Contents

List of Figures

List of Tables

The Design of Multidimensional Sound Interfaces

Michael Cohen
Human Interface Lab
University of Aizu 965-80
Japan
voice: [+81](242)37-2537
fax: [+81](242)37-2549
email: mcohen@u-aizu.ac.jp

Elizabeth M. Wenzel
NASA Ames Research Center, MS 262-2
Moffett Field, CA 94035
USA
voice: [+1](415)604-6290/5198
fax: [+1](415)604-3729
email: beth@aurora.arc.nasa.gov

1 Introduction

1.1 I/O Generations and Dimensions

Early computer terminals allowed only textual I/O. Because the user read and wrote vectors of character strings, this mode of I/O (**c**haracter-based **u**ser **i**nterface, or "CUI") could be thought of as one dimensional, 1D. As terminal technology improved, users could manipulate graphical objects (via a **g**raphical **u**ser **i**nterface, or "GUI") in 2D. Although the I/O was no longer unidimensional, it was still limited to the planar dimensionality of a CRT or tablet. Now there exist 3D spatial pointers and 3D graphics devices; this latest phase of I/O devices [Bla92] [BD92] [Rob92] approaches the way that people deal with "the real world." 3D audio (in which the sound has a spatial attribute, originating, virtually or actually, from an arbitrary point with respect to the listener) and more exotic spatial I/O modalities are under development.

The evolution of I/O devices can be roughly grouped into generations that also correspond to the number of dimensions. Representative instances of each technology are shown in Table 1. This chapter focuses on the italicized entries in the third generation aural sector.

1.2 Exploring the Audio Design Space

Audio alarms and signals of various types have been with us since long before there were computers. But even though music and visual arts are considered sibling muses, a disparity exists between the exploitation of sound and graphics in interfaces. (Most people think that it would be easier to be hearing- than sight-impaired, even though the incidence of disability-related cultural isolation is higher among the deaf than the blind.) For whatever reasons, the development of user interfaces has historically been focused more on visual modes than aural.

This imbalance is especially striking in view of the increasing availability of sound in current technology platforms. Sound is frequently included and utilized to the limits of its availability or affordability in personal computers. However, computer-aided exploitation of audio bandwidth is only beginning to rival that of graphics. General sound capability is slowly being woven into the fabric of applications. Indeed, some of these programs are inherently dependent on sound—

generation/ dimension	mode	input	output
first/1D	textual	keyboard	teletype monaural sound
second/2D	planar	trackball, joystick mouse touchpad light pen	graphical displays stereo sound
third/3D	aural	speech recognition head-tracking	speech synthesis MIDI *spatial sound* *filtears*
	haptic	3D joystick, spaceball DataGlove mouse, bird, bat wand handwriting recognition	tactile feedback: vibrating fingertips force-feedback devices Braille devices, tactor arrays
	olfactory	??	smell emitters
	gustatory	??	?
	visual	head- and eye-tracking	projection systems stereoscopes: **h**ead-**m**ounted **d**isplays holograms vibrating mirrors

Table 1: Generations and dimensions of I/O devices

voicemail, or voice annotation to electronic mail, teleconferencing, audio archiving— while other applications use sound to complement their underlying functionality. Table 2 (extended from [Dea72, p. 124] and [SM87, p. 148]) lists some circumstances in which auditory displays are desirable.

Because of the cognitive overload that results from overburdening other systems (perhaps especially the visual), the importance of exploiting sound as a full citizen of the interface, developing its potential as a vital communication channel, motivates the exploration of both analogues to other modes of expression and also the evolution of models unique to audio. Computer interfaces present special needs and opportunities for audio communication.

This chapter reviews the evolving state of the art of non-speech audio interfaces, driving both spatial and non-spatial attributes. While we discuss both issues, our emphasis is on neither the backend, the particular hardware needed to manipulate sound, nor the frontend, the particular computer conventions used to specify the control. Rather, the chapter is primarily concerned with their integration— crafting effective matches between projected user desires and emerging technological capabilities.

2 Characterization and Control of Acoustic Objects

Part of listening to a mixture of conversation or music is being able to appreciate the overall blend while also being able to hear the individual voices or instruments separately. This synthe-

- when the origin of the message is itself a sound (voice, music)
- when other systems are overburdened (simultaneous presentation)
- when the message is simple and short (status reports)
- when the message will not be referred to later (time)
- when the message deals with events in time ("Your background process is finished.")
- when warnings are sent, or when the message prompts for immediate action ("Your printer is out of paper.")
- when continuously changing information of some type is presented (location, metric, or countdown)
- when speech channels are fully employed
- when a verbal response is required (compatibility of media)
- when illumination or disability limits use of vision (an alarm clock)
- when the receiver moves from one place to another (employing sound as a ubiquitous I/O channel)

Table 2: Motivation for using sound as a display mode

sis/decomposition duality is the opposite effect of masking: instead of sounds hiding each other, they are complementary and individually perceivable. For instance, musical instruments of contrasting color are used against each other. Localization effects contribute to this anti-masking by helping the listener distinguish separate sources, be they instruments in an ensemble or voices in the cacophony of a cocktail party [Bla83, p. 257] [Aro92].

Audio imaging is the creation of sonic illusions by manipulation of stereo channels. For instance, when classical music is recorded, the music from different instruments comes from distinctly different directions. The violins are on the listener's left; the cellos and double basses are on the right; the violas face the listener; and the percussion, woodwinds, and brass are to the rear of the orchestra.

In a stereo system, the sound really comes from only the left and right transducers, whether headphones or loudspeakers. Typical audio systems project only a one-dimensional arrangement of the real or mixed sources. In traditional sound reproduction, the apparent direction from which a sound emanates is typically controlled by shifting the balance of the unmodified sound source between the left and right channels. However, this technique yields images that are diffuse, and located only between the speakers.

Spatial sound involves technology that allows sound sources to have not only a left–right attribute (as in a conventional stereo mix), but up–down and back–forth qualities as well. It is related to, but goes beyond, systems like quadraphonics and surround sound.[1] Augmenting a sound system with spatial attributes opens new dimensions for audio, making spatial sound a

[1] *Surround Sound 360* and THX are two commercial examples of theatrical audio systems, as *Circle Vision 360* and *Omnimax* are examples of analogous visual systems.

potentially rich analogue of 3D graphics.

Clearly sound has many other qualities besides spatial attributes which contribute to its perceptual and cognitive organization. The various widely discussed [PF54] [BB87, p. 396] [Bly87, p. 420] [Man87, p. 422] dimensions of sound generally include the attributes shown in Table 3. Just as with spatial dimensions, such dimensions can be utilized in an information display context to encourage the perceptual segregation and systematic organization of virtual sources within the interface. Following from [Gib79]'s ecological approach to perception, the audible world can be conceived of as a collection of acoustic "objects." In addition to spatial location, 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 can systematically manipulate these features, effectively creating an auditory symbology which operates on a continuum from "literal" everyday sounds, such as the rattling of bottles being processed in a bottling plant [GSO91], to a completely abstract mapping of statistical data into sound parameters [SBG90]. Principles for design and synthesis can also be gleaned from the fields of music [BSG89], psychoacoustics [Pat82], user interface design [BD92], and higher-level cognitive studies of the acoustical determinants of perceptual organization [BGB89] [Bre90].

Another obvious aspect of "everyday listening" [Gav86] is the fact that we live and listen in a three-dimensional world. Thus, a critical advantage of the binaural 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, 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." The auditory system can provide a more coarsely-tuned mechanism to direct the attention of our more finely-tuned visual analyses, as suggested by the effective linkage between direction of gaze (eye and head movements) and localization accuracy [PSBS90] [SMP92]. In fact, Perrott, and his colleagues [PSSS91] have recently 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; for example, in air traffic control displays for the tower or cockpit [BW92].

- harmonic content
  - pitch and register: tone, melody, harmony
  - waveshape (sawtooth, square, triangle, ...)
  - timbre, filtears, vibrato, and equalization
- dynamics
  - intensity/volume/loudness
  - envelope: **a**ttack, **d**ecay, **s**ustain, **r**elease (volume shape)
- timing
  - duration
  - tempo
  - repetition rate
  - duty cycle
  - rhythm and cadence
  - syncopation
- spatial location
  - direction: azimuth, elevation
  - distance/range
- ambience: presence, resonance, reverberance, spaciousness
- representationalism: literal, everyday ("auditory icons") $\leftrightarrow$ abstract ("earcons")

Table 3: Dimensions of sound

Given multiple audio channels, a display[2] system needs a way of perceptually segmenting or distinguishing them from each other. A simple method is by just making the channels of interest louder than their siblings. Spatial sound enhances stream segregation by allowing auditory localization, invoking the 'cocktail party effect.' The cocktail party effect refers to a phenomenon described in the literature [Che53] [Aro92] on binaural hearing in which sound source intelligibility is shown to improve when listening dichotically (with two ears), compared to monotically (with one ear). Thus, at a party with many simultaneous conversations, a mingler can still follow any particular exchange by filtering according to

- position
- speaker voice
- subject matter.

Similarly, someone listening to a song distinguishes the streams (voices, instruments, parts) by

- position
- tone/timbre
- melodic line and rhythm.

2.1 Spatial Dimensions of Sound

The goal of spatial sound synthesis is to project audio media into space by manipulating sound sources so that they assume virtual positions, mapping the source channel into three-dimensional space (the perceptual envelope around the sink[3]). These virtual positions enable auditory localization, a listener's psychological separation in space of the channels, via space-domain multiplexing. The simulation techniques being developed to achieve this goal depend critically on our understanding of the perceptual or psychoacoustical cues used by human listeners when localizing sounds in the real world.

Much of our understanding of human sound localization is based on the classic "duplex theory" [Lor07] which emphasizes the role of two primary cues to location, interaural differences in time of arrival and interaural differences in intensity. The original proposal was that **i**nteraural **i**ntensity **d**ifferences (IIDs) resulting from head-shadowing determine localization at high frequencies, while **i**nteraural **t**ime **d**ifferences (ITDs) were thought to be important only for low frequencies (because of phase ambiguities occurring at frequencies greater than about 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 are used if the signals have relatively slow envelope modulations. The duplex theory also cannot account for the ability of subjects to localize sounds along the median plane where interaural cues are minimal (e.g., see [Bla83]). Further, when subjects listen to sounds over headphones, they are usually perceived as being inside the head even though interaural temporal and intensity differences appropriate to an external source location are present [Ple74]. Many studies

[2]Throughout this chapter, "display" is used in a general sense to denote presentation or output in any medium.

[3]Since the word "speaker" is overloaded, meaning both "loudspeaker" and "talker," "source" is used to mean both, denoting any logical sound emitter. Similarly and symmetrically, "sink" is used to describe a logical sound receiver, a virtual listener.

now suggest that these deficiencies of the duplex theory reflect the important contribution to localization of the direction-dependent filtering which occurs when incoming sound waves interact 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 upon frequency. Such frequency-dependent effects, or filtering, also vary greatly with the direction of the sound source, and it is clear that listeners use such effects to discriminate one location from another. Experiments have shown that spectral shaping by the pinnae is highly direction-dependent [Sha74], that the absence of pinna cues degrades localization accuracy [RB68] [GG73], and that pinna cues are important for externalization or the "outside-the-head" sensation [Ple74] [DRP+92].

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 all of these effects can be coalesced into a single filtering operation, much like the effects of an 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. The acoustic shaping by the two ears is then measured by recording the outputs of small probe microphones placed inside an individual's (or an artificial head's; e.g., the KEMAR [BS75] or Neumann heads) ear canals (Figure 1). If the measurement of the two ears occurs simultaneously, the responses, when taken together as a pair of filters, include estimates 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 1: 3D auditory display: synthesis technique

Filters constructed from these ear-dependent characteristics are examples of **f**inite **i**mpulse **r**esponse (FIR; also known as tapped delay line) filters and are often referred to as **h**ead-**r**elated **t**ransfer **f**unctions (HRTFs). Here, HRTF-filtering in the frequency domain manifests as a point-by-point multiplication operation, while FIR-filtering in the time domain occurs via a somewhat more complex operation known as convolution. By filtering an arbitrary sound with these HRTF-based "earprints," it is possible to impose spatial characteristics on the signal such that it apparently emanates from the originally measured location. Of course, the localizability of a 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 increase the bandwidth of the original signal; it merely transforms frequency components that are already present.

A closely-related issue to spectral content in the localizability of sound sources is their degree of familiarity. A variety of research indicates that (monaural) features such as peaks and valleys in the spectrum of a sound change systematically with location and appear to be the primary cues for elevation [RB68] [Bla70] [BH83] [But87]. Logically, in order for localization to be accurate, spatial cues other than the interaural cues— e.g., cues related to spectral shaping by the pinnae— must be interpreted in light of the original spectrum of the sound source. In effect, the listener must "know" *a priori* what the spectrum of a sound is in order to determine whether a particular feature was "shaped" by the effects of his/her ear structures or was simply present in the source spectrum. In the absence of other disambiguating information, many different spectra could be confused for the same location, and indeed this is often the case [Bla70] [BH83], suggesting that listeners' *a priori* knowledge of source spectra is imperfect. Thus the perception of elevation and relative distance, which both depend heavily on the detection of spectral differences, tend to be superior for familiar signals like speech [Col63] [Bla83, p. 104] [BW93]. Similarly, spectral familiarity can be established through training [Bat67].

It should be noted that the spatial cues provided by HRTFs, especially those derived from simple anechoic (free-field, 'dry,' 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 perceptual consequences of synthesizing spatial cues by using a less complex, and therefore more tractable, 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 stationary sound source in the free-field. One is a relatively small error in resolution ranging from about 1 to 20°, depending upon the experimental paradigm used to estimate localization acuity. In general these paradigms fall into three categories: methods of adjustment [STFJ55] which require the subjects to adjust the position of one source to match that of another; discrimination experiments such as those reported by [Mil58] [Mil72] which ask subjects to detect whether two successive sounds have changed position; and absolute judgment paradigms which simply ask the subjects to identify the source location by methods such as verbal report or pointing [SN36] [OP84a] [OP84b] [OP86] [WK89a] [WK89b] [MM90] [WAKW93]. Discrimination experiments tend to be constrained primarily by peripheral sensory limitations and measure a **j**ust **n**oticeable **d**ifference (JND) or the sensitivity of the subject to various localization cues. Which localization cues are most relevant to a particular JND measurement will depend on the stimulus conditions. For example, discrimination errors are smallest for stimuli in the horizontal plane where the interaural cues are presumably the most important, with a slight auditory "fovea" in that acuity is best directly in front (about 1 to 5°) and worsens for locations out to the side (about 5 to 10°). Absolute judgment paradigms, on the other hand, may be more affected by factors like memory limitations and context effects, and thus are probably more closely related to the conditions that one generally experiences when localizing sounds in a

virtual display (simply deciding "where is it?"). Error measures under these circumstances can be considerably larger (about 10 to 20° or more), especially for sources in the rear. There also seems to be a general tendency for errors in elevation to be somewhat larger than for azimuth, although this may depend upon the region of space relative to the listener that is being probed. Of course, error estimates will also be dependent in fairly complex ways upon the bandwidth, duration, spatial span, etc. of the stimuli being localized. For example, the classic study by [SN36] showed that error magnitudes are dependent on the stimulus frequency, with the greatest errors occurring around 3000 Hz where the interaural phase and level cues are both weakest. For a more complete discussion of the many factors affecting localization acuity and sensitivity measures, see the recent review by [MG91].

