# Cyberspatial Audio Technology

Michael Cohen, Jens Herder, and William L. Martens
Spatial Media Group, University of Aizu
Aizu-Wakamatsu, Fukushima-ken 965-8580
*e-mail:* {mcohen,herder,wlm}@u-aizu.ac.jp
*www*: www.u-aizu.ac.jp/{~mcohen,~herder,~wlm}

Keywords: eartop computing, groupware, interactive sound spatialization, virtual reality.

*Pacs* numbers: 43.66.Pn (Binaural hearing), 43.60.-c (Acoustic signal processing),

43.88.Vk (Stereophonic reproduction, quadriphonics), 43.55.Br (Room acoustics)

## 1 Introduction

Current foci of spatial audio research in recent literature comprise sound localization; lateralization and binaural masking; echoes, precedence, and depth perception; motion perception; sound source segregation and free-field masking; physiology of spatial hearing; models of spatial hearing; (childhood) development of spatial hearing; and applications of binaural technology to auditory displays for human-computer interaction [15]. To cut across these categories in an attempt to outline the current state-of-the-art in spatial auditory displays for a particular range of applications, with an emphasis upon the expected performance of the technology in producing specific user responses required for those applications, this paper considers the value of spatial audio technology in the creation and presentation of virtual environments. The shared synthetic worlds that networked computer users occupy constitute an alternative reality that has come to be termed 'cyberspace.' Auditory display technology that attempts to provide such users with satisfying experiences of virtual acoustical space is termed here "cyberspatial audio" technology.

Cyberspatial audio applications are distinguished from the broad range of spatial audio applications in a number of important ways that help to focus this review. Most significant is that cyberspatial audio is most often designed to be responsive to user inputs. In contrast to non-interactive auditory displays, cyberspatial auditory displays typically allow active exploration of the virtual environment in which users find themselves. Thus, at least some portion of the audio presented in a cyberspatial environment must be selected, processed, or otherwise rendered with minimum delay relative

to user input. Besides the technological demands associated with realtime delivery of spatialized sound, the type and quality of auditory experiences supported are also very different from those associated with displays that support stationary sound localization.

A comprehensive review of the scientific foundations of binaural technology in physical acoustics and psychoacoustics is beyond the scope of this paper. But an informed discussion of the generic applications made possible by cyberspatial audio technology requires an appreciation of what components of auditory response are well controlled using current technology, and what capabilities are not well supported or have not been reduced to practice for widespread application. Current binaural technology is able to create a reasonable match to physical acoustical stimuli, but cannot promise to create the auditory experience specified by the application designer. That cognitive factors play a strong role in determining the user experience is central to this failure of auditory displays to produce specified results in terms of human perception. Jens Blauert (author of the seminal reference "Spatial Hearing" [4]) summarized this state of affairs in the following way [15, p. 593]:

> "It becomes clear that today's binaural technology rests mainly on our knowledge of the physical aspect, with an increasing use of psychoacoustics. Technology exploitation of the cognitive psychology of binaural hearing has barely begun."

| human pilot | representative (projected presence) |
|---|---|
| carbon community | avatar |
| RL (**r**eal **l**ife) | electronic puppet |
| meatspace | synthespian (**synth**etic **thespian**) |
| motion capture | vactor (**v**irtual **actor**) |

Table 1: User and delegate: An exocentric model in which a user is represented by an icon in the context of a virtual space is useful in spatial sound systems; virtual environments with audio can be thought of as graphical mixing consoles.

.

## 2 Psychoacoustic Foundations

The three basic acoustical phenomena that might be simulated by cyberspatial audio systems can be quickly summarized. First, for a sound source located off the listener's median plane, there is the delay in time of arrival at the ear further from the source. Manipulating this ITD (**i**nteraural **t**ime **d**elay) is useful in shifting the auditory image along the interaural (left↔right) axis through the listener's head. Second, there is the ILD

(**i**nteraural **l**evel **d**ifference) between signals arriving at the ears, which manifests as a head shadow only at higher frequencies (above around 1.5 kHz), unless the sound source is located at very close range [13]. (For example, at a range of 10 cm the low-frequency amplitude difference can exceed 25 dB, in contrast to the 2 dB difference at a range of 300 cm.) Introducing amplitude differences at low frequency, despite the deviation from realism for more distant sources, is also very effective in steering the auditory image to the left or right. The third acoustical phenomenon associated with binaural technology is the complex filtering effects of the outer ear. These pinna effects provide the primary means for fixed human listeners (with stationary heads) to localize stationary sound sources above, or below, and to their front or rear (in the real space surrounding them). The relative value of pinna effects for the non-stationary listener and sound source has remained a controversial issue for almost fifty years (see, for example, the 1940 study [25]), and such dynamic situations are exactly those of most interest to cyberspatial applications.

