display." *NASA Technical Memorandum, TM104014,* 1993.

6. Begault, D. R., Stein, N., and Loesche, V. "Advanced Audio Applications in the NASA-Ames Advanced Flight Simulator." (unpublished manuscript): Contact: D. R. Begault, NASA-Ames Research Center, Mail Stop 262-2, Moffett Field, CA 94035-1000.

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

8. Blauert, J. *Spatial Hearing: The Psychophysics of Human Sound Localization.* Cambridge, MA: MIT Press, 1983.

9. Bregman, A. S. *Auditory Scene Analysis.* Cambridge, MA: MIT Press, 1990.

10. Bronkhorst, A. W., and Plomp, R. "The Effect of Head-Induced Interaural Time and Level Differences on Speech Intelligibility in Noise." *Journal of the Acoustical Society of America,* **83** (1988): 1508-1516.

11. Brooks, F. P. "Grasping Reality Through Illusion -- Interactive Graphics Serving Science." *Proceedings of CHI'88, ACM Conference on Human Factors in Computing Systems,* (1988): 1-11.

12. Bryson, S. and Levit, C. "The Virtual Wind Tunnel." *IEEE Computer Graohics and Applications,* **12** (1992): 25-34.

13. Chan, C. J. "Sound Localization and Spatial Enhancement with the Roland Sound Space Processor." In *Cyberarts: Exploring Art and Technology,* edited by L. Jacobson, 95-104, San Francisco, CA: Miller Freeman, Inc, 1992.

14. Cherry, E. C. "Some Experiments on the Recognition of Speech with One and Two Ears." *Journal of the Acoustical Society of America,* **22** (1953): 61-62.

15. Colquhoun, W. P. "Evaluation of Auditory, Visual, and Dual-Mode Displays for Prolonged Sonar Monitoring in Repeated Sessions." *Human Factors,* **17** (1975): 425-437.

16. Deatherage, B. H. Auditory and other sensory forms of information presentation. In H. P. Van Cott & R. G. Kincade (Eds.), *Human Engineering Guide to Equipment Design* (rev. ed.). Washington, DC: U.S. Government Printing Office, (1972): 123-160.

17. Doll, T.J., Gerth, J.M., Engelman, W.R. & Folds, D.J. *Development of simulated directional audio for cockpit applications.* USAF Report No. AAMRL-TR-86-014, 1986.

18. Edwards, A. D. N. "Soundtrack: An Auditory Interface for Blind Users." *Human-Computer Interaction,* **4** (1989): 45-66.

19. Fisher, S. S., Wenzel, E. M., Coler, C., and McGreevy, M. W. "Virtual Interface Environment Workstations." *Proceedings of the Human Factors Society,* **32** (1988): 91-95.

20. Foster, S. H. *Convolvotron^{TM} User's Manual.* Crystal River Engineering, Inc., 12350 Wards Ferry Road, Groveland, CA 95321, 1988.

21. Foster, S. H., Wenzel, E. M., and Taylor, R. M. "Real-Time Synthesis of Complex Acoustic Environments." *ASSP (IEEE) Workshop on Applications of Signal Processing to Audio and Acoustics,* New Paltz, NY (1991).

22. Gardner, M. B., and Gardner, R. S. "Problem of Localization in the Median Plane: Effect of Pinnae Cavity Occlusion." *Journal of the Acoustical Society of America,* **53** (1973): 400-408.

23. Gaver, W. W., Smith, R. B., and O'Shea, T. "Effective Sounds in Complex Systems: The ARKola Simulation." *Proceedings of CHI'91, ACM Conference on Human Factors in Computing Systems,* (1991): 85-90.

24. Gehring, B. *Focal Point^{TM} 3D Sound User's Manual.* Gehring Research Corporation, 189 Madison Avenue, Toronto, Canada, M5R 2S6, 1990.

