Interacting in Shared Reality

Jeremy R. Cooperstock
Centre for Intelligent Machines and
Centre for Interdisciplinary Research in Music Media and Technology McGill University
Montreal, Canada
Tel: +1-514-398-5992
jer@cim.mcgill.ca

Keywords: computer-mediated communication, videoconferencing, telepresence, high-bandwidth networking, multisensory communication

Abstract

Commercial videoconferencing products have begun to reach a level of quality acceptable for many low-intensity interactions. However, these systems fail to deliver true "high-fidelity" that serves as a viable alternative to physical co-presence for more demanding interactions. The solution, we believe, lies in the synergy between high bandwidth networks and the application of information technologies that take advantage of such networks. Specifically, computation can be employed to enrich the communication channel, exploiting an awareness of users' activity in order to better support their needs. In this manner, we are entering an era of communication in which distance need no longer dictate limitations on high quality distributed experiences and interaction.

1 Introduction

Shared Reality refers to a rich sense of co-presence with someone in a remote location. Unlike virtual reality, which synthesizes graphical alternatives to the real world, Shared Reality attempts to convey a true likeness of the participants, through the transmission of life-size video, spatialized audio, and vibro-sensory[1] information. Differing also from conventional videoconferencing, it creates an immersive experience for the participants by performing self-signal isolation on each of the streams: background removal for video, and echo-supression for audio and vibro-sensory data.

Satisfying the requirements of a synchronized, distributed musical performance has been a "Holy Grail" challenge of videoconferencing technology and networks for decades. Music serves as one of the most demanding of tasks from the perspective of both sensory acuity and sensitivity to timing, and is therefore, of greatest interest for study. It is thus these factors of audiovisual quality and low latency that we consider as critical determinants of the success of Shared Reality as a viable and convincing alternative to physical co-presence. Although not all distributed interaction need achieve the quality nor the latency bounds dictated for musical performance in order to succeed, we believe that at least subconsciously, users are aware of degradations and added delay, and alter their interaction behaviour accordingly.

While numerous trials of distributed musical performance, conducting, or teaching, have been carried out previously, these generally entailed either a sacrifice of interactivity (e.g. musicians playing separately and combined via timing tracks) or a forced, unnatural, adaptation of interaction (e.g. musicians deliberately play ahead of the notes heard from the remote end). Such accommodation to the medium was necessitated by virtue of latencies, variously due to signal propagation time, hardware interface delays, and software or codec algorithm details.

Through the remainder of this paper, we explore the challenges of supporting such distributed interaction in which the technology succeeds in bridging the distance gap.

[1] We employ this term to describe the transmission of low frequency stimulus, such as that of floor vibration in response to footsteps or percussion instruments.

2 Previous Studies

Earlier studies have related the delay tolerance of distributed performers in a network scenario to the acoustic delays characteristic of their performing environment in the non-networked case [9]. For multi-party interaction, Rasch's studies [12][13] found that among members of a woodwind instrument ensemble, deviations of onset from 30 to 50ms were perceived as being tightly synchronized. This would correspond to acoustic propagation over a distance of 10-16m, only slightly greater than the size of typical concert stage. Other experiments conducted at CCRMA [3] involving two drummers separated by increasing distances and playing a set of examples of graduated rhythmic complexity, found a critical latency threshold in the vicinity of 100 ms. Below this limit, the performers were able to synchronize well. Regardless, delays arising from typical videoconferencing systems are a minimum of one order of magnitude greater.

At the upper end of the spectrum, in earlier work [16] we cited informal studies by Michael Brook, Bob Adams and Richard Boulanger noting that solo performers find feedback delays from daisy-chained MIDI instruments as low as 2-6 ms to be noticeable and in fact annoying. Others have pointed out that a 0.5 ms feedback delay, for example, from a digital in-ear monitor, will result in a notch at the 1kHz band, highly disruptive for a singer. Chafe [4] provides a detailed study of time delay effects on distributed ensemble performance.

3 Progress toward Shared Reality

A review of the history of networked media performance through the years [5] clearly demonstrates that musicians can learn to compensate for limitations of the audio channel and adapt to the delays associated with network encoding and transmission. Although this compensation may require significant conscious effort, interaction can be achieved through a wide range of latencies, with varying cutoff tolerances depending on the type of music and the training of the performers.