It is surprising how few psychoacoustic studies have focussed solely on pinna effects, given the apparent consensus that these form an important component of the **h**ead-**r**elated **t**ransfer **f**unction (HRTF) which provide the basis for most auditory display technology. Usually, studies of the role of the pinna in directional hearing employ stimuli that include variations in interaural time and level differences that are dominated by the effects of the head, rather than by the pinna. By holding constant the angle of a sound source relative to the interaural axis, head-related variation in interaural time and level differences is minimized, revealing variation that is due primarily to the structure of the pinna. Sound sources arriving at angles spanned by 360° rotation about a point on the interaural axis are said to lie upon a "cone of confusion," due to the front/rear, above/below incidence angle confusion experienced for simple sinusoidal stimuli. For broadband stimuli, accurate directional judgments are apparently enabled by the filtering effects of the pinna. Figure 1 visualizes variation in these pinna effects as an HREF (**h**ead-**r**elated **e**nvelope **f**unction) in the time domain, rather than the more commmonly presented frequency-domain representations (compare [17]).

## 3 Interfaces: Hardware

Precursors to modern spatial audio systems, many of which are still sold with labels like "3D sound" or "multidimensional sound," include spatial enhancers and stereo spreaders [22]. Spatial enhancers filter mono or stereo inputs to add a sense of depth and spaciousness to the signal, allowing for simple operation and backwards compatibility, but precluding placement of individual sounds. Stereo spreaders filter mono inputs, placing the sound

along a linear range to extend the sound field. While such approaches allow for realtime user adjustments, the techniques do not include distance or elevation models.

The most direct approach to spatializing sound is to simply position the sound sources relative to the listener, as in antiphonal concerts. Directly spatialized audio— including attractions like Wisconsin's House on the Rock [18], a museum which features entire rooms lined with orchestral automata— has charm, but is not practical for anything but special-purpose venues and LBE (**l**ocation-**b**ased **e**ntertainment). Fully articulated spatial audio allows dynamic (runtime), arbitrary placement and movement of multiple, separate, sources in a soundscape as well as extra dimensions encoding sound image size, orientation, and environmental characteristics.

Delivery mechanisms for virtual spatial audio can be organized along a continuum of scale. At one end of the spectrum, simple amplitude-panning (balance), perhaps best deployed as constant power cross-fader, can be thought of as a "poor person's spatializer." In conjunction with exocentric visual cues (like a map of sources and sinks, as suggested by Table 1), even such a degenerately simple technique, capable of only lateral (left↔right) effects, can be effective for some applications. In conjunction with egocentric visual cues (like first-person perspective shifts in a large part of the visual field), lateral shifts in the auditory image can disambiguate frontward from rearward sound source incidence angles. Therefore, this simple manipulation can carry surprisingly useful information in applications allowing locomotion through a virtual environment (such as computer games that require targeting of opponents while exploring a 3D-model-based world). Realistic cyberspatial audio, however, involves the simulation of more detailed virtual acoustics, such as the acoustical transformation of sound by the head and pinna. Binaural synthesis techniques typically use **d**irectional **t**ransfer **f**unctions (DTFs), based upon measured **h**ead-**r**elated **t**ransfer **f**unctions (HRTFs), to capture these realistic acoustical effects.