Another class of error observed in nearly all localization studies is the occurrence of front↔back reversals. These are judgements which indicate that a source in the front (rear) hemisphere was perceived by the listener as if it were in the rear (front) hemisphere. Reversal confusions in elevation, with up locations heard as down, and vice versa, have also been observed [WWK91] [WAKW93]. Although the reasons for such reversals are 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" [Mil72]. 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 source, a conical shell describing the locus of all possible sources. Obviously, the true situation is more complicated; the head is not really a simple sphere with two symmetric holes. However, to a first approximation, the model does seem to predict the pattern of interaural cues actually measured for static sources [Kuh77] [MMG89] [MG90]. While the rigid sphere model is not the whole story, the observed ITD and IID data indicate that the interaural characteristics of the stimulus are inherently ambiguous. In the absence of other cues, both front↔back and up↔down reversals (in fact, confusions between any two points along a particular cone) 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 [Bla83, p. 107–116]). For stationary sounds, such cues would essentially be the only clue to disambiguating source location. With dynamic stimuli, however, the situation improves considerably. For example, some studies have shown that allowing or inducing head-motion improves localization ability by substantially reducing the rate of reversals [Bur58] [TR67] [FF68]. With head-motion, the listener can potentially disambiguate front/back locations by tracking changes in the magnitude of the interaural cues over time; for a given lateral head movement, ITDs and IIDs for sources in the front will change in the opposite direction compared to sources in the rear [Wal40].

Another type of localization error is known as **i**n-**h**ead **l**ocalization (IHL). That is, sources sometimes fail to externalize, particularly when the signals are presented over headphones, although IHL has also been observed for real sources [Too69] [Ple74]. The tendency to localize sound sources inside the head is increased if the signals are unfamiliar [Col63] [Gar68] or derived from an anechoic environment [Ple74]. Thus, the use of familiar signals combined with cues that provide a sense of 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 [Col63] [Gar68] [Law73] [Ple74] [MK75], [MB79]. For example, [Beg92] recently investigated the effects of synthetic reverberation on the perceived externalization of static, virtual sound sources. He found that, compared to anechoic stimuli, adding reverberant cues nearly

eliminated IHL but tended to decrease localization accuracy while having no systematic effect on front↔back confusions. There is also some suggestion that head motion may also be a factor in externalization [Wen92, p. 87].

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 which determine distance perception [Col63] [Law73] [MK75] [MB79] [SL93]. Distance judgements 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; i.e., the inverse-square law, which implies a 6 dB decrease in intensity with each doubling of distance. For example, [Beg91] has reported that a 9 dB increase in intensity is required to produce a halving of the apparent distance of a sound source. Also, as noted above, distance perception also depends heavily on factors like stimulus familiarity.

The addition of environmental effects can complicate the perception of location in other ways. [vB60] reports that the spatial image of a sound source grows larger and increasingly diffuse with increasing distance in a reverberant environment, a phenomenon which may tend to interfere with the ability to judge the direction of the source. This problem may be mitigated by the phenomenon known as precedence [WNR49]. 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 which 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 which can confuse the apparent direction of the original source.

However, 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 timbral differences between front and rear sources. By taking advantage of a head-tracker in realtime systems, the loop between the auditory, visual, vestibular, and kinesthetic systems can be closed, and we can study the effects of dynamic interaction with relatively complex, but known, acoustic environments. The specific parameters used in such models must be investigated carefully if localization accuracy is to remain intact. It may be possible to discover an optimal trade-off between environmental parameters which enhance externalization and distance perception while minimizing the impact of the resulting expansion of the spatial image which can interfere with the ability to judge the direction of the source.

The above discussion of the perception of localized sound sources 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. See [MG91], [Mol92], and [Wen92] for somewhat more detailed overviews of localization cues and their synthesis. For an extensive discussion of spatial sound in general, the reader is referred to the in-depth review by [Bla83].

2.2 Implementing Spatial Sound

Perhaps the most direct approach to simulating spatial sound distributes sources by physically locating loudspeakers in the place where each source is located, relative to the listener. These loudspeakers could be statically placed, or perhaps moved around by mechanical means. How-

ever, such an implementation is cumbersome and certainly not portable. Other approaches use analytic mathematical models of the pinna and other body structures [Gen86] in order to directly calculate acoustic responses or, alternatively, provide a simplified model of the essential features of previously measured responses of the ear [KW91]. A third approach to accurate realtime spatialization, which is generally emphasized here, concentrates on **d**igital **s**ignal **p**rocessing (DSP) techniques for synthesizing spatial cues from direct measurements of HRTFs.

By measuring, simulating, or modeling the important cues to localization represented in the HRTFs (usually with DSP), many scientists are developing ways of generating and controlling this multidimensional sound imagery [Cho70] [Cho77] [Mar86] [WWF88a] [WWF88b] [Mar89] [Sco89] [SWKE89] [Fis90] [KMW90] [LHC90] [WSFF90] [BW92] [Wen92] [WAKW93]. The goal of such a sound spatializer is to create the impression that the sound is coming from different sources and different places, just like one would hear "in person." Such a device assigns each source a virtual position with respect to the sink, or listener, and simulates the corresponding auditory positional cues. A display based on this technology exploits the human ability to quickly and subconsciously localize sound sources.

The most frequently used approach to spatial sound generation employs a hardware- or software-based convolution engine that convolves a monaural input signal with pairs of (FIR) digital audio filters to produce output signals for presentation over stereo loudspeakers or headphones. As discussed above, binaural localization cues may be captured by HRTFs, measured for the head and pinna (outer ear) of human or artificial heads in an anechoic environment. For each spherical direction, a left–right pair of these transfer functions is measured, transformed to the time domain, and then stored as FIR filter coefficients [KM84] [Geh87] [ME88] [RP89] [Wen92].

The anechoic implementation of spatial sound described above is often called 'dry'; it includes no notion of a virtual room, and hence no echoes. Conversely, spatial reverberation is a 'wet' technique for simulating the acoustic information used by people listening to sounds in natural environments. A spatial reverberation system creates an artificial ambient acoustic environment by simulating echoes consistent with the placement of both the source and the sink (listener) within a virtual room.

There are two classes of generated echoes: early reflections, which are discretely generated (delayed), and late-field reverberation, which are continuous and statistically averaged. The early reflections are the particular echoes generated by the source, and the late field reverberation is the non-specific ambience of the listening environment. The early reflections off the floor, walls, and ceiling, provide indirect sound to the listener which can have important perceptual consequences. For example, [vB60], [MK75], [Cho77], [MB79], [KM84], and others have demonstrated that the ratio of direct to indirect sound can influence the perceived distance of sound sources.

In practice, room modeling is often limited to rectangular prismatic rooms. This symmetry allows an algorithm such as direct ray-tracing to be used to efficiently determine the propagation delay and direction of individual reflections, which are then spatialized, as if they were separate sources. Each separately spatialized audio source, incident or reflected, requires processing by a separate pair of binaural transfer functions.[4] Late field reverberation reflects the ambience of the virtual auditorium or listening room. Reverberant implementations of spatial sound, as discussed or instantiated by [KM84] [KMF$^+$86a] [KMF$^+$86b] [Mar87], employ a recursive, or **i**nfinite **i**mpulse **r**esponse (IIR) section to yield dense global reverberation effects. A filter that

[4]Often, this implies a hardware implementation which devotes a separate (mono→stereo) DSP channel to each image source. Alternatively, as in the Convolvotron (described in § 2.2.1), an aggregate binaural impulse response composed of the superposition of the direct and reflected images can be computed on-the-fly for the instantaneous configuration of the source and sink in the environment and then rendered in realtime.

combines early reflections with late field reverberation is sometimes called TDR, for **t**apped-**d**elay-plus-**r**ecirculation. The simulation thus models cues to perceived sound direction, sound distance, and room characteristics. This combination of high-order recursive and non-recursive filters enables a spatial reverberation system to implement descriptions of such characteristics as room dimensions and wall absorption, as well as time-varying source and listener positions. Thus, given a monophonic sound source and a specification of source and sink position and motion in a model room, the spatial reverberator approximates the sound field arriving at the model listener's ear drums. In general, however, most of the realtime systems currently available do not implement the full complement of room response characteristics outlined above. Such room modeling requires enormous computational resources and is only beginning to be developed in truly interactive, realtime displays.

2.2.1 Crystal River Engineering Convolvotron

The Crystal River Convolvotron™ is a convolution engine [WWF88b] that spatializes sound by filtering audio channels with transfer functions that simulate positional effects (see [Wen92]). Other recent devices include the Alphatron and Acoustetron II which are based on lower-cost DSP chips and reduced-complexity algorithms and the Snapshot system which allows one to quickly measure individualized HRTFs in any environment. Specifically, HRTFs, in the form of FIRs, are measured using techniques adapted from [MM77]. 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. Small probe microphones are placed near each eardrum of a human listener who is seated in an anechoic chamber [WK89a]. Wide-band test stimuli are presented from one of 144 equidistant locations in the free-field (non-reverberant) environment; 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 [WK89a]. The advantage of this technique is that it preserves the complex pattern of interaural differences over the entire spectrum of the stimulus, capturing the effects of filtering by the pinnae, head, shoulders, and torso. In order to synthesize localized sounds, a map of "location filters" is constructed from all 144 pairs of FIR filters by first transforming them to the frequency domain, removing the spectral effects of the headphones to be used during playback using Fourier techniques, and then transforming back to the time domain. An overview of the perceptual viability of the basic synthesis technique can be found in [Wen92].

In the Convolvotron, designed by Scott Foster of Crystal River Engineering [Fos90], the map of corrected FIR filters is downloaded from a host computer (IBM-compatible PC) to the dual-port memory of a realtime digital signal processor (Figure 2). This set of two printed-circuit boards converts one or more monaural analog inputs to digital signals at a rate of 50 kHz with 16-bit resolution. Each data stream is then convolved with filter coefficients (128 to 512 coefficients/ear; 24-bit integer arithmetic) determined by the coordinates of the desired target locations and the position of the listener's head, '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 simultaneous anechoic sources with an aggregate computational speed of more than 300 million **m**ultiply-accumulate **i**nstructions **p**er **s**econd (MIPS). This processing speed is also sufficient for interactively simulating a single source plus six first-order reflections (28 sources and reflections in the Acoustetron,™ a four-Convolvotron system in a single host computer) with variable sur-

face absorption characteristics in relatively small reverberant environments with head-tracking [FWT91] [SSN93]. The hardware design can also be scaled upward to accommodate additional sources and longer filter lengths required for simulating larger enclosures. The Beachtron, a less costly version of the system for the PC, is capable of spatializing two audio input channels (comparable to Focal Point, described later in § 2.2.2). Currently, this system is anechoic and uses minimum-phase approximations of HRTFs [KW91] [WKA92], which allow a considerable reduction in filter size with minimal perceptual disruption (75 coefficients/ear, 16-bit conversion, 44.1 kHz sampling rate). The Beachtron also includes an onboard Proteus/1XR synthesizer and MIDI control.

Figure 2: The Convolvotron: high-speed realtime **d**igital **s**ignal **p**rocessor

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 [WF93]. The interpolation algorithm effectively computes a new coefficient at the sampling interval (about 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 (like [Pol87]), the listener's head position can be monitored in realtime so that the sources are stabilized in fixed locations or in motion trajectories relative to the user. Again, such head-coupling helps enhance the simulation, since head movements are important for localization [Wal40] [TR67]. This degree of interactivity, especially coupled with smooth motion interpolation and simulation of simple reverberant environments, is apparently unique to the Convolvotron system. In addition, all source code is provided to facilitate the Convolvotron's use as a research tool.

As with any system required to compute data "on the fly," the term 'realtime' is a relative one. The Convolvotron, including the host computer, has a computational delay of about 30–40 ms, depending upon such factors as the number of simultaneous sources, the duration of the HRTFs used as filters, and the complexity of the source geometry. An additional latency of

at least 50 ms is introduced by the head-tracker.[5] 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 Convolvotron updates to a new location about every 90 ms (including a 50 ms delay from a low-cost headtracker). This directional update interval, in turn, corresponds to an angular resolution of about 32° when the relative source–listener speed is 360 degrees/sec, 16° at 180 degrees/sec, and so on. Such delays may or may not result in a perceptible lag, depending upon 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 degrees/sec, minimum audible movement angles ranged from about 4 to 21°, respectively, for a 500 Hz tone-burst [Per82] [PT88]. Thus, slower relative velocities are well within capabilities of the Convolvotron, while speeds approaching 360 degrees/sec may begin to result in perceptible delays, especially when multiple sources or larger filters (e.g., simulations of reverberant rooms) are being generated.

2.2.2 Gehring Research Focal Point

Focal Point™ [Geh87] [Geh90] comprises two different binaural localization technologies, Focal Point Types 1 and 2. In most Focal Point products the audio is 44.1 kHz sampling rate with 16-bit CD quality. Focal Point Type 1 is the original Focal Point technology, utilizing time-domain convolution with HRTF-based impulse responses for anechoic simulation. It performs binaural convolution in realtime on any audio signal and is portable to most DSP and RISC environments; Motorola DSP-based versions for the PC and Macintosh platforms are widely used. Several sets of HRTFs have been demonstrated over the years Focal Point has been available. (One set was measured by the Computer Music Group at Northwestern University using a KEMAR mannequin [BS75].) The current Focal Point HRTFs provide neutral timbre, suitable for music, entertainment, and VR applications.

Typically, Focal Point Type 1 software is downloaded into the DSP upon startup and then updated only when a source is moved; audio processing continues without host CPU interaction, except to reposition a source (by setting three integers). Updating the transfer function in the DSP has a latency of about 3–6 ms, which compares favorably with known visual displays. This update rate, which can be in excess of 300 Hz, is suitable for rapid source motion with respect to the listener.

The Mac and PC versions of Focal Point Type 1 are encapsulated; transfer function synthesis is performed within Focal Point, rather than by the host CPU. This means the entire host resource is available for other applications, such as soundfile playback through Focal Point, direct-to-disc recording concurrently with Focal Point binaural processing, or 3D graphics. The PC version is available as a consumer product.

Focal Point Type 2 (patent pending) is a Focal Point implementation in which sounds are preprocessed offline, creating interleaved soundfiles which can then be positioned in 3D in realtime upon playback. Compared to realtime convolution systems such as Focal Point Type 1,

[5] An estimate of 50 ms for the effective latency of the head-tracker is probably conservative. It also does not reflect the potentially much more problematic issue of positional "jitter" in current tracking systems [MAB93]. For the Convolvotron, and probably all the realtime spatial sound systems described here, latencies through the total system are dominated by the limitations of commercially-available tracking systems. Such latencies and positional jitter are not as critical for the human auditory system, which is less sensitive to changes in angular displacement, as for the human visual system.

Type 2 is very economical; DSP and other high-speed processing are not required. Although sounds must be preprocessed in advance, Type 2 positioning is very fast, since no convolution pipeline delay is involved and positioning latency is measured in microseconds. Such a process also multiplies the storage requirements of the original sound sample, since a new, spatialized, sample is generated (offline) for each possible position. Type 2 can be used on almost any platform or soundcard with stereo audio capability.

Focal Point Types 1 and 2 can also be MIDI-controlled, allowing notes and sounds to be interactively positioned during musical performance using the many MIDI-based software and hardware products. Focal Point development packages typically include source code and sample C projects for several applications, including headtracking, external control via an RS-232 link, and simultaneous soundfile playback.

2.2.3 AKG CAP (Creative Audio Processor) 340M

A kind of binaural mixing console, CAP (Creative Audio Processor) 340M, has been developed by AKG in Austria [AKG91], based partially on work by Blauert [Bla84]. The system is aimed at applications like audio recording, acoustic design, and psychoacoustic research [RP89]. This 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 specification of room response characteristics. The sampling rate of the system is 50 kHz with 16-bit floating point conversion (16-bit mantissa plus 3-bit exponent). FIR filters of 100 coefficients/ear are convolved in the time domain with an aggregate computational speed of 340 MFLOPS on 32-bit floating point arithmetic. The CAP 340M's room simulation algorithm appears to include the ability to impose realtime directional (HRTF) characteristics on the reflected images as well as the direct path, as is the case in the Convolvotron. It also allows simulation of late reverberation using IIR filters. A collection of HRTFs is offered, derived from measurements taken in the ear canals of both artificial heads and individual subjects. A more recent system, the Binaural Audio Processor (BAP 1000), simulates an ideal control room for headphone reproduction using realtime convolution with up to four binaural (HRTF-based) filters (two direct paths plus two mirror-image reflections). The user also has the option of having his/her individual transforms programmed onto a PROM card [Per91]. Interestingly, AKG's literature mentions that best results are achieved with individualized transforms. So far, the system has not been integrated with interactive head-tracking, so data regarding its motional capabilities are currently not available.

