Throwing, Pitching, and Catching Sound: Audio Windowing Models and Modes

Michael Cohen

Human Interface Laboratories
Nippon Telegraph and Telephone Corporation
Yokosuka-shi, Kanagawa-ken 238-03
Japan

voice: [+81](468)55-8503
fax: [+81](468)55-1054
internet: `mcohen@nttspch.ntt.jp` (NeXT mail capable)

after March, 1993:
The University of Aizu
Aza Kami-iawase 90, Oaza Tsuruga
Ikki-machi, Aizu-Wakamatsu-shi
Fukushima-ken 965
Japan

To be published (Spring, 1993) in IJMMS:
The journal of person-computer interaction.
draft: April 25, 2008

Summary

Throwing, Pitching, and Catching Sound:
Audio Windowing Models and Modes

After surveying the concepts of audio windowing, this paper elaborates taxonomies of three sets of its dimensions: spatial audio ("throwing sound"), timbre ("pitching sound"), and gain ("catching sound"), establishing matrices of variability for each, drawing similes, and citing applications. Two audio windowing systems are examined across these three operations: repositioning, distortion/blending, and gain control (state transitions in virtual space, timbre space, and volume space). Handy Sound is a purely auditory system with gestural control, while MAW exploits exocentric graphical control. These two systems motivated the development of special user interface features. (Sonic) piggyback-channels are introduced as filtear manifestations of changing cursors, used to track control state. A variable control/response ratio can be used to map a near-field work envelope into perceptual space. Clusters can be used to hierarchically collapse groups of spatial sound objects. WIMP idioms are reinterpreted for audio windowing functions. Reflexive operations are cast an instance of general manipulation when all the modified entities, including an iconification of the user, are projected into an egalitarian control/display system. Other taxonomies include a spectrum of directness of manipulation, and sensitivity to current position crossed with dependency on some target position.

Keywords: audio windows, CSCW, filtears, groupware, piggyback-channels, spatial sound.

0 Conceptual Overview

This research is concerned with paradigms for audio presentation styles. Audio windowing systems are designed to declutter the cacophony (of a teleconference, say) by introducing axes of audio presentation along which the respective channels may be distributed. They represent user interfaces that perceptually separate the input channels, be they musical voices or teleconference utterances.

Two audio windowing systems have been built by the author. "Handy Sound" is an egocentric system combining a hand-posture interpretation frontend (based on a DataGlove) with an enhanced spatial sound system; it employs a gestural frontend (requiring no keyboard or mouse) driving an auditory backend (requiring no CRT or visual display). "MAW" (acronymic for **m**ultidimensional **a**udio **w**indows) is an exocentric GUI (**g**raphical **u**ser **i**nterface, incorporating visual representations of the entities in a mouse-driven interface; it extends standard idioms for WIMP (**w**indow, **i**con, **m**enu, **p**ointing device) systems to audio window applications, driving a spatial sound backend.

This paper casts audio windowing modes in baseball metaphors, but elaborates several other organizations of user interfaces. After reviewing the evolution of I/O dimensionality and motivating the use of audio in interfaces, including audio imaging via spatial sound and filtears, these two audio windowing systems are examined across three different attributes: virtual position ("throwing"), sound quality ("pitching"), and gain ("catching").

0.1 I/O Generations & 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 touchpad. Now there exist 3D spatial pointers and 3D graphics devices; this latest phase of I/O devices [Blattner 92] [Blattner & Dannenberg 92] 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 paper focuses on the italicized entries in the third generation aural sector.

0.2 Exploring the Design Space

Audio alarms and signals of various types have been with us 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— voicemail, or voice annotation to electronic mail, teleconferencing, audio archiving— while other applications use sound to complement their underlying functionality. Table 2 (extended from [Deatherage 72, p. 124] and [Sanders & McCormick 87, 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 both exploring analogues to other modes of expression and also

generation/ dimension	mode	input	output
first/1D	textual	keyboard	teletype mono 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 trackball, joystick DataGlove/DataSuit flying mouse ("bat")	tactile feedback: vibrating fingertips force-feedback devices Braille devices
	olfactory	??	?
	gustatory	??	?
	visual	head- and eye-tracking	stereoscopic systems: head-mounted displays holograms vibrating mirrors

Table 1: Generations and dimensions of I/O devices

- 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 calls for immediate action (prompts like "Your printer is out of paper.")
- when continuously changing information of some type is presented (e.g., location or metric)
- when speech channels are fully employed
- when illumination limits use of vision (e.g., an alarm clock)
- when the receiver moves from one place to another (employing sound as a ubiquitous I/O channel)
- when a verbal response is required (compatibility of media)

Table 2: Motivation for using sound as a display mode

evolving models unique to audio. Computer interfaces present special needs and opportunities for audio communication.

This paper discusses the evolving state of the art of non-speech audio interfaces, driving both spatial and non-spatial attributes. The 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 paper is primarily concerned with their integration— crafting effective matches between projected user desires and emerging technological capabilities.

Clearly sound has many other qualities besides spatial attributes which contribute to its perceptual and cognitive organization. The various widely discussed [Pollack & Ficks 54] [Baecker & Buxton 87, p. 396] [Bly 87, p. 420] [Mansur 87, 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 [Gibson 79]'s ecological approach to perception, the audible world can be conceived of as a collection of acoustic "objects" [Cohen & Wenzel 95]. In addition to spatial location, various acoustic features such as temporal onsets and offsets, timbre, pitch, intensity, and rhythm, one 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 [Gaver et al. 91], to a completely abstract mapping of statistical data into sound parameters [Smith et al. 90]. Principles for design and synthesis can also be gleaned from the fields of music [Blattner et al. 89], psychoacoustics [Patterson 82], and higher-level cognitive studies of the acoustical determinants of perceptual organization [Buxton et al. 89] [Bregman 90].

0.3 Audio Imaging

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 synthesis/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 [Koenig 50] [Cherry 53].

Audio imaging is the creation of sonic illusions by manipulation of separate channels. For instance, when classical music is recorded, the sound from different instruments comes from distinctly different directions. The violins seem to be 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. Current 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. But this technique yields images that are diffuse, and located only between the speakers.

0.4 Spatial Sound

A display system needs a way of perceptually segmenting multiple audio channels, a way of distinguishing them from each other, a way of segregating the streams. A simple ploy is to just make the channels of interest louder than their siblings. A more sophisticated approach, spatial sound, employs technology that allows sound sources to have not only left–right attributes (as in a conventional stereo mix), but up–down and back–forth qualities as well. It is related, 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 potentially rich analogue of 3D graphics.

Spatial sound encourages auditory localization, a listener's psychological separation in space of the channels, invoking the cocktail party effect. Spatial sound projects audio media into space by manipulating

[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.

sound sources so that they assume virtual positions, mapping them from one-space (the source channel) into multidimensional-space (the perceptual envelope around the listener). In the absence of visual cues, someone listening to music distinguishes the channels (voices, instruments, parts) by position, tone/timbre, and melodic line and rhythm. Similarly, at a party with many simultaneous conversations, a mingler can still follow any particular exchange by filtering according to position, speaker voice, and subject matter. The last two dimensions (tone/voice and melody/subject) are not controlled by spatial sound systems, but by adjusting the virtual position of the sources with respect to the listener, enabling perceptual segmentation of the various channels.

A number of researchers [Chowning 70] [Chowning 77] [Kendall et al. 86] [Martel 86] [Gehring 87] [McKinley & Ericson 88] [Wenzel et al. 88b] [Wenzel et al. 88a] [Martens 89] [Scott 89] [Sorkin et al. 89] [Kendall et al. 90] [Loomis et al. 90] [Wenzel et al. 90] [Wenzel et al. 91] [Wenzel 92] are creating psychoacoustic effects (usually with DSP) and developing ways of generating and controlling this multidimensional sound imagery. They use sound spatializers to create the impression that the sound is coming from different sources and different places, just as one would hear "in person." Spatial hearing can be stimulated by assigning each source a virtual position with respect to the listener and simulating the auditory positional cues. Displays based on sound spatializer technology exploit human ability to quickly and subconsciously localize sound sources.