However, if the latency is sufficiently low, not only is the performance of a musical duet possible *without* such compensation, but the network echo[2] can fall within the tolerance of acoustic echo. This result is particularly exciting for networked music, as current echo-cancellation units are considered to be inadequate for such applications.

In terms of signal fidelity, most videoconferencing systems employ some form of compression in order to reduce bandwidth requirements. However, such compression is typically lossy and may result in readily observable artifacts. Numerous studies have been conducted to understand the effects of such compression on human task performance. However, other factors, often neglected, may be of equal if not greater importance to perception and performance, including video size, gaze awareness, audio spatialization, echo supression, and other modalities of communication. With respect to creating the sensory illusion of a distributed shared environment, one must also consider the need for dynamic background removal and perspective projection so that the video of remote participants appears blended into the local space, possibly rendered to a virtual camera viewpoint as appropriate to the viewer's position.

The observations above motivated us to tackle the prototyping of a high-quality networked immersive audiovisual environment as well as a series of real-time interaction challenges. Perhaps the most demanding, in terms of the timing requirements necessary to preserve a convincing illusion of presence, was our 2001 demonstration of high quality videoconferencing for a distributed violin duet. We note that the tolerance of transmission delay in a duo performance is typically much lower than for larger ensembles, given that the musicians are used to close physical proximity.

The demonstration itself, which took place between two Montreal universities only a few kilometers apart, was accomplished by transmission of both uncompressed audio and video over IP, thereby obviating the undesirable delays associated with standard codecs such as MPEG or H.32x, and enforcing a tight bound on retransmission attempts. Even so, audio and video interface buffers, processor scheduling, routers, retransmission of lost packets, and of course, the physics of light travel all add unavoidable latency, which, in this case, totalled approximately 20ms.

The following year, we conducted a similar trial between Montreal and Stanford, enabling jazz musicians to jam together with an end-to-end delay of less than 50ms. While this delay proved tolerable, the musicians noted that the experience of playing over the network required considerable effort to remain synchronized, an effect more noticeable on certain musical pieces than others. Even without addressing other sources of latency, reducing the network delay

[2] This term refers to the feedback arising from one musician's audio, output through the speakers at a remote location, then picked up by the microphones and fed back into the network, such that it is reproduced as echo at the source.

to that of a dedicated lightpath between the two cities should prove sufficient to alleviate this difficulty. However, another factor noted during the trial was that of gaze awareness and the challenge of supporting multiple independent views on a shared display surface so that each musician maintains an appropriate perspective view of the remote site.

4 Technical and Percpetual Considerations

4.1 Network signal propagation

Our own experiments with low-latency transmissions have been restricted to the research Internet structures of CA*net3 and its successor CA*net4 in Canada and the Internet2's Abilene in the USA. In practice, some of the paths through these networks are more efficient than others, and we were sometimes able to reduce end-to-end latencies by as much as one third by requesting alternate routings, in particular for cross-border events.

While most long distance Internet traffic is carried over a fibre medium, the physics of light propagation accounts for only a small part of Internet delay. A far more significant factor is that of intervening router delays, each of which may buffer incoming packets before relaying them to the next hop along the path to their ultimate destination. Competing traffic through the same network links result in additional delay, a problem exacerbated by some router prioritization rules, which may penalize heavy bursty traffic such as that associated with an unshaped payload of large video frames at (relatively) modest frequency.

The following table compares the latencies of round trip light travel over the network path between a node at McGill University in Montreal and various North American destinations, along with those of actual network ping time. Note that the number of routers is, by itself, a poor indicator of network delay, since only 4ms of the total ping time is spent traversing the first eight routers between our node and the primary CA*net4 backbone.

Destination	**Network distance (km)**	**Light travel time (ms)**	**Ping time (ms)**	**# intervening routers**
Ottawa	130	.8	4	10
New York	452	3	15	12
Vancouver	3198	22	64	13
Los Angeles	4709	32	97	13

The traditional suggestion for coping with competing traffic is to employ quality of service (QoS) mechanisms, which, when enabled, permit high priority packets associated with latency sensitive streams to jump to the front of the queue. However, QoS suffers numerous deployment problems and thus, is only supported to a limited extent through both the research and public Internets. A more exciting prospect for long-term communication trends is the evaluation through CA*net4 of User Controlled LightPaths, which permits users to own and control dedicated wavelengths, thereby permitting more efficient routing mechanisms for high bandwidth traffic and potentially improved latency bounds.