Similar projects in Europe are based on the most recent efforts of Blauert, Poesselt, Lehnert and their colleagues at the Ruhr University at Bochum, Germany [BLB77] [LB89] [PSO+86]. The group at Bochum has been working on a prototype DSP system, again a kind of binaural mixing console, whose proposed features include realtime convolution of HRTFs for up to four sources, interpolation between transforms to simulate motion, and room modeling. The group has also devoted substantial effort to measuring HRTFs for both individual subjects and artificial heads (e.g., the Neumann head), as well as developing computer simulations of transforms.

2.2.4 HEAD Acoustics

Another researcher in Germany, Klaus Genuit, has founded HEAD Acoustics to develop spatial audio systems. Genuit and his colleagues have also produced a realtime, eight-channel binaural mixing console using anechoic simulations as well as a new version of an artificial head [GG89] [GGK92]. The eight binaural channels can also be adapted to simulate simple room character-

istics using a direct path plus up to seven first-order reflections. Genuit's work is particularly notable for his development of a structurally-based model of the acoustic effects of the pinnae (e.g., [Gen86]). That is, rather than using measured HRTFs, Genuit has developed a parameterized, mathematical description (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 is mentioned. Also, since data can be calculated directly, the use of such a parameterized model may obviate the need to be concerned with the required spatial density of measured HRTFs or the nature of the interpolation between measurements needed to achieve smooth motion.

2.2.5 Roland Sound Space (RSS) Processor

Roland has developed a system known as the Roland Sound Space (RSS) processing system which attempts to provide realtime spatialization capabilities for both headphone and stereo loudspeaker presentation [Cha91]. The basic RSS system allows independent placement of up to four sources using time domain convolution (24-bit arithmetic) with FIR filters based on anechoic measurements of HRTFs for an individual human. Details regarding the length of the filters, the spatial density of the HRTF measurements, and methods of interpolation are not given. The sampling rate is switchable between 44.1 and 48 kHz with four 18-bit A/D converters and eight (four stereo pairs) 20-bit D/A converters. Spatial placement is controlled by MIDI input or by two rotary dials per channel which independently control azimuth and elevation.

Since the RSS is aimed primarily at loudspeaker presentation, it incorporates a technique known as transaural processing, or crosstalk cancellation between the stereo speakers. This additional filtering process is required to compensate for the fact that, in effect, speaker presentation causes the sound to be processed by HRTFs 'twice': once by the digital HRTF filters used to manipulate the spatial illusion, and once by the listener's own ears. The transaural technique 'divides out' or equalizes the effects of the speaker crosstalk, so that the transfer functions for the frontal direction become flat at the listener's ear canals. This technique seems to allow an adequate spatial impression to be achieved. As [Cha91] and many others have noted for such speaker presentation systems, localization accuracy suffers as soon as the listener deviates from a listening region near a locus of points equidistant from the two speakers (the "sweet spot"). A sense of increased auditory "spaciousness" (relative to normal stereo techniques) remains relatively intact, however, even for listening positions off the bisector. [Cha91] also notes that elevation was very difficult to discern with the RSS system. The RSS system can also be used for realtime control of spatial sound over headphones in a similar manner to the systems described above by disabling the transaural processing.

2.2.6 Mixels

The realtime systems described above provide a wide range of simulation capabilities which could be employed as the backend of an interface for a spatial auditory display. Thus, the specific attributes of a particular software control interface may or may not be instantiated, depending on which system is used to generate the spatial audio effects. For example, the simplest and least expensive systems available, such as a single-channel version of the Focal Point or Beachtron systems, will allow only the simplest of anechoic simulations for one or two

image sources. Conversely, the large and no doubt expensive array of equipment represented by the CAP 340M could potentially provide a much richer instantiation of a virtual acoustic display.

It is also important to note, however, that the perceptual viability of most of the systems described above (except for the Convolvotron) has not been demonstrated beyond the manufacturers' rather general claims about performance. If such a system is intended for research or other applications that require accurate localization, it is important that the user have access to details about the nature of the simulation techniques and algorithms used in the device. If, on the other hand, simple spatial effects or impression are the primary goal, then a more "black box" type of approach may be sufficient. Alternatively, a researcher may need or wish to independently test the perceptual validity of a particular device. Such testing will probably require at least some degree of access to the internal controls or software of the spatialization system.

Rather than delve more deeply into particular spatial audio systems, which will no doubt develop as rapidly in the future as the rest of computing technology, the remainder of this chapter concentrates on the nature of the control interfaces that will need to be developed to take full advantage of these new capabilities.

The assumption is that many individually spatialized audio channels will become available so that sound can be modeled in a granular fashion to create a "circumambience." The number of channels corresponds to the degree of spatial polyphony, simultaneously spatialized sound sources. By way of analogy to pixels and voxels, we sometimes call these atomic sounds "mixels," acronymic for sound **mix**ing **el**ements, since they form the raster across which a soundscape is projected, defining the granularity of control.

2.3 Non-Spatial Dimensions and Auditory Symbology

Auditory icons [Gav86] are acoustic representations of naturally occurring events that caricature the action being represented. For instance, in the Macintosh SonicFinder [Gav89], a metallic thunk represents a file being tossed into the trashcan upon deletion, and a liquid gurgling signifies a file being copied. "Earcons" [SBJG86] [BSG89] [BG89] are elaborated auditory symbols which compose motifs into artificial non-speech language, phrases distinguished by rhythmic and tonal patterns. Earcons may be combined (by juxtaposing these motifs), transformed (by varying the timbre, register, and dynamics), or inherited (abstracting a property). Infosound [SGH$^+$90] allows the combination of stored musical sequences and sound effects to be associated with application events, like Prokofiev's use of musical themes in *Peter and the Wolf*.

Auditory icons and earcons are classes along a continuum of display styles, from literal event or data representation to dynamic, symbolic representation, which may be more or less abstract. "Filtears" [Coh89] [CL91a] [CL91b] [Coh93], which depend on the distinction between sources and sinks, are one way of spanning this spectrum.

Sound is malleable under an infinite range of manipulation. Voice and music in particular can be gracefully distorted without loss of intelligibility or euphony. Even though audio channels can be perceptually segmented by virtual location, it is also important to have other attribute cues independent of direction and distance. Filtears are a class of such cues, audio filters implemented as separate attribute cues, superimposing information on sound signals by perceptually multiplexing the audio bandwidth.

Imagine a user tele-negotiating with several parties at once, including trusted advisors. Besides whatever spatial array of the various conferees, the user might want to give the advisors' voices a *sotto voce* attribute, perhaps by making their voices sound like whispers, imparting a suggestion of a private utterance, thereby tagging their voices as confidants. If some of the parties (perhaps including some of the advisors) are from outside the user's organization, their

voices might be given an *outside* attribute, perhaps by inhibiting any 'indoors-suggesting' reverberation, so that their voices seem to come from outside the building. These two separate dimensions of control could be used, separately or together (as in an "off-stage" whisper), to sonically label the voice channels, organizing them mnemonically. (Neither of these examples has actually been implemented yet. The filtears that have been deployed are detailed in § 5.2.2.)

Filtears are potentially useful for user interfaces because, unlike an audio zoom feature that simply makes the chosen speaker louder, the extended attributes introduced by spatial sound and filtears are separate from conventional dimensions of control, and they can be adjusted independently. Filtears can be thought of as sonic typography: placing sound in space can be likened to putting written information on a page, with audio emphasis equivalent to *italici*zing or em**bold**ening. Filtears embellish audio channels; they depend on the distinction between source and sink, and warp the channels in some way that is different from parametrizing an original signal.

It is important to note that, while filtears are intended to be perceptually orthogonal to other cues, such independence is difficult to achieve. Sound attributes interact in complex and often unpredictable ways, and such interactions must be taken into account when designing auditory symbologies and implementing them with filtear-type controllers/transformers.

3 Research Applications

Virtual acoustic displays featuring spatial sound can be thought of as enabling two performance advantages:

situational awareness Omnidirectional monitoring via direct representation of spatial information reinforces or replaces information in other modalities, enhancing one's sense of presence or realism.

multiple channel segregation By levering off 'natural noise cancellation' of the previously described cocktail party effect, spatial sound systems improve intelligibility, discrimination, and selective attention among audio sources in a background of noise, voices, or other distractions. Such enhanced stream segregation allows the separation of multiple sounds into distinct 'objects.'

Various application fields that exploit these enhanced capabilities are described below.

3.1 Sonification

Sonification can be thought of as auditory visualization, and has been explored by scientists [Bly82] [MFS84] [SC91] [BD92, Chap. 6] as a tool for analysis, for example, presenting multivariate data as auditory patterns. Because visual and auditory channels can be independent of each other, data can be mapped differently to each mode of perception, and auditory mappings can be employed to discover relationships that are hidden in the visual display. This involves some sort of mapping of the analyzed data to attributes like those outlined in Table 3. Various researchers [KF88] [Ken90] [Ken91] [WSFF90] [Wen94] suggest using spatial sound as a component of sonification, and researchers have designed tools for presenting data as an integrated visual and auditory display, whose stereophonic display correlates the sound with position on the monitor. Exvis [SBG90] interprets a scatterplot as a texture, a dense distribution of data, and then translates that texture into sound.

3.2 Auditory Displays for Visually Disabled Users

There is also increasing interest in providing auditory displays for visually disabled users [Van89] [LHC93]. Some researchers, including [LMCH83] [Man84] [MBJ85], have experimented with mapping x-y graphs to sound, to convey their information to blind users. An "auditory screen" [Edw87] [Edw88] uses a **w**indow, **i**con, **m**enu, **p**ointing device (WIMP) interface to associate musical sound and synthesized speech with tiled screen areas. SeeHear [NMM88] mapped visual signals from optically scanned objects into localized auditory signals. The Sonic Navigator [Set90] localizes synthesized speech to the location of the window being read.

3.3 Teleconferencing

If a voice can be likened to a long arm, with which one can reach across a room or down a flight of stairs to effectively tap someone on the shoulder, then the telephone lengthens that arm even further, stretching one's presence across continents, oceans, and beyond. Many scientists are exploring computer-controlled teleconferencing systems [SS87] [CK91c] [MKT+91] [MK91] [SY91] [TAMS91] [CK92c] [KCA92]. Major thrusts have protocols for invoking a rendezvous [KK86], suitable architectures for deploying such systems [SG85] [Lan86] [Lud89] [Koi91] [TKOK92], and graphical control [SBL+86] [SFB+87] [ADGM88] [SGH+90] [KS93].

3.4 Music

Musical applications [Moo83] [KM84] [BG88] [BO89] [CK91a] [CK93a] [Lez93] feature bouncing and dancing sound. Many spatializers have MIDI interfaces, allowing integration into musical systems. A listener can wander among a marching band or an embracing chord; a composer could program choreography for sonic dancers.

3.5 Virtual Reality and Architectural Acoustics

Virtual **r**eality (VR) systems are computer-generated interactive environments utilizing (typically head-mounted display) 3D graphic scenes and soundscapes, featuring a manual control [Fol87]. They are characterized by an intimate link between display and control, in which the user inhabits the system. Various VR researchers, including [FMHR86] [FWCM88] [WSFF90], have incorporated stereophonic output into their headmounted display. Direct representation of room geometry and absorption/reflection properties allows sonification of architectural acoustics for acoustical CAD/CAM [Ast93] [SSN93].

3.6 Telerobotics and Augmented Audio Reality

"Augmented reality" [CM92] [FMHS93] [FMS93] [WMG93] is used to describe hybrid presentations that overlay computer-generated imagery on top of real scenes. Augmented audio reality extends this notion to include sonic effects, overlaying computer-generated sounds on top of more directly acquired audio signals. Telepresence delegates a robot slave to act on behalf of the human master. Controlled from afar by a pilot wearing effectors corresponding to robot's sensors, the puppet, a surrogate with feedback, could venture into hazardous environments (fires, toxic waste, nuclear power plants, ...). By juxtaposing and mixing 'sampled' and 'synthetic' transmissions, scientists are exploring the relationship between telepresence and VR audio presentations: telepresence manifests as the actual configuration of sources in a sound field, as perceived by a

dummy-head, say; VR as the perception yielded by HRTF-filtering of virtual sources with respect to virtual sinks [CAK93].

team communication If several telerobots are working together, their pilots will likely want to communicate with each other. Instead of equipping each robot with a speaker, and letting each operator speak through the respective robot mouth, it is better to 'short-circuit' the communications path, transmitting a pilot's utterance to the other pilots directly, directionalizing to preserve the spatial consistency of the telepresence [AKKS92] [CKA92].

sonic cursor In telemonitoring, one wants to identify the location of a sound object. Using an augmented audio reality system, one could switch between a telemonitored binaural (dummy-head transmitted) and rendered source (synthesized by binaural DSP) to identify the location.

synesthetic alarm Synesthesia is the act of experiencing one sense modality as another, and can be used to further blur the distinction between actual and artificial worlds. A telerobot equipped to enter hazardous areas might have infrared or radiation meters. These meters could be thresholded, rendering the danger points as auditory alarms, easily superimposed on the auditory soundscape captured by the robot's ears.

4 Interface Control via Audio Windows

"Audio windows" is an auditory-object manager, one potentially powerful implementation of a user interface (frontend) to an audio imaging system. Here, the generalized control model of a window is (by analogy to graphical windows, as in a desktop metaphor) an organizational vehicle in the interface, and has nothing to do with room acoustics. Researchers [LP89] [LPC90] [CL91a] [CL91b] [CK91b] [FKS91] [CK92a] [KC93] have been studying applications and implementation techniques of audio windows for use in providing multimedia communications. The general idea is to permit multiple simultaneous audio sources, such as in a teleconference, to coexist in a modifiable display without clutter or user stress. The distribution of sounds in space is intended to realize some of the same kinds of benefits achieved by distribution of visual objects in graphical user interfaces.

A powerful audio imaging user interface would allow the positions of the audio channels to be arbitrarily set and adjusted, so that the virtual positions of the sinks and sources may be constantly changing as they move around each other and within a virtual room. By using an audio window system as a binaural directional mixing console, a multidimensional pan pot,[6] users can set parameters reflecting these positions. Members of a teleconference altering these parameters may experience the sensation of wandering around a conference room, among the teleconferees. Music lovers at a live or recorded concert could actively focus on a particular channel by sonically hovering over the shoulder of a musician in a virtual concert hall. Minglers at a virtual cocktail party might freely circulate. Sound presented in this dynamically spatial fashion is as different from conventional mixes as sculpture is from painting.

Spatial sound applications can be classified according to source (speaker) and sink (listener) mobility. The simplest spatial sound systems allow neither the sources nor the sinks to move. This kind of configuration is still useful for separating channels and, in fact, offers a good checkpoint to spatial sound applications under development; i.e., the several participants in a

[6] A **pan**oramic **pot**entiometer controls the placement of a channel in a conventional stereo mix.

conference call would project distinct sound images to each other, consistent with their relative virtual (if static) locations. With such a presentation, a user could more easily focus attention on a single speaker or instrument, especially with an audio spotlight (described later in § 5.2.2).

A simple demonstration of this functionality on a conventional system features three users, each with two telephones, calling each other cyclically (Figure 3). Each user's holding the calling and called handsets to different ears demonstrates one application of *stereotelephonics* [Coh87], the use of stereo effects in telephones.

Figure 3: Stereotelephonics and 3-way cyclic conferencing

A system in which the sources are stationary, but the listeners move about (like visitors at a museum) would be useful for displaying orientation, the same way offshore ships get cues from signaling lighthouses, and approaching airplanes use beacons sent from a control tower. The sources might always come from the North, serving as an audio compass, or they might always "point" down, acting like a sonic horizon [Geh88].

If the sources may move around a static listener, it is as if the user were attending a theatre performance or movie. Air traffic controllers looking out of the control tower perceive the circling airplanes this way, as do seated patrons at a restaurant with strolling violinists. Applications of this class might include an *audio cursor* [CL91a] [Coh93] [CAK93], a pointer into 3-space to attract the static user's attention (described later in § 5.2.2).

