Elizabeth M. Wenzel
Human Interface Research Branch,
Aerospace Human Factors Research Division,
NASA Ames Research Center, MS 262-2,
Moffett Field, California 94035

Localization in Virtual Acoustic Displays

Abstract

This paper discusses the development of a particular spatial display medium, the virtual acoustic display. Although the technology can stand alone, it is envisioned ultimately to be a component of a larger multisensory environment and will no doubt find its greatest utility in that context. A general philosophy of the project has been that the development of advanced computer interfaces should be driven first by an understanding of human perceptual requirements, and secondarily by technological capabilities or constraints. In expanding on this view, the paper addresses why virtual acoustic displays are useful, characterizes the abilities of such displays, reviews some recent approaches to their implementation and application, describes the research project at NASA Ames in some detail, and finally outlines some critical research issues for the future.

1 Why Virtual Acoustic Displays?

The recent burgeoning of computing technology requires that people learn to interpret increasingly complex systems of information and control increasingly complex machines. One approach to this problem has been to develop the direct-manipulation, graphic computer interfaces exemplified by the ubiquitous combination of the desktop metaphor and the mouse. Such spatially organized interfaces can provide familiarity and consistency across applications, thus avoiding much of the task-dependent learning of the older text-oriented displays. Lately, a considerable amount of attention has been devoted to a more ambitious type of reconfigurable interface called the virtual display. Despite the oft-touted "revolutionary" nature of this field, the research has many antecedents in previous work in three-dimensional computer graphics, interactive input/output devices, and simulation technology. Some of the earliest work in virtual interfaces was done by Sutherland (1968) using binocular head-mounted displays. Sutherland characterized the goal of virtual interface research, stating that, "The screen is a window through which one sees a virtual world. The challenge is to make that world look real, act real, sound real, feel real." As technology has advanced, virtual displays have adopted a three-dimensional spatial organization, intending to provide a more natural means of accessing and manipulating information. A few projects have taken the spatial metaphor to its limit by directly involving the operator in a data environment (e.g., Furness, 1986; Brooks, 1988; Fisher, Wenzel, Coler, & McGreevy, 1988). For example, Brooks (1988) and his colleagues have worked on a three-dimensional interface in which a chemist can visually and manually interact with a virtual model of a drug compound, attempting to discover the bonding site of a molecule by literally seeing and feeling the interplay of the chemical

Meta
Exhibit 1158
Meta v. 8KHZ
IPR2023-01023
Ex. 1158-001

Downloaded from http://direct.mit.edu/pvar/article-pdf/1/1/80/1622491/pres.1992.1.1.80.pdf by UNIVERSITY OF CENTRAL FLORIDA, GREG WELCH on 04 August 2024

forces at work. It seems that the kind of "artificial reality" once relegated solely to the specialized world of the cockpit simulator is now being seen as the next step in interface development for many types of advanced computing applications (Foley, 1987).

Often the only modalities available for interacting with complex information systems have been visual and manual. Many investigators, however, have pointed out the importance of the auditory system as an alternative or supplementary information channel (e.g., Garner, 1949; Deatherage, 1972; Doll, Geth, Eugelman, & Folds, 1986). Most recently, attention has been devoted to the use of nonspeech audio as an interface medium (Patterson, 1982; Gaver, 1986; Begault & Wenzel, 1992; Blattner, Sumikawa, & Greenberg, 1989; Buxton, Gaver, & Bly, 1989). Useful features of acoustic signals include the fact that they can be heard simultaneously in three dimensions, they tend to produce an alerting or orienting response, and they can be detected more quickly than visual signals (Mowbray & Gebhard, 1961; Patterson, 1982). These 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 (Mowbray & Gebhard, 1961; Kubovy, 1981). 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. Thus audio is particularly suited to monitoring state changes over time, for example, when a car engine suddenly begins to malfunction.

Nonspeech signals have the potential to provide an even richer display medium if they are carefully designed with human perceptual abilities in mind. Just as a movie with sound is much more compelling and informationally rich than a silent film, so could a computer interface be enhanced by an appropriate "sound track" to the task at hand. If used properly, sound need not be distracting or cacophonous or merely uninformative. Principles of design for auditory icons and auditory symbologies can be gleaned from the fields of music (Deutsch, 1982; Blattner et al., 1989), psychoacoustics (Carterette & Friedman, 1978; Patterson, 1982), and higher level cognitive studies of the acoustical determinants of perceptual organization (Bregman, 1981; 1990; Kubovy, 1981; Buxton et al., 1989). For example, following from Gibson's (1979) ecological approach to perception, one can conceive of the audible world as a collection of acoustic "objects." Various acoustic features, such as temporal onsets and offsets, timbre, pitch, intensity, and rhythm, can specify the identities of the objects and convey meaning about discrete events or ongoing actions in the world and their relationships to one another. One could systematically manipulate these features, effectively creating an auditory symbology that operates on a continuum from "literal" everyday sounds, such as the clunk of mail in your mailbox (e.g., Gaver's "Sonic Finder," 1986), to a completely abstract mapping of statistical data into sound parameters (Bly, 1982; Smith, Bergeron, & Grinstein, 1990; Blattner et al., 1989).

An acoustic display could be further enhanced by taking advantage of the auditory system's ability to segregate, monitor, and switch attention among simultaneous streams of sound (Mowbray & Gebhard, 1961). One of the most important determinants of acoustic segregation is an object's location in space (Kubovy & Howard 1976; Bregman, 1981, 1990; Deutsch, 1982).

Such a three-dimensional auditory display can potentially enhance information transfer by combining directional with iconic information in a quite naturalistic representation of dynamic objects in the interface. Borrowing a term from Gaver (1986), an obvious aspect of "everyday listening" is the fact that we live and listen in a three-dimensional world. A primary advantage of the auditory system is that it allows us to monitor and identify sources of information from all possible locations, not just the direction of gaze. In fact, I would like to suggest that a good rule of thumb for knowing when to provide acoustic cues is to recall how we naturally use audition to gain information and explore the environment; that is, "the function of the ears is to point the eyes." Thus the auditory system can provide a more coarsely tuned mechanism to direct the attention of our more finely tuned visual analyses. For example, Perrott, Sadralodabai, Saberi, and Strybel (1991) have recently

Downloaded from http://direct.mit.edu/pvar/article-pdf/1/1/80/1622491/pres.1992.1.1.80.pdf by UNIVERSITY OF CENTRAL FLORIDA, GREG WELCH on 04 August 2024

reported that aurally guided visual search for a target in a cluttered visual display is superior to unaided visual search, even for objects in the central visual field. This omnidirectional characteristic of acoustic signals will be especially useful in inherently spatial tasks, particularly when visual cues are limited and workload is high, as in air traffic control (ATC) displays for the tower or cockpit (Begault & Wenzel, 1992). ATC controllers are being asked to integrate increasingly heavy air traffic into increasingly complex landing patterns, such as the triple parallel approach proposed to maximize the flow of incoming aircraft. Research at NASA Ames, in collaboration with the Federal Aviation Administration, will emphasize two types of acoustic displays because of their conceptual simplicity and the likelihood that they will provide significant benefits to current ATC systems. One example is an ATC display in which the controller hears communications from incoming traffic in positions that correspond to their actual location in the terminal area. In such a display, it should be more immediately obvious to the listener when aircraft are on a potential collision course because they would be heard in their true spatial locations and their routes could be tracked over time. A second example involves alerting systems for ATC. An auditory icon, such as a complex signal with a unique temporal rhythm, could be used as a warning of urgent situations like potential runway incursions. Again, the signal could be processed to convey true directional information and urgency could be emphasized by placing the warning close to the listener's head, for example, within the boundaries of their "personal space" (Begault & Wenzel, 1992).

A related advantage of the binaural system, often referred to as the "cocktail party effect," is that the spatial separation of sounds improves the intelligibility of signals in noise and assists in the segregation of multiple sound streams (Cherry, 1953; Bronkhorst & Plomp, 1988; Bregman, 1990). Segregation enhancement can be critical in applications involving both simultaneous speech channels, as in aviation communication systems, and the kind of encoded nonspeech cues proposed for scientific "visualization." Examples of visualization displays include the acoustic representation of multidimensional data (e.g., Bly, 1982; Blattner et al., 1989; Smith et al., 1990) and the development of alternative interfaces for the visually impaired (Edwards, 1989; Loomis, Hebert, & Cicinelli, 1990).

Another aspect of auditory spatial cues is that, in conjunction with the other senses, they can act as potentiators of information in a display. For example, visual and auditory cues together can reinforce the information content of a display and provide a greater sense of presence or realism in a manner not readily achieved by either modality alone (Colquhoun, 1975; O'Leary & Rhodes, 1984; Warren, Welch, & McCarthy, 1981). Similarly, in direct-manipulation tasks, auditory cues can provide supporting information for the representation of tactile or force-feedback cues (Wenzel, Stone, Fisher, & Foster, 1990), a quite difficult interface problem for multimodal displays that is only beginning to be solved (e.g., Minsky, Ming, Steele, Brooks, & Behensky, 1990). Intersensory synergism will be particularly important in applications involving telepresence, including advanced teleconferencing (Ludwig, Pincaver, & Cohen, 1990), shared electronic workspaces (Fisher et al., 1988; Gaver, Smith, & O'Shea, 1991), and monitoring telerobotic activities in remote or hazardous situations (Wenzel et al., 1990). Similarly, the interaction of the senses will be critical in purely virtual environments for visualization and systems control (Brooks, 1988; Fisher et al., 1988), entertainment (Kendall & Martens, 1984; Kendall & Wilde, 1989), and architectural acoustics (Persterer, 1989; Foster & Wenzel, 1991; Foster, Wenzel, & Taylor, 1991).

The combination of veridical[1] spatial cues with good principles of iconic design could provide an extremely powerful and information-rich display that is also quite easy to use. As noted above, the construction of meaningful acoustic objects or icons involves complex issues in perception, cognition, and synthesis that are beginning to be addressed elsewhere (e.g., Bregman, 1990; Buxton et al., 1989). The remaining sections of this paper will concentrate on past and current techniques for achieving the acoustic display of spatial information per

1. Here, the term veridical is used to indicate that spatial cues are both realistic and result in the accurate transfer of spatial information, for example, the presentation of such cues results in accurate estimates of perceived location by human listeners in psychophysical studies.

Downloaded from http://direct.mit.edu/pvar/article-pdf/1/1/80/1622491/pres.1992.1.1.80.pdf by UNIVERSITY OF CENTRAL FLORIDA, GREG WELCH on 04 August 2024

se. Theoretical and technological antecedents are reviewed and the importance of perceptually validating techniques for simulating spatial cues is discussed in the context of the ongoing research at NASA Ames in virtual acoustic displays. Finally, some critical research problems for the future are outlined. Since these goals are rather ambitious, I apologize in advance for neglecting any important work or issues in an area that seems to be rapidly gaining momentum.

2 Antecedants of Three-Dimensional Virtual Acoustic Displays

The utility of a spatial auditory display depends greatly on the user's ability to localize the various sources of information in auditory space. While compromises obviously have to be made to achieve a practical system, the particular features or limitations of the latest hardware should be considered subservient to human sensory and performance requirements. Thus, designers of such interfaces must carefully consider the acoustic cues needed by listeners for accurate localization and ensure that these cues will be faithfully (or at least adequately, in a human performance sense) transduced by the synthesis device rather than letting current technology drive the implementation. In fact, knowledge about sensory requirements might actually save processing power in some cases and indicate others to which more resources should be devoted.

2.1 Psychoacoustical Antecedents

Much of the research on human sound localization has derived from the classic "duplex theory" (Lord Rayleigh, 1907) that emphasizes the role of two primary cues (Fig. 1), interaural differences in time of arrival and interaural differences in intensity. Because the theory had been based primarily on experiments with single-frequency (sinewave) sounds, the original proposal was that interaural intensity differences (IIDs) resulting from head-shadowing determine localization at high frequencies, while interaural time differences (ITDs) were thought to be important only for low frequencies because of the phase ambiguities occuring at frequencies greater than 1500 Hz. Binaural research over the last few decades, however, points to serious limitations with this approach. For example, it has become clear that ITDs in high-frequency sounds can be used if they have sufficient bandwidth to produce relatively slow modulations in their envelopes (e.g., Henning, 1974).

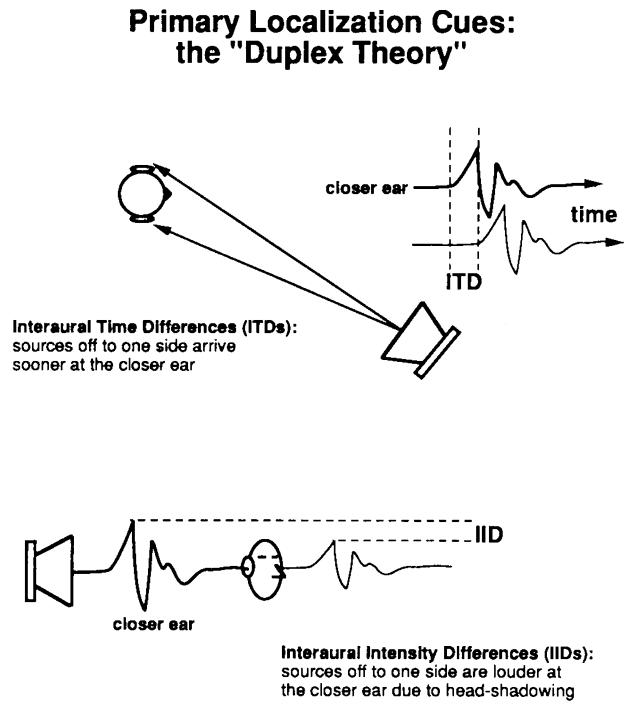

Figure 1. *Illustration of the two primary cues, interaural intensity and interaural time differences (IIDs and ITDs), postulated by the "duplex theory" of sound localization. In particular, the theory proposes that IIDs are particularly important for localization of high frequencies while ITDs are important for low frequencies.*

The duplex theory also cannot account for the ability of subjects to localize sounds on the vertical median plane where interaural cues are minimal (Blauert, 1969; Butler & Belendiuk, 1977; Oldfield & Parker, 1986). Similarly, when subjects listen to stimuli over headphones, they are perceived as being inside the head even though interaural temporal and intensity differences appropriate to an external source location are present (Plenge, 1974). Many studies now suggest that these deficiencies of the duplex theory reflect the important contribution to localization of the direction-dependent filtering that occurs when incoming sound waves inter-

Downloaded from http://direct.mit.edu/pvar/article-pdf/1/1/80/1622491/pres.1992.1.1.80.pdf by UNIVERSITY OF CENTRAL FLORIDA, GREG WELCH on 04 August 2024

act with the outer ears or pinnae. As sound propagates from a source (e.g., a loudspeaker) to a listener's ears, reflection and refraction effects tend to alter the sound in subtle ways and the effect is dependent on frequency. For example, for a particular location a group of high-frequency components centered at, say 8 kHz, may be attenuated more than a different band of components centered at 6 kHz. Such frequency-dependent effects, or filtering, vary greatly with the direction of the sound source. Thus for a different source location, the band at 6 kHz may in turn be more attenuated than the higher frequency band at 8 kHz. It is clear that listeners use these kinds of frequency-dependent effects to discriminate one location from another. Experiments have shown that spectral shaping by the pinnae is highly direction dependent (Shaw, 1974, 1975), that the absence of pinna cues degrades localization accuracy (Gardner & Gardner, 1973; Oldfield & Parker, 1984b), and that pinna cues are primarily responsible for externalization or the "outside-the-head" sensation (Plenge, 1974).