4.2 Interface Hardware

Low-latency videoconferencing remains a fairly small market niche for the manufacturers of audio and video hardware interfaces. Thus, despite tuning our software implementations to efficient, minimal-copy buffer transfers between audio and video peripherals and the computer's network interface [6], our end-to-end signal delivery is nonetheless delayed significantly by the underlying hardware. As but one example, we were surprised to discover that SDI interface cards were adding two video frames of delay[3] as was the video processing circuitry of our plasma displays.[4]

[3] This delay resulted from a decision by the hardware manufacturer to provide access to the video data through a double buffer, as required to ensure clean transitions between multiple video sources. These interfaces are often designed to be used with computer-based video editing systems, where latency is less of an issue. However, we are now investigating with another manufacturer the possibility of extremely fine-grained access to the video buffer as incoming data arrives and hope to achieve latencies on the order of several milliseconds rather than several tens of milliseconds.

[4] These circuit elements are apparently in place to perform scan conversion and de-interlacing; however, it may be possible to bypass such processing if the signal is provided in the exact native format of the display, via digital input, an approach we are currently testing.

4.3 Video Quality

Video quality comprises not only display resolution but also size. The display of one or more videoconference participants on a small video monitor simply cannot convey the subtle visual cues of gaze awareness, facial tension, and other gestures that we take for granted as a part of human communication. Furthermore, as we rely, to a certain extent, on object size to gage distance, a less-than-life-size display of a remote participant immediately violates the intended illusion of *virtual* proximity.

While large displays are often expensive, the ability to see one's counterparts rather than miniature images on a computer monitor, but instead, as life-size, engenders a powerful perceptual effect. We have conducted numerous trials involving life-size and near-life-size displays (see Figure 1) and noted such interesting effects as, for example, one participant, wearing clip-on microphones, *leaning in* toward the screen to repeat her name to the remote participant. Such behavior is, of course, entirely common in everyday co-present interaction, but quite atypical of videoconferencing scenarios.

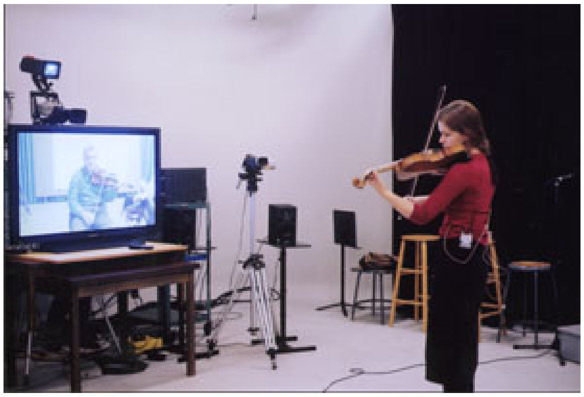

Figure 1: Véronique Mathieu (foreground) plays for Maestro Pinchas Zuckerman (displayed on a 127cm plasma screen in near-life-size).

Pushing further on this sense of immersion, we are also experimenting with background subtraction and depth-based segmentation techniques to render a projected video image of remote participants as if they are physically part of a (virtualized) local environment [1], as illustrated in Figure 2.

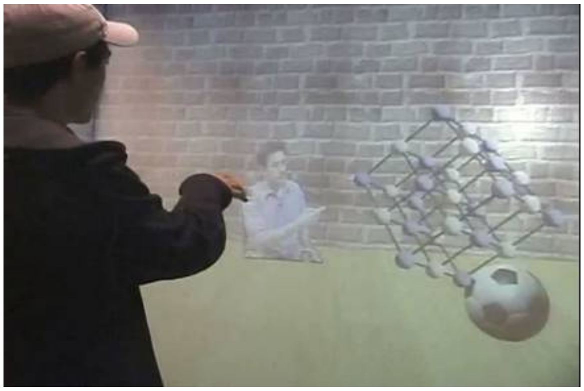

Figure 2: A projected remote participant, rendered as a "2D cardboard cutout" in the virtual museum space. The video display is transformed in real time based on the position of the viewer in order to provide a correct perspective view.

One of the most significant historical contributions to telepresence was the use of a half-silvered mirror to correct for the disparity between camera position and a viewer's gaze, thereby allowing for eye-contact and gaze awareness [2], both integral components of human communication. Unfortunately, it is difficult to scale such a system to a room-sized immersive environment with multiple participants, so we must turn, instead, to computer-mediation in order to construct a view from the vantage point of a virtual camera, as appropriate to the user's position. Numerous techniques have been proposed [11][14][8] [7][17][10] but these either generate obvious reconstruction artifacts or impose computational demands that prevent real-time operation.

