Audio Engineering Society

Convention Paper

Presented at the 117th Convention
2004 October 28–31 San Francisco, CA, USA

A Headphone-Free Head-Tracked Audio Telepresence System

Norman P. Jouppi[1], Subu Iyer[1], and April Slayden[1]

[1] *Hewlett Packard, 1501 Page Mill Rd., Palo Alto, CA 94304 USA*

Correspondence should be addressed to Norman P. Jouppi (norm.jouppi@hp.com)

ABSTRACT

We have developed a headphone-free bidirectional immersive audio telepresence system. The primary user of the system experiences four-channel audio from a remote location while sitting or standing in a 360-degree surround projection display cube. The display cube incorporates numerous acoustic enhancements, including tilted screens, an anechoic ceiling, and speakers ported through slits in the display cube edges. Head tracking based on near-infrared video technology obtains both the user's head position and orientation. Users can then vary the orientation of their projected voice at the remote location merely by rotating their own head. Similarly, the arrival time and volume of sound channels transmitted from the remote location are varied automatically in the display cube based on the position of the user's head, to help maintain proper perceived interaural time and level differences between multiple channels.

1. INTRODUCTION

Telepresense strives to immerse a user in either a remote physical or virtual location. This immersion is accomplished by faithfully recreating the visual and aural environment of the remote location for the user. However, due to limitations of technology, the relative fidelity of the visual and aural environment is usually the result of a compromise.

Many telepresence systems often share much in common with virtual reality systems. The classic virtual reality system, CAVE[5], surrounded the user on three sides with vertical projection screens and also projected onto a white hard floor from above. The screen size was roughly 3 meters by 3 meters, and speakers were positioned in the cube vertices. This placed the speakers at relatively extreme elevation angles with respect to a standing user's ears, however it did not visually detract from the graphics presentation on the screens. Sound localization was reported[5] as being compromised by reflections off the screens (and presumably the hard floor as well).

More recently, collaboration and telepresence have been investigated using more advanced environments similar to CAVEs. blue-c[8] projects on three

sides of a standing user, using 2.24 meter tall vertical glass screens. Again speakers are relegated to the cube vertices. However since there is no projection on the floor, the floor can be covered with carpet, so this reduces reflections.

Many CAVEs have simply used volume panning between channels to indicate direction of audio stimuli[14]. However, other implementations have used more complicated approaches involving Head-Related Transfer Functions (HRTFs)[3] and crosstalk cancellation between loudspeakers[13]. Since the sweet spot may be small in these approaches, head tracking latency can be a problem[18]. Other implementations have used speakers placed far behind the vertical screens with extensive custom filtering electronics[15].

Based on our earlier experience with telepresence[9, 11] we wanted to extend two and three-sided surround projection environments for collaboration and telepresence to a full 360 degrees. At the same time we wanted to significantly improve the audio experience while keeping it relatively simple and robust.

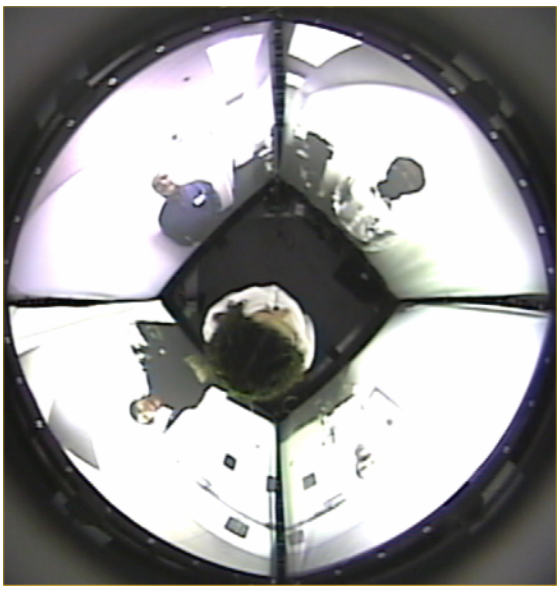

Fig. 1: An overhead fisheye view of a user in the 360-surround display cube.