The best standard of comparison for realism in audio reproduction, however, is the result of binaural recording that includes the natural acoustics of enclosed spaces. Very few simulations attain the realism associated with binaural recordings, since binaural recordings are almost always made in settings that allow indirect sound (such as natural reverberation) to be encoded with the direct sound. Most high-quality simulations intended for headphone reproduction attempt to create naturalistic indirect sound to overcome some of the problems with dry (exclusively direct-sound) simulations, such as the typical outcome that the dry auditory images tend to stay inside the head. Problems associated with loudspeaker reproduction of binaural-quality imagery are perhaps more numerous, primarily because the listener's head may move from an optimal listening location and because acoustics of the reproduction environment can greatly degrade the listener's experience. Cross-talk cancellation techniques, such as the transaural or

| proxemic context | architecture | Display | |
|---|---|---|---|
| | | audio | visual |
| intimate | headset,<br>wearable computers | *eartop*<br>(ex: headphones) | *eyetop*<br>(ex: HMDs,<br>**h**ead-**m**ounted **d**isplays) |
| personal | chair | nearphones | *laptop* display,<br>*desktop* monitor |
| interpersonal | couch or bench | transaural speakers,<br>SDP (stereo dipole) | HDTV |
| multipersonal | automobiles,<br>**s**patially **i**mmersive **d**isplays<br>(ex: Cave$^{TM}$, Cabin) | surround sound<br>(ex: Ambisonics) | projection |
| social | clubs, theaters | speaker array<br>(ex: VBAP) | large-screen displays<br>(ex: IMAX) |
| public | stadia, concert arenas | **p**ublic **a**ddress | (ex: Jumbotron) |

Table 2: Audio and visual displays along a private↔public continuum

SDP (stereo dipole) [20] approaches, can produce binaural-quality imagery from loudspeakers, but exhibit extreme sensitivity to listening position and nearly always involve tradeoffs regarding the effective frequency range of the process.

Adding more speakers can solve some of the problems associated with stereo loudspeaker reproduction, enabling, for example, improved front/rear distinction, but create more problems regarding optimal listening position and synthesis techniques. Standard distribution of 5.1 channels of discrete digital audio (like that provided by DVD, DTS [`www.dtstech.com`], and by PC sound cards supporting the AC-3 standard) is convenient, but synthesis techniques for such speaker array systems assume only rough speaker-placement guidelines. Adding still more speakers makes it possible to have sound arrive from many directions at once, but successful sound spatialization depends on careful calibration common only in specialized installations, such as the *P*ioneer *S*ound *F*ield *C*ontroller or systems based upon Ambisonics (`www.ambisonic.net`) or VBAP (vector-based amplitude panning) [21].[1] Especially intriguing are hybrid presentations which combine multiple backend modalities. For example, an IMAX theatre (like that in Times Square, Shinjuku; `www.imax.com/theatres/tokyo2.html`) presents sound to the audience as a whole via a six-channel multispeaker system simultaneously with two additional channels, individual binaural stereo delivered via nearphones in the eyewear, PSE (for "Personal Sound Environment," `www.imax.com/innovations/theatre/IMAX_PSE/index.html`).

At the University of Aizu, we are exploring the potential of a binaural display driven by pickups from a dummy-head located in speaker array system [1], which can produce indirect (environmental) sound upon which direct sounds can be superimposed. As a perversely round-about approach to spatialization, such a hybrid technique can be used to mix spatial audio with voice-over, as in a groupware situation in which physically present users share an environment, including channels directionalized by a speaker array system as well as their own commentary, with remote listeners.

# 4 Interfaces: Software (Application Programmer Interfaces)

DirectSound 3D (part of DirectX) by Microsoft supports spatial sound, and can be extended to provide special effects like room morphing or obstructions with APIs like A3D 2.0 by Aureal (`www.aureal.com`) and EAX 3.0 by Creative

[1] Leading commercial products and companies include the Ircam Spatialisateur (`www.ircam.fr/produits/logiciels/log-forum/spat-e.html`), Lake DSP Huron (`www.lakedsp.com`), QSound (`www.qsound.com`), Roland RSS-10 (`www.rolandcorp.com/products/AUDIO/digital_processing_equipments/RSS-10.html`), and SRS Labs (`www.srslabs.com`).