4.4 Audio Quality

Notwithstanding the obvious importance of audio to musical applications, high resolution, multichannel audio, long since understood by the audio engineer as a rich and powerful means of conveying *presence*, seems to have been largely neglected, even in supposed "high-end" videoconferencing systems and facilities. It is somewhat ironic that despite the typical allocation of videoconferencing bandwidth two or more orders of magnitude greater than that available to conventional telephony, a simple telephone conversation often delivers a more satisfying interaction experience.

In physical co-presence, we subconsciously exploit binaural audio cues to localize sound sources (i.e. a human speaker) and discriminate between multiple simultaneous conversations. When sound capture and reproduction in a videoconference are stereophonic at best, we are deprived of these important communication cues. In contrast, synthesizing sound reproduction to provide spatial congruence between the apparent audio location and the video representation of remote individuals adds a highly compelling and immersive aspect to the communication. Furthermore, such spatialization allows participants to attend to one of multiple simultaneous conversations possibly being conducted through the medium without confusion (i.e. the "cocktail party effect").

Another critical aspect to audio quality is that of acoustic echo suppression, a pressing problem for loudspeaker – as opposed to handset or headset – interaction. The naive approach employed by inexpensive speakerphones is to mute the far-end signal when a local source of sufficient amplitude is detected, but this is clearly unacceptable for high quality interaction. Careful microphone and loudspeaker placement can alleviate the effect significantly, but achieving a level of echo-supression that convincingly masks the network feedback path, in particular for the multi-channel case, remains an open research problem.

4.5 Codecs

Although computation in the form of audio and video codecs has long been employed to lower the bandwidth requirements for videoconference communication, we argue that this approach is fundamentally unsuitable for the purpose of fostering improved social dynamics in group dialogue and a greater sense of *belonging*. The reason is simple: while conventional compression algorithms substantially reduce the bandwidth required to transmit an acceptable representation of the signal, doing so does not *improve* the quality of communication, rather it introduces excessive latency, anathema in most settings to effective human interaction. Even codecs avoiding a backward prediction phase typically fail to provide robustness to network data loss, which results in either corrupted or dropped frames.

In the ideal of unlimited bandwidth, compression can be avoided entirely, allowing for the devotion of available computational power to active mediation of the communication channel, as required, for example, to support acoustic echo suppression, gaze awareness, high-resolution synthesis of life-size displays, and spatialized audio rendering. Audio or video encoding, if employed, should be concerned more with issues of tolerance of data loss and scalable data representations suitable for multicasting to a pool of heterogeneous clients. In the latter case, these representations should allow for the balancing of conflicting demands of optimal immersive quality, minimal latency and maximal reliability.

In the event that sufficient bandwidth is not available to support uncompressed signal transmission, compression is, of course, necessary, but this can and should provide further benefits to the user, such as that offered by Set Partitioning in Hierarchical Trees (SPIHT) [15] in terms of progressive coding that permits graceful degradation with data truncation.

4.6 Sensory Immersion

We note that while the sensory perception of audio and video quality remains an issue for telepresence communication, the state of the art in current hardware alone is certainly sufficient in this regard to permit effective distance collaboration.[5] However, commercial videoconference systems, for the most part, continue to rely on standard

[5] While far from the norm in conventional videoconferencing environments, high-definition cameras, UXGA resolution displays, and multichannel 24bit/96kHz audio equipment can rival the best of movie theatre quality along both axes of video and audio.

television definition at best, with scant attention to audio quality, by far a more critical factor for effective human conversation.

Similarly, the use of vibro-sensory data remains virtually untapped in distance communication. This modality is arguably of greater importance in entertainment or distributed musical applications than for conventional videoconferencing, but we expect to see it play an increasingly important role in immersive environments. Experiments are underway to measure the effect of the inclusion of vibro-sensory information in various distributed applications.