2. SYSTEM OVERVIEW

We have developed a mutually-immersive telepresence system we call BiReality. Our goal is to recreate to the greatest extent practical, both for a user and people at a remote location, the sensory experience relevant for face-to-face interactions of the user actually being in the remote location. We call the system BiReality since its goal is to create two compelling copies of the real world, one at the remote location minus the user and one at the user's location minus the rest of the remote environment.

As part of BiReality we have developed a headphones-free bidirectional immersive audio telepresence system. The user of the BiReality system experiences audio from a remote location while sitting or standing in a 360-degree surround projection display cube (see Figure 1). They communicate with remote people by using a teleoperated robotic surrogate (shown in Figure 2) at the remote location, connected to the display cube environment via the internet. Live video from the remote location is projected all around the user on the four sides of the display cube. Cameras using pinhole lenses mounted in the vertical seams between projection screens acquire live video of the user for display on the the head of the surrogate. Other aspects of the system, such as its ability to preserve eye contact and gaze, present local and remote participants to each other at life size, preserve the head height of the user, and the high quality multi-stream video can be found in [10].

Multiple speakers are placed along each vertical edge of the display cube and are ported via custom designed exponential horns through a vertical slit between adjacent screens. The speakers do not obstruct the projection since they are located between rear-projection optical beams. We avoid the use of headphones so that we can acquire unobstructed views of the user's head for display at the remote location and avoid any encumbrance on the user.

Head-tracking based on near-infrared video technology is used to obtain both the position and orientation of the user's head. This is used to set the orientation of the user's voice in roughly the direction that the user would be speaking towards at the remote location if they were physically present. The position of the user's head also varies the presentation of the four channel remote sound in the display cube for the user.

Fig. 2: A BiReality surrogate standing.

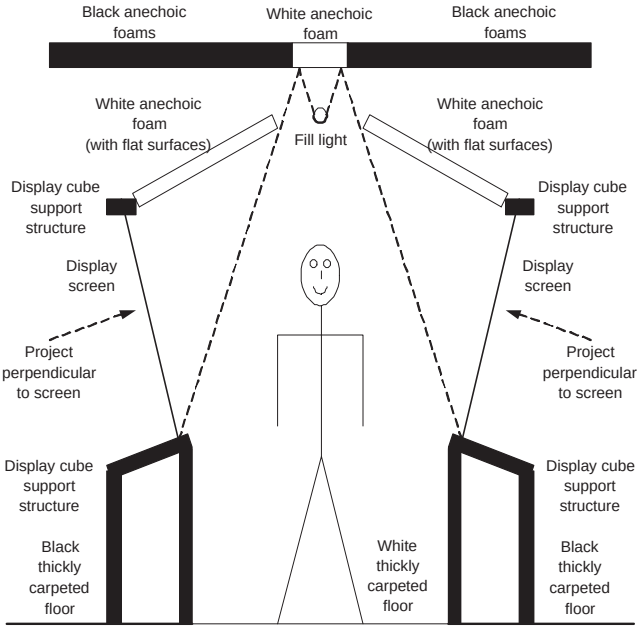

Fig. 3: A cross-section schematic view of a user in the display cube.

3. DISPLAY CUBE DESIGN

Projection screens are quite reflective of sound energy. Our first display cube prototype used parallel vertical screens. With parallel vertical screens the reverberation time was found to be quite long. Moreover, ringing and flutter echoes were also quite apparent. This had several negative consequences. First, users found the audio environment objectionable, and compared it to being inside a bell. Second, the large number of strong close reflections from all sides largely destroyed the directionality of the multi-channel sound reproduced from the remote location. Third, the reverberation exacerbated echoes in the bidirectional sound between the local and remote locations.

