Coupled Auralization and Virtual Video for Full-Room Tele-Performance Environments*

Paul Henderson
Program in Architectural Acoustics
Rensselaer Polytechnic Institute
Troy, NY 12180
hendep2@rpi.edu

Brian Lonsway
School of Architecture
Rensselaer Polytechnic Institute
Troy, NY 12180
lonsway@rpi.edu

Richard Radke
Department of Electrical, Computer and Systems Engineering
Rensselaer Polytechnic Institute
Troy, NY 12180
rjradke@ecse.rpi.edu

ABSTRACT

Most research on distance collaboration using virtual environments treats audio and video problems as essentially independent. However, recent studies on virtual immersive systems have established a significant correlation between the relative qualities of audio and video displays and the overall sense of realism perceived by the participants. This is a logical extension of what we already know about human spatial interaction, independent of any technological mediation: conversants in a room use multi-sensory cues to communicate with others. In this paper, we present a system for synthesizing spatially accurate virtual video to be projected to wall-sized screens, and acoustically accurate audio to drive arrays of loudspeakers. We hypothesize that the strategic coupling of audio and video synthesis techniques within a tele-collaborative system can substantially increase the sense of spatial perception and the fidelity of communication by participants.

Categories and Subject Descriptors

H.5.5 [**Information Interfaces And Presentation**]: Sound and Music Computing—*modeling, signal synthesis*; I.4.8 [**Image Processing And Computer Vision**]: Scene Analysis—*depth cues, motion, object recognition, stereo, time-varying imagery, tracking*; I.6.5 [**Simulation and Modeling**]: Model Development; J.5 [**Arts and Humanities**]: [architecture, performing arts]; K.3.1 [**Computer Uses in Education**]: [collaborative learning, distance learning]

*This work was supported by an internal seed grant from Rensselaer Polytechnic Institute, and the National Science Foundation IUCRC Program under Grant No. EEC-9812706.

ITP2002 Juan Les Pins, France

Keywords

Auralization, virtual video

1. INTRODUCTION

We are concerned with the realization of spatially and acoustically accurate full-room tele-performance environments. We use the word *tele-performance* to connote the types of tele-collaborative activity in which bodily action plays a substantial role in both the completion of tasks and the successful communication of information between individuals. This encompasses not only conventional definitions of "performance," such as music, theater, or dance, but also broader "performative" activities including team-based design project reviews, distance-education lectures by exuberant speakers, or walkthroughs of architectural or archeological sites. These activities all share one common feature in the context of distance collaboration: they pose a number of logistical difficulties for standard tele-conferencing and tele-communication systems.

The majority of work on distance collaboration operates either in the context of a computer screen (or projection) or that of a head-mounted display environment. In terms of tele-performance, computer-screen techniques suffer from a lack of bodily engagement and the presence of distracting peripheral effects that undermine the sense of believability. While larger projection-screen systems (e.g. CAVE-like setups) allow the participants to move more freely, they typically require the use of helmets or goggles that draw attention to the enabling technology. Neither approach is generally known for the inclusion of real-time, acoustically accurate audio that is coupled to the synthesized video environment. Recent research, however, has shown that the perceived sense of realism in virtual immersive systems is a multi-modal effect that requires viable treatment of the acoustical environment, coupled with an accurate video display [2, 11]. Furthermore, from a data fusion perspective, the joint solution of audio and video estimation problems in a virtual environment has more potential to synthesize spatially accurate and effective shared work environments than does treating the problems independently.

Here, we present our development of a novel suite of coupled video tracking, virtual video synthesis, and accurate auralization techniques to create a tele-performance experi-

ence that is both perceptually convincing and spatially and acoustically accurate. We intend to explore how the coupling of the audio and video techniques can create a more realistic connection between two separated sites by accurately representing the spatial characteristics of a shared virtual environment. We expect this work to have immediate impacts in areas including non-intrusive tele-conferencing, distance learning, and computer-mediated design.

2. PRIOR WORK

The fundamental engineering problems to be addressed in our proposal stem from coupled problems of audio and video. Some very pertinent observations have been made on the relationship between audition and vision for aural perception. Larsson et al. [4] provide a good summary.