Such data suggest that perceptually veridical localization over headphones may be possible if this spectral shaping by the pinnae as well as the interaural difference cues can be adequately reproduced. There may be many cumulative effects on the sound as it makes its way to the ear drum, but it turns out that all of these effects can be expressed as a single filtering operation much like the effects of a graphic equalizer in a stereo system. The exact nature of this filter can be measured by a simple experiment in which an impulse (a single, very short sound pulse or click) is produced by a loudspeaker at a particular location.[2] The acoustic shaping by the two ears is then measured by recording the outputs of small probe microphones placed inside an individual's ear canals. If the measurement of the two ears occurs simultaneously, the responses, when taken together as a pair of filters, include an estimate of the interaural differences as well. Thus, this technique allows one to measure all of the relevant spatial cues together for a given source location, a given listener, and in a given room or environment.

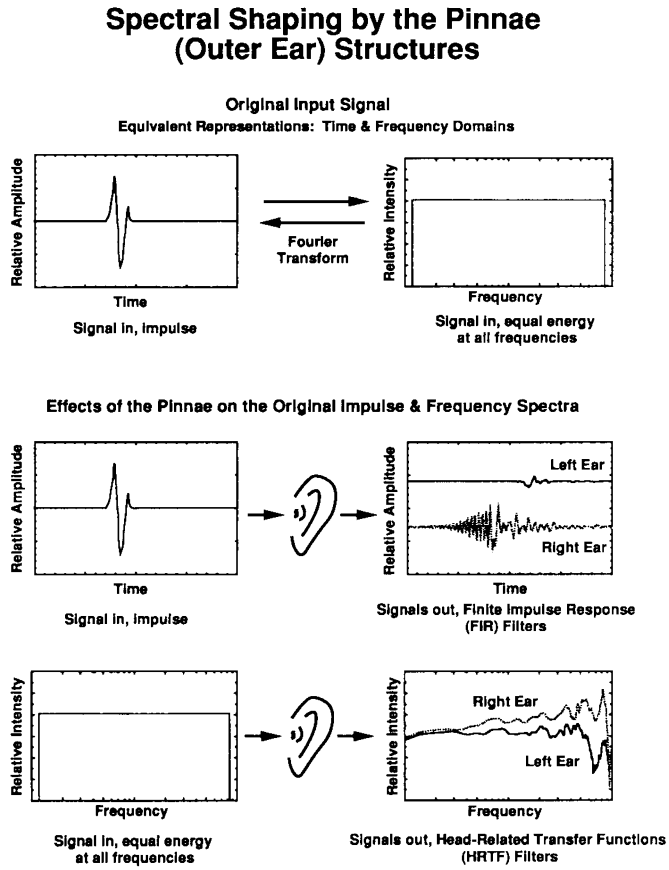

Figure 2. *Illustration of the effects of spectral shaping by the pinnae. The top panels show equivalent representations (via the Fourier transform) in the time (an impulse) and frequency (intensity only shown here) domains of a broadband acoustic signal before interaction with the outer ear structures. The middle panels show what happens to an impulse delivered from a loudspeaker directly to the right (+90° azimuth, 0° elevation) after interaction with the outer ear structures, as measured in the left (solid line) and right (dashed line) ear canals of an individual. The bottom panels show the same interaction, but represented in the frequency domain (spectral intensity only). The transfer characteristics of the measurement system as well as the headphones used during playback have been mathematically removed from the responses plotted here.*

Figure 2 illustrates these effects for the transfer functions of the ears. The top panels show equivalent representations, via a mathematical operation known as the Fourier transform, in the time (an impulse) and frequency (intensity only shown here) domains for an

2. The impulse response technique is analogous to what happens when you strike a bell with a hammer. The hammer strike (an impulse) causes energy at many frequencies to be mechanically transferred to the bell in a very short period of time. The pitch you hear out, the frequencies that are emphasized and get transferred from the bell to the air (its transfer function), depends on the physical structure and resonance properties of the bell.

Downloaded from http://direct.mit.edu/pvar/article-pdf/1/1/80/1622491/pres.1992.1.1.80.pdf by UNIVERSITY OF CENTRAL FLORIDA, GREG WELCH on 04 August 2024

acoustic signal before interaction with the outer ear and other body structures. The middle panels show what happens to an impulse delivered from a loudspeaker located directly to the right after interaction with the outer ear structures, as measured in the left (solid line) and right (dashed line) ear canals of a listener.[3] The bottom panels show the same interaction, but represented in the frequency domain (spectral intensity only). Thus, 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. 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). Here, filtering in the frequency domain is a point-by-point multiplication operation while filtering in the time domain occurs via a somewhat more complex operation known as convolution (see Brigham, 1974, for a useful pictorial discussion of filtering and convolution).

By filtering an arbitrary sound with these HRTF-based filters, it should be possible to impose spatial characteristics on the signal such that it apparently emanates from the originally measured location. Of course, the localizabilty of the sound will also depend on other factors such as its original spectral content; narrowband (pure) tones are generally hard to localize while broadband, impulsive sounds are the easiest to locate. Filtering with HRTF-based filters cannot significantly increase the bandwidth of the original signal, it merely transforms the frequency components that are already present. A closely related issue in the localizability of sound sources is their degree of familiarity. Logically, localization based on spatial cues other than the interaural cues, for example, cues related to spectral shaping by the pinnae, must be largely determined by a listener's a priori knowledge of the spectrum of the sound source. The listener must "know" what the spectrum of a sound is to begin with in order to determine that the same sound at different positions has been differentially "shaped" by the effects of his/her ear structures. Thus both the perception of elevation and relative distance, which depend heavily on the detection of spectral differences, tend to be superior for familiar signals like speech (e.g., Plenge & Brunschen, 1971, in Blauert, 1983, p. 104; Coleman, 1963). Similarly, spectral familiarity can be established through training (Batteau, 1967).

It should be noted that the spatial cues provided by HRTFs, especially those derived from simple anechoic (free-field or echoless) environments, are not the only cues likely to be necessary to achieve veridical localization in a virtual display. Anechoic simulation is merely a first step, allowing a systematic study of the technological requirements and perceptual consequences of synthesizing spatial cues by using a less complex, and therefore more tractible, stimulus. For example, two kinds of error are usually observed in perceptual studies of localization when subjects are asked to judge the position of a static sound source in the free-field. One, which Blauert (1983) refers to as localization blur, is a relatively small error in resolution on the order of about 5 to 20°. Another class of error observed in nearly all localization studies is the occurrence of front-back "reversals." These are judgments that indicate that a source in the front hemisphere, usually near the median plane, was perceived by the listener as if it were in the rear hemisphere. Occasionally, back-to-front confusions are also found (e.g., Oldfield & Parker, 1984a). Recently, we have also observed confusions in elevation, with up locations heard as down, and vice versa (Wenzel, Wightman, & Kistler, 1991).

Although the reason for such reversals is not completely understood, they are probably due in large part to the static nature of the stimulus and the ambiguities resulting from the so-called cone of confusion (Mills, 1972). Assuming a stationary, spherical model of the head and symmetrically located ear canals (without pinnae), a given interaural time or intensity difference will correlate ambiguously with the direction of a sound

3. In addition to distance, azimuth and elevation are the spatial coordinates used to define a sound source's location in space. Azimuth can be thought of as the left–right dimension analogous to longitude on a globe and elevation as the up–down dimension analogous to latitude. These coordinates are usually presented in degrees, with 0° azimuth and 0° elevation defined as directly in front at a listener's ear level. Here, 180 is directly behind in azimuth, −90 is directly left, and +90 is directly right. In elevation, +90 is directly above and −90 is directly below the listener.

Downloaded from http://direct.mit.edu/pvar/article-pdf/1/1/80/1622491/pres.1992.1.1.80.pdf by UNIVERSITY OF CENTRAL FLORIDA, GREG WELCH on 04 August 2024

source, with a conical shell describing the locus of all possible sources (Fig. 3). Intersection of these conical surfaces with the surface of a sphere results in circular projections corresponding to contours of constant ITD or IID (i.e., considering sources at an arbitrary fixed distance). Such projections, known as iso-ITD or iso-IID contours, increase in magnitude with increasing azimuth. Recent studies have attempted to measure the actual values of ITDs and IIDs as a function of signal frequency and position. Obviously, the situation is more complicated than that portrayed in Figure 3; the head is not really a simple sphere with two symmetric holes. However, to a first approximation, the model does seem to predict iso-IDT contours for static sources. Kuhn (1977), for example, has shown that interaural delays can be predicted from the rigid sphere model for frequencies below 4 kHz, while Middlebrooks and Green (1990) observed a similar phenomenon for interaural envelope delays in signals bandpassed between 3 and 16 kHz. The situation for IIDs appears to be more complex. Middlebrooks, Makous, and Green (1989) observed iso-IID contours that increased montonically with increasing azimuth for frequencies below 8 kHz, but for higher frequencies, the regions of constant IID were dependent on both the azimuth and elevation of the source in a complicated manner. Since spectral variations at higher frequencies are known to be critical for elevation perception (Gardner, 1973), this is not particularly surprising. While the rigid sphere model is not the whole story, the observed pattern of iso-IDT and iso-IID contours indicates that the interaural characteristics of the stimulus are inherently ambiguous. In the absence of other cues, both front-back and up-down reversals would appear to be quite likely.

Several cues are thought to help in disambiguating the cones of confusion. One is the complex spectral shaping provided by the HRTFs as a function of location that was described above. For example, presumably because of the orientation and shell-like structure of the pinnae, high-frequencies tend to be more attenuated for sources in the rear than for sources in the front (e.g., see Blauert's, 1983, discussion of "boosted bands," pp. 107–116). For the case of static sounds, such cues would essentially be the only clue to disambiguating source location. With dynamic stimuli, however, the situation improves greatly. A variety of studies have shown that allowing listeners to move their heads substantially improves localization ability and can almost completely eliminate reversals (e.g., Wallach, 1939, 1940; Thurlow, Mangels, & Runge, 1967; Thurlow & Runge, 1967; Fisher & Freedman, 1968). With head motion, the listener can apparently disambiguate front-back locations by tracking changes in the magnitude of the interaural cues over time; for a given lateral head movement, ITDs

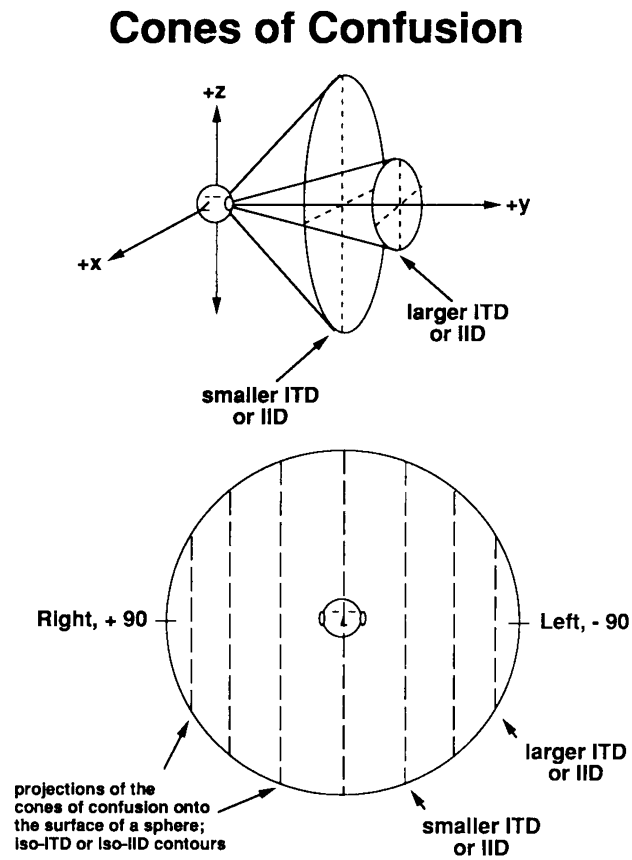

Figure 3. *Illustration of the cone-of-confusion effect for different interaural delays and intensities. Assuming a spherical head and symmetrically located ear canals (without pinnae), all sound sources lying along a conical surface would produce the same interaural time difference (ITD) and interaural intensity difference (IID). Intersection of these conical surfaces with the surface of a sphere results in circular projections corresponding to contours of constant ITD or IID. Such projections, shown in two dimensions in the drawing, are known as iso-ITD or iso-IID contours and increase in magnitude with increasing azimuth.*

Downloaded from http://direct.mit.edu/pvar/article-pdf/1/1/80/1622491/pres.1992.1.1.80.pdf by UNIVERSITY OF CENTRAL FLORIDA, GREG WELCH on 04 August 2024

and IIDs for sources in the front will change in the opposite direction compared to sources in the rear (Wallach, 1939, 1940). Time-varying cues provided by moving sources may also aid in disambiguation, particularly if there is a priori knowledge about the direction of motion, although relatively little research has been done on the topic of source motion.

Another type of localization error is known as in-head localization or IHL. That is, sometimes sources fail to externalize, particularly when the signals are presented over headphones, although IHL has also been observed for real sources (Toole, 1969; Plenge, 1974). The tendency to localize sound sources inside the head is increased if the signals are unfamiliar (Coleman, 1963; Gardner, 1968) or derived from an anechoic environment (Plenge, 1974). Thus the use of familiar signals in the presence of cues that provide a sense of distance and environmental context, such as the ratio of direct to reflected energy and other characteristics specific to enclosed spaces, may help to enhance the externalization of images (Coleman, 1963; Gardner, 1968; Laws, 1972, 1973; Plenge, 1974; Borish, 1984; Begault, 1987). There is some possibility that head motion may also be a factor in externalization. In a rather ingenious experiment, Wallach (1939, 1940) artificially controlled the relationship between head position and the particular loudspeaker producing a sound by means of a switching system. As subjects turned their heads, the signal was always switched to the loudspeaker directly in front of the listener so that interaural cues remained constant as a function of head motion. Consequently, all of the subjects heard the sound source as if it were directly above, that is, at a location where interaural cues would not normally change with lateral head motion. Constant interaural cues would also be predicted for any location along the vertical axis through a listener's head, that is, for internalized source images as well as for sources above or below the listener. Thus, a lack of dynamic interaural cues correlated with head motion could also be interpreted as IHL, especially if pinna cues are weak or unavailable. (Wallach also reports that cues due to head motion tend to dominate pinna cues when the two are put in conflict.) To my knowledge, however, this particular instance of IHL has not been reported in the literature.