25. Gierlich, H. W. "The Application of Binaural Technology." *Applied Acoustics,* **36** (1992): 219-244.

26. HEAD Acoustics. *Binaural Mixing Console.* [product literature.] Contact: Sonic Perceptions, 114A Washington St., Norwalk, CT 06854

27. Jaffe, D., and Smith, J. Extensions of the Karplus-Strong Plucked String Algorithm. *Computer Music Journal,* **7** (1983): 43-55.

28. Kistler, D. K., and Wightman, F. L. "A Model of Head-Related Transfer Functions Based on Principal Components Analysis and Minimum-Phase Reconstruction." *Journal of the Acoustical Society of America,* **91** (1992): 1637-1647.

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

30. Kramer, G., and Ellison, S. "Audification: The Use of Sound to Display Multivariate Data", *Proceedings of the International Computer Music Conference,* (1991): 214-221.

31. Kramer, G. (Ed.) *Auditory Display: Sonification, Audification, and Auditory Interfaces.* Proceedings Volume XVIII, Santa Fe Institute Studies in the Sciences of Complexity, Reading MA: Addison-Wesley, 1994.

32. Lehnart, H., and Blauert, J. "Principles of Binaural Room Simulation." *Applied Acoustics,* **36** (1992): 259-292.

33. Li, X., Logan, R. J., and Pastore, R. E. "Perception of Acoustic Source Characteristics." *Journal of the Acoustical Society of America,* **90** (1991): 3036-3049.

34. Loomis, J. M., Hebert, C., and Cicinelli, J. G. "Active Localization of Virtual Sounds." *Journal of the Acoustical Society of America,* **88** (1990): 1757-1764.

35. Ludwig, L., Pincever, N., and Cohen, M. "Extending the Notion of a Window System to Audio." *Computer,* **23** (1990): 66-72.

36. McKinley, R. L., and Ericson, M. A. "Digital Synthesis of Binaural Auditory Localization Azimuth Cues Using Headphones." *Journal of the Acoustical Society of America,* **88** (1988): S18.

37. Mershon, D. H. & King, L. E. Intensity and reverberation as factors in the auditory perception of egocentric distance. *Perception & Psychophysics,* **18** (1975): 409-415.

38. Mills, A. W. Auditory localization. In J.V. Tobias (Ed.), *Foundations of Modern Auditory Theory, Vol. II* (pp. 301-345), New York: Academic Press, 1972.

39. Minsky, M., Ming, O., Steele, O., Brooks, F. P., and Behensky, M. "Feeling and Seeing: Issues in Force Display." *Computer Graphics,* **24** (1990): 235-243.

40. Morse, P. M., and Ingard, K. U. *Theoretical Acoustics.* New York: McGraw-Hill, 1968.

41. Mowbray, G. H., and Gebhard, J. W. "Man's Senses as Informational Channels." In H.W. Sinaiko (Ed.), *Human Factors in the Design and Use of Control Systems,* edited by H.W. Sinaiko, 115-149. New York: Dover Publications, 1961.

42. Patterson, R. R. "Guidelines for Auditory Warning Systems on Civil Aircraft." Paper No. 82017, London: Civil Aviation Authority, 1982.

43. Perrott, D. R. Studies in the perception of auditory motion. In R.W. Gatehouse (Ed.), *Localization of Sound: Theory and Applications* (pp. 169-193). Groton, CN: Amphora Press, 1982 .

44. Perrott, D. R. Concurrent minimum audible angle: A re-examination of the concept of auditory spatial acuity. *Journal of the Acoustical Society of America,* **75** (1984a): 1201-1206.

45. Perrott, D. R. Discrimination of the spatial distribution of concurrently active sound sources: Some experiments with stereophonic arrays. *Journal of the Acoustical Society of America,* **76** (1984b): 1704-1712.

46. Perrott, D. R. & Tucker, J. Minimum audible movement angle as a function of signal frequency and the velocity of the source. *Journal of the Acoustical Society of America,* **83** (1988): 1522-1527.

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

48. Persterer, A. "A Very High Performance Digital Audio Signal Processing System." *ASSP (IEEE) Workshop on Applications of Signal Processing to Audio and Acoustics,* New Paltz, NY (1989).