Binaural (stereo) localization cues include **i**nteraural **t**ime (phase) **d**elay (ITD), a function of the interaural distance; **i**nteraural **i**ntensity **d**ifference (IID), a consequence of the acoustic head shadow; and complex binaural spectral cues in the spatial response [Blauert 83]. One approach to spatial sound uses a convolution engine, which incorporates **f**inite **i**mpulse **r**esponse (FIR, also known as tapped delay) digital audio filters, whose output signals are presented to stereo loudspeakers or headphones. Binaural localization cues may be captured by **h**ead-**r**elated **t**ransfer **f**unctions (HRTFs), impulse responses measured (or synthesized) for the head and pinna (outer ear) of human or dummy heads. For each spherical direction, a left–right pair of these transfer functions is measured and stored as FIR filter coefficients. Sound can then be spatialized by driving digitized input signals through these HRTFs in a convolution engine, creating psychoacoustic localization effects by introducing binaural spatial cues into an originally monaural channel, "placing" the channel within the perceptual space of the user.

0.5 Audio Windows

"Audio windows" is an auditory-object manager, a potentially powerful implementation of a user interface ('frontend') to an audio imaging system. The generalized control model of a window here 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 [Ludwig & Pincever 89] [Ludwig et al. 90] [Cohen & Ludwig 91a, Cohen & Ludwig 91b] [Cohen & Koizumi 91b] [Cohen & Koizumi 92a] 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,[2] 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. (The style of manipulation of the audio windows can vary; this paper describes two different approaches.)

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 sink 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,[3] 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 [Cohen et al. 92] [Koizumi et al. 92]. 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 [Cohen & Koizumi 91a] [Cohen & Koizumi 93a]. Minglers at a virtual cocktail party might freely circulate. Sound presented in this dynamically spatial fashion is as different

[2] 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.

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

from conventional mixes as sculpture is from painting.

0.6 Filtears

A system with display modality dependent on audio encourages an indication of control state— i.e., augmenting the signal by perceptually multiplexing the audio bandwidth. Even though the channels can be perceptually segmented by virtual location, it is useful to have other attribute cues, independent of direction and distance. Auditory icons [Gaver 86] are sounds of naturally occurring events that caricature the action being represented. For instance, in the Macintosh SonicFinder [Gaver 89], a metallic thunk represents a file being tossed into the trashcan upon deletion, and a liquid gurgling signifies a file being copied. "Earcons" [Sumikawa et al. 86] [Blattner et al. 89] [Blattner & Greenberg 89] are elaborated auditory symbols which compose motives into artificial non-speech language, phrases distinguished by rhythmic and tonal patterns. Earcons may be combined (by juxtaposing these motives), transformed (by varying the timbre, register, and dynamics), or inherited (abstracting a property). Infosound [Sonnenwald et al. 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" [Cohen 89] [Cohen & Ludwig 91a, Cohen & Ludwig 91b], which depend on the distinction between sources and sinks, are one way of spanning this spectrum.

Even though audio channels can be perceptually segmented by virtual location, it is also critical 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. Unlike pure auditory icons, earcons, or realtime sound effects, filtears are conceptually transforming, rather than transformed. If "everyday listening" is contrasted with "musical listening" as attending distal stimulation instead of proximal [Buxton et al. 89, p. 3.1], then filtears can be said to involve decoding medial transformations [Cohen & Wenzel 95].

Imagine a user tele-negotiating with several parties at once, including trusted advisors. Besides whatever spatial arrayal 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 example filtears have actually been implemented yet. The filtears that have been deployed are detailed in § 2.1.)

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 *italic*izing or em**bold**ening. Filtears embolden and italicize 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. Filtears should be transparent unless the user is actively seeking them; perceivable if the user is receptive, and otherwise ignorable.

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 [Cohen & Wenzel 95].

0.7 Research Applications

Auralization, also variously called "audification," "audiolization," and "sonification," can be thought of as auditory visualization, and has been explored by scientists [Bly 82] [Mezrich et al. 84] [Blattner et al. 92] 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 source data to the attributes outlined in Table 3. Various researchers [Kendall & Freed 88] [Kendall 90, Kendall 91] [Wenzel et al. 90] suggest using spatial sound as a component of auralization, 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. For example, Exvis [Smith et al. 90, Smith et al. 92] interprets a scatterplot as a texture, a dense distribution of data, and then translates that texture into sound.

There is also increasing interest in providing auditory displays for visually disabled users. Some researchers, including [Lunney et al. 83] [Mansur 84] [Mansur et al. 85], have experimented with mapping x-y graphs to sound, to convey their information to blind users. An "auditory screen" [Edwards 87, Edwards 88] 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 [Nielsen et al. 88] mapped visual signals from optically scanned objects into localized auditory signals. The Sonic Navigator [Setton 90] localizes synthesized speech to the location of the window being read.

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, allowing one to stretch one's presence across continents, oceans, and beyond. Many scientists are exploring computer-controlled teleconferencing systems [Cohen & Koizumi 91c] [Masaki et al. 91] [Tanigawa et al. 91] [Cohen & Koizumi 92c]. Major thrusts have protocols for invoking a rendezvous [Kraemer & King 86], suitable architectures for deploying such systems [Sarin & Greif 85] [Lantz 86] [Ludwig 89], and graphical control [Stefik et al. 86] [Dayao & Gelman 87] [Stefik et al. 87] [Addeo et al. 88] [Sonnenwald et al. 90].

Musical applications [Moore 83] [Kendall & Martens 84] [Boulez & Gerzso 88] [Bernardini & Otto 89] [Kendall et al. 89] [Cohen & Koizumi 91a] [Cohen & Koizumi 93a] will feature bouncing and dancing sound. A listener could wander among a marching band or an embracing chord; a composer could program choreography for sonic dancers.

Virtual **r**eality (VR) systems [Krueger 82, Krueger 91] [Benedikt 91] [Tachi 91] [Tachi 92] are computer-generated interactive environments utilizing (typically head-mounted display) 3D graphic scenes and soundscapes, featuring a manual (typically DataGloved) control [Foley 87]. They are characterized by an intimate link between display and control, in which the user sometimes inhabits the system [Bricken 92]. Various VR researchers, including [Fisher et al. 86] [Fisher et al. 88] [Wenzel et al. 90], have incorporated stereophonic output into their headmounted display.

1 Throwing: Manipulating Sound Position

Ventriloquism is often described as "throwing one's voice," projecting it to a virtual location. In this spirit, we think of spatial audio systems as throwing sound. While in one of the audio windowing systems described here, Handy Sound, the notion is almost literal (except that the gestures allow no notion of momentum), for the other, MAW, we fall back upon the idiom.

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 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 § 2.1.1).

A simple demonstration of this functionality on a conventional system features three users, each with two telephones, calling each other cyclically (Figure 1). Each user's holding the calling and called handsets to different ears demonstrates one application of *stereotelephonics* [Cohen 87], the use of stereo effects in telephones.

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

- harmonic content
  - pitch and register
  - waveshape (sawtooth, square, triangle, ...)
  - timbre, filtears, and resonance
- dynamics
  - intensity/volume/loudness
  - envelope: attack, decay, sustain, release (volume shape)
- timing
  - repetition rate
  - duty cycle
  - rhythm and cadence
  - tempo
  - syncopation
  - duration
- spatial location

Table 3: Dimensions of sound

Figure 1: Stereotelephonics and 3-way cyclic conferencing

North, serving as an audio compass, or they might always "point" down, acting like a sonic horizon [Gehring 88].

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 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* [Cohen & Ludwig 91a, Cohen & Ludwig 91b], a pointer into 3-space to attract the static user's attention (described later in § 2.1.1).

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. If the environment itself can also be configured by the user, this class enables virtual omnipotence, since the user may manipulate all the entities in the system.

1.1 Manipulating Sound Position in Handy Sound

Handy Sound [Cohen 89] [Cohen & Ludwig 91a, Cohen & Ludwig 91b] explores gestural control of an audio window system. The system is of the "moving sources/stationary sink: *egocentric perspective*" type, allowing a single user to arrange sources around 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 full 3D presentation. Figure 2 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 [VPL 87] which is coupled with a Polhemus 3Space Isotrak [Polhemus 87]. The DataGlove/Polhemus system senses the position and orientation of the wearer's hand,[4] the posture 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. A graphical display module could be interfaced across the same backbone architecture.