Some prior work does address the implications of audio/video coupling for tele-conferencing configurations that extend beyond the desktop, but it focuses on coupling for participant tracking in order to support more conventional tele-conferencing applications. For example, Pingali et al. [7] use audio and video information together to improve tracking estimates. However, this information is not coupled in the sense that there exists real-time feedback between audio and video systems. Microphones are used to provide general location estimates of participants, and cameras are directed toward the sources of these speakers to provide greater-resolution imagery of faces for pose estimation, lip-tracking, and so on. However, people in the room who are not making sounds cannot be tracked. While spatial characteristics of the room are used to perform the first-level acoustic tracking, these spatial (reverberant) characteristics do not play a role in the transmission of spatialized acoustics to a remote location. Through explicitly understanding the auralization/virtual video coupling problem as a spatial problem, we propose a new framework for tele-collaboration that supports non-typical applications, especially required in the arts and design fields that we intend to address.

3. SYSTEM ARCHITECTURE

We are in the process of building the physical facility illustrated in Figure 1. The site consists of two rooms, approximately 7m by 10m, with one wall of each dedicated to video projection. Two wide-angle cameras are placed on either side of the screen and loudspeakers are located around the perimeter of the room, as indicated. Participants are required only to wear lightweight wireless microphones. The system will support multiple users having multiple interactions through the video projection wall as if this wall were an opening between two adjacent rooms, as indicated in Figure 2. Participants will perceive the remote room as an extension both acoustically and visually of their own. Finally, participants at each site will be able to define (acoustically and visually) the virtual spatial environment in which they perceive their remote collaborators.

4. VIRTUAL VIDEO

Since there are only two actual images of each studio available at any time instant, neither camera perspective is likely to match the view that a participant in the other studio expects to see. However, recently developed virtual view synthesis techniques enable us to synthesize physically correct views of the scene from the perspective of a camera

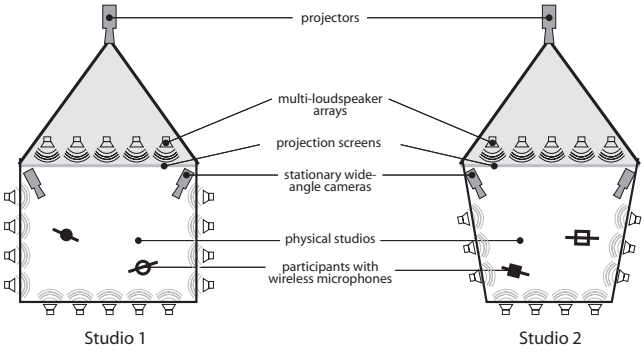

Figure 1: Overhead schematic of physical studios, with input and output devices labeled.

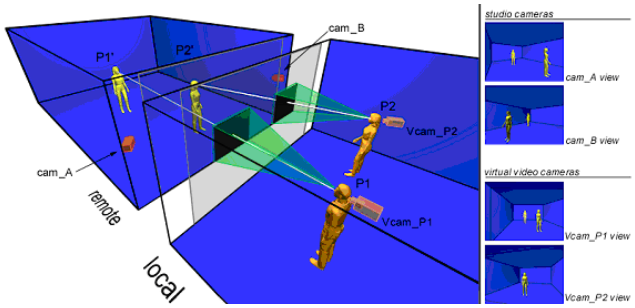

Figure 2: The virtual spatial environment in which the participants interact.

not present in the original environment. This synthesis can be accomplished using image-based techniques that do not require accurate three-dimensional modeling of the scene, which can be very time-consuming for complicated environments. While most of the prior work in the area of virtual view synthesis has been at the still-image level (e.g. [5, 10]), we have recently reported progress on the problem of synthesizing virtual *video* from a pair of input video streams [8, 9]. Our research demonstrates that convincing results can be obtained using only a few real video cameras and processing on normal desktop PCs. A sample result from our existing virtual video algorithm is illustrated in Figure 3.