3.1. Tilted Screens

To enable the accurate reproduction of audio inside the display cube we have made several acoustic enhancements to conventional display cubes (see Figure 3). First, we constructed an angled false ceiling for the display cube made with anechoic foams. Second, we tilt the screens so that the inside of each screen faces upwards. We then measured the reverberation time at various screen angles and found that modest angles (of 5-7 degrees) were sufficient to greatly reduce reverberation. Tilting the screens while keeping them seamed together in the edges results in a trapezoidal screen shape (instead of the typical rectangular shape). This can be easily compensated for by both tilting the projectors at the screen angle and by using keystone correction video processing.

3.2. Speaker Porting

Speakers are placed along each vertical edge of the display cube. In order to minimize the obstruction of the remote location, we designed custom exponential horns with a 25mm horizontal width and 100mm height. The horns are approximately 50mm deep. The speakers and horns are tilted at the same angle as the screens and recessed so they are flush with the screen edges. The speaker horns were constructed with stereolithography. A speaker port is shown in Figure 4. The frequency response of the horn is nulled out with a graphic equalizer.

3.3. Isolation from Local Noise Sources

Any noise from the user's location can serve to reduce their immersion in the remote aural environment. Therefore we have carefully isolated the user's room from externally and internally generated local noises. The user's room is isolated from exterior lo-

cal sounds with a modest amount of conventional sound isolation techniques. We have installed sound barrier materials between the ceiling tiles and the plenum above the room to reduce the noise transmitted into the room from the HVAC equipment in the plenum. The air return contains sound-absorbent materials and exhausts into a quiet hallway outside the room instead of into the plenum. Finally, we have also removed the grille and diffuser in the HVAC supply vent since they generated significant amounts of noise.

The projectors are housed in custom-designed "hush boxes". These boxes have double-pane Plexiglas windows in their front doors allowing light from the projectors to reach the screens. The box tightly encloses the projector except for air vents at the top and bottom of the rear of the hush box. The box is built from 1/2 inch thick plywood and the inside of the box is covered with anechoic foams. The boxes are painted black to increase the contrast of the projected images, and the front of the box (except for the window) is covered in black anechoic foam. If the boxes only have projectors in them, convection cooling through the two openings in the rear panel is sufficient to cool the projector.

The PC driving the projector and the PC running the audio programs are placed in an adjoining room to eliminate their fan noise from the environment. Placing them in the hush boxes can cause the projectors to overheat.

3.4. Reduction of Local Reflections

Local reflections can destroy the ability of the user to experience the ambiance of a remote location. To maximize the immersion in the remote aural environment, we would like the reflection profile to be only that of the remote location. Because of the mobility afforded by the surrogate, the remote location can vary from a small room with hard walls to an open field.

We have covered the walls of the user's room with anechoic foams, so that the user can experience the reflections characteristic of the remote location without distracting reflections from their own room. We have also installed ceiling tiles that absorb most of the incident sound energy for frequencies above 125Hz. Finally, the room is carpeted to reduce sound reflection from the floor. In practice, the user's table and projection screen reflects much of the sound impinging on them, so there are some reflections from the user's location. These are unavoidable since we do not know of a screen or desk material that does not reflect sound.

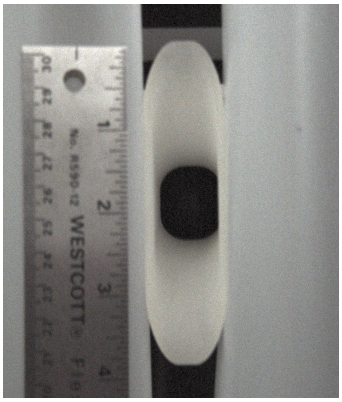

Fig. 4: A speaker port in a vertical edge of the display cube.

4. SURROGATE DESIGN

The surrogate consists of a head, an extensible torso, and a circular base. The bottom of the torso contains two high performance 3.06GHz PCs. This computational capability is required for processing a total of eight high-quality video streams going from the surrogate to the display cube and vice versa.