Giving both sources and sinks full mobility enables a general spatial data management system in which users can browse through a dataworld of movable objects. Teleconferencing applications are perhaps the most obvious example, but more fanciful modes of dance or social intercourse, say, are easily imagined.

5 Interface Design Issues: Case Studies

The issues introduced above are illustrated by three case studies of actual audio imaging systems, described in the following sections: two employ an egocentric metaphor combined with gestural control, the third is exocentric and controlled graphically.

VEOS is a complete VR system, featuring an immersive frontend, via a **h**ead-**m**ounted **d**isplay and wand, driving both stereographic and stereophonic devices. VEOS/FERN, Mercury, and the Sound Renderer and the Audio Browser (described below) were all developed at HITL, the Human Interface Technology Laboratory at the University of Washington.

Handy Sound also has a gestural frontend. It implements an egocentric perspective (in which users arrange spatial sound sources around themselves), and features a purely manual interface (requiring no keyboard or mouse) driving (via direct manipulation through posture and gesture interpretation) a purely auditory backend (requiring no CRT or visual display); it can be used by blind people as well as by sighted.

MAW is implemented as an exocentric GUI in which users can arrange both themselves and spatial sound objects in configurable rooms. MAW extends conventional WIMP idioms to audio windows. Its features include draggably rotating icons and a hierarchical synthesis/decomposition tool, allowing the spatial configuration of audio channels to reflect their logical organization.

5.1 VEOS and Mercury (written with Brian Karr)

VEOS (acronymic for **v**irtual **e**nvironment **o**perating **s**ystem) is a platform-independent distributed processing package which combines many separate computers into a networked virtual multiprocessor. This extensible system handles message passing, pattern matching and program control. Combined with FERN (**f**ractal **e**ntity **r**elativity **n**ode), the system provides distributed database and process management, giving virtual world developers location transparency in the distributed system. FERN is fundamentally a resource administrator for distributed simulation.

Mercury is a self-contained module that interfaces to the virtual environment database, handling both behavior sensing and rendering. The Mercury interface decouples the participant interface from the database, allowing performance to approach the limit of the rendering and position tracking hardware. Similarly, this removes the responsibility for display tasks from the database, allowing both the database (VEOS/FERN) and the renderers to operate at the fastest possible speed.

Mercury maintains an internal representation of the most recent state of the database. The participant experiences this internal representation and is able to move about and interact with entities within it at greater frame rates. Therefore, even though the state of entities in the database may change relatively slowly, the participant can smoothly navigate and interact with objects in the environment. Also, since the renderers are closely coupled with the behavior sensors and the instantaneous state of the external database, aural and visual images are closely registered. Finally, Mercury and its renderers are implemented in a modular fashion, which allows them to be used with almost any VR system. Systems that render new frames only after an entire event loop can benefit most from this approach. Mercury can also be used as an interface to other kinds of software, such as CAD modelers.

5.1.1 Sound Renderer Implementation

The Sound Render is a software package that provides an interface between a VR system and peripheral audio equipment. Such equipment currently includes:

- spatialization hardware such as the Convolvotron
- sound generation devices such as samplers or synthesizers
- effects processors such as reverberation units.

The current implementation allows for the control of four simultaneous, independent audio sources for each participant using serial communications. The Convolvotron (described in

Figure 4: Sound Renderer (VEOS/Mercury)

§ 2.2.1) and geometry software on its host computer are controlled using a serial protocol. Additionally, the Sound Renderer uses the MIDI protocol for control of common audio devices. Some uses for this are:

- triggering raw audio into the spatialization hardware
- manipulating the pitch of the raw audio before convolution to simulate Doppler shift
- controlling effects processors to simulate ambience effects such as reverberation.

The sound sources currently used include an audio sampler (hosted in a computer), speech generation hardware, and microphones. The sampler is configured so that the pitch of all sound sources can be independently manipulated, allowing differently spatialized sources to have separate Doppler shifts. The use of MIDI also easily accommodates other audio spatializers which accept MIDI control data.

The Sound Renderer is not itself an application. Rather it is a programming interface which allows world builders to add sound to their environments at a high level of abstraction. This style configuration is typical of many implementations; the world building abstraction and participant model are otherwise novel up to this point.

Mercury Interface To resolve latency and registration issues, the Sound Renderer was tailored to interface at the C-code level, as in the Mercury Participant System. In this case, a Participant System is a self-contained entity that manages participant behavior (position sensors, etc.) and displays in conjunction with a local database containing the current state of a virtual environment as computed by a distributed system. This allows the participant to smoothly interact with a slowly updating database. The Sound Renderer is provided as a C object library which is linked into Mercury. Because both the visual and audio renderers are running in conjunction with the position sensors on the same computer platform, rather than on two separate machines in the distributed system, network delays are eliminated between them, allowing a closer coupling between visual and audio events.

VEOS/FERN **Interface** For VR systems such as VEOS/FERN, in which environments are coded in Lisp for distributed platforms, the Sound Renderer has been developed with an XLisp interface. All of the functions of the Sound Renderer are available as an object library which is linked into VEOS at appropriate nodes (i.e., machines from which a Sound Renderer may access serial ports). In the sound renderer, reverberation is simulated with MIDI-controlled processors after spatialization. The reverberant signal is mixed with the spatialized signal as would naturally occur in an enclosed space.

5.1.2 The Audio Browser

The Audio Browser is a hierarchical sound file navigation and audition tool [Whi94]. Its intent is to speed up the laborious process of selecting appropriate audio segments from vast archives of sound files, and to help sound designers and foley artists familiarize themselves with new audio sample libraries. Sound is naturally time-linear; we cannot scan sound as we would scan a page of text or images. Informal textual descriptions of sound often do not describe the content accurately. Conversely, we can process many audio streams simultaneously, while we cannot interpret many image streams at once. The Audio Browser takes advantage of the fact that we can naturally monitor many audio streams and selectively focus our attention on any particular one, especially if the sources are spatially separated. Audible transitions from node to node in the database are used to give the listener a feeling that they are "moving" through a tree of nodes.

Inclusive Implementation The Audio Browser makes use of the HITL Sound Renderer and Mercury software systems described above. The sound file samples are prearranged, similar sounds collected into single nodes, and nodes arranged into a quad-tree hierarchy. At each node one can hear representative, looped samples from the four children nodes in each of four front quadrants and the the sample from the parent node behind, as shown in Figure 5. In the inclusive implementation, a graphical representation of the tree is also displayed and the currently auditioned samples are highlighted with a color change.

-bb-error = =

Figure 5: Inclusive Audio Browser

The listener navigates through the sound file hierarchy by choosing the sound representing the node they wish to follow. The selection is accomplished by flying toward the desired sample, at which point it becomes the parent node behind the listener, as its four child samples begin

playing in front. The listener can go back up the tree by flying in reverse. This process is continued until the listener has found the node closest to the desired sound. At this point, the possibly numerous files in the node may be auditioned individually. A more advanced implementation would arrange all files in a node in such a way that they could be inclusively auditioned as well.

5.2 Handy Sound

Handy Sound [Coh89] [Coh90] [CL91a] [CL91b] [Coh93] explores gestural control of an audio window system. The system is of the "moving sources/stationary sink" type and uses an egocentric perspective. It thus allows a single user to arrange sources around him/herself with purely manual manipulation (requiring no keyboard or mouse). Handy Sound is motivated (literally and figuratively) by gestures, i.e., spatial motions that convey information. Gestural recognition via a DataGlove is used as input to a spatial sound system, and virtual sound sources manipulated in a 3D presentation. Figure 6 below illustrates the architecture of the system. Generally in the schematic, digital control data goes down the left, and analog audio signals go up the right.

The user interface of the prototype uses a DataGlove [VPL87] which is coupled with a Polhemus 3Space Isotrak [Pol87]. The system senses the position and orientation of the wearer's hand, the posture[7] of the user's fingers, and the orientation of the user's head. Such tracking is useful for 'soundscape stabilization,' the invariance of the perceived location of the sources under reorientation of the user.

3D tracking products like the coupled Polhemus employ a physically stationary standing wave generator (electromagnetic or ultrasonic) and one or more movable sensors. The resulting systems provide 6 parameters in realtime (the x/y/z of the sensor's physical location and roll/pitch/yaw of the sensor's orientation). Finger posture is calculated by measuring flex-induced leakage in fiber optics laid across the finger joints. With a device like a DataGlove, a user can point and gesticulate using a 3D workspace envelope. In Handy Sound, the DataGlove postures and positions are strobed by a Sun workstation, and integrated into gestures which are used to drive the output.

Sound sources (for simulation) are provided by four samplers [Aka89b], synchronized by a MIDI daisy chain, and cued by a MIDI synthesizer. A digital patch matrix [Aka89a], driven via an RPC-invoked (**r**emote **p**rocedure **c**all) server, is used to switch in the filtears. The *spotlight*, *muffle*, and *highlight filtears* described below are implemented by an aural exciter [Aph89] and a lowpass filter [Ure80]. Since the number of channels in the prototype is fixed, only one channel at a time can be driven through the spotlight or the muffle filtears, and the effects are mutually exclusive (i.e., grabbing an indicated source disables the spotlight as the muffle is enabled), the physical matrix is effectively folded into two logical matrices. The frontend of the system, then, becomes a scheduler, literally handling the dynamic reallocation of the filtear resources.

The backend of the prototype is an enhanced spatial sound system based on the Crystal River Convolvotron (described earlier in § 2.2.1). The control (DataGlove box) and presentation (Convolvotron) processes communicate via internet (UDP) Unix sockets across an Ethernet.[8] The distributed architecture was designed to modularly separate the client (gestural recognition data model) from the server (spatializer and filtear). By using the DataGlove to drive the Convolvotron, virtual sound sources are manipulated in a full 3D auditory display.

[7] This chapter uses the convention of calling (the DataGlove's) recognized static positions "postures," reserving the term "gestures" for the sequential composition of multiple postures.

[8] Ethernet is a trademark of Xerox.

Figure 6: Architecture (Handy Sound). Reproduced with permission of Academic Press, London; from Computer-Supported Cooperative Work and Groupware, 1991; Saul Greenberg, editor

5.2.1 Manipulating Source Position in Handy Sound

By using a posture characterizer to recognize intuitive hand signs along with full-motion arm interpretation, users can gesturally indicate, select, highlight, and relocate sound sources, mapping the reachable work envelope around the user into the much larger perceptual space. Pointing at a source indicates it, as a prelude for selection by grasping. Grasping and releasing are delimiting postures, which respectively enable and disable repositioning. Repositioning is a gesture defined by grasping accompanied by movement.

The Cartesian coordinates of the DataGlove are mapped into spherical coordinates to give the user an egocentric perspective, as shown in eqn. (1). To avoid complications imposed by room geometry, the sound sources are constrained to move spherically: azimuth is adjusted horizontally circularly (as opposed to rectilinearly), elevation is adjusted vertically circularly, and distance is adjusted radially with respect to the user. Azimuth (1a) and elevation (1b) track the user's hand, and distance (1c) (which maps inversely cubically[9] to gain in Handy Sound's dry spatialization) is adjusted proportionally to the radial distance difference between the onset and completion of the relocation, measured from the head to the hand.[10]

$$azimuth = \tan^{-1}\left(\frac{hand_y - head_y}{hand_x - head_x}\right) - \pi/2 \tag{1a}$$

$$elevation = \tan^{-1}\left(\frac{hand_z - head_z}{\sqrt{(hand_x - head_x)^2 + (hand_y - head_y)^2}}\right) \tag{1b}$$

$$distance *= \frac{|\overline{hand(t_2)} - \overline{head(t_2)}|}{|\overline{hand(t_1)} - \overline{head(t_1)}|} \tag{1c}$$

The position of the source is tracked continuously during repositioning. Audio panning and volume control are subsumed by spatial location. For example, if the user indicates an object, grabs, and tugs on it, the object will approach. Figure 7 illustrates pulling a distant source halfway closer: Enamored of a source (represented by concentric rings, whose shading will be explained later) and desiring more intimate proximity, a user repositions it by grasping the proximal projection of its channel, dragging it to a new location, and releasing it.

When the object is released, the azimuth and elevation of the user's hand directly determine the new azimuth and elevation of the object. That is, the azimuthal and elevational control and presentation spaces are the same. Therefore their C/R (**c**ontrol/**r**esponse) ratio $\equiv 1$. For the distance, however, a variable radial C/R ratio is employed, in order to gracefully map the near-field work envelope into the entire perceptual space, finessing issues of scale. In effect, the reachable work envelope is magnified to span the auditory space, giving the user a projected telepresence from physical into perceptual space. The closer an object is to the user, the finer the proximal/distal adjustment (and the higher the radial C/R ratio).

5.2.2 Manipulating Source Quality in Handy Sound

A back-channel is a secondary feedback stream, used to confirm state in control systems. Filtears, which may reflect state information, but do not require a separate display stream, can be used

[9]Gain usually falls off as the inverse of the distance, but Handy Sound deliberately exaggerates distance effects.

[10]The "$*=$" notation means that each new distance value is determined by the product of the old distance and the gesturally determined scaling factor.

Figure 7: Glove at fist site (Handy Sound)

as sonic *piggyback-channel*s, since they are carried by the original source signal. These are audio equivalents of changing cursors, state indicators superimposed on a main channel. Implemented on top of a spatial sound system, sonic piggyback-channels have positional attributes as well as filtear qualities; since repositioning and filtearing are (intended to be) orthogonal, an object may be simultaneously moved and filteared.

Since Handy Sound's main display is purely auditory, modality compatibility motivated the use of sonic piggyback-channels. Rather than create independent sound effects and musical motifs to indicate states, Handy Sound employs filtears for selective transformation of source channels.

The gestural commands (and their feedback filtears) recognized and obeyed by Handy Sound are illustrated by the finite state automaton in Figure 8, and recapitulated in Figure 9, which extends the relocation scenario described earlier (Figure 7). Handy Sound implements three types of (sonic piggyback-channel) filtears, described below:

Spotlight Once audio channels are distributed in space, a telepointer or user-controlled pointer within that space becomes useful. In visual domains, eyegaze selects the focus of attention; there is no direct analogue in audio domains since audition is more omnidirectional than vision. It is easier to detect where someone is looking ('gaze indirection') than to detect what they're listening to. A method of focusing or directing auditory attention is needed to extend the paradigms of graphical indication into audio conferencing.

One could simply instantiate another independent sound source, an *audio cursor*, to superimpose on the selected sources— for instance, a steady or pulsed tone. But this has the disadvantage of further cluttering the auditory space, especially if multiple cursor positions are allowed. In any case, this feature is available intrinsically: user-movable sources can be used as audio cursors "for free" (except for the loss of a channel). User-programmable sources could be the basis of the horizon or compass applications mentioned earlier (in § 4). Like a mythical Siren, sound endowed with the ability to move about can also entice users to follow it. Such a "come-hither" beacon might be used to draw attention to a particular place or workstation window in the office.

Handy Sound explicitly implements a perhaps better solution, an *audio spotlight*, that emphasizes one or more channels [LP89] [CL91a] [CL91b] [BW92]. This emphasis might comprise any combination of the suite of effects used by audio exciters and aural enhancers: equalization,

Figure 8: State transitions and *filtears* (Handy Sound)

pitch shifting, amplitude-dependent harmonic emphasis, and frequency-dependent phase shift. The emphasis augments the source channel's acoustic conspicuousness (variously called brightness, clarity, or presence), making it easier to hear and distinguish, without necessarily making it substantially louder. This emphasis can be likened to sonic italicization, an audio shimmering that draws attention to the emboldened source(s) without overpowering the others. However, a spotlight, unlike a cursor, can only emphasize an active channel, and therefore is less useful as a general pointing device.

The spotlight is used to confirm selection of one or more channels– as a prelude to invoking some action (like amplification, muting, or repositioning), or as an end unto itself, since the emphasis makes the selected objects more prominent. The idea is to create a JND, an acoustic enhancement that is noticeable but ignorable, unambiguous but unintrusive.

In practice, as the hand is swept around the room, pointing at the localized sources, confirmation of direction is achieved by having the indicated source emphasized with a spotlight. An audio spotlight is a way of specifying a subset of the channel mix for special consideration— a way of focusing auditorily, bringing a chosen channel out of background cacophony, and selecting it as the object of a subsequent operation.