Sound sources (for simulation) are provided by four samplers [Akai 89b], synchronized by a MIDI daisy chain, and cued by a MIDI synthesizer. A digital patch matrix [Akai 89a], 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 [Aphex 89] and a lowpass filter [Urei 80]. 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 [Foster 89]. The control (DataGlove box) and presentation (Convolvotron) processes communicate via internet (UDP) Unix sockets across an Ethernet.[5] 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.

[4] Since the DataGlove is a fully three dimensional input device, it is sometimes likened to a bat, a flying mouse.

[5] Ethernet is a trademark of Xerox.

Figure 2: Handy Sound architecture

By using a posture[6] characterizer to recognize intuitive hand signs along with full motion arm interpretation, we can gesturally indicate, select, highlight, and relocate these sound sources, mapping the 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 duals, enabling and disabling repositioning. Repositioning is grasping accompanied by movement.

The Cartesian coordinates of the DataGlove are mapped into spherical coordinates to give the user an egocentric perspective (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) to gain in Handy Sound's dry (reverberationless) spatialization, is adjusted proportionally to the radial distance difference between the onset and completion of the relocation, measured from the head to the hand.[7]

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

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

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

The position of the source is tracked continuously during repositioning. Audio panning and volume control are subsumed by spatial sound. For example, if the user indicates an object, grabs, and tugs on it, the object will approach. Figure 3 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.

Since the azimuthal (1a) and elevational (1b) control and presentation spaces for a (DataGlove-manipulated) spatial sound domain are the same, their C/R (**c**ontrol/**r**esponse) ratio ≡ 1. And with a variable radial (1c) C/R ratio, the near-field work envelope maps gracefully into the entire perceptual space, finessing the 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.

1.2 Manipulating Sound Position in MAW

MAW represents a "moving sources/moving sink: *exocentric perspective*" style system which allows sources and sinks to be arranged in a horizontal plane. Extended from a single-user system [Cohen 90] into an interactive teleconferencing frontend [Koizumi et al. 92], then retrofitted with a batch mode, MAW is suitable for synchronous or asynchronous applications. Its architecture is shown in Figure 4.

The spatialization backend is provided by any combination of the NeXT-platformed Focal Point™ [Gehring 90] 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) driven off a SCSI interface [IOtech 91]. MAW uses 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.

[6] We have adopted the convention of calling the DataGlove's recognized static positions "postures," reserving the term "gestures" for the sequential composition of multiple postures.

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

The graphical representation of MAW's virtual room is an aerial projection, a 2D bird's-eye view. This perspective flattening was implemented partly because of its suitability for visual display on a workstation monitor. Figure 5 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 MAW session. MAW adopts the simplification that all spatial sound objects are at once source and sinks. Spatial sound objects have not only rectangular coordinates, but also angular and focal attributes (described later). 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 her own, blurring the self/other distinction.

For the icons used in Figure 5, 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 this (admittedly mixed) metaphor. Each of the icons is assigned a unique channel number, used to key the spatializing backend. Instead of tracing a strict radius, the radial arm begins away from the center of the iconic circle and projects beyond the perimeter. This style conveys the object's directionality, vectoring the icon without distractingly infringing upon the halo around the user's head.

An alternative iconic representation uses top-down pictures of people's heads, as in Figure 6. Such a view has the advantage of making the birds-eye metaphor consistent, but suffers from making the users more difficult to recognize. Yet another iconic style uses the first-described "head-and-shoulders" pictures, but replaces the radial azimuth-indicating arm with image rotation, as in Figure 7. In fact, any arbitrary combinations of graphics and text may be combined into an icon.

Multiuser systems allow multiple views of the sources and sinks (user tokens). MAW allows private configurations, but also supports a shared model. A simple shared data model, in which the mutual positions are shared among all users, has (arguably) the parsimony of reality. MAW's "relaxed common view" system generalizes WYSIWIS ("what you see is what I see") displays, allowing not only for differently scrolled views of a single visual representation but different iconic representations (supporting, for instance, polyglot representation of names) as well. MAW's salient groupware feature is its support of mutually consistent (graphical and auditory) displays, like that perceived by two users seeing and hearing each other in a single virtual room.

MAW seamlessly extends WIMP interface conventions to manage distributed spatial sound objects with a variety of interaction styles [Cohen & Koizumi 92b]. Draggably rotating icons, which represent non-omnidirectional sources and sinks, answer not only to direct ("grope and grunt") manipulation, but also to arrow keys, chorded with Alternate-,[8] Shift-, and Control-; menu items and Command keys; and numeric panels, all employing the object–command (noun–verb) syntax.

Mode-indicating cursors and dynamic tracking effects contribute to the complete suite of manipulation techniques available. Table 4, crossing graphical window controls (ordered by decreasing degree of direct manipulation) and audio window operations summarizes many of the modes spanned by MAW. The styles of interaction for spatial sound objects go beyond those for the integrated graphical elements (like those used to compose icons). Direct manipulation is the main mode, but other modes have special purpose uses that defy simple categorization. As summarized in Table 5 and elaborated below, adjustment may be by either direct or indirect manipulation, may be rectilinearly constrained or unconstrained, may be continuous or discrete, and may be absolute or relative to the starting position of the icon.

Keyboard arrows can be used to move objects. Normally the keypad arrows translate the selected objects, but when modified with (either of) the Alternate keys, a rotational mode spins any selected spatial sound icons. (This extends the usual NextStep conventions: the Alternate keys are generally used to constrain action, here in the sense that they apply the arrow keys rotationally to the selected objects instead of translationally.) Rather than explicitly specifying a clockwise or counterclockwise rotation, this chorded combination causes a rotation towards the direction of the respective arrow. (This idiom is motivated by the arrangement of the NeXT arrow keys. A different scheme might have used just the left and right arrows, to rotate counter- and clock-wise, respectively, but on the NeXT keyboard the adjacent up and down arrows would have muddied this convention.) Therefore rotation quanta are specified as *up*

[8] This paper uses the notation convention that chording modifier keys and menu commands are shown surrounded by rectangles, suggesting their appearance on the keyboard or screen.

	manipulation				
	direct				indirect
	chair tracker	drag	arrow keys	menu items / Command keys	numeric panels
creation					
instantiation				Iconify & Paste v	
deletion				Deiconify & Cut x	
communication					
selection	√	√			(√)
all				Select All a	
none	√	√		Deselect All A	
extended		Shift			
constrained		Alternate			
control					
relocation		√		(Cluster { & Uncluster })	√
constrained		Alternate	√	√	√
magnified			Shift s		
extended			Control		
quantized		Set Grid...	(√)	√	
reorientation	√	√	Alternate	√	√
magnified			Shift s		
extended			Control		
quantized			(√)	Spin Clockwise (& Spin Counterclockwise)	
refocusing					√
resizing (gain)		√			√
constrained		Alternate		Inflate > & Deflate <	
quantized		Set Grid...			

Table 4: *Creation, communication,* and *control:* audio window operations × graphical window controls (Maw). A check ("√") indicates functionality, elaborated by a menu command where appropriate. Parenthesized items (ex: "(√)") are implicit. Instantiation means the creation of source/sink icons. Extended selection means adding objects individually to the current selection. Constrained selection, applied to dragging, picks objects entirely within (as opposed to merely touched by) the dragged-out area. Constrained relocation is either vertical or horizontal (but not both). Quantized repositioning involves hopping discretely to a new location or orientation.

	direct	*indirect*
continuous	dragging [with Alternate] chair tracker	editing numeric cells in Spatializer or Inspector panels
discrete	dragging across grid [with Alternate]	menu commands and Command keys arrow keys

Table 5: Modes of [constrained] manipulation: {*direct*, *indirect*} × {*continuous*, *discrete*} (Maw)

to a certain absolute $\Delta\theta$; rotating in a given direction when the object is already aligned to that direction is an idempotent no-op. Orthogonally, the Shift keys (⇑ on international keyboards) may be used to magnify the effects of the repositioning, be they translational or rotational.

Further, again independently, the Control key may be used to extend the meaning of the arrow keys, specifying "move to (some) edge." (In textual documents, a Control + arrow combination generally extends a selection.) Control + arrow moves selected objects to the "edge of themselves," leapfrogging over their former position by an amount equal to their diameter. Control + Shift + arrow moves them to the edge of the bounding rectangle circumscribing all the selected items, justifying them against the chosen edge. (Invoked on a singleton selection, this is a no-op.) Control + Shift + Shift + arrow moves selected objects to the edge of the window, justifying them against the inside of the room's wall. Control [+ Shift] + Alternate is an easy chord to push, since the buttons are collocated at the left side of the keyboard; it invokes an immediate alignment with the respective arrow, aligning all selected spatial sound objects with each other as a side effect. Table 6 details the extension and magnification effects of the chorded arrow-key combinations.