Whether distance, the third dimension in a virtual acoustic display, can be reliably controlled beyond mere externalization is more problematic. It appears that humans are rather poor at judging the absolute distance of sound sources and relatively little is known about the parameters that determine distance perception (Coleman, 1963; Laws, 1972). Distance judgments depend at least partially on the relative intensities of sound sources, but the relationship is not a straightforward correspondence to the physical roll-off of intensity with distance (the inverse-square law). For example, as noted above, it also depends heavily on factors such as stimulus familiarity.

The addition of environmental effects can complicate the perception of location in other ways. Blauert (1983) reports that the spatial image of a sound source grows larger and increasingly diffuse with increasing distance in a reverberant environment, a phenomenon that may tend to interfere with the ability to judge the direction of the source. This problem may be mitigated by the phenomemon known as precedence (Wallach, Newman, & Rosenzweig, 1949). In precedence, or the "law of the first wavefront," the perceived location of a sound tends to be dominated by the direction of incidence of the original source even though later reflections could conceivably be interpreted as additional sources in different locations. The impact of the precedence effect is reduced by factors that strengthen the role of the succeeding wavefronts. For example, large enclosed spaces with highly reflective surfaces can result in reflections that are both intense enough and delayed enough (i.e., echoes) to act as "new" sound sources that can confuse the apparent direction of the original source.

The above discussion of possible influences on the perception of localized sound sources is by no means exhaustive. It is meant primarily to give a sense of the potential complexities involved in any attempt to synthesize both accurate and realistic spatial cues in a virtual acoustic display. For a much more extensive discussion of spatial sound, the interested reader is referred to the in-depth review by Blauert (1983).

Downloaded from http://direct.mit.edu/pvar/article-pdf/1/1/80/1622491/pres.1992.1.1.80.pdf by UNIVERSITY OF CENTRAL FLORIDA, GREG WELCH on 04 August 2024

2.2 Approaches to Implementation

Prior to the development of current techniques for synthesizing out-of-head localization, there were some early attempts at creating what we might now call a spatial auditory display. One of these was the rather elaborate pseudophone apparatus (Fig. 4) used during World War I for detecting and locating enemy aircraft. It is an early example of a spatial display for telepresence that actually attempted to enhance localization cues by using large artificial, directional pinnae and an expanded interaural axis. An early example of a simple virtual acoustic display is the FLYBAR system (FLYing By Auditory Reference) developed by Forbes (1946) just after World War II. Rather than transducing and transforming real world sources, this display used only crude left/right intensity panning along with pitch and temporal pattern changes to represent turn, bank, and air speed in a symbolic acoustic display for instrument flying.

Much later, investigators began to think about simulating veridical auditory localization cues as a way of analyzing and enhancing the listening experience in stereo reproduction, and eventually, to display information. In general, the approaches have concentrated on various means for reproducing the effects of the HRTF described above. The specific nature and measurement of HRTFs will be considered later in more detail during the discussion of the NASA Ames project.

One class of simulation techniques derives from binaural recording and the development of normative manikins, such as the KEMAR (Knowles Electronics, Inc.) and Neumann (e.g., Hudde & Schroter, 1981) artificial heads, used for applications such as assessing concert hall acoustics or making spatially realistic recordings of music (see Blauert, 1983). Recent examples of a real-time version of this approach in information display include the work by Doll at the Georgia Institute of Technology (Doll et al., 1986) and the system developed for the Super Cockpit Project at Wright-Patterson Air Force Base (Gehring AL100; see Calhoun, Valencia, & Furness, 1987). In a situation akin to telepresence, these projects used a movable artificial head (KEMAR) to simulate moving sources and correlated head motion. The listener heard headphone signals transduced in the ears of a manikin whose orientation was mechanically coupled to that of the listener's own head.

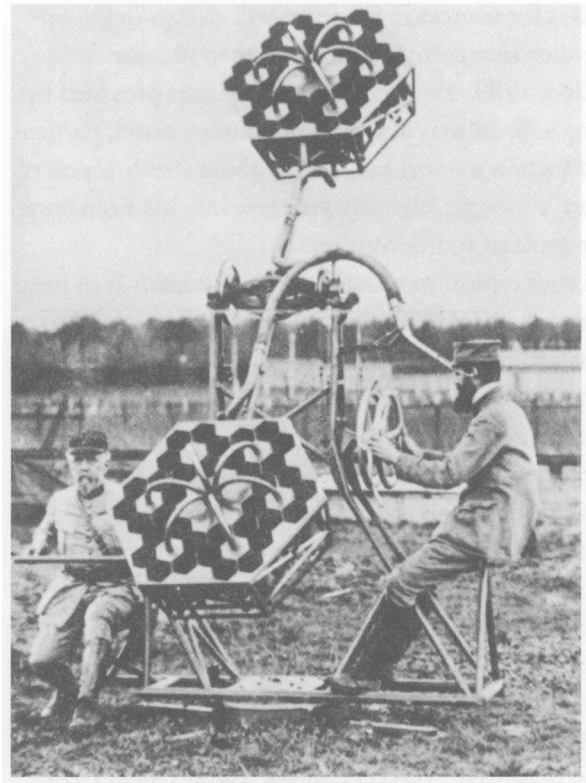

Figure 4. *Photo of the pseudophone apparatus used for detecting and localizing aircraft during World War I. From* Scientists in Power, *Spencer R. Weart, Harvard University Press: Cambridge, MA, with permission from the Neils Bohr Library, American Institute of Physics, New York.*

Another type of real-time virtual display is the work by Loomis et al. (1990) on a navigation aid for the blind. In this analog system, which worked well in an active tracking task, spatial cues were approximated using various types of simple filters with interaural time and intensity differences dynamically linked to head motion. The display also included simple distance and reverberation cues such as an intensity rolloff with distance and a fixed ratio of direct to reflected energy.

Much of the recent work since the early 1980s has been devoted to the measurement and real-time digital synthesis of HRTFs. Techniques for creating digital filters based on measurements of finite impulse responses in the ear canals of either individual subjects or artificial

Downloaded from http://direct.mit.edu/pvar/article-pdf/1/1/80/1622491/pres.1992.1.1.80.pdf by UNIVERSITY OF CENTRAL FLORIDA, GREG WELCH on 04 August 2024

heads have been under development since the late 1970s. But it is only with the advent of powerful new digital signal-processing (DSP) chips that a few real-time systems have appeared in the last few years in Europe and the United States. In general, these systems are intended for headphone delivery and use time-domain convolution to acheive real-time performance.

One example is a kind of binaural mixing console (CAP 340M Creative Audio Processor) developed by AKG in Austria and based at least partially on work by Blauert (1984; personal communication). The system is aimed at applications such as audio recording, acoustic design, and psychoacoustic research (Persterer, 1989; Richter & Persterer, 1989). This particular system is rather large, involving an entire rack of digital signal processors and related hardware, with up to 32 channels that can be independently "spatialized" in azimuth and elevation along with variable simulation of room response characteristics. Figure 5, for example, illustrates the graphic interface of the system for specifying characteristics of the binaural mix for a collection of independently positioned musical instruments. A collection of HRTFs is offered, derived from measurements taken in the ear canals of both manikins and individual subjects. A more recent system simulates an ideal control room for headphone reproduction and the user has the option of having his/her individual transforms programmed onto a PROM card (Persterer, 1991). Interestingly, AKG's literature mentions that best results are achieved with individual transforms. So far, the system has not been integrated with interactive head tracking.

Other projects in Europe derive from the substantial efforts of a group of researchers in Germany. This work includes the most recent efforts of Jens Blauert and his colleagues at the Ruhr University at Bochum (Boerger, Laws, & Blauert, 1977; Lehnert & Blauert, 1989; Posselt, Schroter, Opitz, Divenyi, & Blauert, 1986). The group at Bochum has been working on a prototype PC-based DSP system, again a kind of binaural mixing console, whose proposed features include real-time convolution of HRTFs for up to four sources, interpolation between transforms to simulate motion, and room modeling. The group has also devoted quite a bit of effort to measuring HRTFs for both individual subjects and artificial heads (e.g., the Neumann head), as well as developing computer simulations of transforms.

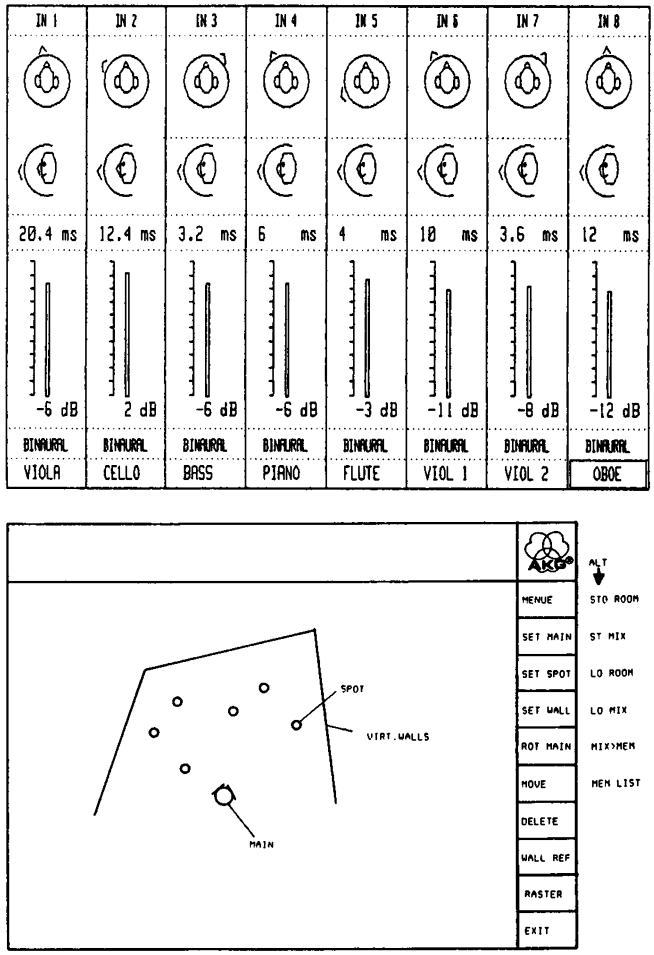

Figure 5. *Illustration of the graphical interface of AKG's Creative Audio Processor for specifying characteristics of the binaural mix for a collection of independently positioned musical instruments. Reproduced from product literature for the CAP 340M, with permission.*

Another researcher in Germany, Klaus Genuit, worked at the Institute of Technology of Aachen and later went on to found HEAD Acoustics. Genuit and his colleagues have also produced a real-time, four-channel binaural mixing console using anechoic simulations as well as a new version of an artificial head (Gierlich & Genuit, 1989). Genuit's work is particularly notable for his development of a structurally based model of the acoustic effects of the pinnae (e.g., Genuit, 1986). That is, rather than use individualized HRTFs, Genuit has developed a parameterized, mathematical description

Downloaded from http://direct.mit.edu/pvar/article-pdf/1/1/80/1622491/pres.1992.1.1.80.pdf by UNIVERSITY OF CENTRAL FLORIDA, GREG WELCH on 04 August 2024

(based on Kirchhoff's diffraction integrals) of the acoustic effects of the pinnae, ear canal resonances, torso, shoulder, and head. The effects of the structures have been simplified; for example, the outer ears are modeled as three cylinders of different diameters and length. The parameterization of the model adds some flexibility to this technique and Genuit states that the calculated transforms are within the variability of directly measured HRTFs, although no data on the perceptual viability of the model are mentioned.

In the United States, similar projects are currently in progress. For example, at Wright-Patterson Air Force Base, McKinley and Ericson (1988) developed a prototype system that synthesizes a single source in azimuth in real-time. The system uses HRTFs based on measurements from a KEMAR manikin made at 1° intervals in azimuth with a head-tracker to achieve source stabilization. Kendall's group, currently at Northwestern University, has also been working on a real-time system aimed at spatial room modeling for recording and entertainment (e.g., Kendall & Martens, 1984). Recently, Gehring (1990) has offered a software application for anechoic simulation using an off-the-shelf DSP card with one card for each sound source. So far, at least two sets of HRTFs have been demonstrated, with the filters substantially reduced in resolution (relative to the original measurements) to conform to the limitations of the DSP chip (Motorola 56001). One set was from a KEMAR manikin measured by Kendall's group and the other was from an individual subject measured by Wightman at the University of Wisconsin, Madison (Wightman & Kistler, 1989a). Apparently, however, the Wightman data are not included with Gehring's software package.

3 The NASA Ames 3-D Auditory Display Project

Since 1986, our group at NASA Ames has been working on a real-time system for use in both basic research in human sound localization and applied studies of acoustic information display in advanced human–computer interfaces (Wenzel, Wightman, & Foster, 1988a,b). The research began as part of the Ames VIrtual Environment Workstation (VIEW) project (Fisher et al., 1988). To achieve our objective, we have taken a four-part approach: (1) develop a technique for synthesizing localized, acoustic stimuli based on psychoacoustic principles, (2) in parallel, develop the signal-processing technology required to implement the synthesis technique in real time, (3) perceptually validate the synthesis technique with basic psychophysical studies, and (4) use the real-time device as a research tool for evaluating and refining the approach to synthesis in both basic and applied contexts. The research has been a collaborative effort between myself as project director, Scott Foster of Crystal River Engineering (Groveland, CA), Fred Wightman and Doris Kistler of the University of Wisconsin, Madison, and since 1988, Durand Begault (National Research Council) and Philip Stone (Sterling Software) at NASA Ames.

