



# Audio Engineering Society Convention Paper

Presented at the 113th Convention  
2002 October 5–8 Los Angeles, California, USA

*This convention paper has been reproduced from the author's advance manuscript, without editing, corrections, or consideration by the Review Board. The AES takes no responsibility for the contents. Additional papers may be obtained by sending request and remittance to Audio Engineering Society, 60 East 42<sup>nd</sup> Street, New York, New York 10165-2520, USA; also see [www.aes.org](http://www.aes.org). All rights reserved. Reproduction of this paper, or any portion thereof, is not permitted without direct permission from the Journal of the Audio Engineering Society.*

## Mutually-Immersive Audio Telepresence

Norman P. Jouppi<sup>1</sup> and Michael J. Pan<sup>2</sup>

<sup>1</sup> HP Labs, Palo Alto, CA 94304, USA.

<sup>2</sup> UCLA Department of Computer Science, Los Angeles, CA 90095, USA.  
Correspondence should be addressed to corresponding author ( [norm.jouppi@hp.com](mailto:norm.jouppi@hp.com) )

### ABSTRACT

Mutually-immersive audio telepresence attempts to create for the user the audio perception of being in a remote location, as well as simultaneously create the perception for people in the remote location that the user of the system is present there. The system provides bidirectional multi-channel audio with relatively good fidelity, isolation from local sounds, and a reduction of local reflections. The system includes software for reducing unwanted feedback and joystick control of the audio environment. We are investigating the use of this telepresence system as a substitute for business travel. Initial user response has been positive, and future improvements in networking quality of service (QOS) will improve the interactivity of the system.

### INTRODUCTION AND CONTEXT

Mutually-immersive audio telepresence attempts to create for the user the audio perception of being in a remote location, as well as simultaneously create the perception for people in the remote location that the user of the system is present there. This work is part of a larger system, mutually-immersive mobile telepresence, which also strives to achieve immersive visual telepresence, physical presence, and mobility. We are investigating the use of this telepresence system as a substitute for business travel.

### Overview of the First-Generation System

The user sits in a room with a large display wall (see Figure 1). Cameras around the user acquire live

video of the user's head. The user wears a lapel microphone for acquiring their voice. Speakers reproducing sounds from the remote location are mounted to the user's left, front, right, and rear.



Figure 1. The user sitting in front of the display wall.

The user controls a teleoperated robotic surrogate at the remote location (see Figure 2). The surrogate has a “head” built from LCD panels on which it displays live video of the front and both sides of the user’s head. Directly above the LCD panels are 8 cameras for acquiring a high-resolution video mosaic of the remote location for transmission to and display for the user. Above the video cameras are four short shotgun microphones for acquiring the remote sound field. Speakers are also mounted under the head of the surrogate (although this is not shown in Figure 2, which shows an earlier version of the surrogate).



Figure 2. The surrogate at the remote location.

The audio and video at both ends of the system are acquired, processed, and output by PCs. The PCs at the local and remote locations communicate over high-speed internet connections.

## AUDIO TELEPRESENCE

This audio component of this work addresses many challenges, discussed below.

### Directionality of the Remote Sound Field

Recreating the directionality of the remote sound field is key to creating a sense of aural presence. For example, this enables users to utilize the “cocktail party effect”, where a directional sound field can provide up to a 20dB gain in signal-to-noise ratio when listening to a sound source in a specific direction[1]. We currently acquire the directionality of the remote sound field with four short shotgun microphones at right angles to each other in the horizontal plane. Since most useful sounds in an office environment are in a horizontal plane, and human vertical direction perception is weaker than horizontal direction perception, we have only addressed horizontal directionality to date.

We believe binaural techniques are not applicable for several reasons. First, we do not want to detract from the view of the user that would normally be present at the remote location if the user was physically present. Thus, the user cannot wear headphones (even earbuds) or anything else that would change the appearance of their head.

Second, because the user can rotate their head, any binaural representation would need to change with time and track the movement of the user’s head. The most reliable means of head tracking encumber the user’s head with a tracking device. Besides marring the user’s appearance, even the best devices have latencies that can cause motion sickness when used for modification of the visual presentation in virtual reality applications. We do not know of any recorded instances of motion sickness due to lags in audio stimuli, so any audio lags would probably be just an annoyance. Nevertheless, in this application we believe reproducing multichannel sound at the user’s location to be superior to binaural approaches.

