SHAKE, RATTLE, AND ROLL: GETTING IMMERSED IN MULTISENSORY, INTERACTIVE MUSIC VIA BROADBAND NETWORKS

Wieslaw Woszczyk, AES Fellow, **Jeremy Cooperstock**, AES Member,
John Roston, **William Martens**, AES Member
Centre for Interdisciplinary Research in Music Media and Technology (CIRMMT), Faculty of Music, McGill University, Montreal, Québec, Canada H3A 1E3

Broadband Internet (transmission rates more than a gigabit per second) enables bidirectional real-time transmission of multiple streams of audio, video, and motion data with latency dependent on distance plus network and processing delays. In this article we describe a new immersive multisensory environment recently constructed at McGill University, designed for network-based communication for music performance coordinated between remote sites, potentially over great distance. The system's architecture allows participants to experience the music with greatly enhanced presence through the use of multiple sensors and effectors and high-resolution multimodal transmission channels. Up to 24 channels of audio, digital video, and four channels of vibration can be sent and received over the network simultaneously, allowing a number of diverse applications such as remote music teaching, student auditions, jam sessions and concerts, recording sessions, and postproduction for remotely-captured live events. The technical and operational challenges of this undertaking are described, as well as potential future applications.

INTRODUCTION

Our four-year research project, funded by the Valorisation-Recherche Québec of the Government of Québec, aims to develop software, hardware, and methods enabling "real-time communication of high-resolution multisensory content using broadband networks." An interdisciplinary group of researchers composed of audio engineers, computer scientists, network specialists, psychologists, acousticians, video technologists, music producers, and electrical and mechanical engineers has created a laboratory for the creation of an immersive multisensory environment. Important nontechnical members of the development team include classical and jazz musicians, music teachers and students, and renowned world-class performers. Using the laboratory, our goal is to develop a superior-quality communication system that will be transparent to the users and will satisfy their most demanding sonic and behavioral requirements. Artists, such as Pinchas Zukerman, who rely on distance learning tools to teach music have praised McGill technology for its quality. The system allows recording and reproduction and real-time bidirectional transmission of uncompressed multisensory, multichannel music with subtlety and detail afforded by high-resolution.

During the last two years the system design team has faced many technical and performance considerations in their effort to match the limitations of current technology with the high musical sensitivity and expectations of the users. In this article we review historical progress on this project and present critical issues that were addressed, as well as those that still need to be resolved. A picture of a new type of interactive studio/music environment based on networked communication emerges as a model for the future. This project also reminds us that successful networked music collaboration requires practitioners in diverse fields—music studios, teaching labs, television, live-concert production, and telephony—to share expertise and techniques to achieve the level of quality sought by both technologists and artists. The challenges of quality networked communication was first discussed in 1999 in the AES Technical Council White Paper "Networking Audio and Music Using Internet2 and Next-Generation Internet Capabilities" (PDF available at http://www.aes.org/technical/documents/i2.html). The topic was also addressed by the AES Technical Committee on Networked Audio Systems in an AES *Journal* paper, "Real-Time Streaming of Multichannel Audio Data over Internet" (2000 July/August, p. 627).

TELEPRESENCE, SHARED REALITY, AND IMMERSIVE REALITY

For high-fidelity networked communication we must create telepresence. The goal of telepresence is to connect physically separated spaces using multiple sensory links so that participants feel like they are in a single, unified space (see Fig. 1). A sophisticated multimodal display system should recreate in each space a comprehensive, visual portal into the remote site through which light, sound, and structural vibrations can be exchanged (de Bruijn, 2004).

In the virtual representation of a distant space and the activity within it we try to approximate physical reality, but often this representation will not be

enough. Musicians practicing in the same room keep time by relying on temporal cues, subtle tonal and spatial details of the sound, and occasional quick looks at body movement and instrument handling. They may even subconsciously rely on vibrations transmitted through the floor. A musician may choose to review any of these elements in any modality at any time, and he expects them always to be present and accurate in time, location, and magnitude. For our purposes, only when these multimodal events become perceptually fused through sensory integration into one construct in consciousness can we consider the technology facilitating the communication to be transparent.

At the same time, the re-creation of only perceptually genuine reality is not good enough to satisfy the requirements of advanced communication. To justify the use of such technology, it would be nice to hear and see better than musicians practicing in the same room, with greater detail or larger perspective or simply with more volume. So the challenge is to make all options available to the participants without calling particular attention to any one of them, because the lack of balance would inhibit perceptual fusion. This requires tremendous resolution and a large canvas to create lifelike auditory, visual, and haptic channels in an immersive display. The information provided for the participants must be laid out fully and openly and must be easily accessible without constraints. To achieve this we need a wide-angle (extending into peripheral vision) video screen presenting life-size objects; immersive multichannel sound with height, width, and depth in high-resolution digital audio; and full-body motion with multiple degrees of freedom.