A numeric panel, opened via the Spatializer command, represents sound spatializer modules. It reflects the state of the currently active sources and sinks, and can be used, along with an Inspector, in an inverted fashion to drive the graphical display. Users can numerically edit spatial sound object attributes through the spatializer panel cells. Besides this feature's immediate usefulness, such an action might correspond to external (from a MIDI sequencer, a joystick or other controller) manipulation of the spatial sound objects, to be reflected graphically as well as auditorily.

Table 7 illustrates another way of organizing these translation and rotation operators: classification by sensitivity to current position (*absolute* vs. *relative*) and dependency on some target position (*non-convergent* vs. *convergent*). For instance, the unextended translation arrow key operators are *relative/non-convergent*, since they invoke an vector with respect to the current location, and holding down the key moves the selected objects arbitrarily distant from the original locations. Rotationally, the unextended arrow key operators are *relative/convergent*, since they invoke an orientation dependent on both the initial position and the target azimuth, repeated invocation of such a rotation converging on the target. Dragging invokes an *absolute/convergent* translation, but a *relative/convergent* rotation, since the icons swing around gradually. The extended magnifications of the translation operators span the meaningful range of {*non-convergent*, *convergent*} × {*absolute*, *relative*} crossings: "edge of self" and "edge of selection" are both *relative* operators, while "edge of window" is *absolute*. "Edge of self" is *nonconvergent*, while "edge of window" and "edge of selection" are idempotent (degeneratively *convergent*). The Spin Clockwise and Spin Counter-clockwise menu commands have no notion of a target, and so are *non-convergent/relative*, while the arrow alignment operators, invoked with extended rotation, have no dependency on original position, and so are *convergent/absolute*. The Spatializer numeric panel defies exclusive characterization: it causes the respective objects to jump to a newly specified position with no memory of the former, but since the new position may actually be edited in a numeric field (instead of simply overwritten), there is a type of sensitivity to initial position. (For example, the user might reflect the object by prepending a minus sign to the azimuth, or nudge the horizontal position from "123." to "123.4".) Therefore these menu commands are characterized *absolute+relative/convergent*.

1.2.1 Chair Tracker

The *absolute/relative* distinction becomes especially relevant in the context of MAW's chair tracker [Cohen & Koizumi 92b], crafted with a Polhemus 3Space Isotrak [Polhemus 87], 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 an egocentric display with ego- and exocentric control.

As illustrated by Figure 8, 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

Figure 3: Glove at fist site (Handy Sound)

Alternate	Control	Shift s ("magnify")		
		nor ⇒ *fine*	xor ⇒ *medium*	and ⇒ *coarse*
no ⇒ translate	no ⇒ *normal*	move 1 pixel	move 10 pixels	move 100 pixels
	yes ⇒ *extend*	move to edge of self	justify with edge of selection	justify with edge of window
yes ⇒ rotate	no ⇒ *normal*	spin up to 1°	spin up to 10°	spin up to 90°
	yes ⇒ *extend*	align with arrow	align with arrow	align with arrow

Table 6: The effect of the shift and control buttons in arrow-key translation and rotation (Maw)

Figure 4: Maw architecture

Figure 5: Screen shot (Maw)

Figure 6: Top-down icons (Maw)

Figure 7: Rotating head-and-shoulders icons (Maw)

	dependent on target position	sensitive to current position: no ⇒ *absolute*	sensitive to current position: yes ⇒ *relative*
translation	no ⇒ *non-convergent*		arrow keys, edge of self
	yes ⇒ *convergent*	numeric Spatializer and Inspector panels edge of window drag	edge of selection
rotation	no ⇒ *non-convergent*		Spin menu commands chair tracker
	yes ⇒ *convergent*	numeric Spatializer and Inspector panels align with arrow	arrow keys drag

Table 7: Sensitivity to current position and dependency on target position: {*absolute*, *relative*} × {*non-convergent*, *convergent*} for translation and rotation (Maw)

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).

2 Pitching: Manipulating Sound Quality

Sound is malleable and robust under an infinite range of manipulation. Voice and music can be gracefully mixed and distorted without loss of comprehensibility, recognizability, or euphony. Systematic blending and warping of audio channels, "pitching" the sound by manipulating its timbre, can be a useful technique for enriching the user interface. For this paper, the difference between throwing and pitching is analagous to that in baseball: in throwing, the object is to get a projectile from one place to another, but in pitching, an equally important goal is to put some spin on it.

2.1 Manipulating Sound 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 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 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 motives 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 9, and recapitulated in Figure 10, which extends the relocation scenario described earlier in Figure 3. In the example, the user points at a source, which casts an audio *spotlight* on it (top left of Figures 9 and 10); grabs it, which *muffle*s its sound (top center); uses fingers extended from a still-clenched fist to *highlight* it (bottom right); pulls it closer (top right); and finally releases it (bottom left). These three types of (sonic piggyback-channel) filtears are detailed in following subsections.

2.1.1 Spotlight

Once audio channels are distributed in space, a telepointer, a user-controlled pointer within that space, becomes useful. In visual domains, eyegaze selects the focus of attention; there is no direct analog in audio domains since audition is more omnidirectional than vision. 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. 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 § 1). 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. A related class of monitor tools are "eyecon" applications– eyes that follow the cursor (or, by extension, an arbitrary object) around the screen, allowing gaze indirection.

Handy Sound explicitly implements a perhaps better solution, an *audio spotlight*, that emphasizes one or more channels [Ludwig & Pincever 89] [Cohen 89] [Cohen & Ludwig 91a, Cohen & Ludwig 91b] [Begault

Figure 8: Chair tracker geometry (Maw)

Figure 9: State transitions and *filtears* (Handy Sound)

Figure 10: Gestured state transitions (Handy Sound)

& Wenzel 92]. This emphasis might comprise any combination of the suite of effects used by audio exciters and aural enhancers: echoes and reverberation, equalization, 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 making it 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 the selection can be an end unto itself, since the emphasis makes the selected objects more prominent. The idea is to create a "just noticeable difference," an acoustic enhancement that is noticeable but ignorable, unambiguous but unintrusive.

In practice, as the user sweeps her hand around the room, pointing at the localized sources, she gets confirmation of the direction 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.

2.1.2 Muffle

A *muffle* filtear [Cohen 89] [Cohen & Ludwig 91a, Cohen & Ludwig 91b] 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 [Sanders & McCormick 87]. The muffling effect is accomplished with a lowpass filter, as a covering hand tends to attenuate the high-frequency components of a sound source. Again, the filtear quality of just-noticeability is important in order to avoid loss of intelligibility[9] in the selected channel.

2.1.3 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. 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 can't 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 [Ludwig & Pincever 89]. 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.

2.2 Manipulating Sound Quality in MAW

Clusters are hierarchically collapsed groups 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 spatializer 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 are not displayed, but as it approaches, a group appears 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.

[9] A filtear to mute the channels of boring or boorish speakers might be called a *muzzle*. Perhaps for this type of filtear, loss of intelligibility wouldn't be so bad.

logical organization of hierarchical structure Individually recording (or mic-ing or synthesizing) the individual instruments, presenting each of the channels to MAW, and mixing them at audition (or performance) 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 8), 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 orchestra → section → instrument and actually break down the instrument itself. This super decomposition aspect of the *cluster* feature could allow, for example, separating the strings of an instrument, as if the user were shrunk and listening out the hole of a guitar, say.

MAW features such a Cluster utility. When associated, the clustered elements merge into one another: aligning and converging until they are spatially and graphically indistinguishable. Unlike members of a group that maintain their mutual distance, members of a cluster temporarily lose their identity. The last-selected[10] icon is designated as the model: when the cluster operation is invoked, the other selected icons dynamically approach it, converging on its position, orientation, and display attributes.

Once clustered, a bundle of sources moves and resizes together. The semantics of the manipulations of the cluster reflect the philosophy that operations which apply to the spatializer parameters— location, gain (size), orientation, etc.— get passed along by the cluster to its constituents, but that purely graphical attributes— color, line thickness, fill mode, outline mode etc.— are not transmitted.