As noted above, one technique for capturing both pinnae and interaural difference cues involves binaural recording of program material with microphones placed in the ears of a manikin (Plenge, 1974; Doll et al., 1986) or the ear canals of a human (Butler & Belendiuk, 1977). When sounds recorded this way are presented over headphones, there is an immediate and veridical perception of three-dimensional (3-D) auditory space (Plenge, 1974; Butler & Belendiuk, 1977; Blauert, 1983; Doll et al., 1986). Our procedure is closely related to binaural recording, but rather than record and play back stimuli directly, we measure the acoustical transfer functions, from free-field to eardrum, at many source positions, and use these HRTFs as the basis of filters with which we synthesize stimuli. Specifically, HRTFs, in the form of FIRs, are measured using techniques adapted from Mehrgardt and Mellert (1977). Although similar in principle to the impulse response method described earlier, the measurement is actually made with trains of pseudo-random noisebursts to improve the signal-to-noise ratio of the responses. Figure 6 illustrates the technique. Small probe microphones are placed near each eardrum of a human listener who is seated in an anechoic chamber (Wightman & Kistler, 1989a). Wideband test stimuli are presented from one of 144 equidistant locations in the free-field (nonreverberant) environ-

Downloaded from http://direct.mit.edu/pvar/article-pdf/1/1/80/1622491/pres.1992.1.1.80.pdf by UNIVERSITY OF CENTRAL FLORIDA, GREG WELCH on 04 August 2024

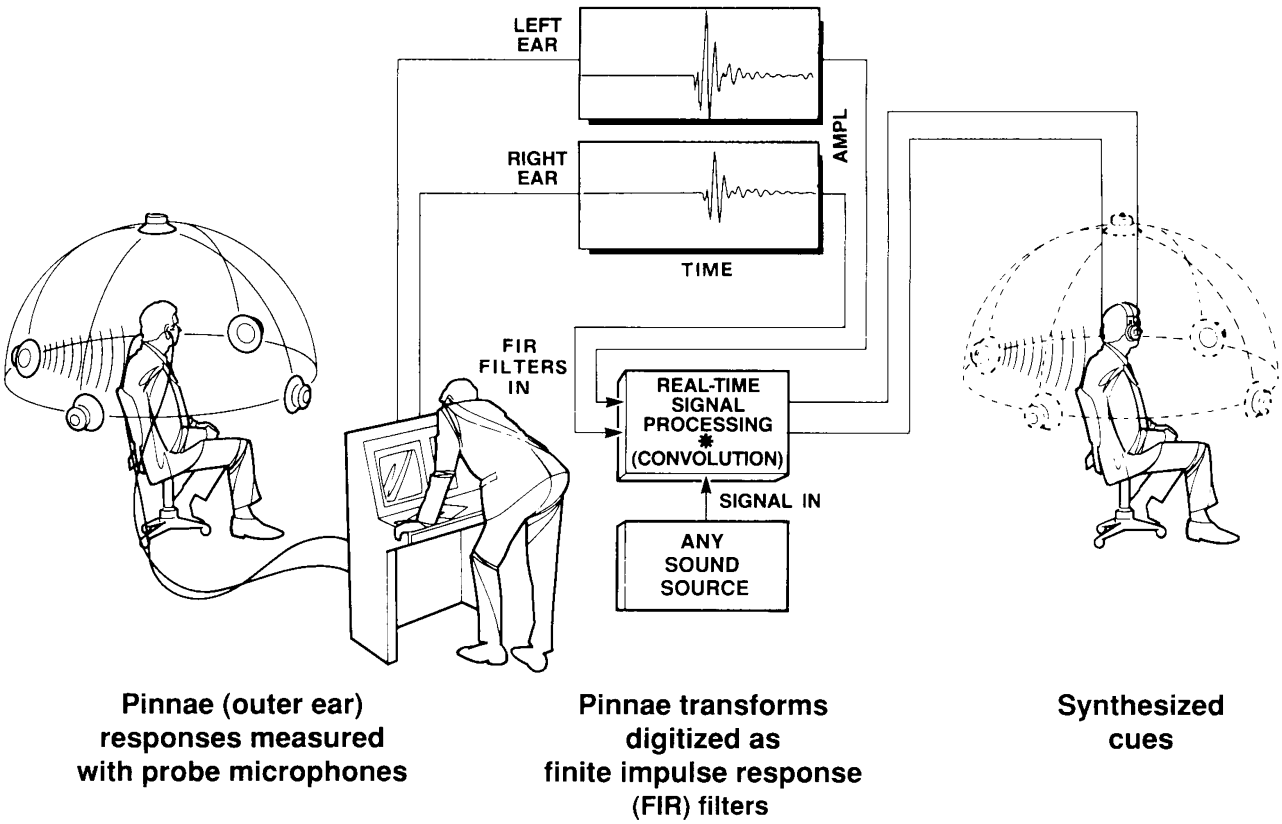

Figure 6. *Illustration of the technique for synthesizing virtual acoustic sources with measurements of the head-related transfer function. An example of a pair of finite impulse responses measured for a source location at 90° to the left and 0° elevation (at ear level) is shown in the insets for the left and right ears. The placement of the loudspeakers in the drawing is illustrative only.*

ment. A different pair of impulse responses is measured for each location in the spherical array at intervals of 15° in azimuth and 18° in elevation (elevation range: −36 to +54°). HRTFs are estimated by deconvolving (mathematically dividing out) the effects of the loudspeakers, test stimulus, and microphone responses from the recordings made with the probe microphones (Wightman & Kistler, 1989a). The advantage of this technique is that it preserves the complex pattern of interaural differences over the entire spectrum of the stimulus, thus capturing the effects of filtering by the pinnae, head, shoulders, and torso.

For example, the center insets in Figure 6 show a pair of FIR filters measured for one subject for a speaker location directly to the left and at ear level, that is, at −90° in azimuth and 0° in elevation. As one would expect, the waveform from this source arrived first and was larger in the left ear than the response measured in the right ear. As explained in Figure 2, the frequency-dependent effects can be analyzed by applying the Fourier transform to these temporal waveforms.

Figure 7 shows how both interaural amplitude and phase (or equivalently time) varies as a function of frequency for four different locations in azimuth, all at 0° in elevation. For example, the top left panels show that for 0° in azimuth or directly in front of the listener, there is very little difference in either the amplitude or phase responses between the two ears (the IIDs and ITDs). On the other hand, in the top right panels for +90° or directly to the listener's right, one can see that across the frequency spectrum, the amplitude and phase responses for the right ear are larger and lead in time with respect to the left ear.

To synthesize localized sounds, a map of "location

Downloaded from http://direct.mit.edu/pvar/article-pdf/1/1/80/1622491/pres.1992.1.1.80.pdf by UNIVERSITY OF CENTRAL FLORIDA, GREG WELCH on 04 August 2024

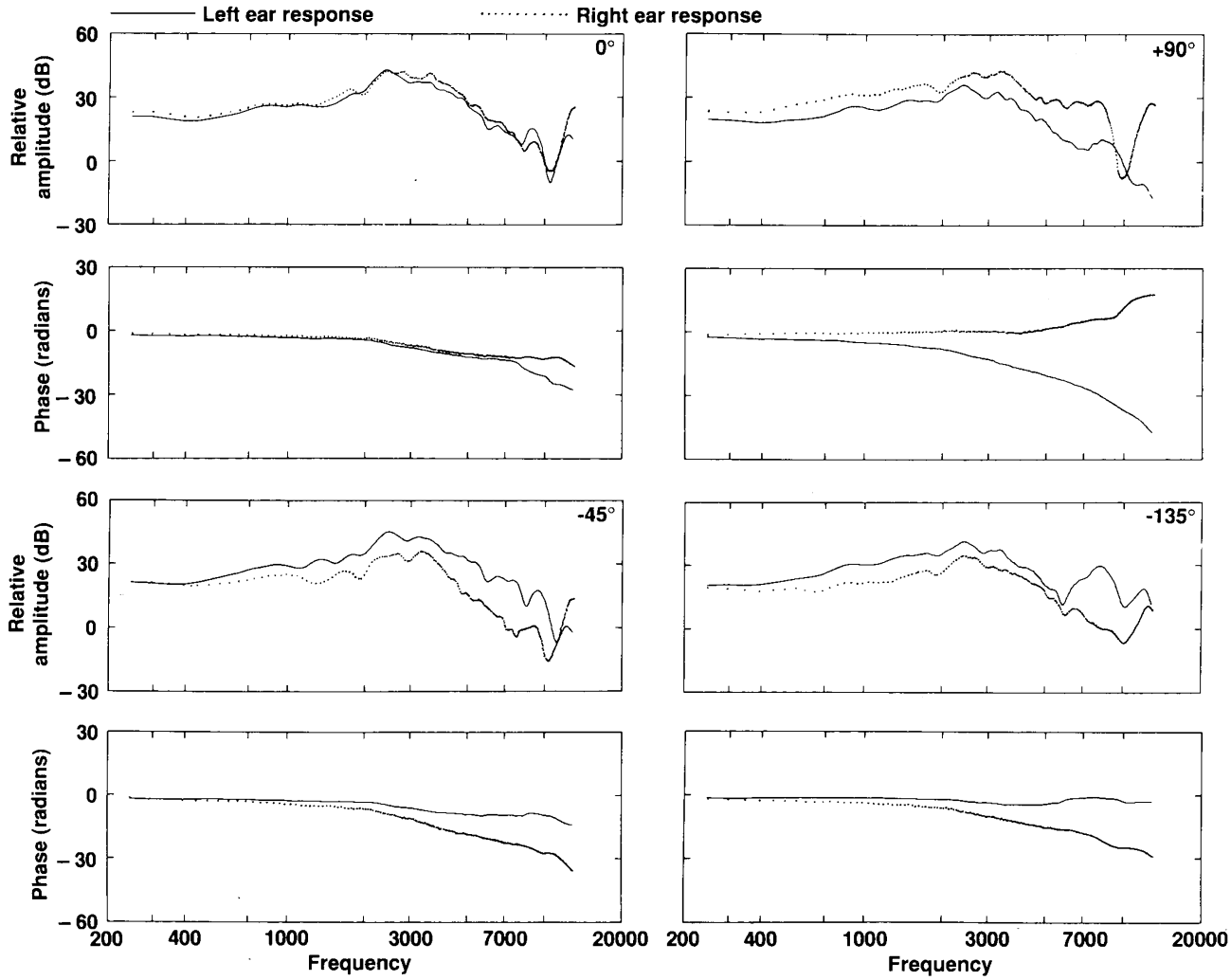

Figure 7. *Examples of HRTFs in a frequency domain representation based on the Fourier transform of pairs of FIRs. Magnitude and phase responses are plotted as a function of frequency for the two ears of a single subject. Four different azimuths (0, −45, −135, +90 deg) at 0° elevation are shown.*

filters" is constructed from all 144 pairs of FIR filters by first transforming them to the frequency domain, dividing out the spectral effects of the headphones to be used during playback using Fourier techniques, and then transforming back to the time domain.

4 The Real-Time System: The Convolvotron

In the real-time system, designed by Scott Foster of Crystal River Engineering (Foster, 1988), the map of corrected FIR filters is downloaded from a host computer to the dual-port memory of a real-time digital signal processor known as the Convolvotron (Fig. 8). This set of two printed circuit boards converts one or more monaural analog inputs to digital signals at a rate of 50 kHz (16-bit resolution). Each data stream is then convolved with filter coefficients determined by the coordinates of the desired target locations and the position of the listener's head, thus "placing" each input signal in the perceptual 3-space of the listener. The resulting data streams are mixed, converted to left and right analog signals, and presented over headphones. The current configuration allows up to four independent and simul-

Downloaded from http://direct.mit.edu/pvar/article-pdf/1/1/80/1622491/pres.1992.1.1.80.pdf by UNIVERSITY OF CENTRAL FLORIDA, GREG WELCH on 04 August 2024

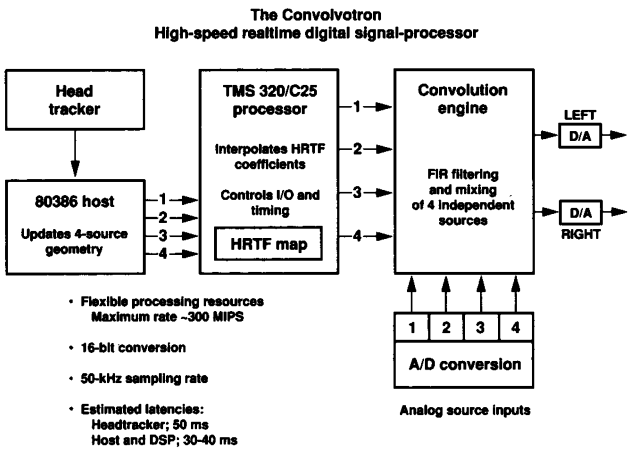

Figure 8. *Block diagram of the Convolvotron system designed by Scott Foster for digital filtering of signals with head-related transfer functions in real-time.*

taneous anechoic sources with an aggregate computational speed of more than 300 million multiply-accumulates per second. This processing speed is also sufficient for interactively simulating a single source plus six early reflections in relatively small reverberant environments (i.e., with head tracking; Foster & Wenzel, 1991; Foster et al., 1991). The hardware design can also be scaled upward to accommodate additional sources and the longer filter lengths required for simulating larger enclosures.

Motion trajectories and static locations at greater resolution than the empirical measurements are simulated by selecting the four measured positions nearest to the desired target location and interpolating with linear weighting functions. The interpolation algorithm effectively computes a new coefficient at the sampling interval (every 20 μsec) so that changes in position are free from artifacts like clicks or switching noises. When integrated with a magnetic head-tracking system (e.g., Polhemus 3-Space Isotrack), the listener's head position can be monitored in real time so that the four simultaneous sources are stabilized in fixed locations or in motion trajectories relative to the user. Again, such head coupling should help to enhance the simulation since previous studies suggest that head movements are important for localization (e.g., Wallach, 1940; Thurlow et al., 1967; Thurlow & Runge, 1967). This degree of interactivity, especially coupled with smooth motion interpolation and simulation of simple reverberant environments, is apparently unique to the Convolvotron system.

A pilot study conducted at the University of Wisconsin suggests that the interpolation approach is generally valid perceptually. In an absolute judgment task, localization performance was compared for static sources (non-real time) synthesized from empirical measurements of a subject's own HRTFs versus stimuli synthesized from HRTFs based on simple two-way linear interpolations in either azimuth or elevation. The general experimental paradigm was similar to the study by Wightman and Kistler (1989b) described below. Interpolation of the temporal waveforms was computed at separations of either 30 or 60° in azimuth and 36° in elevation; again, the smallest separation for the measured HRTFs for azimuth and elevation was 15 and 18°, respectively. Location judgments of four subjects indicated that stimuli derived from interpolations as far apart as 60° in azimuth were perceptually indistinguishable from stimuli synthesized from measured coefficients at the same target locations. For elevation, location judgments for the 36° interpolation showed increased variability compared to sources synthesized from the empirical HRTFs but remained monotonic with respect to the target location. These data suggest that the HRTF map of a real-time display could tolerate interpolation separations as large as 60° in azimuth (currently 30 to 60° in the Convolvotron) and still maintain perceptual viability. On the other hand, interpolations of 36° in elevation (18° in the Convolvotron) are more problematic. Such results are in apparent contradiction to the observation that perceptual resolution is greater in the dimension of azimuth than in elevation (e.g., Oldfield & Parker, 1984a). Since perceptual resolution presumably reflects the rate of change of discriminable features in the stimulus space, one might expect that the interpolation of HRTFs would have the greatest impact in the stimulus dimension with the greatest rate of change. However, the behavioral effects of interpolation may be more the result of the exact nature of the errors necessarily introduced by the averaging process. That is, the overall increase in the magnitudes of the interaural cues that correspond to increasing azimuth may be relatively

Downloaded from http://direct.mit.edu/pvar/article-pdf/1/1/80/1622491/pres.1992.1.1.80.pdf by UNIVERSITY OF CENTRAL FLORIDA, GREG WELCH on 04 August 2024

unaffected by interpolation, while the more subtle patterns of spectral coloration thought to determine elevation are probably much more easily disrupted, especially in the critical high-frequency regions. More comprehensive evaluations of the perceptual consequences of interpolation are underway at NASA Ames.