4.1. Surrogate Head

The head of the surrogate contains integrated speakers and microphones. This reduces their visual profile and makes them less distracting. Similar to the display cube, but at a smaller scale, the surrogate head contains a camera and one speaker between each pair of displays. The speaker is ported through a speaker grille below the camera, since people's mouths are below their eyes. The frequency response of the porting is nulled out with a graphic equalizer at the user's location. By placing the equalizer at the user's location we reduce the amount of hardware that must fit in the space and power constrained surrogate. The surrogate speaker porting is shown in Figure 5. Inside the surrogate head the speakers are wrapped in sound barrier material to reduce transmission into the surrogate head and from there into the microphones.

Fig. 5: A speaker port at the bottom of a vertical edge of the surrogate head. The speaker is covered with a black metallic speaker grille.

In the display cube the speakers along the vertical screen edges are offset by 45 degrees from the projection axis. Since the projection axis corresponds to the surrogate camera optical axis, the microphones in the head of the surrogate must be offset by 45 degrees from the surrogate cameras in order to maintain the correct geometric relationship between the audio and video streams. This means that the microphones need to be centered in the middle of each LCD panel. We have placed the microphones above the LCD panel (rather than below) to increase the separation with the speakers. Unlike gaze preservation issues with video, a small vertical offset has relatively little effect on the audio signals. Supercardioid directional lapel microphones are embedded in the top of the surrogate head, pointing slightly down to the expected positions of remote participants. Supercardioid microphones have been chosen because they make each channel directional, yet the overall response over the four channels remains fairly flat.

Figure 6 shows the top of the surrogate head and several microphones. The microphones are camouflaged relatively well by the flat black color of the surrogate's head structure. The microphones are most visible on the left and right sides where the wind screens of the microphones are visible in profile.

Fig. 6: A directional microphone is placed above each display in the surrogate head.

4.2. Surrogate Noise Issues

To minimize PC case fan noise in the surrogate we have designed it so that all airflow exhausts down through the interior of the base ring. This prevents any direct path to remote participants for fan noise. The gap between the upper and lower portions of the surrogate torso serves both as an air intake and isolation to reduce direct transmission of fan noise. Carpeting in typical modern office environments also aids by attenuating the noise level of sound reflected off the floor. We have also taken care to minimize the fan noise from various sources in the PCs themselves, such as the CPU fan, chipset fan, and graphics accelerator fan. This lowers the noise floor heard by the surrogate user. Also, carefully reducing surrogate noise generation is also important so that remote participants sitting near the surrogate in quiet conference rooms are not distracted.

4.3. Directional Output of Surrogate Audio

Directional audio output has many uses. For example, directional output enables a person to whisper in another person's ear. However, note that there is no difference in hardware between the four sides of the surrogate's head, and each edge has a speaker. The front of the surrogate is merely the side currently displaying the front of the user's head. In order to provide directional output capabilities for the user, we have implemented electronically-controlled directional audio output on the surrogate. This is discussed in the section on head tracking.

5. INPUT, OUTPUT, AND NETWORKING

For both the surrogate and display cube we use high performance real-time data acquisition PCI cards to acquire and output the audio data. We are currently sampling the channels at 40KHz using 16-bit resolution. The sample data is compressed via ADPCM with as little buffering as possible and sent over the internet via UDP[16]. We include a timestamp and sequence number (similar to RTP[16]), so that lost packets can be detected and concealed.

Audio is output through a high-performance 16-bit analog output PCI card with 4X oversampling in software in both the surrogate and display cube. The output of the card is filtered with analog filter circuits and drives studio amplifiers connected to multiple high-quality 4.1 speaker systems. Each of these speaker systems has a subwoofer driven from its the primary audio channels via its own crossover network. This ensures the frequency response is flat with a minimum of effort. It also reduces the number of channels that must be output from the analog output PCI card. Subwoofers are placed under the vertical edges of the display cube, since their low frequency output is less directional.