Muffle A *muffle* filtear is used to suggest the grasping of a source. Grabbing a channel, as a prelude to moving or highlighting, muffles its sound, imitating the effect of a hand closed around it. This aural confirmation of a gesture fulfills the user interface principle of conceptual compatibility. The muffling effect is accomplished with a lowpass filter, as a covering hand tends to attenuate the high-frequency components of a sound source. The filtear must be subtle to avoid loss of intelligibility in the selected channel.

Highlights *Highlights* are a way of emphasizing audio channels, of endowing them with a perceptual prominence, of promoting and demoting them along a hierarchy of conspicuousness.

Figure 9: Gestured state transitions (Handy Sound)

MAW's highlighting gesture comprises grasping accompanied by a hierarchical specification, represented by extended fingers. Highlights are like an ordered ladder of spotlight-like effects that can be associated with channels. Since they are linked to pointing direction, spotlights cannot be locked on a source, but highlights, which are closely related, may be. Unlike spotlights or muffles, highlights persist beyond the pointing or grasping. They are used to impose a perceptual hierarchical organization on an ensemble of channels. Spotlighting is meant as an immediate feedback feature, guiding selection of a source in a manner analogous to emboldening of the window title bar for graphical interfaces. Highlights are meant as longer-term mnemonic aids, perhaps comparable to choice of font for textural graphical windows.

In a gestural interface like Handy Sound, pointing and grasping are natural postures for indicating and securing. The exact style of extending digits to indicate promotion/demotion in a perceptual hierarchy is culturally sensitive, but the idea of counting on the fingers and thumb is global, and the particular configurations are easily programmed or learned.

5.2.3 Manipulating Sound Volume in Handy Sound

In order to maintain the purity of the simplified gestural interface, the only way to adjust gain in Handy Sound is to bring a source closer. (Alternatively, additional postures could have been defined to raise or lower the volume of a selected source.) Volume is controlled by closeness/distance effects; gain is set inversely proportional to the virtual distance from the source. While the user might simply adjust the volume of the headphone mixer, the only way to make everyone louder via the gestural interface is by pulling everyone closer individually. The only way to turn a sound off is to push it away until it vanishes, thus making it difficult to retrieve.

5.2.4 Summary

Handy Sound demonstrates the general possibilities of gesture recognition and spatial sound in a multi-channel conferencing system. The technology employed, however, is better suited for a concept demonstration than for day-to-day use. The number of spatialized sources is limited to four, with no easy way to scale up. The hardware is complicated, expensive, and unwieldy: The glove itself does not interfere with many other tasks (including writing and typing), but the cables are cumbersome. Further, ambient electromagnetic noise, reflected by metal surfaces in a typical lab environment, make reliable operation of the Polhemus tracker difficult beyond a short range, and measurement of orientation (which direction the hand, as opposed to the arm, is pointing) impractical.

The response of the system is somewhat restricted by the speed of the processors and the high bandwidth requirements, forcing the user to be deliberate in manipulating sound objects. The system is tuned using a choke, a parameter specifying how many postural events to coalesce before transmission. Averaging, debouncing, and hysteresis (to clean up noisy data) must be adjusted to match the environment.

There are two sets of data which should be individually calibrated for each user: the ear maps, modeling the HRTFs of the user, and the posture characteristics, calibrating the various hand positions. In practice, without individually-tailored HRTFs the ear maps are not always perceptually precise [WWK91] [WAKW93] [ACK94]. Further, active control of (especially multiple) sources is difficult with only an auditory display, even with spotlights, making a visual display useful for confirmation of source placement.

5.3 MAW

MAW (acronymic for **m**ultidimensional **a**udio **w**indows) [Coh90] [CK92c] [Coh92] [Coh93] represents a "moving sources/moving sink: *exocentric perspective*" style system which allows sources and sinks to be arranged in a horizontal plane. Developed as an interactive teleconferencing frontend, MAW was retrofitted with a batch mode, making it also suitable for automatic, single-user invocation. Its architecture, shown in Figure 10, is appropriate for both synchronous and asynchronous applications.

The spatialization backend is provided by any combination of MAW's native Focal Point™ [Geh90] and external convolution engines, including the Stork and Digital Audio Processor SIM**2 [acronymic for **s**ound **im**age **sim**ulator], in-house DSP modules. The ellipses below the convolution engines in the schematic indicates that any number of these external convolution engines may be deployed, daisy-chained together on a GPIB (**g**eneral **p**urpose **i**nterface **b**us, or IEEE 488) driven off a SCSI interface [IOt91]. MAW uses configuration files, dynamic maps of virtual spatial sound spaces, to calculate the gain control and HRTF selection for this scalable heterogeneous backend, assigning logical channels to physical devices via a preferences (control) panel. The outputs of all spatialization filters are combined into a stereo pair presented to the user.

The graphical representation of MAW's virtual room is a plan view. This perspective flattening was implemented partly because of its suitability for visual display on a workstation monitor. Figure 11 shows a typical view of such an overhead representation (along with the border, buttons, and scrollers that make it a window) as part of a snapshot of a typical session. MAW adopts the simplification that all spatial sound objects are at once 2D sources and sinks. Spatial sound objects have not only rectangular coordinates, but also angular and focal attributes (described later). Visual icons for sources and sinks indicate their orientation by pointing in the direction that the object is facing. Since all the participants are represented by separate icons, a user can adjust another's virtual position as easily as his/her own.

5.3.1 Manipulating Source and Sink Positions in MAW

For the icons used in Figure 11, the pictures are clipped to the interior of the circle, so the face of the respective user is like the face of a clock, the single hand pointed in the direction the user is "facing" in its admittedly mixed (frontal/aerial) metaphor. Each of the icons is assigned a unique channel number, used to key the spatializing backend.

An alternative iconic representation uses top-down pictures of people's heads, as in Figure 12. Such a view has the advantage of making the bird's-eye metaphor consistent, but suffers from making the users more difficult to recognize. Another iconic representation uses the first-described "head-and-shoulders" pictures, but replaces the radial azimuth-indicating arm with image rotation, as in Figure 10. These various representations may be concatenated and selected "on the fly," allowing users to change the iconic views according to whim.

As in Handy Sound, the notion of a changing cursor to indicate mode is employed by MAW. In Handy Sound, this feedback role is assumed by filtears, which reflect that an audio source is being indicated, relocated, accented, etc. In MAW, the use of a hand to indicate repositioning is elaborated to distinguish an open hand, suggesting rectilinear translation, from a hand with a pointed pivot finger, suggesting rotation thereabout.

MAW extends WIMP interface conventions to manage spatial sound objects with a variety of interaction styles. Draggably rotating icons, which represent non-omnidirectional sources and sinks, are controlled not only by direct manipulation, but also by arrow keys, chorded with

Figure 10: System schematic (Maw)

Figure 11: Screen shot (Maw)

Figure 12: Top-down icons (MAW)

Alternate-, Shift-, and Control-keys; menu items and Command-keys; and numeric panels, all employing the object–command (noun–verb) syntax. Various commands move the icons relative to themselves, each other, and the virtual room.

MAW has a chair tracker [CK92b], crafted with a Polhemus 3Space Isotrak [Pol87], which automatically offsets the azimuth of a particular sink from a (perhaps moving) datum, established via explicit iconic manipulation. The chair tracker blurs the distinction between egocentric and exocentric systems by integrating the egocentric display with ego- and exocentric control, as well as providing the dynamic cues discussed in § 2.1.

As illustrated by Figure 13, the virtual position of the sink, reflected by the (graphically exocentric) orientation of its associated graphical icon, pivots ($\mp\delta$) in response to (kinesthetically egocentric) sensor data around the datum/baseline (θ) established by WIMP (exocentric) iconic manipulation. Symmetrically, the system can be thought of as a user of MAW arbitrarily adjusting a (static or moving) orientation established by the chair tracker. Users may exploit both modes interleaved or simultaneously, adjusting or amplifying their physical position virtually, like setting flaps and trim tabs on an airplane. The WIMP-based operations of MAW can set absolute positions; the chair tracker's reporting of absolute positions has been disabled to allow graphical adjustment. With only WIMP-based rotational initialization, the system behaves as a simple tracker, consistent with proprioceptive sensations. Both MAW's WIMP-based functions and the chair tracker send positional updates to a multicasting conferencing server, so that everyone in a conference or concert may observe the respective sink spinning (to face a source, for instance, enabling 'gaze awareness' [IK92]).

Figure 13: Chair tracker geometry (MAW): exocentric θ, egocentric δ

5.3.2 Organizing Acoustic Objects in MAW

MAW features a cluster utility. Clusters are hierarchically collapsed groups [SZB+93] of spatial sound objects. By bundling multiple channels together, a composite timbre is obtained. Clusters have two main purposes:

conservation of spatializer resources Postulating a switching matrix on either side of the spatial sound processor, along with dynamic allocation of spatializer channels, a cluster feature organizes separate input streams that share a single spatializing channel. One application might involve zooming effects. Distant sources would not be displayed; but as it approaches, a cluster would appear as a single point; only to disassociate and distribute spatially as it gets closer. This focus allows navigation in arbitrarily large space, assuming a limited density of point sources. Alternatively, with limited spatializing resources, a user might chose to group a subset of the (less important or less pleasant) channels together, stacking them in a corner or closet.

logical organization of hierarchical structure For example, in the context of a concert, individually recording (or mic-ing or synthesizing) the separate instruments, presenting each of the channels to MAW, and mixing them at audition time, rather than in "post-production," allow the instruments to be rearranged by the listener. With the appropriate interface, one could grab onto an orchestral cluster, for instance (shown as part of the concert in Table 4), shake it to separate the different instruments, grab one of those instruments and move it across the room. This successive differentiation could go right through concert $\rightarrow$ orchestra $\rightarrow$ section $\rightarrow$ instrument and actually break down the instrument itself. This super decomposition aspect of the cluster feature could allow, for example, the user to listen to spatially separate strings of a violin.

Unclustering can be likened to viewing the sources through a generalized fish eye lens [Fur86] [SB94], which spatially warps the perception of the localized sources to enlarge an area of focus and shrink everything else. That is, when the user indicates a direction of special interest, the sources in that direction effectively approach the user and recede from each other in perspective. While the other objects do not get pushed into the background, the idea is the same: to effect an external rearrangement of sources that complements an internal reordering.

Table 4: Concert decomposition

5.3.3 Manipulating Sound Volume in Maw

In exocentric systems like Maw, it is possible to positionally adjust perceived gain in two different ways: sidle a sink up to a speaker or group of sources, or move the sources nearer to a sink. As in Handy Sound, there is no "volume knob" in Maw; the notion of volume adjustment has been folded into the spatial metaphor.

Maw also provides a more direct way of adjusting gain. The user can resize a selected object by dragging one of the resize handles (knobs) on its bounding rectangle (as in the top right icon of Figure 11). The size of a source corresponds to individual gain (amplification); the size of a sink corresponds to general gain (sensitivity). For the sake of parsimony, iconic size (along one linear[11] dimension) is used as a determinant of both metaphorical ear and mouth size. Gain is proportional to the size of the source's mouth (amplification) and the sink's ear (sensitivity), so enlarging an icon makes its owner both louder and more acute. Thus, to make a *single channel* louder or softer, a user simply resizes the respective icon, but to make *everyone* louder or softer, the user need only resize his/her own icon. Gain is also inversely proportional to the distance between the sink and source, so another way to change perceived volume is to have the source and sink approach or recede from each other. A modified frequency-independent cardioidal pattern is used to model the sound field radiation of non-omnidirectional sources. The chosen relationship specifies an azimuth-dependent beaming of the speaker, an idealized directional pattern, with exaggeratable distance effects. Therefore, the overall amplitude of a source→sink transmission is independent of the sink's transfer function, and can be specified [CK91b] [CK92c] [CK93b] as a function of focus and mutual position and orientation. Focus represents the dispersion, or beaming, of the sound. For a focus of zero, the radiation pattern is omnidirectional. A focus of greater than zero enables a cardioidal pattern, as if the source were using a megaphone. The icon shown in the top left of Figure 14 has focus = 0.1, and a corresponding radiation pattern (sound field density, in which lighter areas indicate greater intensity) in the top right is almost omnidirectional. In contrast, the icon in the bottom left has focus = 0.9 (as indicated by its thicker arm), and its radiation pattern in the bottom right is almost perfectly quiet in the shadow behind the head.

Some systems support multiple visual windows, each featuring a different perspective on a scene. In flight simulators, for example, these might be used to display (egocentric) views out cockpit windows, and/or views from a completely different location— high above the airplane, for example, looking down (exocentrically): a virtual "out-of-body" experience. Since audition is (biasedly) omnidirectional, perhaps audio windows can be thought of as implicitly providing this multiperspective capability, audio sources being inherently superimposed. Maw also features a 'schizophrenic' mode, allowing multiple sinks in the same or different conference rooms, explicitly overlaying multiple audio displays.

A simple teleconferencing configuration typically consists of several icons, representing the distributed users, moving around a shared conference space. Each icon represents a source, the voice of the associated user, as well as a sink, that user's ears. However, Maw allows users to have multiple sinks designated (through a preferences panel), effectively increasing their attendance in the conference, enhancing the *quantity* (and not the quality) of presence. Such a feature might be used to pay close attention to multiple sources, even if those sources are not repositionable; just as in ordinary settings, social conventions might inhibit dragging someone else around a shared space. One could pay close attention to multiple instruments in a concert without rearranging

[11]The linear dimension that actually determines the gain is (arbitrarily) the width, but since icons are almost always resized while holding the aspect ratio fixed (to avoid graphical distortion), height (or diagonal, or circumference...) would have worked equally well.

Figure 14: Icons and radiation patterns (Maw)

the ensemble. One could leave a pair of ears in one conference, while sending another pair to a side caucus, even if the side caucus happens to be in the same room. Such distilled ubiquity, the ability to be anywhere, is better than being everywhere, since it is selective.

The apparent paradoxes of one's being in multiple places simultaneously are resolved by partitioning the sources across the sinks. If the sinks are distributed in separate conference rooms, each source is localized only with respect to the sink in the same room. If multiple sinks share a single conference room, an 'autofocus' mode is employed by anticipating level difference localization, the tendency to perceive multiple identical sources in different locations as a single fused source. (This is related to the precedence effect, mentioned earlier in § 2.1.) Rather than adding or averaging the contribution of each source to the multiple sinks, Maw localizes each source only with respect to the best (loudest, as a function of distance and mutual gain, including focus and orientation effects) sink.

Figure 15 illustrates this behavior for a conference h,1,4,binary,12,cmr (top row) with two sinks, represented by top-down icons, and two different sources, represented by a square and a triangle. In the absence of room acoustics, multiple sinks perceiving a single source is equivalent, via "reciprocity" or symmetry, to a single sink perceiving multiple identical sources. Therefore the exampled scene can be decomposed source-wise into two additive scenes h,2,4,binary,12,cmr

(second row), each single sink combining the parent sinks' perceptions of the respective sources. These configurations reduce h,4,4,binary,12,cmr (third row), via the 'autofocus' level difference anticipation, to the respective sinks and only the loudest source. The loudest source is typically the closest, since the respective pairs of sources are identical, the chorus of phantom sources being a manifestation of the multiple sinks. Finally h,8,4,binary,12,cmr (bottom row), the additive scenes are recombined, yielding the overall simplified percept.

5.3.4 Summary

Unlike Handy Sound, but like VEOS, MAW is designed to be expandable to a large number of channels, and features two techniques for spatializing multiple channels. One is the ability to locally drive multiple alternate external spatializers, assigning logical channels to heterogeneous physical devices. Further, an arbitrary number of workstations may execute a single conference as a 'distributed whiteboard,' each using its own spatializer(s). The outputs of each workstation can then be mixed to form multiple, spatial audio channels.