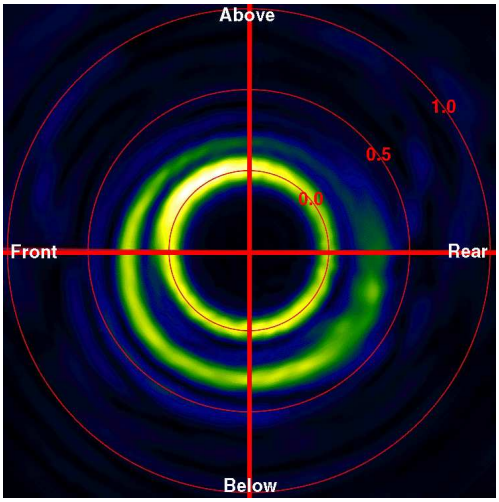



Figure 1: Pinna Effects: **H**ead-**R**elated **E**nvelope **F**unctions (HREFs) for sound sources located 60° from the subject's median plane, visualized over a 360° range of directions defining the so-called "Cone of Confusion." The arrival of the first wavefront is aligned in time (the radial dimension of the plot) with the smallest red circle. Other red circles mark time in .5 ms intervals extending outward from the origin. The second significant wavefront arrives at minimal delay of around .1 ms for elevated angles of incidence. The maximal delay of around .3 ms for the second significant wavefront occurs for sources that arrive from below, and a nearly monotonic transition between these two extreme values is observed as sound sources move between extremes of elevation.

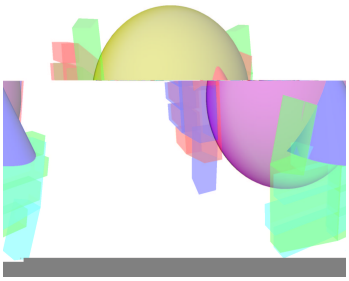


Figure 2: Figurative Avatar Interdigitation: A source representing a human teleconferee denotes mutedness with an iconic hand clapped over its mouth, oriented differently (thumb up or thumb down) depending on whether the source was muted by its owner (or one of its owners) or another user. To distinguish between deafness self-imposed (invoked by a user whose attention is directed elsewhere) vs. distally imposed (invoked by a user desiring selective privacy), hands clasped over the ears orient differently depending on the agent of deafness [9]. Being both virtual and conceptually orthogonal, these various hands interpenetrate.

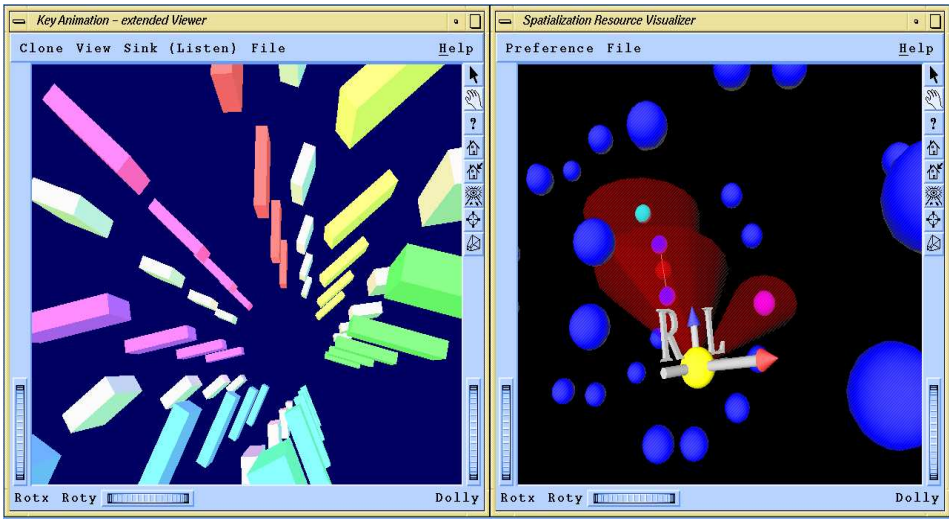



Figure 3: Visualization of clusters in the Helical Keyboard: The left side shows the interactive "Helical Keyboard" application [16], which visualizes and spatializes MIDI streams. The spatialization resource visualizer on the right side shows, for the scene on the left, the soundscape with sound sources (blue) and one sound sink (yellow). The red translucent cones visualize localization errors used by a clustering algorithm [**?**] to decide which sources can be coalesced.