The participants in each studio are segmented from the background as part of the view synthesis algorithm. Since we assume the cameras are calibrated beforehand, we can track the participants in three dimensions. Hence, there is little conceptual or computational difficulty in composing them into a virtual, computer-generated environment instead of rendering an actual image of the physical room in which they are present.

We note that the location *and* pose of each participant must be estimated in order to ensure the most realistic overall experience. For example, when a participant in the receiving studio turns to face a different direction, both the projected video and the signals sent to the loudspeaker array should change accordingly.

5. AURALIZATION

An integral and somewhat new area of research in the implementation of telepresence environments is the high-accuracy rendering of coupled acoustic spaces. The overall

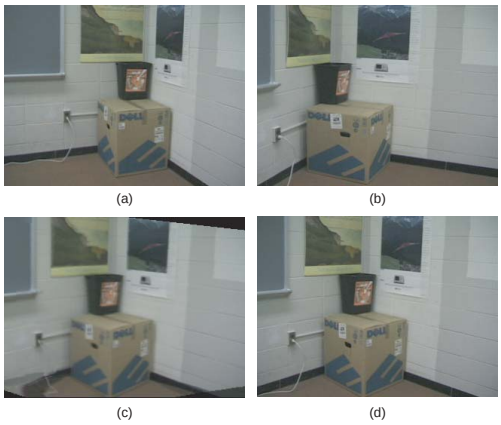

Figure 3: Pictures (a) and (b) are images from two real source cameras observing the same scene at the same time. Picture (c) is a synthetic image constructed from the perspective of a camera whose optical center lies between the optical centers of the two original cameras. Picture (d) is an actual image from that position, demonstrating the accuracy of the virtual view synthesis algorithm.

goal of this project is to reproduce, in real-time, an immersive, realistic sonic environment that is accurate over a wide listening area. While methods exist for creating spatial audio renderings using large numbers (i.e. hundreds) of loudspeakers [1], we aim to develop an optimized method for achieving a perceptually accurate sound field over a large listening area using a practical number of devices.

Conventional methods for computing room impulse responses for virtual reality typically neglect non-specular scattering and rely on simple geometrical acoustics, e.g. the image-source model, ray tracing, or some combination. A geometrical acoustics model, however, erroneously predicts a discontinuous sound field and introduces inaccuracies in the scattering strength when the acoustic wavelength is comparable to the dimensions of the surfaces [13]. We utilize a time-domain wedge-assemblage method to model edge diffraction and scattering [3] based on Svensson's extensions to the Biot-Tolstoy-Medwin (BTM) solution [12], allowing first- and second-order scattering to be calculated. Finally, since reverberation is perceived less sensitively than the early part of the room impulse response, the reverberation tail is either measured or computed using statistical means. The total pressure field (a superposition of (1) the direct sound and specular reflections, (2) non-specular surface scattering, and (3) reverberation) is then rendered over a discrete number of loudspeakers.

To investigate the accuracy of the acoustic rendering technique over a finite loudspeaker array, we compared a reproduced acoustic wavefront with that of a equivalent computed continuous sound field. To simplify the comparison of results, we present data for a direct sound path only; a complete reflected field would simply be the superposition of multiple simple wavefronts. A single virtual acoustic point source was introduced at a location just outside of the listening plane. In the artificially coupled environment, this source introduces a specific acoustic field into the receiving room, which can be described at a single point in time by a map of pressure deviations versus spatial position on a measurement plane (the instantaneous pressure wavefront). The instantaneous pressure wavefront induced on the measurement plane by sinusoidal (monochromatic) radiation at a frequency of 500Hz is shown in Figure 4a, and exhibits the ideal spherical spreading inherent from the source.