MAW is intended to interpolate between conventional telephony and VR, but cannot be said to do more than suggest actual acoustic environments. For instance, simulating distance cues is a difficult and not-yet-solved problem which goes beyond MAW's simple gain changes. Besides peoples' natural inability to estimate distance with precision and MAW's distortion of distance effects, an inverse relation does not perfectly capture real effects [Law73] [Bla83] [Beg91] [Wen92] [SL93] [Beg94]. MAW's modeling of source directionality is also not veridical: the selection of a cardioid is somewhat arbitrary, and a flat (frequency-independent) attenuation of gain is not the best model of a rotating source, which should change timbre as well as loudness. It would be more accurate to have a second set of transfer functions that capture these shadow effects, and convolve the digitized source thrice: once for source rotation, and twice (left and right ears) for sink revolution. MAW further over-simplifies reality by neglecting occlusion, the obstruction of a source's sound by other objects in the virtual room; doppler shifts, the pitch bending exhibited by moving sources; indirect early reflections (discrete echoes), the ratio of whose energy to that of direct sounds is another cue for estimating distance; and late reverberation, statistically averaged room ambience, which enhances externalization and auditory "spaciousness." The absence of some of these these cues is sometimes associated with perceptual errors like front↔back reversals [FWT91], as mentioned in § 2.1.

6 Conclusions

As sound technology matures, and more and more audio and multimedia messages and sessions are sent and logged, the testimony of sound may come to rival that of the written word. Audio windows and other multidimensional sound interfaces organize and control virtual acoustic environments. New media spend their early years recapitulating the modes of older media [MF67]; the research described by this chapter hopes to abbreviate this phase for audio interfaces by accelerating their conceptual development. Recurrent themes in the design of multidimensional sound systems include perspective, multiuser capability, C/R mapping and feedback (control state) mechanisms, dimensionality, and integration with other modalities.

Both egocentric and exocentric displays are effective paradigms for virtual audio systems. Egocentric displays like VEOS and Handy Sound are most compatible with inclusion-style VR systems: In such an inside-out display, gestural interpretation control is parsimonious, a natural extension of our normal mode of rearranging the world. An exocentric paradigm like MAW's

Figure 15: Schizophrenic mode with autofocus (Maw)

blurs the self/other distinction by iconifying all users with similar tokens. A mouse- and monitor-driven GUI allows manipulation of all the system entities; the metaphorical universe is projected onto an external and egalitarian medium. This is especially important when the user may have multiple existence (as in MAW's schizophrenic mode).

Groupware applications require a permission system to avoid mutex (**mut**ual **ex**clusion) violation on shared entities. Because of its egocentric nature, Handy Sound features individual data models; no notion of synchronized models is imposed. Handy Sound (which was never deployed in a multiuser environment) decouples the control commands from conferencing users, decreeing a null permission system. With such a data model, unconstrained by a physical analog (i.e., the virtual layout may be inconsistent among the users), two users could sit mutually on each other's laps. VEOS allows multiple participants to share a distributed universe, the database cached through Mercury. MAW's exocentric paradigm is more subtle: it requires an abstraction, and depends on social conventions to establish its implicit permission system.

The ability to rearrange objects is important for mnemonic spatial data organization, since a user is most likely to know where something is if he/she put it there. VEOS, Handy Sound and MAW share several features, including the use of a direct manipulation object–command (noun–verb) syntax, continuous feedback and dynamic tracking.

Potential dimensions for an audio windowing system include not only spatial dimensions, but also qualities of orientation, focus, gain, and other features controllable by filtear mechanisms such as those outlined in Table 3. Further, in order to support an individually-configurable teleconferencing system, a large[12] number of audio channels must be channeled through audio imaging processors. Other applications, including voicemail, hypermedia, and music, require an arbitrarily large number of separately spatialized sonic channels. For all of these, and any task involving terminal audio management, spatial data organization, or scientific sonification, a grouping mechanism is useful, both as a way of imposing a logical hierarchy on many sources, and in conjunction with an audio switching matrix, as a way of conserving channels. The Audio Browser in VEOS (§ 5.1.2) and clusters in MAW (§ 5.3.2) provide a way of selectively collapsing dimensions.

Auditory localization, especially distance perception, is difficult. But visual and acoustic displays complement each other; a glance at a map can disambiguate auditory cues [MM76] [WWM81]. Audio windowing systems can be likened to sonic (analytic) cubism: they present several audio perspectives (on an assembled conference or concert) simultaneously. Multidimensional sound interfaces organize acoustic space, and the interpretation of gestures and the reinterpretation of WIMP conventions seem natural frontends to such systems. Such systems should be designed to exploit innate localization abilities, perception of both spatial and non-spatial attributes, and intuitive notions of how to select and manipulate objects distributed in space. When sound has physical manifestation, it can become an icon for anything imaginable.

[12]Since, in a full-duplex conference, every user must spatialize every other user's voice, the total number of mixels, or channels to spatialize simultaneously, grows quadratically, or as $O(|users|^2)$.

A Acronyms and Initials

A/D: analog → digital

C/R: **c**ontrol/**r**esponse [ratio]

CAD: **c**omputer-**a**ided **d**esign

CAM: **c**omputer-**a**ided **m**anufacturing

CPU: **c**entral **p**rocessing **u**nit

CRT: **c**athode **r**ay **t**ube

CUI: **c**haracter-based **u**ser **i**nterface

D/A: digital → analog

DSP: **d**igital **s**ignal **p**rocessing

FIR: **f**inite **i**mpulse **r**esponse (also known as tapped delay line)

GPIB: **g**eneral **p**urpose **i**nterface **b**us (IEEE 488)

GUI: **g**raphical **u**ser **i**nterface

HRTF: **h**ead-**r**elated **t**ransfer **f**unction

IHL: **i**n-**h**ead **l**ocalization

IID: **i**nteraural **i**ntensity **d**ifference

IIR: **i**nfinite **i**mpulse **r**esponse

ITD: **i**nteraural **t**ime **d**ifference

JND: **j**ust **n**oticeable **d**ifference

MFLOPS: **m**illions of **fl**oating-point **o**perations **p**er **s**econd

MIDI: **m**usical **i**nstrument **d**igital **i**nterface

MIXEL: [sound] **mix**ing **el**ement

MIPS: **m**ultiply-accumulate **i**nstructions **p**er **s**econd

PROM: **p**rogrammable **r**ead-**o**nly **m**emory

RISC: **r**educed **i**nstruction **s**et **c**omputer

RPC: **r**emote **p**rocedure **c**all

SCSI: **s**mall **c**omputer **s**erial **i**nterface

TDR: **t**apped-**d**elay-plus-**r**ecirculation

UDP: [internet] **u**ser [unreliable] **d**atagram **p**rotocol

VR: **v**irtual **r**eality

WIMP: **w**indow, **i**con, **m**enu, **p**ointing [device]

References

Note: Inspired by the notion of a cross-reference as a back-traversable hyperlink, this bibliography uses *zebrackets* to indicate the sections in which each entry was cited. Each of the six delimiter slots represents a section of this document, ordered top h,1,6,binary,12,cmr [§1] → bottom h,32,6,binary,12,cmr [§6]. Citations in each section manifest as ticks in the respective zebracket slots here in the bibliography: the left delimiter points to any explicitic references and the right indicates the invisible ones.

ACK94 Shigeaki Aoki, Michael Cohen, and Nobuo Koizumi. Design and control of shared conferencing environments for audio telecommunication. *Presence: Teleoperators and Virtual Environments*, 3(1):60–72, Winter 1994. www.mitpressjournals.org/loi/pres, ISSN 1054-7460.

ADGM88 E. J. Addeo, A. B. Dayao, A. D. Gelman, and V. F. Massa. An experimental multi-media bridging system. In Robert B. Allen, editor, *Proc.* **C***onf. on* **O***ffice* **I***nformation* **S***ystems*, pages 236–242, Palo Alto, CA, March 1988.

Aka89a Akai. *DP3200 Audio Digital Matrix Patch Bay.* Akai Digital, P.O. Box 2344; Fort Worth, TX 76113, 1989.

Aka89b Akai. *S900* MIDI *Digital Sampler Operator's Manual.* Akai Professional, P.O. Box 2344; Fort Worth, TX 76113, 1989.

AKG91 AKG. CAP 340M Creative Audio Processor. AKG Akustischer und Kino-Geräte GmbH; A-1150 Vienna; Austria, 1991.

AKKS92 Shigeaki Aoki, Nobuo Koizumi, Yusuke Kusumi, and Kiyoshi Sugiyama. Virtual conferencing environment using HRTFs of listeners (in Japanese). In *Proc.* IEICE *Spring Conf.*, Noda, Chiba; Japan, March 1992.

Aph89 Aphex. *Aural Exciter Type C Model 103A Operating Guide.* Aphex Systems Ltd., 13340 Saticoy St.; North Hollywood, CA 91605, 1989.

Aro92 Barry Arons. A review of the cocktail party effect. *J. of the American Voice I/O Society*, 12:35–50, July 1992.

Ast93 Peter Astheimer. What you see is what you hear – acoustics applied in virtual worlds. In *VR93: Proc.* IEEE *Symp. on Research Frontiers in* **V***irtual* **R***eality (in conjunction with* IEEE *Visualization)*, pages 100–107, San Jose, CA, October 1993.

Bat67 D. W. Batteau. The role of the pinna in human localization. In *Proc. Royal Society of London*, volume B168, pages 158–180, 1967.

BB87 Ronald M. Baecker and William A.S. Buxton. *The Audio Channel*, chapter 9, pages 393–399. Morgan Kaufmann Publishers, Inc., 1987. ISBN 0-934613-24-9.

BD92 Meera M. Blattner and Roger B. Dannenberg, editors. *Multimedia Interface Design.* ACM Press: Addison-Wesley, 1992. ISBN 0-201-54981-6.

Beg91 D. R. Begault. Preferred sound intensity increase for sensation of half distance. *Perceptual and Motor Skills*, 72:1019–1029, 1991.

Beg92 D. R. Begault. Perceptual effects of synthetic reverberation on three-dimensional audio systems. *J. Aud. Eng. Soc.*, 40:895–904, 1992.

Beg94 Durand R. Begault. *3-D Sound for Virtual Reality and Multimedia*. Academic Press, Boston, 1994. ISBN 0-12-084735-3.

BG88 Pierre Boulez and Andrew Gerzso. Computers in music. *Scientific American*, 258(4):44–50, April 1988.

BG89 Meera M. Blattner and Robert M. Greenberg. Communicating and learning through non-speech audio. In Alistair D. N. Edwards, editor, *Multi-media Interface Design in Education*. Springer-Verlag, August 1989.

BGB89 W. Buxton, W. Gaver, and S. Bly. The use of non-speech audio at the interface. ACM/SIGCHI Tutorial No. 10, ACM Conference on Human Factors in Computing Systems, New York, 1989.

BH83 R. A. Butler and C. C. Helwig. The spatial attributes of stimulus frequency in the median sagittal plane and their role in sound localization. *American J. of Otolaryngology*, 4:165–173, 1983.

Bla70 Jens Blauert. Sound localization in the medial plane. *Acustica*, 22:205–213, 1969/1970.

Bla83 Jens Blauert. *Spatial Hearing: The Psychophysics of Human Sound Localization*. MIT Press, 1983. ISBN-10 0-262-02190-0.

Bla84 Jens Blauert. Psychoakustik des binauralen horens [the psychophysics of binaural hearing]. In *Proc. DAGA*, Darmstadt, Germany, 1984. Invited plenary paper.

Bla92 Meera M. Blattner. Messages, Models, and Media. *Multimedia Review*, 3(3):15–21, Fall 1992.

BLB77 G. Boerger, P. Laws, and Jens Blauert. Stereophonic headphone reproduction with variation of variation of various transfer factors by means of rotational head movements. *Acustica*, 39:22–26, 1977.

Bly82 Sara A. Bly. Presenting information in sound. In *Proc. CHI: Conf. on Computer-Human Interaction*, pages 371–375, New York, NY, 1982. ACM.

Bly87 Sara A. Bly. Communicating with sound. In Ronald M. Baecker and William A. S. Buxton, editors, *Readings in Human-Computer Interaction: A Multidisciplinary Approach*, chapter 9, pages 420–421. Morgan Kaufmann, 1987. ISBN 0-934613-24-9.

BO89 Nicola Bernardini and Peter Otto. Trails: An interactive system for sound location. In *Proc. ICMC: Int. Comp. MusicConf.*, pages 29–33. Computer Music Association, November 1989.

Bre90 Albert S. Bregman. *Auditory Scene Analysis*. MIT Press, Cambridge, MA, 1990. ISBN 0-262-02297-4.

BS75 M. D. Burkhardt and R. M. Sachs. Anthropomorphic manikin for acoustic research. *J. Acous. Soc. Amer.*, 58:214–222, 1975.

BSG89 Meera M. Blattner, Denise A. Sumikawa, and Robert M. Greenberg. Earcons and Icons: Their Structure and Common Design Principles. *Human-Computer Interaction*, 4(1):11–44, 1989.

Bur58 J. F. Burger. Front-back discrimintation of the hearing system. *Acustica*, 8:301–302, 1958.

But87 R. A. Butler. An analysis of the monaural displacement of sound in space. *Perception and Psychophysics*, 41:1–7, 1987.

BW92 D. R. Begault and E. M. Wenzel. Techniques and applications for binaural sound manipulation in man-machine interfaces. *Int. J. of Aviation Psychology*, 2(1):1–22, 1992.

BW93 Durand R. Begault and Elizabeth M. Wenzel. Headphone localization of speech. *Human Factors*, 35(2):361–376, 1993.

CAK93 Michael Cohen, Shigeaki Aoki, and Nobuo Koizumi. Augmented audio reality: Telepresence/VR hybrid acoustic environments. In *Proc. Ro-Man:* 2^{nd} IEEE *Int. Wkshp. on Robot and Human Communication*, pages 361–364, Tokyo, November 1993. ISBN 0-7803-1407-7.

Cha91 C. J. Chan. Sound localization and spatial enhancement realization of the roland sound space processor. Unpublished handout from a presentation at Cyberarts, Pasadena, CA, 1991.

Che53 E. Colin Cherry. Some experiments on the recognition of speech, with one and with two ears. *J. Acous. Soc. Amer.*, 25(5):975–979, September 1953.

Cho70 John M. Chowning. The simulation of moving sound sources. In *AES: Audio Engineering Society Conv.*, May 1970. Preprint 726 (M-3).

Cho77 John M. Chowning. The simulation of moving sound sources. *Computer Music J.*, 1(3):48–52, June 1977.

CK91a Michael Cohen and Nobuo Koizumi. Audio window. In *Den Gaku*. Tokyo Contemporary Music Festival: Music for Computer, December 1991.

CK91b Michael Cohen and Nobuo Koizumi. Audio windows for binaural telecommunication. In *Proc. Joint Meeting of Human Communication Committee and Speech Technical Committee*, pages 21–28, Tokyo, September 1991. **I**nstitute of **E**lectronics, **I**nformation and **C**ommunication **E**ngineers. Vol. 91, No. 242; SP91-51; HC91-23; CS91-79.

CK91c Michael Cohen and Nobuo Koizumi. Audio windows for sound field telecommunication. In *Proc. Seventh Symp. on Human Interface*, pages 703–709, Kyoto, October 1991. SICE (Society of Instrument and Control Engineers). 2433.

CK92a Michael Cohen and Nobuo Koizumi. Audio windows: User interfaces for manipulating virtual acoustic environments. In *Proc. ASJ: Acoustical Society of Japan Spring Meeting*, pages 479–480, Tokyo, March 1992. Special Session on Virtual Reality, 2-5-12.

CK92b Michael Cohen and Nobuo Koizumi. Iconic control for audio windows. In *Proc. Eighth Symp. on Human Interface*, pages 333–340, Kawasaki, Japan, October 1992. SICE (Society of Instrument and Control Engineers). 1411.

CK92c Michael Cohen and Nobuo Koizumi. *Exocentric Control of Audio Imaging* in Binaural Telecommunication. IEICE *Trans. on Fundamentals of Electronics, Communications and Computer Sciences*, E75-A(2):164–170, February 1992. ISSN 0916-8508, search.ieice.org/bin/index.php?category=Alang=Ecurr=1.