Uncluster is the inverse of Cluster. When invoked, the top level of nested components replaces its selected parents. The clustered objects retain a memory of their former characteristics, including relative displacement from their (possibly relocated) cluster icon, so that unclustering restore these attributes, in the inverse, dynamically animated, operation. The animation comprises a spatiotemporal interpolation between the original and target states, restoring the objects to their natural attributes.

Unclustering can be likened to viewing the sources through a generalized multifocal fisheye lens [Furnas 86], 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 (in perspective) approach the user (and recede from each other). While the other objects do not get pushed into the background, the idea is the same: to effect a external rearrangement of sources that compliments an internal reordering.

Since clusters can be thought of as modeling actual objects, which are indeed composed of sound-producing particles of molecular fineness, this approach seems intuitive. Eventually, individually spatialized audio channels will be cheap enough that we can think of sound in this granular fashion. By way of analogy to pixels and voxels, we sometimes call these atomic sounds "mixels," acronymic for sound **mix**ing **el**ements.

3 Catching: Manipulating Sound Volume

In between throwing and pitching is the issue of gain control. In a dry system, gain adjustment is as indistinguishable from distance effects as it is from simply talking, singing, or playing louder. In the spirit of idioms like "catch what someone said" and "catch someone's ear," along with "broadcatch" (meaning selective reception of broadcast [Brand 87]), gain control can be thought of as catching sound.

3.1 Manipulating Sound Volume in Handy Sound

In Handy Sound, the only way to adjust gain is to bring a source closer. Volume is controlled by closeness/distance effects; gain is set inversely proportional to the cube of the virtual distance from the source. While the user may simply adjust the volume of the headphone mixer, the only pure way to make everyone louder is by pulling everyone closer individually, by grabbing each of their voices and pulling.

[10] The last-selected icon was chosen as the paragon so that the user can easily reposition the bundle. That is, the new cluster is created in the same place as the last-selected component, so that, if the Command key equivalent is used (instead of the menu item), the cursor is already in a position to relocate the new cluster.

Table 8: Concertal decomposition

3.2 Manipulating Sound Volume in MAW

In exocentric systems like MAW, however, it is possible to positionally adjust perceived gain in two different ways: sidle (the sink) up to a speaker or group of sources, or move the sources nearer to the 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 left icon of Figure 5). The user may also shrink or grow an atomic spatial sound icon by editing the numeric cell corresponding to its size in the Spatializer panel. 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, icon size is used as a determinant of both metaphorical ear and mouth size. Overall 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 her own icon. Gain is also inversely proportional to the square of 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 the non-omnidirectional sources. The chosen relationship specifies an azimuth-dependent beaming of the speaker. The overall *gain* of a source→sink transmission, independent of the sink's transfer function, can be calculated [Cohen & Koizumi 92c] as:

$$gain = \begin{cases} \dfrac{1 + (focus_{\text{source}})(\cos(\theta))}{1 + (focus_{\text{source}})} \cdot \dfrac{(size_{\text{source}})(size_{\text{sink}})}{distance^2} & \text{far-field} \\ \\ \dfrac{1 + (focus_{\text{source}})(\cos(\theta))}{1 + focus_{\text{source}}} & \text{near-field} \end{cases} \tag{3}$$

where $0 \leq focus < 1$,
θ is the rotated angle of inclination of the sink with respect to the source,
and *distance* is between the source and the sink,
in the same units as the respective *size* s.

The $focus$ represents the dispersion, or beaming, of the sound. The first term of the extended eqn. (3) captures the rotational effects of the source; the normalized fraction uses the $focus$ coefficient to scale the contribution of the azimuth. For a $focus$ of zero, the radiation pattern is omnidirectional (as the first terms of the extended expression reduce to unity). A $focus$ of greater than zero enables a cardioidal pattern, as if the source were using a megaphone. As two objects approach each other, the second far-field fraction factor approaches unity, and is rectified to unity as the objects osculate and overlap, invoking near-field behavior. The icon shown in the top left of Figure 11 has $focus = .1$, and a corresponding radiation pattern (sound field density) in the top right is almost omnidirectional. In contrast, the icon in the bottom left has $focus = .9$ (as indicated by its thicker arm), and its radiation pattern in the bottom right is almost perfectly quiet in the shadow behind the head.

4 Conclusion

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 are a way of organizing and controlling sound. Handy Sound was conceived as a feasibility study, testing purely auditory presentation and purely gestural control. Deployed only briefly in an expensive lab environment, only a few users had a chance to play with it (before support for the project was terminated). MAW was designed to exploit more I/O modalities, and is consequently more accesible; it is deployed in an office environment, where it currently enjoys frequent demos. New media spend their early years recapitulating the modes of older media [McLuhan & Fiore 67]; the research described by this paper hopes to abbreviate this phase for audio windows by accelerating its conceptual development.

Neologisms introduced by this research include *piggyback-channels*, whose sonic instances employ filtears to reflect control state, and *mixel*s, atomic sound **mix**ing **el**ements. Taxonomies introduced include

Figure 11: Icons and radiation patterns (Maw)

analogs between I/O device generations and dimensionality; throwing/pitching/catching metaphors, capturing state transitions in virtual space, timbre space, and volume space; audio window operations crossed with graphical window controls; and sensitivity to current position crossed with dependency on some target position. The two audio windowing systems described by this paper can be compared and analyzed in the context of global design issues.

dimensionality While Handy Sound manipulates objects in three spatial dimensions, MAW has only two (not counting rotation). MAW's virtual room "gods' eye" orthographic projection was implemented because of its suitability for visual display on a workstation monitor, and the planar slice was chosen partly for maximal positional distinction: azimuthal discrimination is easier than elevation discrimination [Wenzel et al. 88b]. While the engineering required to upgrade from 2D to 3D audio is straightforward, graphical manipulation is simpler than corresponding 3D techniques, and MAW's operation set is consequently richer than Handy Sound's.

feedback mechanisms and C/R ratio When one gives sound a physical manifestation, it can become an icon for anything imaginable. Audio windowing systems offer a way to organize acoustic space, and the interpretation of gestures and the reinterpretation of WIMP conventions seem natural frontends to such systems. Audio windowing 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. Everything in the systems should be manifest, and objects and actions should be understandable purely in terms of their effect on the respective displays.

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. Handy Sound and MAW share several features, including the use of a direct manipulation object–command (noun–verb) syntax and continuous feedback (dynamic tracking). Handy Sound features a variable radial C/R ratio to map the work envelope into perceptual space. MAW's extended arrow key combinations, amplifiable with the Shift keys, can also be thought of as invoking a variable C/R ratio. Both systems use a Polhemus, deployed as a head- or a chair-tracker, for soundscape stabilization.

The notion of a changing cursor to indicate mode or control state was also employed by both systems. In MAW, selected objects sprout knobbies, providing visual feedback about the selection set as well as handles for resizing. During object relocation, the default arrow cursor changes to a hand. An open hand suggesting rectilinear translation is used for non-spatial sound objects, while a hand with a pointed pivot finger is used for spatial sound icons, suggesting rotation around the hotspot at the tip of the finger. Handy Sound's filtears, deployed as *piggyback-channels*, are sonic analogs of these cues: the spotlight can be compared to the arrow cursor, indicating the direction of interest, and the muffle is like the hand cursor, indicating relocation.

gain determination and distance modeling Handy Sound used an inverse cube relationship, while the release of MAW described in this paper used an inverse squared falloff. (Subsequent research [Cohen & Koizumi 93b] has generalized gain even further.) As loudness perception is logarithmic, the choice of the exponent is relatively unimportant, and can be thought of as a tunable parameter.