49. Plenge, G. "On the Difference Between Localization and Lateralization." *Journal of the Acoustical Society of America,* **56** (1974): 944-951.

50. Poesselt, C., Schroeter, J., Opitz, M., Divenyi, P., and Blauert, J. "Generation of Binaural Signals for Research and Home Entertainment." Paper B1-6, *Proceedings of the 12th International Congress on Acoustics, Toronto,* 1986.

51. Lord Rayleigh [Strutt, J. W.] "On Our Perception of Sound Direction. *Philosophical Magazine,* **13** (1907): 214-232.

52. Richter, F., and Persterer, A. "Design and Applications of a Creative Audio Processor. Preprint 2782 (U-4). *86th Convention of the Audio Engineering Society, Hamburg,* 1989.

53. Scaletti, C., and Craig, A. B. "Using Sound to Extract Meaning from Complex Data." *Proceedings of the SPIE, San Jose, CA,* **1459** (1991): 207-219.

54. Schroeder, M. R. "Digital Simulation of Sound Transmission in Reverberant Spaces." *Journal of the Acoustical Society of America,* **47** (1970): 424-431.

55. Shaw, E. A. G. "The External Ear." *Handbook of Sensory Physiology, Vol. V/1, Auditory System,* edited by W.D. Keidel and W.D. Neff, 455-490. New York: Springer-Verlag, 1974.

56. Smith, J. O. (1992) Physical modeling using digital waveguides, *Computer Music Journal,* **16** (1992): 74-98.

57. Smith, S., Bergeron. R. D., and Grinstein, G. G. "Stereophonic and Surface Sound Generation for Exploratory Data Analysis." *Proceedings of CHI'91, ACM Conference on Human Factors in Computing Systems,* (1991): 125-132.

58. Warren, D. H., Welch, R. B., and McCarthy, T. J. "The Role of Visual-Auditory "Compellingness" in the Ventriloquism Effect: Implications for Transitivity Among the Spatial Senses." *Perception and Psychophysics,* **30** (1981): 557-564.

59. Welch, R. B. *Perceptual modification: Adapting to altered sensory environments.* New York: Academic Press, 1978.

60. Wenzel, E.M., Wightman, F.L., & Foster, S.H. A virtual display system for conveying three-dimensional acoustic information. *Proc. Hum. Fac. Soc.,* 32 (1988), 86-90.

61. Wenzel, E. M. "Localization in Virtual Acoustic Displays." *Presence: Teleoperators and Virtual Environments,* **1** (1992): 80-107.

62. Wenzel, E. M., Arruda, M., Kistler, D, J., and Wightman, F. L. "Localization of Non-Individualized Head-Related Transfer Functions." *Journal of the Acoustical Society of America,* **94** (1993): 111-123.

63. Wenzel, E. M., and Foster, S. H. "Perceptual Consequences of Interpolating Head-Related Transfer Functions During Spatial Synthesis." *Proceedings of the ASSP (IEEE) Workshop on Applications of Signal Processing to Audio and Acoustics,* New Paltz, NY, Oct. 17-20 (1993).

64. Wenzel, E. M., Stone, P. K., Fisher, S. S., and Foster, S. H. "A System for Three-Dimensional Acoustic "Visualization" in a Virtual Environment Workstation." *Proceedings of the IEEE Visualization '90 Conference, San Francisco,* (1990): 329-337.

65. Wenzel, E. M., Gaver, W. W., Foster, S. H., Levkowitz, H., and Powell, R. "Perceptual vs. Hardware Performance in Advanced Acoustic Interface Design." *Proceedings of INTERCHI'93, ACM Conference on Human Factors in Computing Systems,* Amsterdam (1993): 363-366.

66. Wightman, F. L., and Kistler, D. J. "Headphone Simulation of Free-Field Listening I: Stimulus Synthesis. *Journal of the Acoustical Society of America,* **85** (1989): 858-867.

67. Wightman, F. L., and Kistler, D. J. "Headphone Simulation of Free-Field Listening II: Psychophysical Validation. *Journal of the Acoustical Society of America,* **85** (1989): 868-878.

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