Often we also have the challenge of blending two different environments. The two remotely connected yet independent spaces should appear to be present at each site. However, depending on the character of each, these spaces can be quite different from one another. If, for example, one site is an office room and the other is a recital hall, the sense of a common acoustic space should be created by combining the two unlike environments into a unified one. In a case such as this we could expect that when the office room is connected to the recital hall, people speaking in the office would want to hear the reverberant response of the recital hall because it would indicate to them that the recital hall is a part of their own space. Persons speaking in the hall would most likely not be required to hear the acoustic response of the office room, as it would likely become masked by the hall's reverberation. Some order of priority and plausibility needs to be established for each case, requiring a synthesis of a shared ambiance.

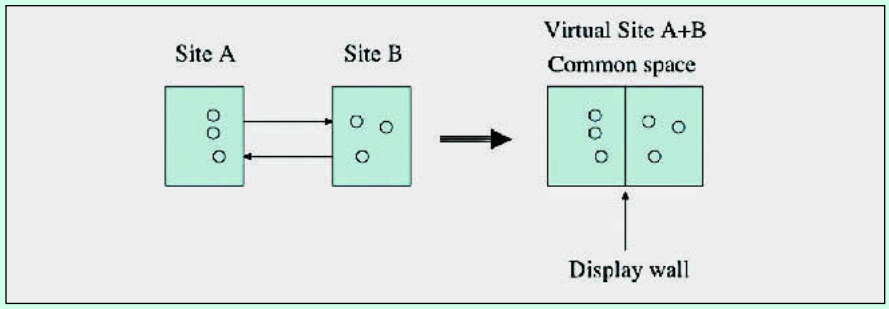

Fig. 1. The concept of telepresence: two physically remote sites become connected electronically via a common display wall that recreates for each the presence of the other site using audio, video, and haptic channels transmitted via broadband networks.

SYNTHESIS OF SHARED AMBIANCE

Thus far we have found that each connected site must have the means to create shared ambiance that approximates some plausible balance between the two component spaces or creates an imaginary idealized space altogether. There could be a situation in which a common virtual space has been created from artificial sets of visual backgrounds and from imaginary acoustic environments not representing the actual spaces used. But it could be that we must reproduce an enhanced (actively responding) version of the actual space that is familiar to those present. So we need to have the means to capture that space and use it in the creation of a composite space. One solution is based on capturing meaningful multichannel impulse responses of each space and convolving them with close-captured source signals. The benefits include quiet yet responsive environments from sampled and processed spaces and better immunity from echoes by virtue of closely positioned microphones.

In all these cases microphone signals acquired at the connected sites must be triggering the common synthetic space. This means that when a user walks into one space and speaks, sings, or plays an instrument, the synthesized shared ambiance must respond. The ambiance of that one space is modified electronically to reflect the connection to the apparently adjacent site. Acoustical and visual harmony also must be presented clearly to all participants to invoke the sense of being in the same space. For example, dynamic response of the space to the source plus directional cues provided by the space and the sources must match one another.

AUDIO ENVIRONMENT

To achieve the perceptual goals of unimpeded (transparent) bidirectional high-fidelity communication, a sophisticated multisensory research and development environment was created at McGill University to test the experience in music participation between remote locations. The auditory display system is configured (see Fig. 2) as a spherical loudspeaker array consisting of 6 low-frequency drivers (ranging from 25 to 300 Hz) and 96 mid- and high-frequency drivers (ranging from 300 to 30,000 Hz). The lower-frequency drivers are placed at standard locations for the 6 main speakers in surround sound reproduction (the speaker angles in degrees relative to the median plane are 0°, ±30°, ±110°, and 180°). The upper-frequency drivers are dipole-radiating, full-range electrodynamic ribbon transducers featuring PFT (planar focus technology), and these 96 loudspeakers are placed 4 units wide in 24 panels (two loudspeakers per audio channel, two channels per panel) in 24 locations lying on the surface of an imaginary sphere of 4-meter diameter. Besides 6 locations at extremely high elevation, the spatial organization of the upper-frequency drivers is defined by 3 planes at elevation angles of –15°, +25°, and +45° relative to the horizontal plane. ➥