As with any system required to compute data "on the fly," the term real time is a relative one. The Convolvotron, including the host computer, has a computational delay of about 30–40 msec, depending on such factors as the number of simultaneous sources, the duration of the HRTFs used as filters, and the complexity of the source geometry.[4] An additional latency of at least 50 msec is introduced by the head tracker. This accumulation of computational delays has important implications for how well the system can simulate realistic moving sources or realistic head motion. At the maximum delay, the system can only update to a new location about every 90 msec. This directional update interval, in turn, corresponds to an angular resolution of about 32° or greater when the relative source-listener speed is 360°/sec, 16° or greater at 180°/sec, and so on. Such delays may or may not result in a perceptible lag, depending on how sensitive humans are to changes in angular displacement (the minimum audible movement angle) for a given source velocity. Recent work on the perception of auditory motion by Perrott and others using real sound sources (moving loudspeakers) suggests that these computational latencies are acceptable for moderate velocities. For example, for source speeds ranging from 8 to 360°/sec, minimum audible movement angles ranged from about 4 to 21°, respectively, for a 500-Hz tone burst (Perrott, 1982; Perrott & Tucker, 1988). Thus, slower relative velocities are well within the capabilities of the Convolvotron, while speeds approaching 360°/sec should begin to result in perceptible delays, especially when multiple sources or larger filters (e.g., simulations of reverberant rooms) are being generated.

4. There is a trade-off in the Convolvotron between the number of sources and the length of the HRTFs. For example, up to 512-, 256-, and 128-point filters/ear can be used for one, two, and four sources, respectively. Pilot data suggest that a minimum of 128-point, and preferably 256-point, filters are advisable to maintain good localization accuracy.

Currently, the Convolvotron is being used in a variety of government, university, and industry research labs besides ours, including the NASA Ames Crew Station Research and Development Facility; Naval Ocean Systems Center, San Diego; the Psychoacoustics Lab at the University of Wisconsin, Madison; the Psychoacoustics Lab at the Research Laboratory of Electronics at MIT; Matsushita Electric Works; Bellcore (Ludwig et al., 1990); and the Human Interface Technology Lab, University of Washington. The system also forms part of VPL Research's "Audiosphere" component of their virtual reality system.

5 Psychophysical Validation of the Synthesis Technique

The working assumption of our synthesis technique is that if, using headphones, we could reproduce ear canal waveforms identical to those produced by a free-field source, we would duplicate the free-field experience. Presumably, synthesis using individualized HRTFs would be the most likely to replicate the free-field experience for a given listener. Both the measurement of HRTFs and the synthesis of stimuli will always be subject to some error. It is also possible that higher level cognitive factors, such as the subjects' knowledge that they are using headphones, may affect the simulation. Thus, the only conclusive test of the adequacy of the simulation is an operational one in which free-field and synthesized, free-field localization are directly compared in psychophysical studies. Although researchers have been developing simulation techniques using HRTFs for some time, there is surprisingly little behavioral data available on their perceptual validity.

5.1 Validation for Static Sources Using Individualized HRTFs

A recent study by Wightman and Kistler (1989b) confirmed the perceptual adequacy of the basic approach for static sources. The stimuli were spectrally scrambled noisebursts transduced either by loudspeakers in an anechoic chamber or by headphones. In both free-field and

Downloaded from http://direct.mit.edu/pvar/article-pdf/1/1/80/1622491/pres.1992.1.1.80.pdf by UNIVERSITY OF CENTRAL FLORIDA, GREG WELCH on 04 August 2024

headphone conditions, the subjects indicated the apparent spatial position of 72 different target locations by calling out numerical estimates of azimuth and elevation (in degrees) using a modified spherical coordinate system. Thus, the subjects were essentially asked to imagine that they were pointing an arrow or vector that started at the origin, the center of their head, and intersected with the apparent location of the target stimulus. For example, a sound heard directly in front and at ear level would produce a response of "0, 0," a sound heard directly to the left and somewhat elevated might produce "−90 azimuth, +15 elevation," while one far to the rear on the right and below might produce "+170 azimuth, −30 elevation." Subjects were blindfolded and no feedback was given. Detailed explanations of the procedure and results can be found in the original paper.

The data analysis of localization experiments is complicated by the fact that the stimuli and responses are represented by points in three-dimensional space, in particular, as points on the surface of a sphere since distance was constant in this experiment. For these spherically organized data, the usual statistics of means and variances are potentially misleading. For example, an azimuth error of 15° at ear level is much larger in terms of absolute distance than a 15° error at a much higher or lower elevation. Thus, it is more appropriate to apply the techniques of spherical statistics to summarize the data (Fisher, Lewis, & Embleton, 1987). The spherical statistic reported here, the centroid, is defined by an azimuth and an elevation and can be thought of as the "average direction" of a set of judgment vectors for a given target location. Two indicators of variability, K^{-1} and the average angle of error, were also computed but these results will not be discussed here. The reader is referred to the original paper.

As discussed above, another type of error observed in nearly all localization studies is the presence of front-back reversals as well as reversals in elevation (Wenzel et al., 1991). It is difficult to know how to treat these errors fairly in the data analysis. Since the reversal rate is often low (e.g., Oldfield & Parker, 1984a), reversals have generally been resolved when computing descriptive statistics; that is, the responses are coded as if the subjects had indicated the correct hemisphere (as in the analyses of Table 1 and Fig. 9). Otherwise, estimates of error would be greatly inflated. On the other hand, if we assume that subjects' responses correctly reflect their perceptions, resolving such reversals could be misleading. Thus, the usual procedure is to resolve the judgments when computing parametric estimates like the centroid and to report the rate of reversals as a separate statistic.

Table 1. *Summary Statistics Comparing Resolved Localization Judgments of Free-Field (**Boldface Type**) and Virtual Sources (in Parentheses) for Eight Subjects*[a]

Subject	Goodness of fit	Azimuth correlation	Elevation correlation	Front–back reversals (%)	
SDE	**.93** (.89)	**.98** (.97)	**.68** (.43)	**12**	(20)
SDH	**.95** (.95)	**.96** (.95)	**.92** (.83)	**5**	(13)
SDL	**.97** (.95)	**.98** (.98)	**.89** (.85)	**7**	(14)
SDM	**.98** (.98)	**.98** (.98)	**.94** (.93)	**5**	(9)
SDO	**.96** (.96)	**.99** (.99)	**.94** (.92)	**4**	(11)
SDP	**.99** (.98)	**.99** (.99)	**.96** (.88)	**3**	(6)
SED	**.96** (.95)	**.97** (.99)	**.93** (.82)	**4**	(6)
SER	**.96** (.97)	**.99** (.99)	**.96** (.94)	**5**	(8)
Mean				**5.6**	(11)

[a]Adapted from Wightman and Kistler (1989b).

Table 1 provides a general overview of the results of Wightman and Kistler (1989b). Summary statistics comparing the eight subjects' resolved judgments of location for real (free-field) and synthesized stimuli are shown; the numbers in boldfaced type are for the free-field data and the numbers in parentheses are for the synthesized conditions. Note that overall goodness of fit between the actual and estimated source coordinates is quite comparable, 0.89 or better for the synthesized stimuli and 0.93 or better for free-field sources. The two correlation measures indicate that while source azimuth appears to be synthesized nearly perfectly, synthesis of source elevation is more problematic, particularly for SDE who also has difficulty judging elevation in the free field. Examples of the range of patterns of localization behavior for resolved judgments can be seen in Figure 9. Actual source azimuth (and, in the insets, elevation) versus the judged azimuth are plotted for subjects SDO and

Downloaded from http://direct.mit.edu/pvar/article-pdf/1/1/80/1622491/pres.1992.1.1.80.pdf by UNIVERSITY OF CENTRAL FLORIDA, GREG WELCH on 04 August 2024

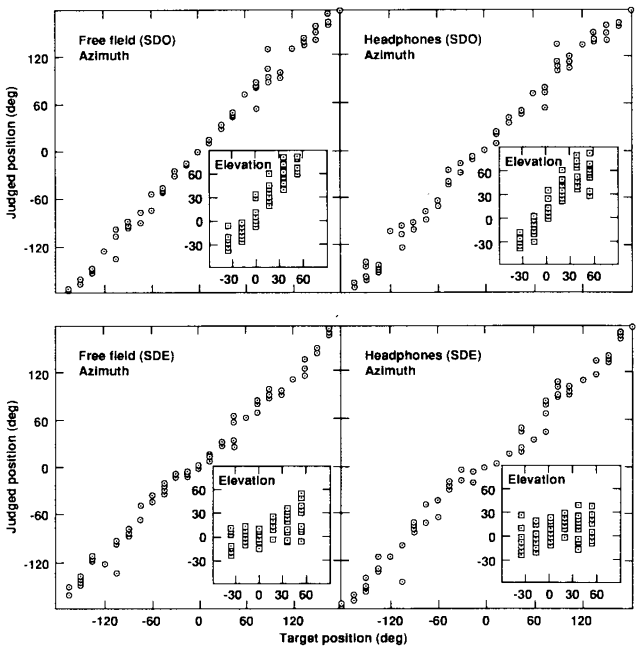

Figure 9. *Scatterplots of actual source azimuth (and, in the insets, elevation) versus judged source azimuth for subjects SDO and SDE in both free-field and headphone conditions. The plot on the left shows free-field judgments and the plot on the right shows judgments for the stimuli synthesized from the subjects' own transfer functions. Each data point represents the centroid of at least 6 judgments. Seventy-two source positions are plotted in each panel. Data from 6 different source elevations are combined in the azimuth plots and data from 24 different source azimuths are combined in the elevation insets. Note that the scale is the same in the azimuth and elevation plots. After Wightman & Kistler (1989b).*

SDE (both female) of Wightman and Kistler (1989b). The panel on the left plots free-field judgments and the panel on the right shows judgments for the stimuli synthesized from the subjects' own transfer functions. On each graph, the positive diagonal, or a straight line with a slope of 1.0, corresponds to perfect performance.

The reversal rates (Table 1) were relatively low, with average rates of about 6 and 11% for free-field and synthesized sources, respectively. Similar to the location centroids, reversal rates for the synthesized stimuli tended to be greatest for subjects who also had higher rates in the free field. Thus, while individual differences did occur, the pattern of results across synthesized and free-field conditions was generally consistent for a given subject; it appears that Butler and Belendiuk's (1977) observation of "good" and "poor" localizers is supported by these data.

5.2 Acoustic Determinants of Performance

Individual differences in localization behavior suggest that there may be acoustic features peculiar to each subject's HRTFs that influence performance. Thus, the use of averaged transforms, or even measurements derived from normative manikins such as the KEMAR, may or may not be an optimum approach for simulating free-field sounds for all listeners. An example of the degree of the between-subjects variability in acoustic features observed in HRTFs is illustrated in Figure 10, which plots the left- and right-ear magnitude responses for a single source location for eight different subjects (after Wenzel et al., 1988b). Obviously, any straightforward averaging of these functions would tend to smooth the peaks and valleys, thus removing potentially significant features in the acoustic transforms.

On the other hand, it may be possible to identify specific features of HRTFs that result in good or poor localization for most listeners. The psychophysical data of Wightman and Kistler (1989b) indicate that elevation is particularly difficult to judge, especially for subject SDE. A preliminary analysis of elevation coding suggests that there is indeed an acoustic basis for this poor performance.

Figure 11 plots "interaural elevation dependency" functions for four subjects' interaural amplitude data. The computational derivation of these functions can be found in Wightman and Kistler (1989b). Essentially, the six functions on each graph show how interaural intensity changes for different elevations normalized to zero elevation (the flat function) when the magnitude responses are collapsed across all azimuths. In spite of the large intersubject variability illustrated in Figure 10, the dependency functions for the better localizers (shown in the top three graphs) are quite similar to each other and show clear elevation dependencies above about 5 kHz. SDE's functions, on the other hand, are different from the other subjects and show little change with elevation.

Downloaded from http://direct.mit.edu/pvar/article-pdf/1/1/80/1622491/pres.1992.1.1.80.pdf by UNIVERSITY OF CENTRAL FLORIDA, GREG WELCH on 04 August 2024

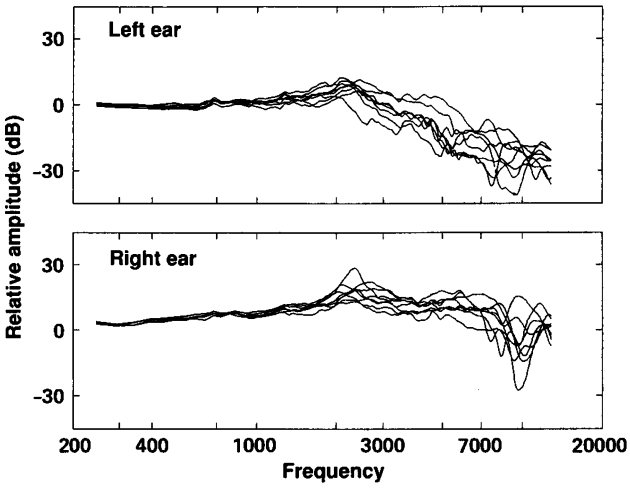

Figure 10. *Magnitude responses for a single source position are shown for eight subjects. The left and right ears are plotted separately. After Wenzel et al. (1988b).*

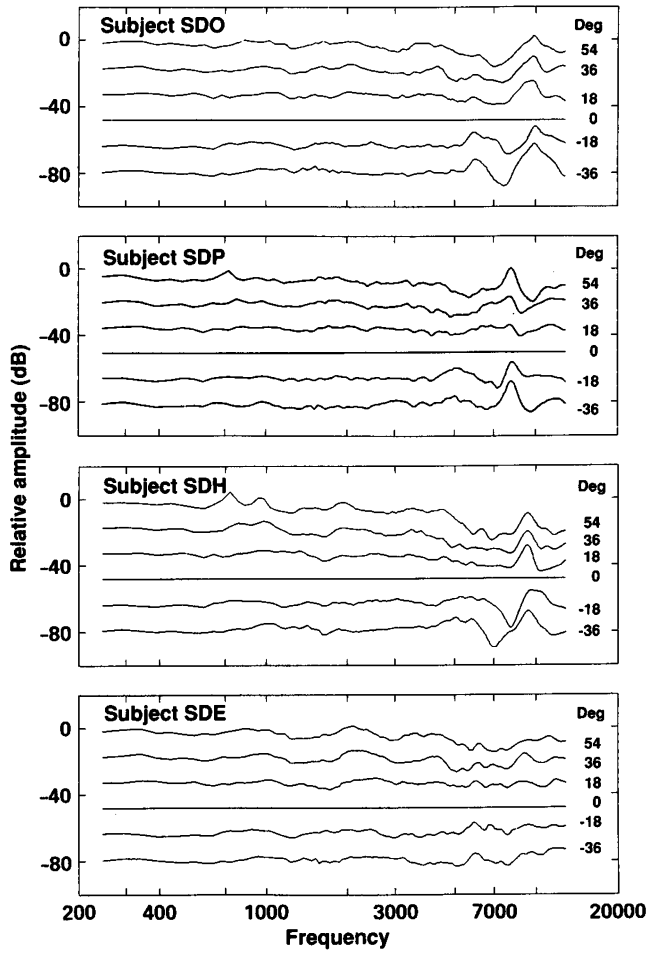

Figure 11. *Interaural elevation dependency functions are plotted for four subjects. From top to bottom, the functions within a panel represent elevations of +54, +36, +18, 0 (the reference elevation), −18, and −36°.*

Thus, it appears that SDE's poor performance in judging elevation for both real and synthesized stimuli may be due to a lack of distinctive acoustic features in her HRTFs. That is, the structure of SDE's pinnae may be such that sources at different elevations do not produce discriminable changes in the spectral coloration of the sound that vary reliably with elevation.