CK93a Michael Cohen and Nobuo Koizumi. Audio windows for virtual concerts. In *Proc. JMACS: Japan Music and Computer Science Society Meeting*, pages 27–32, Tokyo, February 1993. No. 47.

CK93b Michael Cohen and Nobuo Koizumi. Virtual gain for audio windows. In *VR93: Proc.* IEEE *Symp. on Research Frontiers in* **V***irtual* **R***eality (in conjunction with* IEEE *Visualization)*, pages 85–91, San Jose, October 1993.

CKA92 Michael Cohen, Nobuo Koizumi, and Shigeaki Aoki. Design and control of shared conferencing environments for audio telecommunication. In *Proc. ISMCR: Int. Symp. on Measurement and Control in Robotics*, pages 405–412, Tsukuba Science City, Japan, November 1992. SICE (Society of Instrument and Control Engineers).

CL91a Michael Cohen and Lester F. Ludwig. Multidimensional audio window management. *IJMMS: J. of Person-Computer Interaction*, 34(3):319–336, March 1991. ISSN 0020-7373.

CL91b Michael Cohen and Lester F. Ludwig. Multidimensional audio window management. In Saul Greenberg, editor, *Computer Supported Cooperative Work and Groupware*, chapter 10, pages 193–210. Academic Press, London, 1991. ISBN-10 0-12-299220-2.

CM92 T. P. Caudell and David W. Mizell. Augmented reality: An application of heads-up display technology to manual manufacturing processes. In *Proc. Hawaii Int. Conf. on Systems Sciences*. IEEE, January 1992.

Coh87 Michael Cohen. Stereotelephonics. Internal Memorandum IM-000-21460-87-04, Bell Communications Research, October 1987.

Coh89 Michael Cohen. Multidimensional audio window management. Technical Memorandum TM-NPL-015362, Bell Communications Research, October 1989.

Coh90 Michael Cohen. *Multidimensional Audio Windows: Extending User Interfaces through the Use of Spatial Auditory Information*. PhD thesis, Northwestern University, December 1990.

Coh92 Michael Cohen. Integrating graphical and audio windows. *Presence: Teleoperators and Virtual Environments*, 1(4):468–481, Fall 1992. www.mitpressjournals.org/loi/pres, ISSN 1054-7460.

Coh93 Michael Cohen. Throwing, pitching, and catching sound: Audio windowing models and modes. *IJMMS: J. of Person-Computer Interaction*, 39(2):269–304, August 1993. ISSN 0020-7373; www.u-aizu.ac.jp/ mcohen/welcome/publications/tpc.pdf.

Col63 P. D. Coleman. An analysis of cues to auditory depth perception in free space. *Psychological Bulletin*, 60:302–315, 1963.

CW95 Michael Cohen and Elizabeth M. Wenzel. The design of multidimensional sound interfaces. In Woodrow Barfield and Thomas A. Furness III, editors, *Virtual Environments and Advanced Interface Design*, chapter 8, pages 291–346. Oxford University Press, New York, 1995. ISBN-10 0-19-507555-2, ISBN-13 978-0195075557.

Dea72 B. H. Deatherage. Auditory and other sensory forms of information presentation. In H. P. Van Cott and R. G. Kinkade, editors, *Human Engineering Guide to Equipment Design*. U.S. Government Printing Office, Washington, DC, 1972.

DRP+92 N. I. Durlach, A. Rigpulos, X. D. Pang, W. S. Woods, A. Kulkarni, H. S. Colburn, and E. M. Wenzel. On the externalization of auditory images. *Presence: Teleoperators and Virtual Environments*, 1(2):251–257, Spring 1992. ISSN 1054-7460.

Edw87 Alistair D. N. Edwards. Modeling blind users' interactions with an auditory computer interface. Report 25, Centre for Information Technology in Education, The Open University, Milton Keynes, England, 1987.

Edw88 Alistair D. N. Edwards. The design of auditory interfaces for visually disabled users. In *Proc. CHI: Conf. on Computer-Human Interaction*, pages 83–88, 1988.

FF68 H. Fisher and S. J. Freedman. The role of the pinna in auditory localization. *J. Aud. Res.*, 8:15–26, 1968.

Fis90 Scott Fisher. Virtual Environments, Personal Simulation & Telepresence. *Multidimensional Media*, pages 229–236, 1990.

FKS91 Yuichi Fujino, Naobumi Kanemaki, and Kazunori Shimamura. An Audio Window System (in Japanese). In *Proc.* IEICE *Spring Mtg.*, March 1991. D-255.

FMHR86 S. S. Fisher, M. McGreevy, J. Humpries, and W. Robinett. Virtual Environment Display System. ACM *Workshop on* 3D *Interactive Graphics*, pages 77–87, October 1986.

FMHS93 Steven Feiner, Blair MacIntyre, Marcus Haupt, and Eliot Solomon. Windows on the World: 2D Windows for 3D Augmented Reality. In *Proc.* UIST'93 *(*ACM *Symp. on User Interface Software and Technology)*, Atlanta, November 1993.

FMS93 Steven Feiner, Blair MacIntyre, and Dorée Seligmann. Knowledge-based augmented reality. *Communications of the* ACM, 36(7):52–62, July 1993.

Fol87 James D. Foley. Interfaces for advanced computing. *Scientific American*, 257(4):126–135, October 1987.

Fos90 Scott Foster. *Convolvotron*™. Crystal River Engineering, 1990.

Fur86 George W. Furnas. Generalized fisheye views. In *Proc. CHI: Conf. on Computer-Human Interaction*, pages 16–23, Boston, April 1986.

FWCM88 S. S. Fisher, E. M. Wenzel, C. Coler, and M. W. McGreevy. Virtual interface environment workstations. In *Proc.* **H***uman* **F***actors* **S***oc. 32nd Mtg.*, pages 91–95, Santa Monica, 1988.

FWT91 Scott H. Foster, Elizabeth M. Wenzel, and R. Michael Taylor. Real-time synthesis of complex acoustic environments. In *Proc.* (IEEE) *ASSP Wkshp. on Applications of Signal Processing to Audio and Acoustics*, New Paltz, NY, October 1991. Summary.

Gar68 M. B. Gardner. Proximity image effect in sound localization. *J. Acous. Soc. Amer.*, 43:163, 1968.

Gav86 W. W. Gaver. Auditory icons: Using sound in computer interfaces. *Human-Computer Interaction*, 2(2):167–177, 1986.

Gav89 William W. Gaver. The SonicFinder: An Interface that Uses Auditory Icons. *Human-Computer Interaction*, 4(1):67–94, 1989.

Geh87 Bo Gehring. *Auditory Localizer Model AL-201 Product Description.* Gehring Research Corporation, 189 Madison Avenue, Toronto, Ontario M5R 2S6, October 1987.

Geh88 Bo Gehring. U.S. Patent 4774515: Attitude Indicator. 189 Madison Avenue, Toronto, Ontario M5R 2S6, September 1988.

Geh90 Bo Gehring. *Focal Point*^{TM} *3-D Sound User's Manual.* Gehring Research Corporation, 1402 Pine Avenue, #127; Niagara Falls, NY 14301, 1990. (716)285-3930 or (416)963-9188.

Gen86 K. Genuit. A description of the human outer ear transfer function by elements of communication theory. In *Proc. 12th Int. Congress on Acoustics*, Toronto, 1986. Paper B6-8.

GG73 M. B. Gardner and R. S. Gardner. Problem of localization in the median plane: Effect of pinnae cavity occlusion. *J. Acous. Soc. Amer.*, 53:400–408, 1973.

GG89 H. W. Gierlich and K. Genuit. Processing artificial-head recordings. *J. Aud. Eng. Soc.*, 37:34–39, 1989.

GGK92 K. Genuit, H. W. Gierlich, and U. Künzli. Improved possibilities of binaural recording and playback techniques. In *AES: Audio Engineering Society Conv.*, Wien, Austria, March 1992. 3PS1.06.

Gib79 J. J. Gibson. *The ecological approach to visual perception.* Houghton Mifflin, Boston, 1979. ISBN 0-89859-959-8.

GSO91 W. W. Gaver, R. B Smith, and T. O'Shea. Effective sounds in complex systems: The ARKola simulation. In *Proc. CHI: Conf. on Computer-Human Interaction*, pages 85–90, 1991.

Han89 Stephen Handel. *Listening: An Introduction to the Perception of Auditory Events.* MIT Press, 1989. ISBN 0-262-08179-2.

IK92 Hiroshi Ishii and Minoru Kobayashi. Clearboard: A Seamless Medium for Shared Drawing and Conversation with Eye Contact. In *Proc.* CHI *'92*, pages 525–532, New York, 1992. ACM Press. ISBN 0-89791-513-5.

IOt91 IOtech. SCSI*488/N Bus Controller*. IOtech, Inc., 25971 Cannon Rd.; Cleveland, OH 44146, 1991.

KC93 Nobuo Koizumi and Michael Cohen. Audio Windows: Graphical User Interfaces for Manipulating Virtual Acoustic Environments (in Japanese). In *Proc. 18th Meeting, Society of* **C***omputer-***A***ided* **I***nstruction*, pages 19–22, Tokyo Rika Daigaku, Noda, Tokyo, August 1993. A-5-4.

KCA92 Nobuo Koizumi, Michael Cohen, and Shigeaki Aoki. Design of virtual conferencing environments in audio telecommunication. In *AES: Audio Engineering Society Conv.*, Wien, Austria, March 1992. 4CA1.04, preprint 3304.

Ken90 Gary S. Kendall. Visualization by ear: Auditory imagery for scientific visualization and virtual reality. In Amnon Wolman and Michael Cohen, editors, *Proc. Dream Machines for Computer Music*, pages 41–46, School of Music, Northwestern University, November 1990.

Ken91 Gary S. Kendall. Visualization by ear: Auditory imagery for scientific visualization and virtual reality. *Computer Music J.*, 15(4):70–73, Winter 1991.

KF88 Gary S. Kendall and Daniel J. Freed. Scientific visualization by ear. Technical report, Northwestern Computer Music, Northwestern University; Evanston, IL 60208, 1988.

KK86 Kenneth L. Kraemer and John Leslie King. Computer-based systems for cooperative work and group decisionmaking: Status of use and problems in development. Technical report, University of California, Irvine, CA 92717;, September 1986.

KM84 Gary S. Kendall and William L. Martens. Simulating the cues of spatial hearing in natural environments. In *Proc. ICMC: Int. Comp. MusicConf.*, pages 111–126, Paris, 1984. Computer Music Association.

KMF+86a Gary S. Kendall, William L. Martens, Daniel J. Freed, M. Derek Ludwig, and Richard W. Karstens. Image model reverberation from recirculating delays. In *AES: Audio Engineering Society Conv.*, New York, 1986.

KMF+86b Gary S. Kendall, William L. Martens, Daniel J. Freed, M. Derek Ludwig, and Richard W. Karstens. Spatial Processing Software at Northwestern Computer Music. In *Proc. ICMC: Int. Comp. MusicConf.*, pages 285–292. Computer Music Association, October 1986.

KMW90 Gary S. Kendall, William L. Martens, and Martin D. Wilde. A spatial sound processor for loudspeaker and headphone reproduction. In *AES: Audio Engineering Society Conv.*, Washington, D.C., May 1990. ISBN 0-937803-15-4.

Koi91 Nobuo Koizumi. A review of control technology for sound field synthesis (in Japanese). *J. Inst. Telev. Eng. of Jap.*, 45(4):474–479, 1991. 0386-6831.

Kru91 Myron W. Krueger. *Artificial Reality II.* Addison-Wesley, Reading, MA, 1991. ISBN 0-201-52260-8.

KS93 Makoto Kobayashi and Itiro Siio. Virtual conference room: A metaphor for multi-user real-time conferencing systems. In *Proc. Ro-Man: 2nd* IEEE *Int. Wkshp. on* **Ro***bot and Hu****man*** *Communication*, pages 430–435, Tokyo, November 1993.

Kuh77 G. F. Kuhn. Model for the interaural time differences in the azimuthal plane. *J. Acoust. Soc. Amer.*, 62:157–167, 1977.

KW91 Doris J. Kistler and Frederic L. Wightman. A model of head-related transfer functions based on principal components analysis and minimum-phase reconstruction. *J. Acous. Soc. Amer.*, 91:1637–1647, 1991.

Lan86 Keith A. Lantz. An experiment in integrated multimedia conferencing. Technical report, Department of Computer Science, Stanford University, Stanford, CA 94305, December 1986.

Law73 P. Laws. Entfernungshoeren und das Problem der Im-Kopf- Lokalisiertheit von Hoereignissen [Auditory distance perception and the problem of 'in-head' localization of sound images]. *Acustica*, 29:243–259, 1973.

LB89 H. Lehnert and Jens Blauert. A concept for binaural room simulation [summary]. In *Proc.* (IEEE) *ASSP Wkshp. on Applications of Signal Processing to Audio and Acoustics*, New Paltz, NY, October 1989.

Lez93 Fernando López Lezcano. A four channel dynamic sound location system. In *Proc. JMACS: Japan Music and Computer Science Society Meeting*, pages 23–26, Tokyo, February 1993. No. 47.

LHC90 Jack M. Loomis, Chick Hebert, and Joseph G. Cicinelli. Active localization of virtual sounds. *J. Acous. Soc. Amer.*, 88(4):1757–1763, October 1990.

LHC93 Jack M. Loomis, Chick Hebert, and Joseph G. Cicinelli. Personal guidance system for the visually impaired using GPS, GIS, and VR technologies. In Harry J. Murphy, editor, *Proc. Virtual Reality and Persons with Disabilities*, San Francisco, CA, 1993.

LMCH83 D. Lunney, R. C. Morrison, M. M. Cetera, and R. V. Hartness. A microcomputer-based laboratory aid for visually impaired students. IEEE *Micro*, 3(4), 1983.

Lor07 Lord Rayleigh (Strutt, J. W.). On our perception of sound direction. *Philosophical Magazine*, 13:214–232, 1907.

LP89 Lester F. Ludwig and Natalio C. Pincever. Audio windowing and methods for its realization. Technical Memorandum TM-NPL-015361, Bell Communications Research, October 1989.

LPC90 Lester F. Ludwig, Natalio C. Pincever, and Michael Cohen. Extending the notion of a window system to audio. *(*IEEE*) Computer*, 23(8):66–72, August 1990. ISSN 0018-9162, www.computer.org/csdl/mags/co/1990/08/index.html.

Lud89 Lester F. Ludwig. Real-time multi-media teleconferencing: Integrating new technology. Technical Report, Bell Communications Research **I**ntegrated **M**edia **A**rchitecture **L**aboratory, Red Bank, NJ 07746, 1989.

MAB93 Kenneth Meyer, Hugh L. Applewhite, and Frank A. Biocca. A survey of position trackers. *Presence: Teleoperators and Virtual Environments*, 1(2):173–200, 1993.

Man84 Douglass L. Mansur. Graphs in sound: A numerical data analysis method for the blind. Report UCRL-53548, Lawrence Livermore National Laboratory, June 1984.

Man87 Douglass L. Mansur. Communicating with sound. In Ronald M. Baecker and William A. S. Buxton, editors, *Readings in Human-Computer Interaction: A Multidisciplinary Approach*, chapter 9, pages 421–423. Morgan Kaufmann, 1987. ISBN-10 0-934613-24-9.

Mar86 Alain Martel. The SS-1 sound spatializer: A real-time MIDI spatialization processor. In *Proc. ICMC: Int. Comp. MusicConf.*, pages 305–307. Computer Music Association, October 1986.

Mar87 William L. Martens. Principal components analysis and resynthesis of spectral cues to perceived direction. In *Proc. ICMC: Int. Comp. MusicConf.*, San Francisco, 1987. Computer Music Association.

Mar89 William Martens. Spatial image formation in binocular vision and binaural hearing. In *Proc.* 3D **M***edia* **T***echnology Conf.*, Montréal, Québec, May 1989.

MB79 Donald H. Mershon and John N. Bowers. Absolute and relative cues for the auditory perception of egocentric distance. *Perception*, 8:311–322, 1979.