Studies have shown that perception of directionality is enhanced if the angle between the channels as seen by the user is 60 degrees or less[3]. Using this criterion, for optimum audio directionality we would need a ring of six speakers around the user. Unlike theater applications where the audience is spread over a large percentage of the area surrounded by the speakers, the location of the user in our application is central and relatively fixed. This reduces the need for additional channels to maintain proper directionality for audience members in outlying positions.

In our first-generation system we have compromised the reproduction of directionality somewhat by reducing the number of channels to 4 in order to make the implementation simpler and reduce the required bandwidth. This increases the angular spacing between channels to 90 degrees. The sensitivity of the short shotgun microphones that we use drops off 3dB at 45 degrees away from center. If we provided more channels, we would need either longer shotgun microphones or a phased-array microphone system. Longer shotgun microphones would protrude beyond the LCD panels and detract from the appearance of the user at the remote location. In the future a phased-array microphone

system would enable this, but at the price of higher cost and increased complexity.

### **Directionality of the User's Voice**

To provide mutually-immersive audio telepresence, the directionality of the user's voice output also needs to be preserved. In a free-field environment, the volume in front of a speaking person can be 20dB higher than that behind their head[1]. People instinctively use this characteristic in certain situations. For example, when a user desires to whisper to another meeting attendee at the remote location, they turn their head towards the other person's ear. In this situation users would not want other people to hear their voice at the same level.

In our first-generation system we preserve the directionality of the user's voice by mounting two small forward-facing speakers under the LCD panels of the surrogate's head on both sides of the surrogate's neck. (Since the user may tilt the surrogate's head up and down, we could not mount a single speaker under the center of the surrogate's chin. A speaker there would hit the top of the surrogate's body when the head was tilted down.) A subwoofer is placed inside the main body of the surrogate.

Users can control the direction of their voice at the remote location by controlling the orientation of their surrogate head and the position of their surrogate body. Users can pan their surrogate head 50 degrees to either side of the front of the surrogate. Since the speakers are mounted on the surrogate's head, when the user turns it their voice turns as well. The user can move the surrogate around the remote location by use of teleoperated controls. This gives them some measure of mobility at the remote location. While this is currently far less mobility than if they were physically present, it is much better than being stuck in a speakerphone at the center of a conference room table.

### **Frequency and Dynamic Range**

Recreating the remote sound field with good fidelity is also important to creating a sense of aural presence. Ideally we would like to reproduce signals spanning the entire range of human hearing in terms of both frequency perception as well as dynamic range.

We use high performance real-time data acquisition PCI cards to acquire and output the audio data. We are currently sampling the channels at 22KHz using 16-bit resolution. This lower than usual sampling rate was chosen in order to reduce the bandwidth requirements. As higher bandwidth wireless networks become available, we plan to use higher sampling rates (e.g., 44 or 48KHz). The sample data is compressed via ADPCM with as little buffering as possible and sent over a TCP/IP connection (we have experimented with other network protocols as well). At the user's location the sound field is reproduced with a high-performance 16-bit analog output PCI card and 4X oversampling in software. The output of the card is filtered with analog filter circuits and drives studio amplifiers connected to a high-quality 4.1 speaker system. The speakers are positioned to the left, front, right, and rear of the user. We use a speaker system with a subwoofer driven from each of 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.

We initially used a high-quality consumer 5.1 channel receiver for driving the speakers. However, when people were not speaking at the remote location users could hear noise from the receiver. Hearing a hiss from the inadequate noise floor reduces the immersiveness of the audio. We were able to lower the noise floor in the system by approximately 13dB by switching to studio amplifiers.

The shotgun microphones we use on the surrogate contain microphone preamplifiers. However, these did not reach the levels required by our data acquisition card, so we built an additional level of amplification into the surrogate. This amplification also filters the input to prevent aliasing at the sampling frequency. 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 such that a shout or clap several feet from the surrogate peaks at the maximum signal level. Sounds louder than this are unlikely to occur in an office environment. With attention to detail we have reduced the noise floor of

the whole system such that surrogate fan noise is now the limiting factor to dynamic range on the low end.

### Isolation from Local Sounds

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 (formerly a standard walled office) from externally and internally generated local noises. The user's room is isolated from exterior local 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. We removed the air return entirely, because it opened directly onto the plenum. We rely on air leakage to complete the HVAC circuit. (A separate air return duct would be a better long-term solution.) 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 ½ 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.

### Reduction of Local Reflections

Local reflections can destroy the ability 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.