The observation of both behavioral and acoustical individual differences brings up a topic that has often been conjectured about but rarely directly tested (see Butler & Belendiuk, 1977, for an early example). That is, can one manipulate localization performance simply by listening "through" another person's ears? Or put another way, can we learn to take advantage of a set of good HRTFs even if we are a poor localizer? The following data from Wenzel, Wightman, Kistler, and Foster (1988c) illustrate the kind of "cross-ear listening" paradigm that is possible using our synthesis technique. Again, the subjects provided absolute judgments of location as in the experiment by Wightman and Kistler (1989b).

Figure 12 shows what happens to resolved azimuth and elevation judgments when a good localizer listens to stimuli synthesized from another good localizer's pinna transforms. Azimuth is plotted in the top panels and elevation is on the bottom. The left and far-right graphs plot centroids for SDP's and SDO's azimuth judgments versus the target locations when the stimuli were synthesized from their own HRTFs. Front-back reversals have been resolved as described above. As can be seen, both SDP and SDO localize the synthesized stimuli based on their own HRTFs quite well. The center graphs show what happens when SDP listens "through" SDO's pinnae. Localization of azimuth degrades somewhat, but not a great deal. Elevation performance degrades further, suggesting that elevation cues are not as robust as azimuth cues across different individuals, but an overall correspondence between real and perceived locations remains intact.

Downloaded from http://direct.mit.edu/pvar/article-pdf/1/1/80/1622491/pres.1992.1.1.80.pdf by UNIVERSITY OF CENTRAL FLORIDA, GREG WELCH on 04 August 2024

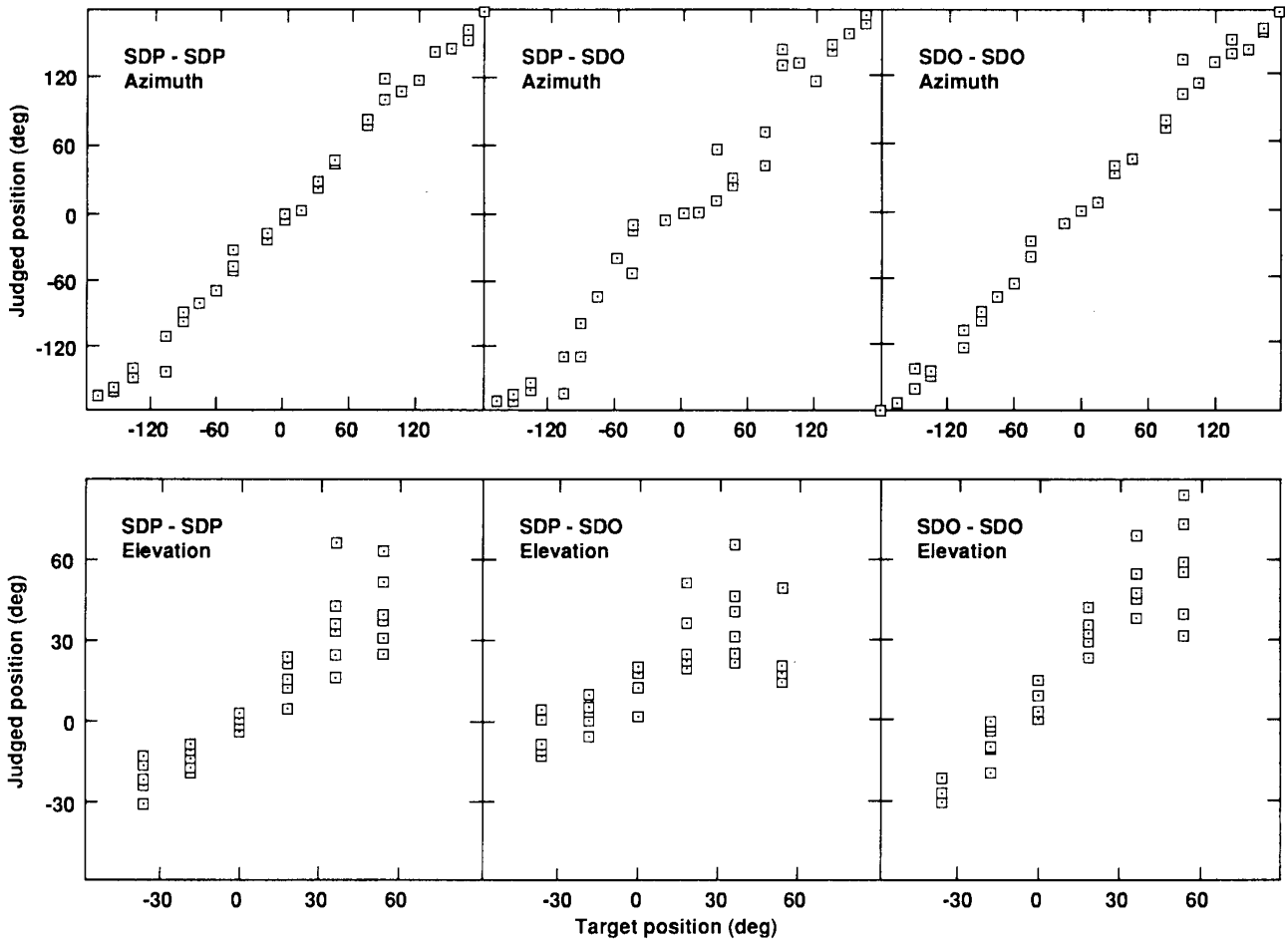

Figure 12. *Scatterplots of actual source azimuth (and, in the insets, elevation) versus judged source azimuth for three different headphone conditions. The panels on the far left and right show SPD's and SDO's judgments for stimuli synthesized from their own transfer functions. The center plot shows SDP's judgments for stimuli synthesized from SDO's HRTFs. Each data point represents the centroid of at least 6 judgments. Thirty-six source positions are plotted in each panel. Data from 6 different source elevations are combined in the azimuth panels and data from 18 different source azimuths are combined in the elevation insets. Note that the scale is the same in the azimuth and elevation plots.*

Figure 13 compares performance when a good localizer, SDO, listens to stimuli synthesized from the HRTFs of poor localizer, SDE. Again for azimuth there is little degradation. However, for elevation, it seems that SDE's pinnae provide poor elevation cues for SDO as well, supporting the notion that spectral features of the acoustic transforms determine localization.

If acoustic features do determine localization, one might conclude that the reciprocal case is true; that SDE could actually improve her performance if she could listen "through" SDO's ears. Figure 14 plots these data. Again, SDE whose azimuth judgments are accurate for stimuli synthesized from her own HRTFs, performs nearly as well when listening to SDO's azimuth cues. However, it appears that cross-ear listening is not a symmetrical effect for elevation. Even after about 30 hr of additional testing (including about 6 hr with verbal feedback), compared to only 2 hr for the good localizers, SDE still could not take advantage of the presumably better cues provided by SDO's pinnae. This result contradicts Butler and Belendiuk's (1977) study of elevation in the median plane; even their poorest localizer was able to improve localization scores in only 100 trials without feedback by listening to stimuli simulated from the bet-

Downloaded from http://direct.mit.edu/pvar/article-pdf/1/1/80/1622491/pres.1992.1.1.80.pdf by UNIVERSITY OF CENTRAL FLORIDA, GREG WELCH on 04 August 2024

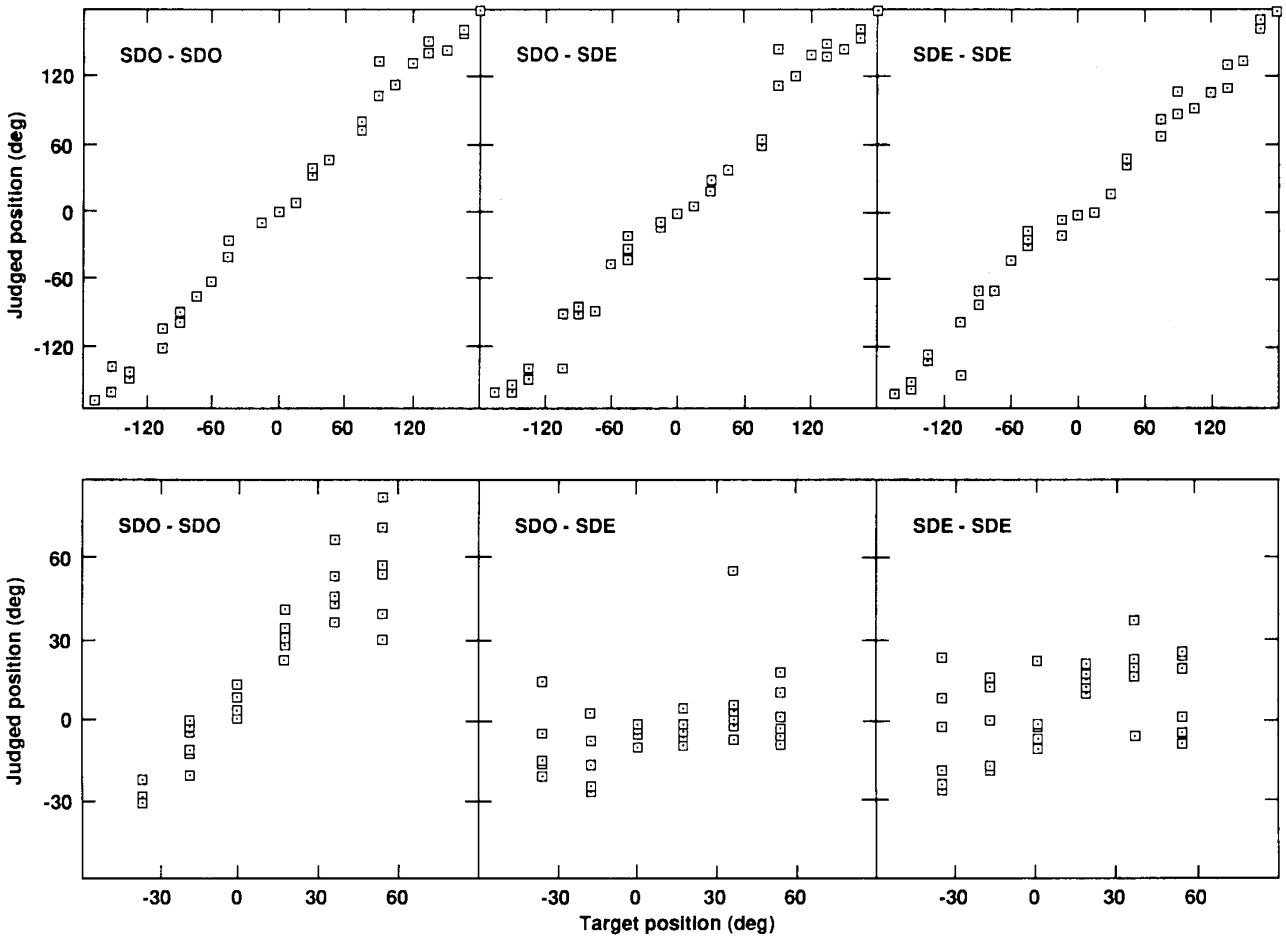

Figure 13. *Same as Figure 12 except for subjects SDO and SDE on the left and right panels, respectively.*

ter localizers' cues. Conversely, poor elevation performance was "transferred" to the better localizers using the poor localizer's cues. Butler and Belendiuk's result for the poor localizer is rather puzzling; it seems somewhat unlikely that elevation cues could be learned almost immediately and without feedback. Perhaps this subject was not totally deficient in a general ability to judge elevation. Since only eight positions in the median plane were tested, this is rather difficult to assess.

On the other hand, the data of Wenzel et al. (1988c) are hardly conclusive since they are based on a sample size of one; only SDE of the eight subjects in Wightman and Kistler (1989b) showed such poor elevation performance to begin with. Such data suggest that there may be a critical period for learning localization cues that, once past, can never be regained. Perhaps more likely is that simple exposure to the unfamiliar cues for elevation was not enough. SDE may have needed prolonged and consistent experience with SDO's HRTFs to learn to discriminate the subtle acoustic cues she does not normally hear. Apparently a few hours of exposure a day with static stimuli, especially in the absence of correlated information from the other senses, are not enough to allow learning to occur.