Labs (`www.soundblaster.com`). The latest VRML [2] specification has only a sound node to support positional sound, and does not define soundscape attributes to describe reverberation, although a browser could infer the size of a space and then set important reverberation parameters. The Java3D [23] specification is in that regard more advanced, supporting a notion of a soundscape, an application area with aural attributes capturing delay times and reflections. These APIs, and the new MPEG-4 object-oriented multimedia standard (`www.cselt.it/mpeg`), are adequate for multimedia content, but are not suitable for room acoustics, which are much more complicated and require a more physically-based approach — specification of the material and transfer functions of all sound objects in a simulated space [24] [12] — as is required for auralization applications.

# 5 Applications

As suggested by Table 3, following the advent of desktop computing there was movement towards laptop and even eyetop form factors. We have identified a number of applications for which an "eartop" computer seems appropriate:

1) telecommunication (for example, audio-only teleconferencing) 2) navigational aids for two populations: a) those with severe visual deficits b) those for whom the visual channel is otherwise occupied 3) entertainment (such as computer-aided interactive musical performance, like that suggested by Figure 3) 4) voicemail browsing and synthetic-speech-based browsing of textual e·mail (such applications will probably involve speech-recognition for commands and may include voicemail response entry).

In the visual domain, "augmented," "mediated," or "mixed reality" refers to compositing sampled and synthesized images, like overlaying CG sprites on top of video frames. An audio analog aligns artificial and actual [8], putting synthesized sounds into the real world. For instance, one might want one's teleconversants' voices directionalized to their isomorphic (modeled after actual) locations, reinforcing situational awareness and disambiguating even similar-sounding voices via segregation akin to the cocktail party effect [**?**] that leverages against natural geographic intuition. A driver or a (perhaps blind) pedestrian might want to hear way-finding directions and cues spatialized to relevant directions or locations, like the destination, North, or the next turn.

Binaural cues can be generated by stereo microphones. For example, a telerobot open to the public at Fujita Venté [14, 6] in Sendagaya is not autonomous, but piloted by visitors through an adjacent deconstructive microcosmos, the human seeing and hearing what the drone senses, including collision alarms, through stereo sight and sound. Such technology might be deployed in hazardous environments like nuclear power plants, fires, toxic

waste dumps, and deep mining explorations. Further likely deployments are in telemedicine applications, such as in laproscopic surgery. A human pilot, projecting hirself to a robot's location, might want to hear captured stereo signals mixed with synthetically generated cues, like the directionalized voices of the pilots of other robots, or synesthetically generated cues, like an infrared meter displayed as an appropriately-localized audio alarm.

## 6 Social Aspects

Given the potential power of cyberspatial applications, it is important to begin to anticipate what the social consequences of the shift towards more immersive media might be for the human user. As we may be spending more and more time immersed in virtual environments, distinguishable from "ordinary" multimedia by realtime interactivity, and a sense of presence, supported through maps and avatars, they may begin to determine how we think, speak, and act. The idioms suggested by Figure 2 may be most effective in social situations, enabling synchronous interaction with other users. Text-based societies (like IRC [**i**nternet **r**elay **c**hat] and MUDs [**m**ulti-**u**ser **d**ungeons]) suggest the potential of online communities. Social, multiuser systems, like ActiveWorlds' (`www.activeworlds.com`) Ultimate 3D Chat, Moove's Roomancer (`www.moove.com`), NTT InterSpace (`www.nttinteractive.com`), Sony's Community Place (`vs.sony.co.jp`), and Ubique Virtual Places (`www.vplaces.com`) represent the next generation of multiuser VEs.

Usually one thinks of one's perspective as residing in a single place—namely, behind one's eyes, between one's ears, etc.— but telepresence enables such points of attendance to be distributed and non-singular, by replicating perceiving subject instead of perceived object. Protocols and methods will have to be defined for such systems to enable narrow- and multi-casting idioms for selective privacy, selective attendance, scalability and LoD (**l**evel **o**f **d**etail, as suggested by the clusters in the right side of Figure 3), and side- and back-channels.