5 Conclusions

Multisensory data transmission over broadband networks promises to revolutionize distributed human interaction. However, we must be cognizant of several factors, often overlooked in conventional videoconferencing, in developing and deploying these systems, if they are to be more than simply tolerated by the user community. Notably, due attention must be paid to modalities other than video, in particular, high-quality audio, the physical extent, both in terms of image size and audio spatialization, and end-to-end latency of the system. By ensuring that these factors are raised now and addressed in current research, we may soon be able to realize the full potential of distributed *Shared Reality* interaction, especially as the underlying technology becomes increasingly affordable and “broadband connectivity” begins to provide the necessary levels of bandwidth to distribute the data with sufficiently high quality and low latency.

Acknowledgments

The author wishes to thank Canarie Inc., Valorisation Recherche Québec, and the Natural Sciences and Engineering Research Council of Canada, whose generous funding has supported the research project described in this paper. Additional thanks must be given to our numerous research partners who have offered valuable feedback throughout the development and evolution of our systems. The Ultra-Videoconferencing software may be downloaded from `http://ultravideo.mcgill.edu`.

References

[1] Boussemart, Y., Rioux, F., Rudzicz, F., Wozniewski, M. and Cooperstock, J.R. A Framework for Collaborative 3D Visualization and Manipulation in an Immersive Space using an Untethered Bimanual Gestural Interface. Virtual Reality Systems and Techniques, Hong Kong, November 10–12, 2004.

[2] Buxton, W. and Moran, T. EuroPARC’s Integrated Interactive Intermedia Facility (iiif): Early Experience, In S. Gibbs & A.A. Verrijn-Stuart (Eds.). Multiuser interfaces and applications, Proceedings of the IFIP WG 8.4 Conference on Multi-user Interfaces and Applications, Heraklion, Crete. Amsterdam: Elsevier Science Publishers B.V. (North-Holland), 11-34.

[3] Chafe, C. SoundWIRE Group at CCRMA: What Seems to be the Delay? http://www-ccrma.stanford.edu/groups/soundwire/delay_p.html

[4] Chafe, C., Gurevich, M., Leslie, G., and Tyan, S. Effect of Time Delay on Ensemble Accuracy. Proceedings of the International Symposium on Musical Acoustics, March 31 – April 3, 2004, Nara, Japan.

[5] Cooperstock, J.R. Milestones in Real Time Networked Media. http://www.cim.mcgill.ca/ jer/research/rtnm/history.html

[6] Cooperstock, J.R. and Spackman, S. “The Recording Studio that Spanned a Continent,” Proc. *IEEE International Conference on Web Delivering of Music (WEDELMUSIC)*, Florence, 161–167, 2001.

[7] Fitzgibbon, A.W., Wexler, Y. and Zisserman, A. Image-based rendering using image-based priors. International Conference on Computer Vision, 2003.

[8] Hartley, R. and Zisserman, A. Multiple View Geometry in Computer Vision. Cambridge University Press, 2000.

[9] Lazzaro, J. and Wawrzynek, J. A Case for Network Musical Performance. NOSSDAV ’01, June 25-26, 2001, Port Jefferson, NY.

[10] Paris, S. Sillion, F. and Quan, L. A Surface Reconstruction Method Using Global Graph Cut Optimization. International Journal of Computer Vision, 2005.

[11] Rander, P., Narayanan, PJ, and Kanade, T. Virtualized Reality: Constructing Time-Varying Virtual Worlds From Real World Events. IEEE Visualization '97.

[12] Rasch, R. The perception of simultaneous notes as in polyphonic music. Acustica, 40, 21-33, 1978.

[13] Rasch, R. Synchronization in performed ensemble music. Acustica, 43, 121-131, 1979.

[14] Raskar, R., Welch, G., Cutts, M., Lake, A., Stesin, L. and Fuchs, H. The Office of the Future: A Unified Approach to Image-Based Modeling and Spatially Immersive Displays. SIGGRAPH '98.

[15] Said, A. and Pearlman, W.A. A New Fast and Efficient Image Codec Based on Set Partitioning in Hierarchical Trees. IEEE Transactions on Circuits and Systems for Video Technology, vol. 6, pp. 243–250, June 1996.

[16] Xu, A., Woszczyk, W., Settel, Z., Pennycook, B., Rowe, R., Galanter, P., Bary, J., Martin, G., Corey, J., and Cooperstock, J.R. Real-Time Streaming of Multichannel Audio Data over Internet. Journal of the Audio Engineering Society, July-August, 2000.

[17] Zitnick, C.L., Kang, S.B., Uyttendaele, M., Winder, S. and Szeliski, R. High-quality video view interpolation using a layered representation. SIGGRAPH 2004.