### System and Network Latency

We have carefully optimized our system to reduce communication latencies, however this is still a limitation for highly interactive conversations. Our current one-way latency using networking within our campus is about one half second. It is dominated by buffering just before audio output, which is necessary to prevent gaps in audio reproduction during times of poor wireless network quality of service (QOS). Future wireless networking technologies with better QOS guarantees will allow us to reduce the amount of buffering just before output. When the surrogate is operated in a tethered (non-mobile) mode, the buffering can be greatly reduced, lowering the latency to well under a half second.

DMA buffers in the analog input and output cards currently set a lower limit on latency. The analog input card buffer contains 16K samples and the analog output board buffer contains 32K samples. The program transfers data when the input buffer reaches half full and when the output buffer is half empty. Assuming transfers to the board are much faster than the input and output rates, and given a 22KHz sampling rate, 4 channels, and 4X oversampling on the output, we can calculate the latency solely due to the DMA board buffers. Considering transmission from the surrogate to the user, up to  $16K/2 \cdot 4 \cdot 22KHz = 93ms$  of data may be in the input board DMA buffer just before it is emptied, and  $32K/4 \cdot 4 \cdot 22KHz = 93ms$  of data may be in the output buffer just after it is filled. This

currently sets a lower bound on worst-case one-way latency of  $93+93 = 186$ ms. However, when the sampling rate is doubled to 44KHz, the latency lost due to board buffering will be halved to 93ms total. Finally, we acquire four copies of the user's lapel microphone channel and output four copies of it on the surrogate, in order to reduce on-board buffering latencies on that path to be the same as in the reverse direction.

The audio latency also depends on the electrical and optical latencies in the network. Long-distance latencies (e.g. halfway around the world via undersea fiber optic cables) can increase the latency by up to a few hundred milliseconds. The higher latencies of geosynchronous satellites are unacceptable.

### Feedback

Since the system is bidirectional, feedback is obviously a key concern. We employ several techniques to reduce feedback. First we try to increase the gain of local sounds at each end while reducing the gain of remote sounds. We also employ software feedback reduction techniques.

We acquire the user's voice from a wireless lapel microphone. We considered several alternatives to a lapel microphone, but a lapel microphone is best for several reasons. First, the gain of the user's voice relative to the remote signal from the speakers with a lapel microphone is good. Second, although the sound level from the lapel microphone does vary as the user turns their head relative to their torso, this level variation tends to be less than when using directional microphones mounted somewhere other than on the user's body. Omnidirectional microphones are not an option due to resulting feedback problems between the local and remote sites.

On the surrogate side, the speakers are mounted under the surrogate's head, while the shotgun microphones are mounted on top of it. This shields the microphones as much as possible from direct radiation from the speakers, while having both mounted on the head so that they mechanically track the surrogate's head orientation.

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 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 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 at the remote location. It also lets the user know the precise order of audio events relative to 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.

### Enhanced Capabilities

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 speaker, making the speaker 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.

## RESULTS

We have only recently completed the implementation of our first-generation audio telepresence system, so we have not yet had time to conduct extensive user studies.

### Initial Results

To date we have informally evaluated our system in a limited number of business 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. Some people have even described the system as "spooky", since with their eyes closed the perception of the presence of the remote meeting participants was so real. The high dynamic range of the system combined with the directionality even allowed users to accurately auralize the position of a ticking clock on the remote conference room wall. This ability was further enhanced by the ability to steer one's hearing with the joystick.

### Future Work

In future work we plan to test the intelligibility of spoken words in the presence of other speakers and/or noises using the system with and without joystick steering, in comparison to actually being present at the remote location. We also plan to work on further reducing the latency. Low latency is crucial for supporting active interactive conversations with turn taking.

## ACKNOWLEDGEMENTS

Wayne Mack built the Model 1 surrogate and the user environment. We would like to thank Jason Wold for his help in building and maintaining the computer systems and networks used in this work. Subu Iyer also helped with the software and system development. Keith Farkas helped us interface to the data acquisition and output cards.

## REFERENCES

- [1] Blauert, Jens, and Allen, John S., "Spatial Hearing: The Psychophysics of Human Sound Localization", MIT Press, 1996.
- [2] Gilkey, Robert H., and Anderson, Timothy R., (editors) "Binaural and Spatial Hearing in Real and Virtual Environments", Lawrence Erlbaum Assoc., 1997.
- [3] Talbot-Smith, Michael, "Audio Engineer's Reference Book", Focal Press, 1999.