## References


[1] Katsumi Amano, Fumio Matsushita, Hirofumi Yanagawa, Michael Cohen, Jens Herder, William Martens, Yoshiharu Koba, and Mikio Tohyama. A Virtual Reality Sound System Using Room-Related Transfer Functions Delivered Through a Multispeaker Array: the PSFC at the University of Aizu Multimedia Center. *TVRSJ:* **T***rans.* **V***irtual* **R***eality* **S***ociety of* **J***apan*, 3(1):1–12, March 1998. ISSN 1344-011X; `www.u-aizu.ac.jp/~mcohen/welcome/publications/PSFC.ps`.

[2] Gavin Bell, Rikk Carey, and Chris Marrin. ISO/IEC 14772-1:1997: The Virtual Reality Modeling Language (VRML97), 1997. `http://www.vrml.org/Specifications/VRML97/`.

[3] Jens Blauert. Acoustical simulation and auralization for VR and other applications. In *Proc. ASVA: Int. Symposium on* **S***imulation,* **V***isualization and* **A***uralization for* **A***coustic Research and Education*, pages 261–268, Tokyo, April 1997.

[4] Jens Blauert. *Spatial Hearing: The Psychophysics of Human Sound Localization.* MIT Press, revised edition, 1997. ISBN 0-262-02413-6.

[5] Albert S. Bregman. *Auditory Scene Analysis: The Perceptual Organization of Sound.* MIT Press, 1990. ISBN 0-262-02297-4.

[6] Michael Cohen. Cybertokyo: a Survey of Public VRtractions. *Presence: Teleoperators and Virtual Environments*, 3(1):87–93, Winter 1994. ISSN 1054-7460.

[7] Michael Cohen. Quantity of presence: Beyond person, number, and pronouns. In Tosiyasu L. Kunii and Annie Luciani, editors, *Cyberworlds*, chapter 19, pages 289–308. Springer-Verlag, Tokyo, 1998. ISBN 4-431-70207-5; `www.u-aizu.ac.jp/~mcohen/welcome/publications/QuantityOfPresence.pdf`.

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

[9] Michael Cohen and Jens Herder. Symbolic representations of exclude and include for audio sources and sinks. In Martin Göbel, Jürgen Landauer, Ulrich Lang, and Matthias Wapler, editors, *Proc. VE:* **V***irtual* **E***nvironments*, pages 235–242, Stuttgart, 1998. IEEE, Springer-Verlag Wien. ISSN 0946-2767; ISBN 3-211-83233-5.

[10] Michael Cohen and Nobuo Koizumi. Virtual gain for audio windows. *Presence: Teleoperators and Virtual Environments*, 7(1):53–66, February 1998. ISSN 1054-7460.

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

[12] Bengt-Inge Dalenbäck, Mendel Kleiner, and Peter Svensson. Auralization, virtually everywhere. In *the 100th Convention of the AES*, Copenhagen, May 1996. Preprint 4228 (M-3).

[13] Richard O. Duda and William L. Martens. Range dependence of the response of a spherical head model. **J**. **A***cous.* **S***oc.* **A***mer.*, 104(5):3048–3058, 1998.

[14] Fujita Venté. 4-6-15 Sendagaya, Shibuya-ku; Tokyo. `www2.fujita.co.jp/vente/amusment/`.

[15] Robert H. Gilkey and Timothy R. Anderson, editors. *Binaural and Spatial Hearing in Real and Virtual Environments.* Lawrence Erlbaum Associates, Mahway, NJ, 1997. ISBN 0-8058-1654-2.

[16] Jens Herder and Michael Cohen. Project Report: Design of a Helical Keyboard. In *Proc. ICAD:* **I***nt.* **C***onf. on* **A***uditory* **D***isplay*, pages 139–142, Palo Alto, CA, November 1996. `www.santafe.edu/~icad/ICAD96/proc96/herder.htm`.

[17] Y. Hiranaka and H. Yamasaki. Envelope representations of pinna impulse responses relating to three-dimensional localization of sound sources. **J**. **A***cous.* **S***oc.* **A***mer.*, 73:291–296, 1983.

[18] The House on the Rock. 5754 Highway 23; Spring Green, WI 53588; USA. `www.thehouseontherock.com`.

[19] Jean-Marc Jot. Real-time spatial processing of sounds for music, multimedia and interactive human-computer interfaces. *Multimedia Systems*, 7(1):55–69, 1999.