5.1. Microphone Preamplifiers

In order to get high performance in a small package, we built our own quad fixed gain surrogate microphone preamplifier using commercially available integrated circuits. We use a commercially available mixer on the user's side for conditioning the output of the wireless lapel microphone for input to its data acquisition card. We have set signal levels on both ends such that a shout or clap peaks at the maximum signal level. Sounds louder than this are unlikely to occur in an office environment.

5.2. Feedback Reduction

During normal operation, we reproduce the remote location at its actual sound level, and attenuate the user's microphone if the output volume is greater than a critical value. The microphone attenuation is in proportion to the amount that the output volume is greater than the critical value. Similarly, when the user is speaking, we attenuate the microphones on the surrogate if the surrogate's output is greater than a critical value. This attenuation is also proportional to the amount that the user's voice output is greater than the critical output value. Doing this on both the user and remote sides is enough to prevent oscillation and unwanted feedback, but users can still hear themselves speaking (at an attenuated level) at the remote location. This serves as a useful confirmation that they are being heard by remote participants. It also lets the user know the precise order of audio events relative to the remote location. The critical values are chosen so that feedback is generally not audbile at the remote location.

We compute the volume for the feedback suppression calculations by averaging recent values sent to the analog output card. Input channel attenuation in the case of output levels greater than the critical value is performed with different attack and decay time constants. Exponential attack and decay profiles are used to reduce the discontinuities in the reproduced audio channel. We use an attack time constant that is roughly 5 times faster than the decay time constant for quicker feedback suppression.

5.3. Audio Joystick

Because the audio is mediated, we have an opportunity to provide "super-human" capabilities to the user. In some cases this can make remote interactions better than being physically present. As an example of this, we provide the user with a joystick for adjusting their audio environment. The position of the joystick handle adjusts the relative volume of each output channel by $+/-$ 10dB. This allows the user to steer their hearing around the remote room. For example, imagine the situation where a noisy projector is placed to the left of the surrogate and a presenter at the front of the remote location is not speaking loud enough to be clearly understood. The user can increase the forward speaker gain while reducing the left speaker gain by pushing the joystick forward and to the right. This reduces the noise from the projector at the user's location while increasing the volume of the presenter, making them more intelligible.

The "thrust wheel" control of the joystick has been programmed to adjust the overall volume of all the channels by $+/-$ 10dB. Two buttons on the joystick have been programmed to lock or unlock the gain settings implied by the joystick position. This frees the user's hand once the desired audio setting has been specified.

We used other buttons on the joystick to facilitate control experiments in user studies by providing features roughly equivalent to land line phone systems.

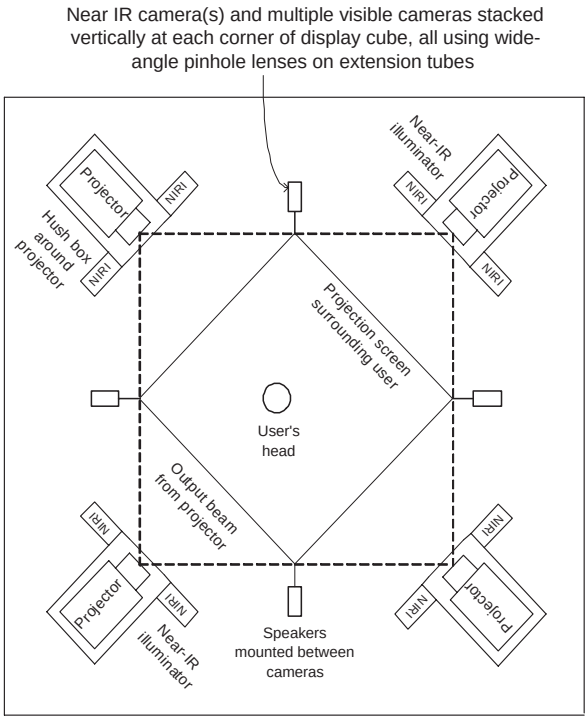

Fig. 7: An overhead schematic view of a user in the display cube.