MBJ85 D. L. Mansur, M. M. Blattner, and K. I. Joy. Sound-graphs: A numerical data analysis method for the blind. In *Proc. Hawaii Int. Conf. on Systems Sciences*, January 1985.

ME88 Richard L. McKinley and Mark A. Ericson. Digital synthesis of binaural auditory localization azimuth cues using headphones. *J. Acous. Soc. Amer.*, 83:S18, Spring 1988.

MF67 Herbert Marshall McLuhan and Quentin Fiore. *The Medium is the Message*. Random House, 1967.

MFS84 J. J. Mezrich, S. Frysinger, and R. Slivjanovski. Dynamic representation of multivariate time series data. *J. of the American Statistical Association*, 79(385):34–40, 1984.

MG90 John C. Middlebrooks and David M. Green. Directional dependence of interaural envelope delays. *J. Acous. Soc. Amer.*, 87:2149–2162, 1990.

MG91 John C. Middlebrooks and David M. Green. Sound localization by human listeners. *Annual Review of Psychology*, 42:135–59, 1991.

Mil58 W. Mills. On the minimum audible angle. *J. Acous. Soc. Amer.*, 30:237–46, 1958.

Mil72 A. W. Mills. Auditory localization. In J. V. Tobias, editor, *Foundations of Modern Auditory Theory, Vol. II*, pages 301–345. Academic Press, New York, 1972.

MK75 D. H. Mershon and L. E. King. Intensity and reverberation as factors in the auditory perception of egocentric distance. *Perception and Psychophysics*, 18:409–15, 1975.

MK91 Masato Miyoshi and Nobuo Koizumi. New transaural system for teleconferencing service. In *Proc. Int. Symp. on Active Control of Sound and Vibration*, pages 217–222, Tokyo, April 1991.

MKT+91 Shigeki Masaki, Naobumi Kanemaki, Hiroya Tanigawa, Hideya Ichihara, and Kazunori Shimamura. Personal multimedia-multipoint teleconference system for broadband ISDN. In *Proc. IFIP TC6/WG6.4 Third Int. Conf. on High Speed Networking*, pages 215–230. Elsevier Science Publishers B.V. (North-Holland), March 1991.

MM76 H. McGurk and J. McDonald. Hearing lips and seeing voices. *Nature*, (264):746/748, 1976.

MM77 S. Mehrgardt and V. Mellert. Transformation characteristics of the external human ear. *J. Acous. Soc. Amer.*, 61:1567–1576, 1977.

MM90 J. C. Makous and J. C. Middlebrooks. Two-dimensional sound localization by human listeners. *J. Acous. Soc. Amer.*, 87:2188–2200, 1990.

MMG89 J. C. Middlebrooks, J. C. Makous, and D. M. Green. Directional sensitivity of sound-pressure levels in the human ear canal. *J. Acous. Soc. Amer.*, 86:89–108, 1989.

Mol92 H. Moller. Fundamentals of binaural technology. *Applied Acoustics*, 36:171–218, 1992.

Moo83 F. Richard Moore. A general model for spatial processing of sounds. *Computer Music J.*, 7(3):6–15, Fall 1983.

NMM88 Lars Nielsen, Misha Mahowald, and Carver Mead. SeeHear. Technical report, California Institute of Technology, 1988.

OP84a S. R. Oldfield and S. P. A. Parker. Acuity of sound localisation: a topography of auditory space I. Normal hearing conditions. *Perception*, 13:601–617, 1984.

OP84b S. R. Oldfield and S. P. A. Parker. Acuity of sound localisation: a topography of auditory space II. Pinna cues absent. *Perception*, 13:601–617, 1984.

OP86 S. R. Oldfield and S. P. A. Parker. Acuity of sound localisation: a topography of auditory space III. Monaural hearing conditions. *Perception*, 15:67–81, 1986.

Pat82 R. R. Patterson. Guidelines for auditory warning systems on civil aircraft. Paper No. 82017, Civil Aviation Authority, London, 1982.

Per82 D. R. Perrott. Studies in the perception of auditory motion. In R.W. Gatehouse, editor, *Localization of Sound: Theory and Applications*, pages 169–193. Amphora Press, Groton, CN, 1982.

Per91 A. Persterer. Binaural simulation of an 'ideal control room' for headphone reproduction. (preprint 3062 (k-4)). In *90th Convention of the AES*, Paris, February 1991.

PF54 I. Pollack and L. Ficks. Information of elementary multidimensional auditory displays. *J. Acous. Soc. Amer.*, 26(2):155–158, 1954.

Ple74 G. Plenge. On the difference between localization and lateralization. *J. Acous. Soc. Amer.*, 56:944–951, 1974.

Pol87 Polhemus. 3SPACE ISOTRAK™ *User's Manual*. Polhemus Navigation Science Division, McDonnell Douglas Electronic Company, Colchester, VT, May 1987.

PSBS90 D. R. Perrott, K. Saberi, K. Brown, and T. Z. Strybel. Auditory psychomotor coordination and visual search performance. *Perception and Psychophysics*, 48:214–226, 1990.

PSO+86 C. Poesselt, J. Schroeter, M. Opitz, P. Divenyi, and Jens Blauert. Generation of binaural signals for research and home entertainment (paper b1-6). In *Proc. 12th Int. Congress on Acoustics*, Toronto, 1986.

PSSS91 D. R. Perrott, T. Sadralodabai, K. Saberi, and T. Z. Strybel. Aurally aided visual search in the central visual field: Effects of visual load and visual enhancement of the target. *Human Factors*, 33:389–400, 1991.

PT88 David R. Perrott and J. Tucker. Minimum audible movement angle as a function of signal frequency and the velocity of the source. *J. Acous. Soc. Amer.*, 83:1522–1527, 1988.

RB68 S. K. Roffler and R. A. Butler. Factors that influence the localization of sound in the vertical plane. *J. Acous. Soc. Amer.*, 43:1255–1259, 1968.

Rob92 Warren Robinett. Synthetic experience: A proposed taxonomy. *Presence: Teleoperators and Virtual Environments*, 1(2):229–247, 1992. ISSN 1054-7460.

RP89 F. Richter and A. Persterer. Design and applications of a creative audio processor. In *AES: Audio Engineering Society Conv.*, Hamburg, March 1989. Preprint 2782 (U-4).

SB94 Manojit Sarkar and Marc H. Brown. Graphical fisheye views. *Communications of the* ACM, 37(12):73–84, December 1994.

SBG90 Stuart Smith, R. Daniel Bergeron, and Georges G. Grinstein. Stereophonic and surface sound generation for exploratory data analysis. In Jane Carrasco Chew and John Whiteside, editors, *Proc. CHI: Conf. on Computer-Human Interaction*, pages 125–132, Seattle, WA, April 1990. Addison-Wesley.

SBJG86 Denise A. Sumikawa, Meera M. Blattner, K. I. Joy, and Robert M. Greenberg. Guidelines for the syntactic design of audio cues in computer interfaces. In *Proc. Hawaii Int. Conf. on Systems Sciences*, Honolulu, 1986.

SBL+86 M. Stefik, D.G. Bobrow, S. Lanning, D. Tatar, and G. Foster. WYSIWIS revised: Early experiences with multi-user interfaces. In *Conf. on* **C***omputer-***S***upported* **C***ooperative* **W***ork*, pages 276–290, Austin, TX, December 1986.

SC91 C. Scaletti and A. B. Craig. Using sound to extract meaning from complex data. In E. J. Farrell, editor, *Extracting Meaning from Complex Data: Processing, Display, Interaction II, Proc. SPIE 1459*, 1991.

Sco89 Douglas Scott. A processor for locating stationary and moving sound sources in a simulated acoustical environment. In *Proc. ICMC: Int. Comp. MusicConf.*, pages 277–280. Computer Music Association, November 1989.

Set90 Mark Setton. Sonic Navigator™. Project Report, Berkeley Systems, Inc., 1990.

SFB+87 Mark Stefik, Gregg Foster, Daniel G. Bobrow, Kenneth Kahn, Stan Lanning, and Lucy Suchman. Beyond the chalkboard: Computer support for collaboration and problem solving in meetings. *Communications of the* ACM, 30(1):32–47, January 1987.

SG85 Sunil Sarin and Irene Greif. Computer-based real-time conferencing systems. *(*IEEE*) Computer*, 18(10):33–45, October 1985.

SGH+90 Diane H. Sonnenwald, B. Gopinath, Gary O. Haberman, William M. Keese, III, and John S. Myers. Infosound: An audio aid to program comprehension. In *Proc. Hawaii Int. Conf. on Systems Sciences*, Honolulu, January 1990.

Sha74 E. A. G. Shaw. The external ear. In *Handbook of Sensory Physiology, Vol. V/1, Auditory System*, pages 455–490. Springer-Verlag, New York, 1974.

SL93 Jon M. Speigle and Jack M. Loomis. Auditory distance perception by translating observers. In *VR93: Proc.* IEEE *Symp. on Research Frontiers in* **V***irtual* **R***eality (in conjunction with* IEEE *Visualization)*, pages 92–99, San Jose, CA, October 1993.

SM87 Mark S. Sanders and Ernest J. McCormick. *Human Factors in Engineering and Design.* McGraw-Hill, New York, sixth edition, 1987. ISBN 0-07-044903-1.

SMP92 Thomas Z. Strybel, Carol L. Manlingas, and David R. Perrott. Minimum audible movement angle as a function of azimuth and elevation of the source. *Human Factors*, 34(3):267–275, 1992.

SN36 S. S. Stevens and E. B. Newman. The localization of actual sources of sound. *American J. of Psychology*, 48:297–306, 1936.

SS87 Shoji Shimada and J. Suzuki. A new talker location recognition through sound image localization control in multipoint teleconference system. IEICE *Trans. on Fundamentals of Electronics, Communications and Computer Sciences*, J70-B(9):491–497, 1987.

SSN93 Youichi Shinomiya, Kazuya Sawada, and Junji Nomura. Development for real-time acoustic simulation in virtual realities (in Japanese). *J. Acous. Soc. Jap.*, 49(7):515–521, 1993.

STFJ55 T. T. Sandel, D. C. Teas, W. E. Feddersen, and L. A. Jeffress. Localization of sound from single and paired sources. *J. Acous. Soc. Amer.*, 27:842–52, 1955.

SWKE89 Robert D. Sorkin, Frederic L. Wightman, Doris J. Kistler, and Greg C. Elvers. An exploratory study of the use of movement-correlated cues in an auditory head-up display. *Human Factors*, 31(2):161–166, April 1989.

SY91 Shoji Shimada and Yoshio Yamasaki. Evolution of digital signal processing in communication networks (in Japanese). *J. Acous. Soc. Jap.*, 47(7):491–497, 1991.

SZB+93 Doug Schaffer, Zhengping Zuo, Lyn Bartram, John Dill, Shelli Dubs, Saul Greenberg, and D. Roseman. Comparing fisheye and full-zoom techniques for navigation of hierarchically clustered networks. In *Proc. Graphics Interface*, Toronto, 1993. Morgan-Kaufmann.

TAMS91 Hiroya Tanigawa, Tomohiko Arikawa, Shigeki Masaki, and Kazunori Shimamura. Personal multimedia-multipoint teleconference system. In *Proc.* IEEE *InfoCom'91*, pages 1127–1134, April 1991.

TKOK92 Haruo Takemura, Yasuichi Kitamura, Jun Ohya, and Fumio Kishino. Distributed processing architecture for virtual space teleconferencing. In Susumu Tachi, editor, *Proc. ICAT: Int. Conf. on Artificial Reality and Telexistence*, pages 27–32, Tokyo, July 1992.

Too69 F. E. Toole. In-head localization of acoustic images. *J. Acous. Soc. Amer.*, 48:943–949, 1969.

TR67 W. R. Thurlow and P. S. Runge. Effects of induced head movements on localization of direction of sounds. *J. Acous. Soc. Amer.*, 42:480–488, 1967.

Ure80 Urei. *Dual Parametric Equalizer Model 546 Operating Instructions*. Urei (United Recording Electronics Industries), 8460 San Fernando Rd.; Sun Valley, CA 91352, 1980.

Van89 Gregg C. Vanderheiden. Nonvisual alternative display techniques for output from graphics-based computers. *Journal of Visual Impairment & Blindness*, pages 383–390, October 1989.

vB60 G. von Békésy. *Experiments in Hearing*. McGraw-Hill, New York, 1960.

VPL87 VPL. *DataGlove Model 2 Operating Manual*. VPL (Visual Programming Language) Research, Inc., 656 Bair Island Rd.; Suite 304; Redwood City, CA 94063, 1987.

WAKW93 Elizabeth M. Wenzel, Marianne Arruda, Doris J. Kistler, and Frederic L. Wightman. Localization using nonindividualized head-related transfer functions. *J. Acous. Soc. Amer.*, 94(1):111–123, July 1993.

Wal40 H. Wallach. The role of head movements and vestibular and visual cues in sound localization. *J. Exp. Psych.*, 27:339–368, 1940.

Wen92 Elizabeth M. Wenzel. Localization in virtual acoustic displays. *Presence: Teleoperators and Virtual Environments*, 1(1):80–107, 1992. ISSN 1054-7460.

Wen94 Elizabeth M. Wenzel. Spatial sound and sonification. In Greg Kramer, editor, *Proc. First Int. Conf. on Auditory Display*, pages 127–150, Santa Fe, NM, 1994. ISBN 0-201-62603-9.

WF93 Elizabeth M. Wenzel and Scott H. Foster. Perceptual consequences of interpolating head-related transfer functions during spatial synthesis. In *Proc.* (IEEE) *ASSP Wkshp. on Applications of Signal Processing to Audio and Acoustics*, New Paltz, NY, October 1993.

Whi94 John F. Whitehead. An audio database navigation tool in a virtual environment. In *Proc. ICMC: Int. Comp. MusicConf.*, pages 280–283, Århus, Denmark, 1994. Computer Music Association.

WK89a F. L. Wightman and D. J. Kistler. Headphone simulation of free-field listening I: stimulus synthesis. *J. Acous. Soc. Amer.*, 85:858–867, 1989.

WK89b F. L. Wightman and D. J. Kistler. Headphone simulation of free-field listening II: psychophysical validation. *J. Acous. Soc. Amer.*, 85:868–878, 1989.

WKA92 F. L. Wightman, D. J. Kistler, and M. Arruda. Perceptual consequences of engineering compromises in synthesis of virtual auditory objects. *J. Acous. Soc. Amer.*, 92:2332, 1992.

WMG93 Pierre Wellner, Wendy Mackay, and Rich Gold. Communications of the ACM, July 1993.

WNR49 H. Wallach, E. B. Newman, and M. R. Rosenzweig. The precedence effect in sound localization. *American J. of Psychology*, 57:315–336, 1949.

WSFF90 Elizabeth M. Wenzel, Philip K. Stone, Scott S. Fisher, and Scott H. Foster. A system for three-dimensional acoustic "visualization" in a virtual environment workstation. In *Proc. First* IEEE *Conf. on Visualization*, pages 329–337, San Francisco, October 1990.

WWF88a Elizabeth M. Wenzel, Frederic L. Wightman, and Scott H. Foster. Development of a three-dimensional auditory display system. In *Proc. CHI: Conf. on Computer-Human Interaction*, Washington, DC, May 1988.

WWF88b Elizabeth M. Wenzel, Frederic L. Wightman, and Scott H. Foster. A virtual display system for conveying three-dimensional acoustic information. In *Human Factors Society 32nd Annual Meeting*, pages 86–90, Santa Monica, CA, 1988.

WWK91 Elizabeth M. Wenzel, Frederic L. Wightman, and Doris J. Kistler. Localization of non-individualized virtual acoustic display cues. In Scott P. Robertson, Gary M. Olson, and Judith S. Olson, editors, *Proc. CHI: Conf. on Computer-Human Interaction*, New Orleans, LA, May 1991. Addison-Wesley. ISBN 0-201-51278-5.

WWM81 David H. Warren, Robert B. Welch, and Timothy J. McCarthy. The role of visual-auditory "compellingness" in the ventriloquism effect: Implications for transitivity among the spatial senses. *Perception and Psychophysics*, 30(6):557–564, 1981.