[20] Ole Kirkeby, Philip A. Nelson, and Hareo Hamada. The stereo dipole — binaural sound reproduction using two closely spaced loudspeakers. In *AES:* **A***udio* **E***ngineering* **S***ociety Conv.*, Munich, March 1997. Preprint 4463 (I6).

[21] Ville Pulkki. Virtual source positioning using vector base amplitude panning. *J.* **A***ud.* **E***ng.* **S***oc.*, 45(6):456–466, June 1997.

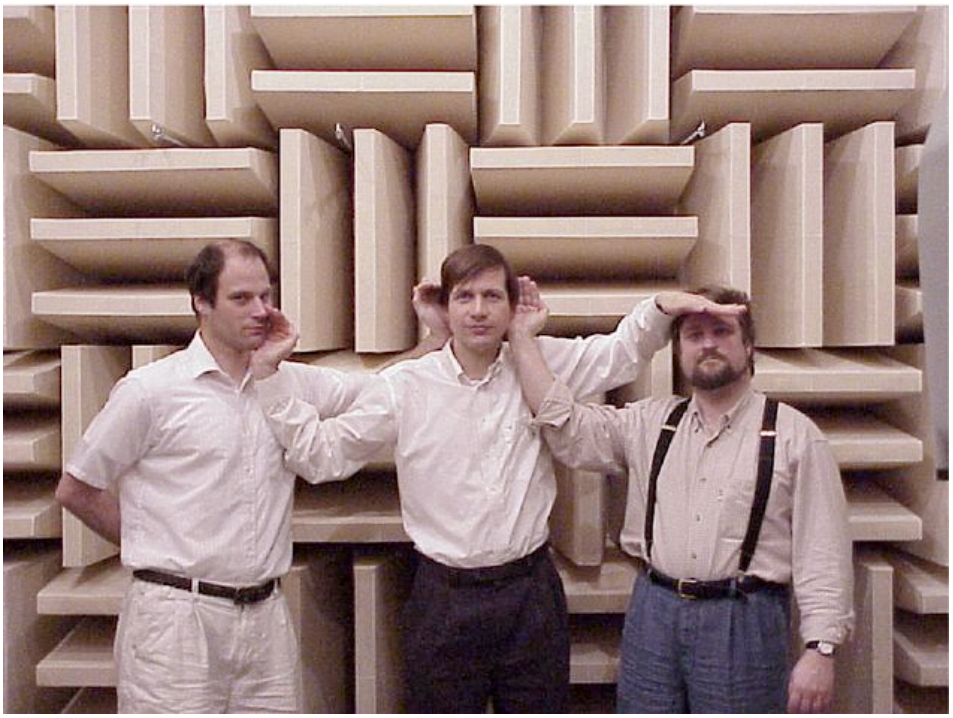
[22] Toni Schneider. Virtual audio. *VR News*, pages 38–41, April 1996.

[23] Henry Sowizral, Kevin Rushforth, Michael Deering, Warren Dale, and Daniel Petersen. Java$^{TM}$ 3D API Specification. Sun Microsystems, August 1997. `www.javasoft.com/products/java-media/3D/forDevelopers/3Dguide/j3dTOC.doc.html`.

[24] Mikio Tohyama, Hideo Suzuki, and Yoichi Ando. *The Nature and Technology of Acoustic Space.* Academic Press, 1995. ISBN 0-12-692590-9.

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

## Authors' Information



Cohen, Herder, and Martens: Mutual Directional Amplification of Auditory (spoken and heard) and Visual Modalities. Michael Cohen and William Martens received Ph.D.s from Northwestern University (near Chicago), in EE/CS and Psychology, respectively, in 1991, and Jens Herder received his Dr. Eng. from Tsukuba Daigaku in July, 1999. Besides the interest in spatial audio manifested by this paper, Cohen has research interests in telecommunication semiotics and hypermedia; Herder has interests in computer graphics, software engineering, and object-oriented programming; and Martens is a psychological scientist researching spatial perception, Together they comprise the Spatial Media Group at the University of Aizu, investigating virtual reality, human-machine interfaces and advanced multimedia systems.

# Contents

# List of Figures



# List of Tables