For example, one button switches the multi-channel audio to a monophonic average of all four channels. Another button reduces the frequency range, while a third button reduces the dynamic range of the system. Each joystick button can be used either individually or in conjunction with the other buttons.

6. HEAD TRACKING

We track the position and orientation of the user in the display cube with techniques based on the analysis and triangulation of near-infrared video images (see Figure 7). Cameras acquiring the near-infrared video images use pinhole lenses along each vertical edge of the display cube.

6.0.1. Tracking X and Y

Signals from four directional microphones in the surrogate's head are transmitted to the user's location and output over speakers in each vertical edge of the display cube. As the user moves around the display cube, the distance from their head to each speaker will change. To keep the user's perception of arrival times and volume of the four channels as accurate as possible, we vary the volume and arrival time of each channel based on the distance from the user's head to the corresponding active speaker.

6.0.2. Tracking Z

The user may also sit down and stand up in the display cube. To accommodate this the display cube has multiple small speakers ported through each vertical display cube edge, one at a typical standing person's ear height, and one at a typical sitting person's ear height. Based on the tracked height of the user's head, we primarily drive the speaker closer to the user's ear height to maintain a near-horizontal perceived elevation for sound sources at the remote location.

6.0.3. Tracking User Head Rotation

The user's voice is captured with a wireless lapel microphone. The orientation of the user's head is also used to automatically vary the output of the user's voice at the remote location. We vary the volume of each of the four surrogate head speakers such that the user's voice is primarily output in the direction of the front of their face (as displayed on the screens of the surrogate's head). This allows a user to direct their voice at the remote location just by turning their head in the display cube. This is a key feature for enabling natural private and semi-private conversations with people at the remote location.

7. USER EXPERIENCES

We have informally evaluated our system in a series of staff meetings. Participants using the system have found that it was a significant improvement over traditional audio conferencing technology, primarily due to the increased dynamic range and directionality. The ambiance of the remote location is also preservered quite well by the system.

One of the more challenging environments in which we used the system were staff meetings held in our site's cafeteria (see Figure 8). The cafeteria is quite noisy and has many hard surfaces which cause a lot of reverberation. Nevertheless, users reported being able to easily identify the position of sound sources around the their remote presence in the cafeteria. Furthermore, the timbre of various noise sources (e.g., clinking silverware on plates and unwrapping cracker packets) was reported as being surprising realistic and immersive. Based on comparisons with

Fig. 8: A surrogate in use in a noisy cafeteria. Conversations in this environment would be difficult or impossible without multichannel audio.

more conventional inferior sound systems (as selected by the audio joystick buttons), users informally reported much higher intelligibility for conversations using the BiReality audio telepresence system.

We also evaluated the surrogate at the cafeteria both with and without directional output. We quickly noticed that without directional output, people at other tables around the surrogate would look up at the user whenever they spoke, since the perceived sound levels were similar to when someone was addressing them. The implementation of directional output was a significant milestone in reducing disruption to other people at the remote location and in allowing the surrogate user to feel natural and not self-conscious when using the system.

One compromise in the system is that we only provide four audio channels from the remote location. As a consequence, when users are asked to identify the direction of a sound from the remote location with their eyes closed, they may report it as being centered on a speaker in the cube edge when the remote source is actually between the edge and the center of the screen. However, when users are also presented with visual imagery from the remote location, the "ventriloquist effect" results in less perceptual direction error. In the current system design, increasing the number of channels in the horizontal plane would require breaking the visual field up into more segments, which is undesirable. Thus the current directional accuracy provided by the system seems to be a relatively good compromise.