Next, we used a linear array of six loudspeakers spaced 35 cm apart to recreate the effective acoustic field over the measurement plane. A truncated Helmholtz-Huygens relationship [6] was used to compute the signals produced by each loudspeaker. Figure 4b illustrates a simulation of the wavefront radiated by the six-loudspeaker array. While the spatial discretization and truncation of the field by a linear array creates errors in the overall magnitude of the reproduced field, the wavefront curvature and apparent acoustic origin remain intact. In order to evaluate the performance of an actual loudspeaker array, we performed a full wave field measurement with a microphone array in a hemianechoic chamber. Figure 4c shows that the actual wavefront as measured in the real environment agrees with the simulation. To demonstrate the qualitative difference from a "standard" videoconferencing auralization, we simulated the field reproduced by a spaced pair of loudspeakers using a conventional stereophonic technique, shown in Figure 4d. Clearly, the wavefront is poorly reconstructed compared to the six-loudspeaker array. While stereophonic reproduction may produce an approximately correct localization at a small set of positions in the receiving room, only a multiple-element array is able to maintain correct acoustic imaging over a large listening area. In this experiment, we demonstrated accurate reconstruction at 500Hz using only six loudspeakers. While parametrically correct reconstruction of a higher-frequency signal calls for an increase in loudspeaker array density [14], this may be unnecessary for correct localization of speech sources.

6. AURAL AND VISUAL COUPLING

An understanding of the perceptual benefits of coupling can be used to develop more robust and immersive interactive environments. For example, when a participant moves closer to the screen in the receiving room, he or she expects more of the scene to be revealed at the borders of the image, but it may be difficult for projected video to achieve this effect at close range. On the other hand, the simulated sound field will be most accurate close to the screen. Previous research in auditory-visual interaction [4] suggests that the accurate audio may improve the perceived quality of the video. Conversely, at positions further back in the receiving room where the auralization may be less accurate, realistic virtual video may act to improve the perceived quality of the sound field.

As an initial test of perceptual effectiveness, we computed and rendered the complete acoustic field for a virtual rectangular room with specularly reflecting walls, and coupled this with a (non-tracked) moving video image of a person speaking. Initial subjective tests indicate that even when the loudspeaker spacing is less than the theoretical optimum relative to wavelength, the perception of the sound field localization is somewhat compensated for by the video image. More thorough and formal subject testing in this environment is required to qualify these results. Nevertheless, these findings seem to support the hypothesis that the requirements for array density and rendering accuracy depend

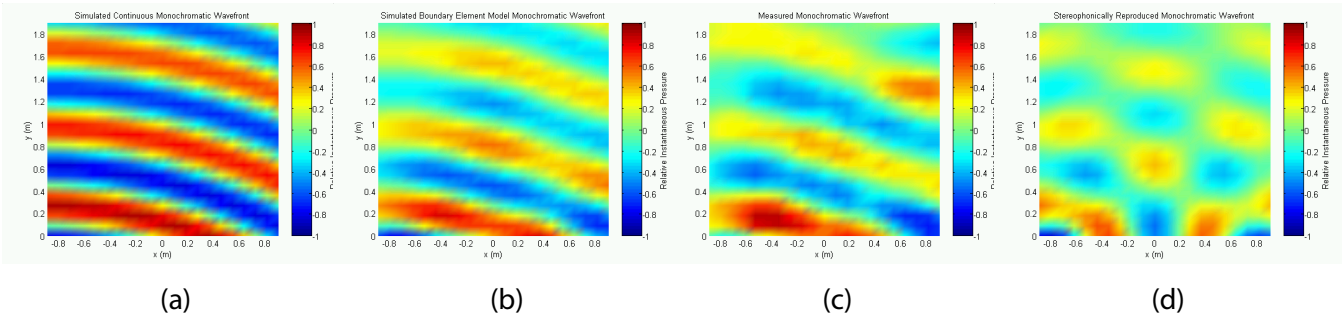

Figure 4: Synthesis of a monochromatic (500Hz) virtual acoustic field on a listening plane from an external source. (a) instantaneous pressure wavefront from an ideal continuous field by a single source; (b) simulated wavefront as reproduced through a 6-loudspeaker array; (c) actual measured wavefront as reproduced in a hemi-anechoic chamber using a 6-loudspeaker array; (d) simulated wavefront as reproduced using a conventional stereophonic technique using 2 loudspeakers spaced at 1.75m.

on the perceptual coupling between the visual and auditory sensory cues [11].