5.3 Inexperienced Listeners and Nonindividualized HRTFs

In practice, measurement of each potential listener's HRTFs may not be feasible. It may also be the case that the user of a spatial auditory display will not have the opportunity for extensive training. Thus, a critical

Downloaded from http://direct.mit.edu/pvar/article-pdf/1/1/80/1622491/pres.1992.1.1.80.pdf by UNIVERSITY OF CENTRAL FLORIDA, GREG WELCH on 04 August 2024

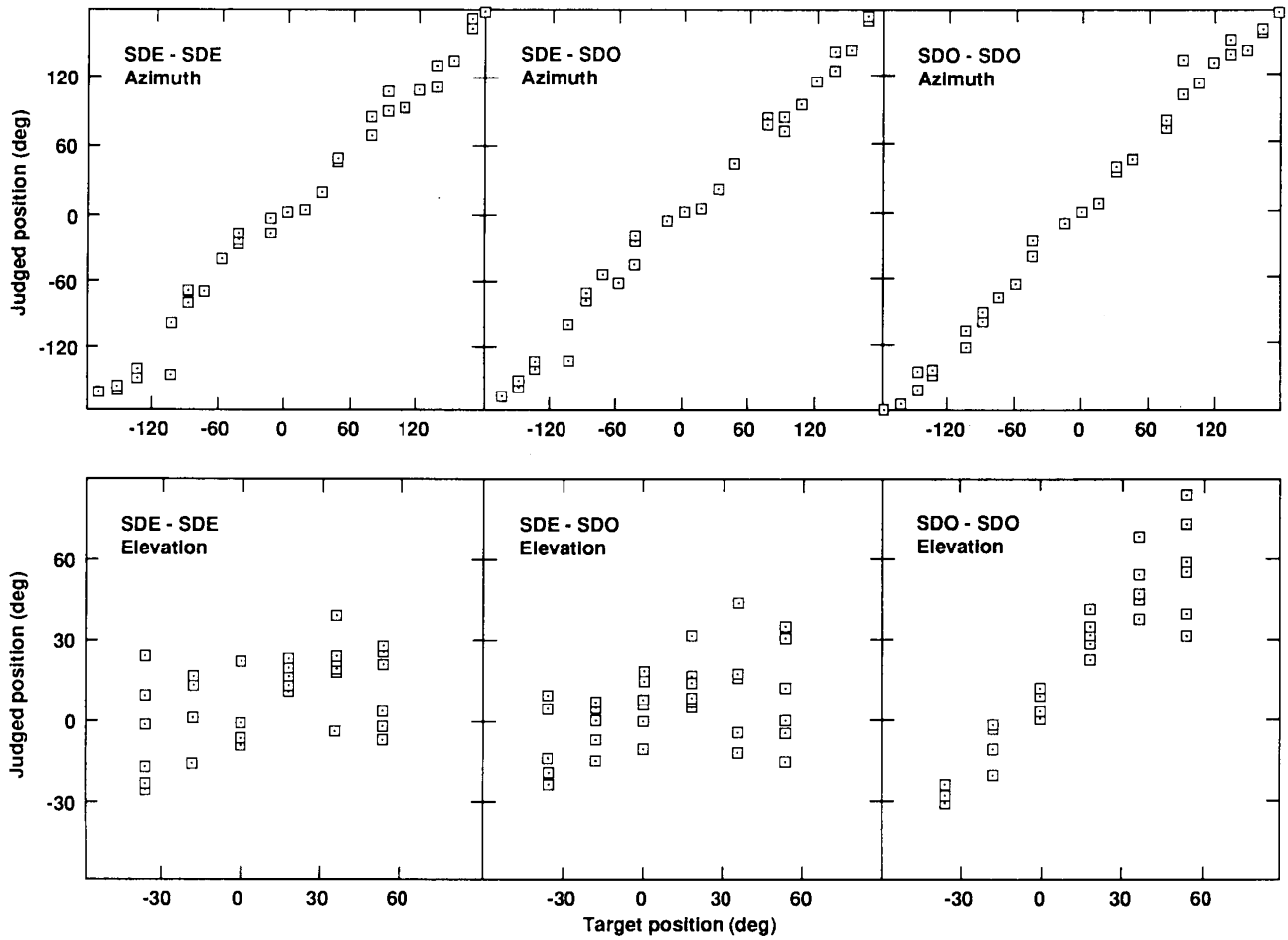

Figure 14. *Same as Figure 12 except for subjects SDE and SDO on the left and right panels, respectively.*

research issue for virtual acoustic displays is the degree to which the general population of listeners can readily obtain adequate localization cues from stimuli based on nonindividualized transforms. The individual difference data of Figures 12–14 suggest that, even in the worst case, using nonindividualized transforms does not degrade localization accuracy much more than the listener's inherent ability. In general, then, even inexperienced listeners may be able to use a particular set of HRTFs as long as they provide adequate cues for localization. A reasonable approach is to use the HRTFs from a subject whose measurements have been "behaviorally calibrated" and are thus correlated with known perceptual ability in both free-field and headphone conditions. Recently, Wenzel et al. (1991) completed a more extensive study using a variant on the cross-ear listening paradigm; 16 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, SDO, a "good localizer" from the experiment by Wightman and Kistler (1988b).

Figure 15 illustrates the behavior of 12 of the 16 subjects. When reversals are resolved, localization performance is quite good, with judgments for the nonindividualized stimuli nearly identical to those in the free field. Like SDE in Wenzel et al. (1988c), two of the subjects show poor elevation performance in both free-field and headphone conditions, a response pattern that is at least consistent across the free-field and virtual source conditions (Fig. 16). The third pattern is illustrated in Figure

Downloaded from http://direct.mit.edu/pvar/article-pdf/1/1/80/1622491/pres.1992.1.1.80.pdf by UNIVERSITY OF CENTRAL FLORIDA, GREG WELCH on 04 August 2024

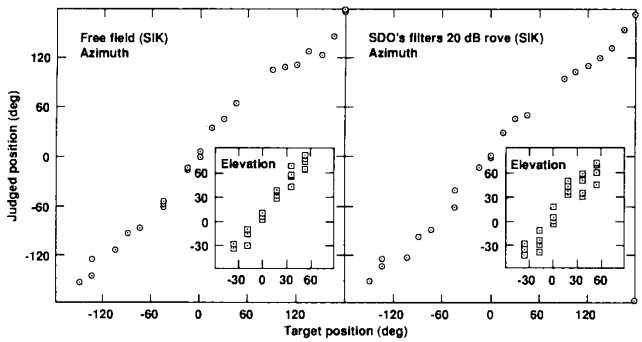

Figure 15. *Scatterplots of actual source azimuth (and, in the insets, elevation) versus judged source azimuth for subject SIK in both free-field and headphone conditions. The plot on the left plots free-field judgments and the plot on the right shows judgments for the stimuli synthesized from nonindividualized transfer functions. Each data point represents the centroid of 9 judgments. Twenty-four source positions are given in each plot. Data from 6 different source elevations are combined in the azimuth plots and data from 18 different source azimuths are combined in the elevation insets. Note that the scale is the same in the azimuth and elevation plots.*

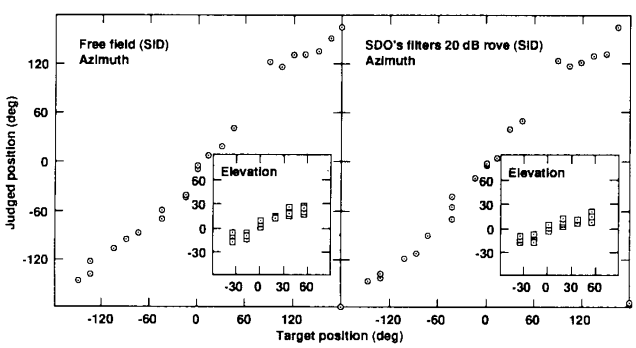

Figure 16. *Same as Figure 15, except for subject SID.*

17; here, two subjects show inconsistent behavior with poor elevation accuracy in only the synthesized conditions. The latter phenomenon, if it turns out to be common, would be a particular problem for virtual displays.

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. Again, the results plotted in

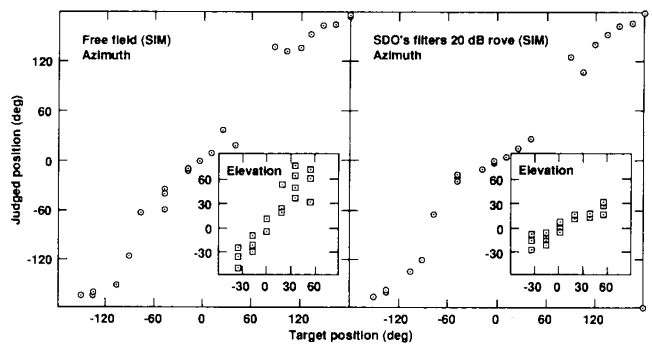

Figure 17. *Same as Figure 15, except for subject SIM.*

Figures 9 and 12–17 are based on analyses in which errors due to front-back reversals are resolved. For free-field versus simulated free-field stimuli, the eight experienced listeners in the Wightman and Kistler study exhibit front-back reversal rates of about 6 vs. 11% while the 16 inexperienced listeners using nonindividualized HRTFs show average rates of about 19 vs. 31% (Wenzel et al., 1991; average elevation reversal rates in this study were 6 and 18% for free-field vs. synthesized sources). In the data of Wenzel et al. (1988c), front-back confusion rates averaged across the three subjects were 6, 12, and 35% for the free-field stimuli, the stimuli synthesized from individualized HRTFs, and the stimuli synthesized from nonindividualized transforms, respectively.

Note that the existence of free-field reversals indicates that these confusions are not strictly the result of the simulation. Rather, as discussed in section 2.1, they are probably caused by inherent ambiguities in the stimuli due to cone-of-confusion effects. From the listener's perspective, both errors in the synthesis process and the use of nonindividualized pinna cues may simply exacerbate these effects by adding "jitter" to the subtle spectral features in the HRTFs, particularly with respect to the cues they are used to hearing. Such effects could act as a noise process that effectively changes the relative size or shifts the relative location of spectral peaks and valleys, resulting in higher reversal rates because of a mismatch between actual and expected spectral patterns. It is possible, as Asano, Suzuki, and Sone (1990) have claimed, that reversals tend to diminish as subjects gain experience (without feedback) with the impoverished stimulus

Downloaded from http://direct.mit.edu/pvar/article-pdf/1/1/80/1622491/pres.1992.1.1.80.pdf by UNIVERSITY OF CENTRAL FLORIDA, GREG WELCH on 04 August 2024

conditions provided by static anechoic sources, whether real or simulated. That is, even "on their own," subjects may adapt to unfamiliar spectral cues, relearning the rather fine discriminations of location-dependent spectral differences that eventually allow them to reliably resolve the cone of confusion. The higher reversal rates for the inexperienced listeners of Wenzel et al. (1991) compared to the more experienced subjects of Wightman and Kistler (1989b) tend to support this view. Thus, it may be that some form of adaptation and/or task-dependent training will usually be required to take full advantage of a virtual acoustic display.

6 Improving Virtual Acoustic Displays: Problem Areas and Research Issues

So far, the data 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. As suggested above, although the reason for reversals is not completely understood, they are probably due in large part to the static nature of the stimulus and the ambiguity resulting from the cone of confusion. Cone-of-confusion effects alone, however, cannot explain the observation of 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 (see Wallach, 1940; Warren et al., 1981; Welch, 1978). That is, given an ambiguous acoustic stimulus in the absence of an obvious visual correlate, it may be that the perceptual system tends to resolve the ambiguity with a heuristic that assumes the source is behind the listener where it cannot be seen. No doubt, the addition of correlated visual cues and dynamic acoustic cues coupled with head motion will do much toward restoring the ability to resolve these ambiguities in virtual acoustic displays. Systematic investigations of such effects will be a next step in the work at NASA Ames.

Again, another problem in synthesizing veridical acoustic images over headphones is the fact that such stimuli sometimes fail to externalize, particularly when the signals are unfamiliar (e.g., the spectrally scrambled noisebursts used here) and simulated from anechoic measurements of HRTFs. Environmental cues such as the ratio of direct to reflected energy, reverberation time, and other characteristics specific to particular enclosed spaces may help to enhance the externalization of images (Coleman, 1963; Gardner, 1968; Laws, 1972, 1973; Plenge, 1974; Borish, 1984; Begault, 1987). 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 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. It may be possible to discover an optimal tradeoff between environmental parameters that enhance externalization while minimizing the impact of the resulting expansion of the spatial image that may interfere with the ability to judge the direction of the source.

Whether distance, the third dimension in a virtual acoustic display, 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 and relatively little is known about the parameters that determine distance perception (Coleman, 1963; Laws, 1972). 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 such as stimulus familiarity and reverberation (Gardner, 1968; Mershon & King, 1975). Attempting to enhance 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. For example, the Convolvotron system allows real-time, interactive gain control of independent sources that goes

Downloaded from http://direct.mit.edu/pvar/article-pdf/1/1/80/1622491/pres.1992.1.1.80.pdf by UNIVERSITY OF CENTRAL FLORIDA, GREG WELCH on 04 August 2024

some way toward fulfilling the impression of relative distance by allowing dynamic comparisons of relative source intensities. Further understanding of the role of environmental cues, and the ability to synthesize such cues interactively (Foster et al., 1991), may eventually improve the reliable discrimination of source distances. Additionally, the success of any reasonably complex spatial display will depend on our understanding of localization masking, or the stimulus parameters that affect the identification, segregation (e.g., Bregman, 1990), and discrimination (e.g., Perrott, 1984a,b) of multiple sources. Surprisingly, little or no research has been done on the localization of more than two simultaneous sources.

Another critical area is the further specification of the role of individual differences and perhaps the development of efficient techniques for training or adaptation to nonindividualized transforms. The fact that individual differences in performance are apparently correlated with acoustical idiosyncracies in the HRTFs suggests that the systematic analysis and manipulation of HRTF characteristics may provide a means for counteracting individual difference effects. In addition to appropriate adaptation techniques, it may eventually be possible to construct a set of "universal transforms" using parametric techniques such as Genuit's structural model (1986), data reduction techniques such as specialized averaging models and principal components analysis (Asano et al., 1990; Kistler & Wightman, 1990), or perhaps even enhancing the features of empirically derived transfer functions (Durlach & Pang, 1986; Durlach, 1991; Van Veen & Jenison, 1991).

Other important research will be related to further refinements in the techniques for the accurate measurement, manipulation, and perceptual validation of HRTFs, including practical signal-processing issues such as determining optimal techniques for interpolation between measured or modeled transforms to ensure veridical motion.

The simulation techniques investigated here provide both a means of implementing a virtual acoustic display and the ability to study features of human sound localization that were previously inaccessable due to a lack of control over the stimuli. The availability of real-time control systems (e.g., Wenzel et al., 1988a,b) further expand the scope of the research, allowing the study of dynamic, intersensory aspects of localization that may do much toward alleviating the problems encountered in producing the reliable and veridical perception that is critical for many applied contexts.

Acknowledgments

This work has been supported by NASA, the California Department of Commerce, the Office of Naval Research, and the Federal Aviation Authority. A briefer version of this paper will appear as a chapter in R. Dannenberg and M. Blattner (Eds.), *Interactive Multimedia Computing*. New York: ACM Press. Many thanks to my past and present colleagues on the project: Scott Foster, Fred Wightman, Doris Kistler, Durand Begault, Phil Stone, and Scott Fisher. Thanks also to Nathaniel Durlach and Steven Colburn for their valuable comments on an earlier draft of the paper.

References

Asano, F., Suzuki, Y., & Stone, T. (1990). Role of spectral cues in median plane localization. *Journal of the Acoustical Society of America, 88,* 159–168.

Batteau, D. W. (1967). The role of the pinna in human localization. *Proceedings of the Royal Society of London, B168,* 158–180.

Begault, D. R. (1987). *Control of auditory distance.* Unpublished doctoral thesis. University of California, San Diego, CA.

Begault, D. R., & Wenzel, E. M. (1992). Techniques and applications for binaural sound manipulation in man-machine interfaces. *International Journal of Aviation Psychology,* in press.

Blattner, M. M., Sumikawa, D. A., & Greenberg, R. M. (1989). Earcons and icons: Their structure and common design principles. *Human-Computer Interaction, 4,* 11–44.

Blauert, J. (1969). Sound localization in the median plane. *Acustica, 22,* 205–213.

Blauert, J. (1983). *Spatial hearing: The psychophysics of human sound localization.* Cambridge, MA: The MIT Press.

Blauert, J. (1984). *Psychoakustik des binauralen Horens.* [*The*

Downloaded from http://direct.mit.edu/pvar/article-pdf/1/1/80/1622491/pres.1992.1.1.80.pdf by UNIVERSITY OF CENTRAL FLORIDA, GREG WELCH on 04 August 2024

psychophysics of binaural hearing.] Invited plenary paper presented at DAGA'84. Darmstadt, Germany.