8. SUMMARY

We have developed a headphone-free bidirectional immersive audio telepresence system as part of a system we call BiReality. The primary user of the system experiences four-channel audio from a remote location while sitting or standing in a 360-degree surround projection display cube. The display cube incorporates numerous acoustic enhancements, including tilted screens, an anechoic ceiling, and speakers ported through slits in the display cube edges. Head tracking based on near-infrared video technology obtains both the user's head position and orientation. Users can then vary the orientation of their projected voice at the remote location merely by rotating their own head. Similarly, the arrival time and volume of sound channels transmitted from the remote location are varied automatically in the display cube based on the position of the user's head, to help maintain proper perceived interaural time and level differences between multiple channels.

ACKNOWLEDGEMENTS

The authors would like to thank Jacob Augustine, Shivarama Rao Kokrady, and Deepa Kuttipparambil for their help with the software. We would also like to thank Stan Thomas and Wayne Mack for their design and implementation of the BiReality surrogate and display cube.

9. REFERENCES

[1] J. Blauert. *Spatial Hearing: The Psychophysics of Human Sound Localization.* MIT Press, 2nd edition, 1997.

[2] J. Borwick, editor. *Loudspeaker and Headphone Handbook.* Focal Press, third edition, 2001.

[3] C. Cheng and G. Wakefield. Introduction to Head-Related Transfer Functions (HRTFs): Represen tations of HRTFs in Time, Frequency, and Space. *Journal of the Audio Engineering Society*, 49(4):231–249, April 2001.

[4] C. Cherry. Some Experiments on the Reception of Speech with One and Two Ears. In *Journal*

of the Acoustical Society of America, pages 975–979, 1953.

[5] C. Cruz-Neira, D. J. Sandin, and T. A. DeFanti. Surround-Screen Projection-Based Virtual Reality. In *Proc. of ACM SIGGRAPH*, pages 135–142, 1993.

[6] F. A. Everest. *Master Handbook of Acoustics*. McGraw Hill, fourth edition, 2001.

[7] R. H. Gilkey and T. R. Anderson. *Binaural and Spatial Hearing in Real and Virtual Environments*. Lawrence Erlbaum Associaites, 1997.

[8] M. Gross and et. al. blue-c: A Spatially Immersive Display and 3D Video Portal for Telepresence. In *Proc. of ACM SIGGRAPH*, pages 819–827, 2003.

[9] N. P. Jouppi. First Steps Towards Mutually-Immersive Mobile Telepresence. In *Proc. of the ACM Conference on Computer Supported Cooperative Work*, pages 354–363, 2002.

[10] N. P. Jouppi, S. Iyer, S. Thomas, and A. Slayden. BiReality: Mutually-Immersive Telepresence. In *Proc. of ACM Multimedia*, 2004.

[11] N. P. Jouppi and M. J. Pan. Mutually-Immersive Audio Telepresence. In *Proc. the 113th Audio Engineering Society Convention*, October 2002.

[12] M. W. Matlin and H. J. Foley. *Sensation and Perception*. Allyn and Bacon, 4th edition, 1997.

[13] A. Mouchtaris, P. Reveliotis, and C. Kyriakakis. Inverse Filter Design for Immersive Audio Rendering Over Loudspeakers. *IEEE Transactions on Multimedia*, pages 77–87, June 2000.

[14] M. Naef, O. Staadt, and M. Gross. Spatialized Audio Rendering for Immersive Virtual Environments. In *Proc. of ACM VRST*, November 2002.

[15] T. Ogi, T. Kayahara, M. Kato, H. Asayama, and M. Hirose. Immersive Sound Field Simulation in Multi-Screen Projection Displays. In *Proc. of the Eurographics International Immersive Projection Technologies Workshop*, pages 135–142, 2003.

[16] L. L. Peterson and B. S. Davie. *Computer Networks: A Systems Approach*. Morgan Kaufmann, third edition, 2003.

[17] M. Talbot-Smith, editor. *Audio Engineer's Reference Book*. Focal Press, 2nd edition, 1999.

[18] J.-R. Wu, C.-D. Duh, and M. Ouhyoung. Head Motion and Latency Compensation on Localization of 3D Sound in Virtual Reality. In *Proc. of ACM VRST*, 1997.