7. FUTURE DIRECTIONS

Our short-term goal for this project is to produce the following demonstration in real time, expanding upon the acoustic coupling experiment described above and our previous experience in view synthesis. A pair of cameras in one room (the "sending room") is calibrated beforehand and used to track the three-dimensional position of a moving person in a real environment. This position estimate is passed via TCP/IP to a computer in a second room (the "receiving room") where the source position and signal are used to synthesize an approximation to the continuous sound field that is synthesized using a linear loudspeaker array. To date, we have each of the pieces integrated into a near real-time system. The next step is to extend the demonstration by projecting dynamic virtual video in the receiving room that changes with respect to both source and receiver position, first with one participant in each space, and then with multiple participants in each. Drawing on the previous acoustic perception tests, we are beginning to conduct perceptual studies with real users to investigate the coupling between the visual and acoustic representations.

8. ADDITIONAL AUTHORS

Rendell Torres (School of Architecture, Rensselaer Polytechnic Institute, email: `rrtorres@rpi.edu`), Yasushi Shimizu (School of Architecture, Rensselaer Polytechnic Institute, email: `shimiy@rpi.edu`), and Seema Jaisinghani (School of Architecture, Rensselaer Polytechnic Institute, email: `jaisis@rpi.edu`).

9. REFERENCES

[1] Carrouso project homepage. `http://emt.emt.iis.fhg.de/projects/carrouso/`.

[2] S. Fussel, R. Kraut, and J. Siegel. Coordination of communicational effects of shared visual context on collaborative work. In *Proc. ACM Computer Supported Collaborative Work 2000*, pages 21–30, 2000.

[3] R. S. Keiffer and J. Novarini. A time domain rough surface scattering model based on wedge diffraction: Application to low-frequency backscattering from two-dimensional sea surfaces. *J. Acoust. Soc. Am.*, 107(1):27–39, January 2000.

[4] P. Larsson, D. Västfjäll, and M. Kleiner. Auditory-visual interaction in virtual reality: Spatial auditory cues and presence in virtual environments. In *Submitted to the 22nd Conference of the Audio Eng. Soc.*, June 2002. Espoo, Finland.

[5] L. McMillan and G. Bishop. Plenoptic modeling: An image-based rendering system. In *Computer Graphics (SIGGRAPH '95)*, pages 39–46, August 1995.

[6] A. D. Pierce. *Acoustics: An Introduction to its Physical Principles and Applications*. Acoustical Society of America, 1991.

[7] G. Pingali, G. Tunali, and I. Carlbom. Audio-visual tracking for natural interactivity. In *Proc. ACM Multimedia '99*, pages 373–382, October 1999.

[8] R. Radke, P. Ramadge, S. Kulkarni, and T. Echigo. Using view interpolation for low bit-rate video. In *Proc. ICIP 2001*, October 2001. Thessaloniki, Greece.

[9] R. Radke, P. Ramadge, S. Kulkarni, T. Echigo, and S. Iisaku. Recursive propagation of correspondences with applications to the creation of virtual video. In *Proc. ICIP 2000*, September 2000. Vancouver, Canada.

[10] S. Seitz and C. Dyer. View morphing. In *Computer Graphics (SIGGRAPH '96)*, pages 21–30, August 1996.

[11] R. Storms. *Auditory-Visual Cross-Modal Perception Phenomena*. PhD thesis, Naval Postgraduate School, 1998.

[12] U. P. Svensson, R. I. Fred, and J. Vanderkooy. An analytical secondary source model of edge diffraction impulse responses. *J. Acoust. Soc. Am.*, 106(5):2331–2344, 1999.

[13] R. Torres, U. Svensson, and M. Kleiner. Computation of edge diffraction for more accurate room acoustics auralization. *J. Acoust. Soc. Am.*, 109(2):600–610, 2001.

[14] D. Ward and T. D. Abhayapala. Reproduction of a plane-wave sound field using an array of loudspeakers. *IEEE Transactions on Speech and Audio Processing*, 9:697, September 2001.