Both Handy Sound and MAW are intended to interpolate between conventional telephony and VR, but cannot be said to do more than suggest actual acoustic environments. Simulating distance cues is a difficult and not-yet-solved problem which goes beyond their simple gain changes. Besides peoples' natural inability to estimate distance with precision and whatever deliberate exaggeration of distance effects, an inverse (exponential) relation does not perfectly capture real effects [Blauert 83] [Begault 91] [Wenzel 92]. 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 each digitized source thrice: once for source rotation, and twice (left and right ears) for sink revolution. Both systems further over-simplify 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 reflections (discrete echoes), the ratio of whose energy to that of direct sounds is another cue for estimating distance; and reverberation, statistically averaged room ambience.

modality integration Auditory localization, especially distance perception, is difficult. Experience with Handy Sound indicates that active control of (especially multiple) sources is difficult with only an auditory display, even with filtears, making a visual display useful for confirmation of source placement. Without a debugging monitor, manipulation of a spatial environment was awkward. In Maw, visual and acoustic displays complement each other; a glance at a map can disambiguate auditory cues.

permission system 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 actually deployed in multiple-user 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. Each user controls the half-duplex of sounds that she hears, while her utterances are considered public multicasts, available for arbitrary spatializing and filtearing by other users in the conference. Maw's exocentric paradigm depends on social conventions to establish its implicit permission system. Maw also has a token-passing scheme to ensure mutex on non-audio windowing elements, shared graphical objects in its distributed whiteboard paradigm.

perspective Both egocentric and exocentric displays can be effective paradigms for audio windowing systems. In egocentric displays like Handy Sound, the user has no explicit representation. Such an egocentric model is compatible with immersive VR-style systems: gestural interpretation control is parsimonious, a natural extension of our normal mode of rearranging the world. Without reliance on visual media, it is especially suitable for visually impaired users. An exocentric paradigm like Maw's blurs the self/other distinction by iconifying sources and sinks with similar tokens. The other–self difference manifests as that between transitive and reflexive actions, which are syntactically indistinguishable. 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 forked presence, existing in multiple virtual places simultaneously.)

resource allocation In order to support an individually-configurable teleconferencing system, a large number of audio channels must be channeled through audio imaging processors. Since, in a full-duplex conference, every user must spatialize every other user's voice, the total number of channels to spatialize grows quadratically, or as $O(|users|^2)$. 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 auralization, a clustering mechanism like Maw's is useful, both as a way of imposing a logical hierarchy on many sources, and in conjunction with an audio switching matrix like Handy Sound's, as a way of conserving channels.

Handy Sound and Maw instantiate audio windowing systems via elaborations of gestural and WIMP conventions, respectively. Some of their features find close approximation in visual or physical media; others elude analogy. The explored models give users a telepresence into metaphorical space. Using systems like Handy Sound and Maw for audio windowing can be likened to sonic (analytic) cubism: they present multiple simultaneous audio perspectives on a conference or concert.

5 Acknowledgements

Thanks to Elizabeth M. Wenzel, A. Toni Cohen, my collaborators in the IMAL group at Bellcore, the HITLab at the University of Washington, and the Speech and Acoustics Human Interface Lab at NTT, and anonymous referees for their thoughtful and valuable suggestions.

References

[Addeo et al. 88] E. J. Addeo, A. B. Dayao, A. D. Gelman, and V. F. Massa. An experimental multimedia bridging system. In R. B. Allen, editor, *Proc.* **C***onf. on* **O***ffice* **I***nformation* **S***ystems*, pages 236–242, Palo Alto, CA, March 1988.

[Akai 89a] Akai. *DP3200 Audio Digital Matrix Patch Bay.* Akai Digital, P.O. Box 2344; Fort Worth, TX 76113, 1989.

[Akai 89b] Akai. *S900* MIDI *Digital Sampler Operator's Manual.* Akai Professional, P.O. Box 2344; Fort Worth, TX 76113, 1989.

[Aphex 89] Aphex. *Aural Exciter Type C Model 103A Operating Guide.* Aphex Systems Ltd., 13340 Saticoy St.; North Hollywood, CA 91605, 1989.

[Baecker & Buxton 87] R. M. Baecker and W. A. Buxton. *The Audio Channel*, chapter 9, pages 393–399. Morgan Kaufmann Publishers, Inc., 1987. ISBN 0-934613-24-9.

[Begault & Wenzel 92] D. R. Begault and E. M. Wenzel. Techniques and applications for binaural sound manipulation in man-machine interfaces. **I***nt.* **J***. of* **A***viation* **P***sychology*, 2(1):1–22, 1992.

[Begault 91] D. R. Begault. Preferred sound intensity increase for sensation of half distance. *Perceptual and Motor Skills*, 72:1019–1029, 1991.

[Benedikt 91] M. Benedikt. *Cyberspace: First Steps.* MIT Press, Cambridge, MA, 1991. ISBN 0-262-02327-X.

[Bernardini & Otto 89] N. Bernardini and P. Otto. Trails: An interactive system for sound location. In *Proc. ICMC:* **I***ntnl.* **C***omp.* **M***usic* **C***onf.*, pages 29–33. Computer Music Association, November 1989.

[Blattner & Dannenberg 92] M. M. Blattner and R. B. Dannenberg, editors. *Multimedia Interface Design.* ACM Press: Addison-Wesley, 1992. ISBN 0-201-54981-6.

[Blattner & Greenberg 89] M. M. Blattner and R. M. Greenberg. Communicating and learning through non-speech audio. In A. D. N. Edwards, editor, *Multi-media Interface Design in Education.* Springer-Verlag, August 1989.

[Blattner 92] M. M. Blattner. Messages, Models, and Media. *Multimedia Review*, 3(3):15–21, Fall 1992.

[Blattner et al. 89] M. M. Blattner, D. A. Sumikawa, and R. M. Greenberg. Earcons and Icons: Their Structure and Common Design Principles. *Human-Computer Interaction*, 4(1):11–44, 1989.

[Blattner et al. 92] M. M. Blattner, R. M. Greenberg, and M. Kamegai. Listening to Turbulence: An Example of Scientific Audiolization. In M. M. Blattner and R. B. Dannenberg, editors, *Multimedia Interface Design*, chapter 6, pages 87–102. Addison-Wesley, 1992. ISBN 0-201-54981-6.

[Blauert 83] J. Blauert. *Spatial Hearing: The Psychophysics of Human Sound Localization.* MIT Press, 1983. ISBN 0-262-02190-0.

[Bly 82] S. A. Bly. Presenting information in sound. In *Proc. CHI:* ACM *Conf. on* **C***omputer-***H***uman* **I***nteraction*, pages 371–375, New York, NY, 1982. ACM.

[Bly 87] S. A. Bly. Communicating with sound. In R. M. Baecker and W. 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.

[Boulez & Gerzso 88] P. Boulez and A. Gerzso. Computers in music. *Scientific American*, 258(4):44–50, April 1988.

[Brand 87] S. Brand. *The Media Lab.* Viking Press, 1987.

[Bregman 90] A. S. Bregman. *Auditory Scene Analysis.* MIT Press, Cambridge, MA, 1990. ISBN 0-262-02297-4.

[Bricken 92] W. Bricken. Virtual Reality: Directions of Growth. In *Proc. Imagina*, pages I27–I40, Monte Carlo, January 1992.

[Buxton et al. 89] 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.

[Cherry 53] E. C. Cherry. Some experiments on the recognition of speech, with one and with two ears. **J**. **A***cous.* **S***oc.* **A***mer.*, 25(5):975–979, September 1953.

[Chowning 70] J. M. Chowning. The simulation of moving sound sources. In *AES:* **A***udio* **E***ngineering* **S***ociety Conv.*, May 1970. Preprint 726 (M-3).

[Chowning 77] J. M. Chowning. The simulation of moving sound sources. *Computer Music J.*, 1(3):48–52, June 1977.

[Cohen & Koizumi 91a] M. Cohen and N. Koizumi. Audio window. In *Den Gaku.* Tokyo Contemporary Music Festival: Music for Computer, December 1991.

[Cohen & Koizumi 91b] M. Cohen and N. 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.

[Cohen & Koizumi 91c] M. Cohen and N. Koizumi. Audio windows for sound field telecommunication. In *Proc. Seventh Symp. on Human Interface*, pages 703–709, Kyoto, October 1991. SICE (**S**ociety of **I**nstrument and **C**ontrol **E**ngineers). 2433.

[Cohen & Koizumi 92a] M. Cohen and N. Koizumi. Audio windows: User interfaces for manipulating virtual acoustic environments. In *Proc. ASJ:* **A***coustical* **S***ociety of* **J***apan Spring Meeting*, pages 479–480, Tokyo, March 1992. Special Session on Virtual Reality, 2-5-12.