Bly, S. (1982). *Sound and computer information presentation.* Unpublished doctoral thesis (No. UCRL-53282). Lawrence Livermore National Laboratory and University of California, Davis, CA.

Boerger, G., Laws, P., & Blauert, J. (1977). Stereophonic headphone reproduction with variation of variation of various transfer factors by means of rotational head movements. *Acustica, 39,* 22–26.

Borish, J. (1984). Extension of the image model to arbitrary polyhedra. *Journal of the Acoustical Society of America, 75,* 1827–1836.

Bregman, A. S. (1981). Asking the "What for" question in auditory perception. In M. Kubovy & J. R. Pomerantz (Eds.), *Perceptual organization* (pp. 99–118). Hillsdale, NJ: Lawrence Erlbaum.

Bregman, A. S. (1990). *Auditory scene analysis.* Cambridge, MA: The MIT Press.

Brigham, E. O. (1974). *The fast Fourier transform.* Englewood Cliffs, NJ: Prentice-Hall.

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

Brooks, F. P. (1988). Grasping reality through illusion—Interactive graphics serving science. *Proceedings of CHI'88. ACM Conference on Human Factors in Computing Systems,* 1–11.

Butler, R. A., & Belendiuk, K. (1977). Spectral cues utilized in the localization of sound in the median sagittal plane. *Journal of the Acoustical Society of America, 61,* 1264–1269.

Buxton, W., Gaver, W., & Bly, S. (1989). *The use of non-speech audio at the interface.* (Tutorial No. 10). Presented at CHI'89, ACM Conference on Human Factors in Computing Systems. New York: ACM Press.

Calhoun, G. L., Valencia, G., & Furness, T., III (1987). Three-dimensional auditory cue simulation for crew station design/evaluation. *Proceedings of the Human Factors Society, 31,* 1398–1402.

Carterette, E., & Friedman, M. (Eds.) (1978). *Hearing, handbook of perception* (Vol. IV). New York: Academic Press.

Cherry, E. C. (1953). Some experiments on the recognition of speech with one and two ears. *Journal of the Acoustical Society of America, 22,* 61–62.

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

Colquhoun, W. P. (1975). Evaluation of auditory, visual, and dual-mode displays for prolonged sonar monitoring in repeated sessions. *Human Factors, 17,* 425–437.

Deatherage, B. H. (1972). 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.; pp. 123–160). Washington, DC: U.S. Government Printing Office.

Deutsch, D. (Ed.) (1982). *The psychology of music.* New York: Academic Press.

Doll, T. J., Gerth, J. M., Engelman, W. R., & Folds, D. J. (1986). *Development of simulated directional audio for cockpit applications* (Report No. AAMRL-TR-86-014). Dayton, OH: Wright-Patterson Air Force Base.

Durlach, N. I. (1991). Auditory localization in teleoperator and virtual environment systems: Ideas, issues, and problems. *Perception, 20,* 543–554.

Durlach, N. I., & Pang, X. D. (1986). Interaural magnification. *Journal of the Acoustical Society of America, 80,* 1849–1850.

Edwards, A. D. N. (1989). Soundtrack: An auditory interface for blind users. *Human-Computer Interaction, 4,* 45–66.

Fisher, H., & Freedman, S. J. (1968). The role of the pinna in auditory localization. *Journal of Auditory Research, 8,* 15–26.

Fisher, N. I., Lewis, T., & Embleton, B. J. J. (1987). *Statistical analysis of spherical data.* Cambridge, United Kingdom: Cambridge University Press.

Fisher, S. S., Wenzel, E. M., Coler, C., & McGreevy, M. W. (1988). Virtual interface environment workstations. *Proceedings of the Human Factors Society, 32,* 91–95.

Foley, J. D. (1987). Interfaces for advanced computing. *Scientific American, 257,* 126–135.

Forbes, T. W. (1946). Auditory signals for instrument flying. *Journal of the Aeronautical Society,* May, 255–258.

Foster, S. H. (1988). Convolvotron™ User's Manual. Crystal River Engineering. IAC, 12350 Ward's Ferry Road, Groveland, CA 95321.

Foster, S. H. & Wenzel, E. M. (1991). *Virtual acoustic environments: The Convolvotron.* Demonstration system presented at the "Tomorrow's Realities Gallery," SIGGRAPH '91, 18th ACM Conference on Computer Graphics and Interactive Techniques, Las Vegas, NV.

Foster, S. H., Wenzel, E. M., and Taylor, R. M. (1991). Real-time synthesis of complex acoustic environments [Summary]. *Proceedings of the ASSP (IEEE) Workshop on Applications of Signal Processing to Audio & Acoustics,* New Paltz, NY.

Furness, T. A. (1986). The super cockpit and its human factors

Downloaded from http://direct.mit.edu/pvar/article-pdf/1/1/80/1622491/pres.1992.1.1.80.pdf by UNIVERSITY OF CENTRAL FLORIDA, GREG WELCH on 04 August 2024

challenges. *Proceedings of the Human Factors Society, 30,* 48–52.

Gardner, M. B. (1968). Proximity image effect in sound localization. *Journal of the Acoustical Society of America, 43,* 163.

Gardner, M. B. (1973). Some monaural and binaural facets of median plane localization. *Journal of the Acoustical Society of America, 54,* 1489–1495.

Gardner, M. B., & Gardner, R. S. (1973). Problem of localization in the median plane: Effect of pinnae cavity occlusion. *Journal of the Acoustical Society of America, 53,* 400–408.

Garner, W. R. (1949). Auditory signals. In *A survey report on human factors in undersea warfare* (pp. 201–217). Washington, D.C.: National Research Council.

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

Gaver, W. W., Smith, R. B., & O'Shea, T. (1991). Effective sounds in complex systems: The ARKola simulation. *Proceedings of CHI'91. ACM Conference on Computer-Human Interaction,* 85–90.

Gehring, B. (1990). Focal Point™ 3D Sound User's Manual. Gehring Research Corporation, 189 Madison Avenue, Toronto, Ontario, Canada, M5R 2S6.

Genuit, K. (1986). A description of the human outer ear transfer function by elements of communication theory (Paper B6-8). *Proceedings of the 12th International Congress on Acoustics,* Toronto.

Gibson, J. J. (1979). *The ecological approach to visual perception.* Boston: Houghton Mifflin.

Gierlich, H. W., & Genuit, K. (1989). Processing artificial-head recordings. *Journal of the Audio Engineering Society, 37,* 34–39.

Henning, G. B. (1974). Detectability of inetraural delay in high-frequency complex waveforms. *Journal of the Acoustical Society of America, 55,* 84–90.

Hudde, H., & Schroter, J. (1981). [Improvements in the Neumann artificial head system]. In German, In *Runkfunktechnische Mitteilungen [Radio Technology Reports],* Federal Republic of Germany.

Kendall, G. S., & Martens, W. L. (1984). Simulating the cues of spatial hearing in natural environments. *Proceedings of the 1984 International Computer Music Conference,* Paris.

Kendall, G. S., & Wilde, M. D. (1989). Production and reproduction of three-dimensional spatial sound [abstract]. *Journal of the Audio Engineering Society, 37,* 1066.

Kistler, D. J., & Wightman, F. L. (1990). Principal components analysis of head-related transfer functions [abstract]. *Journal of the Acoustical Society of America, 88,* S98.

Kubovy, M. (1981). Concurrent pitch-segregation and the theory of indispensable attributes. In M. Kubovy & J. R. Pomerantz (Eds.), *Perceptual organization* (pp. 55–98), Hillsdale, NJ: Lawrence Erlbaum.

Kubovy, M., & Howard, F. P. (1976). Persistence of a pitch-segregating echoic memory. *Journal of Experimental Psychology: Human Perception and Performance, 2,* 531–537.

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

Laws, P. (1972). *Zum Problem des Entfernungshorens und der Im-Kopf-Lokalisiertheit von Horeregnissen* [*On the problem of distance hearing and the localization of auditory events inside the head*]. Unpublished doctoral thesis, Technische Hochschule, Aachen, Federal Republic of Germany.

Laws, P. (1973). Entfernungshoeren und das Problem der Im-Kopf-Lokalisiertheit von Hoerereignissen. [Auditory distance perception and the problem of "in-head localization of sound images]. *Acustica, 29,* 243–259.

Lehnert, H., & Blauert, J. (1989, October). A concept for binaural room simulation [Summary]. *Proceedings of the IEEE ASSP Workshop on Applications of Signal Processing to Audio and Acoustics,* New Paltz, NY.

Loomis, J. M., Hebert, C., & Cicinelli, J. G. (1990). Active localization of virtual sounds. *Journal of the Acoustical Society of America, 88,* 1757–1764.

Ludwig, L., Pincever, N. & Cohen, M. (1990). Extending the notion of a window system to audio. *Computer,* 66–72.

McKinley, R. L., & Ericson, M. A. (1988). Digital synthesis of binaural auditory localization azimuth cues using headphones. *Journal of the Acoustical Society of America, 83,* S18.

Mehrgardt, S., & Mellert, V. (1977). Transformation characteristics of the external human ear. *Journal of the Acoustical Society of America, 61,* 1567–1576.

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

Middlebrooks, J. C., Makous, J. C., & Green, D. M. (1989). Directional sensitivity of sound-pressure levels in the human ear canal. *Journal of the Acoustical Society of America, 86,* 89–108.

Mills, A. W. (1972). Auditory localization. In J. V. Tobias (Ed.), *Foundations of modern auditory theory, Vol. II* (pp. 301–345). New York: Academic Press.

Minsky, M., Ming, O., Steele, O., Brooks, F. P., & Behensky, M. (1990). Feeling and seeing: Issues in force display. *Computer Graphics, 24,* 235–243.

Downloaded from http://direct.mit.edu/pvar/article-pdf/1/1/80/1622491/pres.1992.1.1.80.pdf by UNIVERSITY OF CENTRAL FLORIDA, GREG WELCH on 04 August 2024

Mowbray, G. H., & Gebhard, J. W. (1961). Man's senses as informational channels. In H. W. Sinaiko (Ed.), *Human factors in the design and use of control systems* (pp. 115–149). New York: Dover.

O'Leary, A., & Rhodes, G. (1984). Cross-modal effects on visual and auditory object perception. *Perception and Psychophysics, 35,* 565–569.

Oldfield, S. R., & Parker, S. P. A. (1984a). Acuity of sound localisation: A topography of auditory space. I. Normal hearing conditions. *Perception, 13,* 581–600.

Oldfield, S. R., & Parker, S. P. A. (1984b). Acuity of sound localisation: A topography of auditory space. II. Pinna cues absent. *Perception, 13,* 601–617.

Oldfield, S. R., & Parker, S. P. A. (1986). Acuity of sound localisation: A topography of auditory space. III. Monaural hearing conditions. *Perception, 15,* 67–81.

Patterson, R. R. (1982). *Guidelines for auditory warning systems on civil aircraft.* (Paper No. 82017), London: Civil Aviation Authority.

Perrott, D. R. (1982). 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.

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

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

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

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

Persterer, A. (1989, October). A very high performance digital audio signal processing system. [Summary]. *Proceedings of the IEEE ASSP Workshop on Applications of Signal Processing to Audio and Acoustics,* New Paltz, NY.

Persterer, A. (1991). Binaural simulation of an "ideal control room" for headphone reproduction. (Preprint 3062 (K-4)). *90th Convention of the Audio Engineering Society,* Paris.

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

Posselt, C., Schroter, J., Opitz, M., Divenyi, P., & Blauert, J. (1986). Generation of binaural signals for research and home entertainment. (Paper B1-6). *Proceedings of the 12th International Congress on Acoustics,* Toronto.

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

Richter, F., & Persterer, A. (1989). Design and applications of a creative audio processor. (Preprint 2782 (U-4)). *86th Convention of the Audio Engineering Society,* Hamburg.

Shaw, E. A. G. (1974). The external ear. In W. D. Keidel & W. D. Neff (Eds.), *Handbook of sensory physiology, Vol. V/1, Auditory system* (pp. 455–490). New York: Springer-Verlag.

Shaw, E. A. G. (1975). The external ear: New knowledge. In S. C. Dalsgaard (Ed.), *Earmolds and Associated Problems. Proceedings of the 7th Danavox Symposium, Scandinavian, Audiology, Suppl. 5,* 24–50.

Smith, S., Bergeron, R. D., & Grinstein, G. G. (1990). Stereophonic and surface sound generation for exploratory data analysis. *Proceedings of CHI'90. ACM Conference on Human Factors in Computing Systems,* Seattle, 125–132.

Sutherland, I. E. (1968). Head-mounted three-dimensional display. *Proceedings of the Fall Joint Computer Conference, 33,* 757–764.

Thurlow, W. R., & Runge, P. S. (1967). Effects of induced head movements on localization of direction of sound sources. *Journal of the Acoustical Society of America, 42,* 480–488.

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

Toole, F. E. (1969). In-head localization of acoustic images. *Journal of the Acoustical Society of America, 48,* 943–949.

Van Veen, B. D., & Jenison, R. L. (1991). Auditory space expansion via linear filtering. *Journal of the Acoustical Society of America, 90.*

Wallach, H. (1939). On sound localization. *Journal of the Acoustical Society of America, 10,* 270–274.

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

Wallach, H., Newman, E. B., & Rosenzweig, M. R. (1949). The precedence effect in sound localization. *American Journal of Psychology, 57,* 315–336.

Warren, D. H., Welch, R. B., & McCarthy, T. J. (1981). The role of visual-auditory "compellingness" in the ventrilo-

Downloaded from http://direct.mit.edu/pvar/article-pdf/1/1/80/1622491/pres.1992.1.1.80.pdf by UNIVERSITY OF CENTRAL FLORIDA, GREG WELCH on 04 August 2024

quism effect: Implications for transitivity among the spatial senses. *Perception and Psychophysics, 30,* 557–564.

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

Wenzel, E. M., Wightman, F. L., & Foster, S. H. (1988a). Development of a three-dimensional auditory display system. *SIGCHI Bulletin, 20,* 52–57.

Wenzel, E. M., Wightman, F. L., & Foster, S. H. (1988b). A virtual display system for conveying three-dimensional acoustic information. *Proceedings of the Human Factors Society, 32,* 86–90.

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

Wenzel, E. M., Stone, P. K., Fisher, S. S., & Foster, S. H. (1990). A system for three-dimensional acoustic "visualization" in a virtual environment workstation. *Proceedings of the IEEE Visualization '90 Conference,* San Francisco, 329–337.

Wenzel, E. M., Wightman, F. L., & Kistler, D. J. (1991). Localization of nonindividualized virtual acoustic display cues. *Proceedings of the CHI'91. ACM Conference on Computer-Human Interaction,* New Orleans, 351–359.

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

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

Downloaded from http://direct.mit.edu/pvar/article-pdf/1/1/80/1622491/pres.1992.1.1.80.pdf by UNIVERSITY OF CENTRAL FLORIDA, GREG WELCH on 04 August 2024