[Cohen & Koizumi 92b] M. Cohen and N. Koizumi. Iconic control for audio windows. In *Proc. Eighth Symp. on Human Interface*, pages 333–340, Kawasaki, Japan, October 1992. SICE (**S**ociety of **I**nstrument and **C**ontrol **E**ngineers). 1411.

[Cohen & Koizumi 92c] M. Cohen and N. Koizumi. *Exocentric Control of Audio Imaging* in Binaural Telecommunication. IEICE *Trans. on Fundamentals of Electronics, Communications and Computer Sciences* (Special Section on Fundamentals of Next Generation Human Interface), E75-A(2):164–170, February 1992. 0916-8508.

[Cohen & Koizumi 93a] M. Cohen and N. Koizumi. Audio windows for virtual concerts. In *Proc. JMACS:* **J***apan* **M***usic* **A***nd* **C***omputer Science* **S***ociety Meeting*, pages 27–32, Tokyo, February 1993. No. 47.

[Cohen & Koizumi 93b] M. Cohen and N. Koizumi. Virtual gain for audio windows. In *HCI: Proc.* **H***uman-***C***omputer* **I***nteraction*, page 283, Orlando, FL, August 1993. Poster.

[Cohen & Ludwig 91a] M. Cohen and L. F. Ludwig. Multidimensional audio window management. *IJMMS: J. of Person-Computer Interaction*, 34(3):319–336, March 1991. Special Issue on Computer Supported Cooperative Work and Groupware. ISSN 0020-7373.

[Cohen & Ludwig 91b] M. Cohen and L. F. Ludwig. Multidimensional audio window management. In S. Greenberg, editor, *Computer Supported Cooperative Work and Groupware*, chapter 10, pages 193–210. Academic Press, London, 1991. ISBN 0-12-299220-2.

[Cohen & Wenzel 95] M. Cohen and E. M. Wenzel. The design of multidimensional sound interfaces. In W. Barfield and T. A. Furness III, editors, *Virtual Environments and Advanced Interface Design*, chapter 8, pages 291–346. Oxford University Press, 1995. ISBN 0-19-507555-2.

[Cohen 87] M. Cohen. Stereotelephonics. Internal Memorandum IM-000-21460-87-04, **Bell Co**mmunications **Re**search, October 1987.

[Cohen 89] M. Cohen. Multidimensional audio window management. Technical Memorandum TM-NPL-015362, **Bell Co**mmunications **Re**search, October 1989.

[Cohen 90] M. Cohen. *Multidimensional Audio Windows: Extending User Interfaces through the Use of Spatial Auditory Information.* PhD dissertation, Northwestern University, December 1990.

[Cohen et al. 92] M. Cohen, N. Koizumi, and S. Aoki. Design and control of shared conferencing environments for audio telecommunication. In *Proc. ISMCR:* **I***nt.* **S***ymp. on* **M***easurement and* **C***ontrol in* **R***obotics*, pages 405–412, Tsukuba Science City, Japan, November 1992. SICE (**S**ociety of **I**nstrument and **C**ontrol **E**ngineers).

[Dayao & Gelman 87] A. B. Dayao and A. D. Gelman. Graphic user control interface for multi-point conferencing. Technical memorandum, **Bell Co**mmunications **Re**search, 1987.

[Deatherage 72] B. H. Deatherage. Auditory and other sensory forms of information presentation. In H. P. V. Cott and R. G. Kinkade, editors, *Human Engineering Guide to Equipment Design.* U.S. Government Printing Office, Washington, DC, 1972.

[Edwards 87] A. 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.

[Edwards 88] A. D. N. Edwards. The design of auditory interfaces for visually disabled users. In *Proc. CHI:* ACM *Conf. on* **C***omputer-***H***uman* **I***nteraction*, pages 83–88, 1988.

[Fisher et al. 86] 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.

[Fisher et al. 88] 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.

[Foley 87] J. D. Foley. Interfaces for advanced computing. *Scientific American*, 257(4):126–135, October 1987.

[Foster 89] S. Foster. *Convolvotron.* Crystal River Engineering, 1989.

[Furnas 86] G. W. Furnas. Generalized fisheye views. In *Proc. CHI:* ACM *Conf. on* **C***omputer-***H***uman* **I***nteraction*, pages 16–23, Boston, April 1986.

[Gaver 86] W. W. Gaver. Auditory icons: Using sound in computer interfaces. *Human-Computer Interaction*, 2(2):167–177, 1986.

[Gaver 89] W. W. Gaver. The SonicFinder: An Interface that Uses Auditory Icons. *Human-Computer Interaction*, 4(1):67–94, 1989.

[Gaver et al. 91] W. W. Gaver, R. B. Smith, and T. O'Shea. Effective sounds in complex systems: The ARKola simulation. In *Proc. CHI:* ACM *Conf. on* **C***omputer-***H***uman* **I***nteraction*, pages 85–90, 1991.

[Gehring 87] B. Gehring. *Auditory Localizer Model AL-201 Product Description.* Gehring Research Corporation, 189 Madison Avenue, Toronto, Ontario M5R 2S6, October 1987.

[Gehring 88] B. Gehring. U.S. Patent 4774515: Attitude Indicator. 189 Madison Avenue, Toronto, Ontario M5R 2S6, September 1988.

[Gehring 90] B. Gehring. *Focal Point*™ *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.

[Gibson 79] J. J. Gibson. *The ecological approach to visual perception.* Houghton Mifflin, Boston, 1979. ISBN 0-89859-959-8.

[IOtech 91] IOtech. SCSI*488/N Bus Controller.* IOtech, Inc., 25971 Cannon Rd.; Cleveland, OH 44146, 1991.

[Kendall & Freed 88] G. S. Kendall and D. J. Freed. Scientific visualization by ear. Technical report, Northwestern Computer Music, Northwestern University; Evanston, IL 60208, 1988.

[Kendall & Martens 84] G. S. Kendall and W. L. Martens. Simulating the cues of spatial hearing in natural environments. In *Proc. ICMC:* **I***ntnl.* **C***omp.* **M***usic* **C***onf.*, pages 111–126, Paris, 1984. Computer Music Association.

[Kendall 90] G. S. Kendall. Visualization by ear: Auditory imagery for scientific visualization and virtual reality. In A. Wolman and M. Cohen, editors, *Proc. Dream Machines for Computer Music*, pages 41–46, School of Music, Northwestern University, November 1990.

[Kendall 91] G. S. Kendall. Visualization by ear: Auditory imagery for scientific visualization and virtual reality. *Computer Music J.*, 15(4):70–73, Winter 1991.

[Kendall et al. 86] G. S. Kendall, W. L. Martens, D. J. Freed, M. D. Ludwig, and R. W. Karstens. Image model reverberation from recirculating delays. In *AES:* **A***udio* **E***ngineering* **S***ociety Conv.*, New York, 1986.

[Kendall et al. 89] G. S. Kendall, W. L. Martens, and S. L. Decker. Spatial reverberation: Discussion and demonstration. In M. V. Mathews and J. R. Pierce, editors, *Current Directions in Computer Music Research.* MIT Press, 1989.

[Kendall et al. 90] G. S. Kendall, W. L. Martens, and M. D. Wilde. A spatial sound processor for loudspeaker and headphone reproduction. In *AES:* **A***udio* **E***ngineering* **S***ociety Conv.*, Washington, D.C., May 1990. ISBN 0-937803-15-4.

[Koenig 50] W. Koenig. Subjective effects in binaural hearing. **J**. **A***cous.* **S***oc.* **A***mer.*, 22(1):61–62, January 1950.

[Koizumi et al. 92] N. Koizumi, M. Cohen, and S. Aoki. Design of virtual conferencing environments in audio telecommunication. In *AES:* **A***udio* **E***ngineering* **S***ociety Conv.*, Wien, Austria, March 1992. 4CA1.04, preprint 3304.

[Kraemer & King 86] K. L. Kraemer and J. L. 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.

[Krueger 82] M. W. Krueger. *Artificial Reality.* Addison-Wesley, Reading, MA, 1982.

[Krueger 91] M. W. Krueger. *Artificial Reality II.* Addison-Wesley, Reading, MA, 1991. ISBN 0-201-52260-8.

[Lantz 86] K. A. Lantz. An experiment in integrated multimedia conferencing. Technical report, Department of Computer Science, Stanford University, Stanford, CA 94305, December 1986.

[Loomis et al. 90] J. M. Loomis, C. Hebert, and J. G. Cicinelli. Active localization of virtual sounds. **J**. **A***cous.* **S***oc.* **A***mer.*, 88(4):1757–1763, October 1990.

[Ludwig & Pincever 89] L. F. Ludwig and N. C. Pincever. Audio windowing and methods for its realization. Technical Memorandum TM-NPL-015361, **Bell Co**mmunications **Re**search, October 1989.

[Ludwig 89] L. F. Ludwig. Real-time multi-media teleconferencing: Integrating new technology. Technical Report, **Bell Co**mmunications **Re**search **I**ntegrated **M**edia **A**rchitecture **L**aboratory, Red Bank, NJ 07746, 1989.

[Ludwig et al. 90] L. F. Ludwig, N. C. Pincever, and M. Cohen. Extending the notion of a window system to audio. *(*IEEE*) Computer* (Special Issue on Voice in Computing), 23(8):66–72, August 1990. ISSN 0018-9162.

[Lunney et al. 83] 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.

[Mansur 84] D. L. Mansur. Graphs in sound: A numerical data analysis method for the blind. Report UCRL-53548, Lawrence Livermore National Laboratory, June 1984.

[Mansur 87] D. L. Mansur. Communicating with sound. In R. M. Baecker and W. A. S. Buxton, editors, *Readings in Human-Computer Interaction: A Multidisciplinary Approach*, chapter 9, pages 421–423. Morgan Kaufmann, 1987. ISBN 0-934613-24-9.

[Mansur et al. 85] D. L. Mansur, M. M. Blattner, and K. I. Joy. Sound-graphs: A numerical data analysis method for the blind. In *Proc.* **H***awaii* **I***nt.* **C***onf. on* **S***ystems* **S***ciences*, January 1985.

[Martel 86] A. Martel. The SS-1 sound spatializer: A real-time MIDI spatialization processor. In *Proc. ICMC:* **I***ntnl.* **C***omp.* **M***usic* **C***onf.*, pages 305–307. Computer Music Association, October 1986.

[Martens 89] W. Martens. Spatial image formation in binocular vision and binaural hearing. In *Proc.* 3D **M***edia* **T***echnology Conf.*, Montréal, Québec, May 1989.

[Masaki et al. 91] S. Masaki, N. Kanemaki, H. Tanigawa, H. Ichihara, and K. 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.

[McKinley & Ericson 88] R. L. McKinley and M. A. Ericson. Digital synthesis of binaural auditory localization azimuth cues using headphones. **J***.* **A***cous.* **S***oc.* **A***mer.*, 83:S18, Spring 1988.

[McLuhan & Fiore 67] H. M. McLuhan and Q. Fiore. *The Medium is the Message.* Random House, 1967.

[Mezrich et al. 84] 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.

[Moore 83] F. R. Moore. A general model for spatial processing of sounds. *Computer Music J.*, 7(3):6–15, Fall 1983.

[Nielsen et al. 88] L. Nielsen, M. Mahowald, and C. Mead. SeeHear. Technical report, California Institute of Technology, 1988.

[Patterson 82] R. R. Patterson. Guidelines for auditory warning systems on civil aircraft. Paper No. 82017, Civil Aviation Authority, London, 1982.

[Polhemus 87] Polhemus. 3SPACE ISOTRAK^{TM} *User's Manual.* Polhemus Navigation Science Division, McDonnell Douglas Electronic Company, Colchester, VT, May 1987.

[Pollack & Ficks 54] I. Pollack and L. Ficks. Information of elementary multidimensional auditory displays. **J***.* **A***cous.* **S***oc.* **A***mer.*, 26(2):155–158, 1954.

[Sanders & McCormick 87] M. S. Sanders and E. J. McCormick. *Human Factors in Engineering and Design.* McGraw-Hill, New York, sixth edition, 1987. ISBN 0-07-044903-1.

[Sarin & Greif 85] S. Sarin and I. Greif. Computer-based real-time conferencing systems. *(*IEEE*) Computer*, 18(10):33–45, October 1985.

[Scott 89] D. Scott. A processor for locating stationary and moving sound sources in a simulated acoustical environment. In *Proc. ICMC:* **I***ntnl.* **C***omp.* **M***usic* **C***onf.*, pages 277–280. Computer Music Association, November 1989.

[Setton 90] M. Setton. Sonic Navigator™. Project Report, Berkeley Systems, Inc., 1990.

[Smith et al. 90] S. Smith, R. D. Bergeron, and G. G. Grinstein. Stereophonic and surface sound generation for exploratory data analysis. In J. C. Chew and J. Whiteside, editors, *Proc. CHI:* ACM *Conf. on* **C***omputer-***H***uman* **I***nteraction*, pages 125–132, Seattle, WA, April 1990. Addison-Wesley.

[Smith et al. 92] S. Smith, R. D. Bergeron, and G. G. Grinstein. Stereophonic and Surface Sound Generation for Exploratory Data Analysis. In M. M. Blattner and R. B. Dannenberg, editors, *Multimedia Interface Design*, chapter 11, pages 173–182. Addison-Wesley, 1992. ISBN 0-201-54981-6.

[Sonnenwald et al. 90] D. H. Sonnenwald, B. Gopinath, G. O. Haberman, W. M. Keese, III, and J. S. Myers. Infosound: An audio aid to program comprehension. In *Proc.* **H***awaii* **I***nt.* **C***onf. on* **S***ystems* **S***ciences*, Honolulu, January 1990.

[Sorkin et al. 89] R. D. Sorkin, F. L. Wightman, D. J. Kistler, and G. 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.

[Stefik et al. 86] M. Stefik, D. 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.

[Stefik et al. 87] M. Stefik, G. Foster, D. G. Bobrow, K. Kahn, S. Lanning, and L. Suchman. Beyond the chalkboard: Computer support for collaboration and problem solving in meetings. *Communications of the* ACM, 30(1):32–47, January 1987.

[Sumikawa et al. 86] D. A. Sumikawa, M. M. Blattner, K. I. Joy, and R. M. Greenberg. Guidelines for the syntactic design of audio cues in computer interfaces. In *Proc.* **H***awaii* **I***nt.* **C***onf. on* **S***ystems* **S***ciences*, Honolulu, 1986.

[Tachi 91] S. Tachi, editor. *Proc. ICAT:* **I***nt.* **C***onf. on* **A***rtificial Reality and* **T***ele-Existence*, Tokyo, July 1991.

[Tachi 92] S. Tachi, editor. *Proc. ICAT:* **I***nt.* **C***onf. on* **A***rtificial Reality and* **T***ele-Existence*, Tokyo, July 1992.

[Tanigawa et al. 91] H. Tanigawa, T. Arikawa, S. Masaki, and K. Shimamura. Personal multimedia-multipoint teleconference system. In *Proc.* IEEE *InfoCom'91*, pages 1127–1134, April 1991.

[Urei 80] Urei. *Dual Parametric Equalizer Model 546 Operating Instructions.* Urei (United Recording Electronics Industries), 8460 San Fernando Rd.; Sun Valley, CA 91352, 1980.

[VPL 87] VPL. *DataGlove Model 2 Operating Manual.* VPL (Visual Programming Language) Research, Inc., 656 Bair Island Rd.; Suite 304; Redwood City, CA 94063, 1987.

[Wenzel 92] E. M. Wenzel. Localization in virtual acoustic displays. *Presence: Teleoperators and Virtual Environments*, 1(1):80–107, 1992. ISSN 1054-7460.

[Wenzel et al. 88a] E. M. Wenzel, F. L. Wightman, and S. H. Foster. Development of a three-dimensional auditory display system. SIGCHI *Bulletin*, pages 52–57, 1988. 20.

[Wenzel et al. 88b] E. M. Wenzel, F. L. Wightman, and S. 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.

[Wenzel et al. 90] E. M. Wenzel, P. K. Stone, S. S. Fisher, and S. 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.

[Wenzel et al. 91] E. M. Wenzel, F. L. Wightman, and D. J. Kistler. Localization of non-individualized virtual acoustic display cues. In S. P. Robertson, G. M. Olson, and J. S. Olson, editors, *Proc. CHI:* ACM *Conf. on* **C***omputer-***H***uman* **I***nteraction*, New Orleans, LA, May 1991. Addison-Wesley. ISBN 0-201-51278-5.

Contents

List of Figures

List of Tables