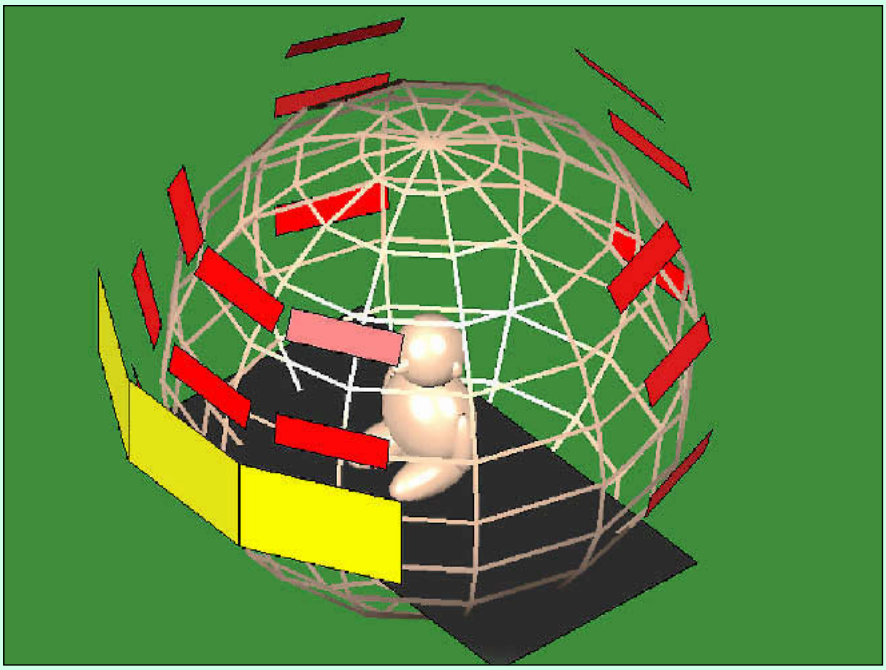

Fig. 2. Graphic depicting the 3-D configuration for 18 of the 24 loudspeakers (shown as red panels, some hidden from view) positioned on the surface of an imaginary sphere of 2-meter radius. The listener is situated on a 4-ft by 8-ft motion platform (show in grey) facing three video screens (shown in yellow) which subtend 90 degrees of horizontal angle and 17 degrees of vertical angle.

Within each plane of differing elevation angle, 6 speakers are placed at azimuth angles matching those of the 6 lower-frequency drivers (again, 0°, ±30°, ±110°, 180° relative to the median plane). The goal here is to do more than create the sense of listener envelopment available in conventional surround sound reproduction; instead, we attempt to simulate a more comprehensive soundfield in which components of captured and/or synthesized reverberation are presented from angles of incidence that remain spatially stabilized as listeners turn their heads relative to the spherical loudspeaker array. By encircling listeners with 6 low-frequency drivers, even with 90° head turning, there can always be a reproduction of low-frequency incoherence naturally associated with room acoustics for binaural listeners (Martens et al., 2004).

Fig. 3. Pinchas Zukerman in Ottawa giving a master class to violin students in Montreal (photo courtesy Owen Egan).

VIDEO ENVIRONMENT

Our goal is to create an environment for the participants as close as possible to what they would experience if they were sharing the same physical space. In this particular case, the essence of that experience is human interaction with another person, whether as teacher, student, or performer. The role of the video component is to provide the visual information and cues essential for human interaction. Therefore high-quality video is used to reproduce the necessary visual detail and fidelity of movement on a life-size display. Objects should appear as they would if present in the same space as the viewer, neither smaller nor larger.

In the initial experiments in distance music teaching, broadcast standard digital video (SDI at 270 Mbps) was used and displayed on a 50-inch (127-cm) diagonal plasma display. We are now moving to broadcast standard high-definition video (HD-SDI at 1.5 Gbps) on a 65-inch (165-cm) plasma display. Since low latency is important, the progressive scan standard 720p60 will be used to match as closely as possible the native resolution of the plasma display. This will minimize video processing by eliminating the need for de-interlacing that would be necessary if the competing 1080i60 standard were used. Some recent model plasma displays have direct SDI/HD-SDI inputs that also minimize video processing and reduce latency.

High-resolution video enables the viewer to be as close as 5 feet (1.5 m) from the display without seeing video scan lines. This more closely approximates the customary distance for one-to-one human interaction. Since there is a camera above the display sending an image of the viewer to the remote location, moving the viewer closer to the display makes it obvious that the viewer is looking below the camera, not directly into it. This is disconcerting to the viewer at the remote location since the lack of eye contact detracts from effective human interaction. Putting life-size images of remote participants at a customary one-to-one distance from the viewer contributes to his increased sense of presence within the shared virtual space.

WHOLE-BODY HAPTIC STIMULATION

Whole-body haptic stimulation from motion and vibration signals, captured from live performance using accelerometers or synthesized via automatic analyses (Martens and Woszczyk, 2004c), are presented via a commercially available motion platform (the Odyssée system from the Quebec-based company D-BOX Technology; http://www.d-box.com). This platform uses four coordinated actuators to vertically displace a wooden platform of 4 by 8 feet on which chairs are fixed. When all four actuators move together, participants can be displaced linearly up or down. The movement can be very quick and forceful (the feedback-corrected linear system frequency response is flat from DC to 50 Hz). The controller can move the platform with two types of angular motion. Thus, the potential mismatch between audiovisual stimulation and

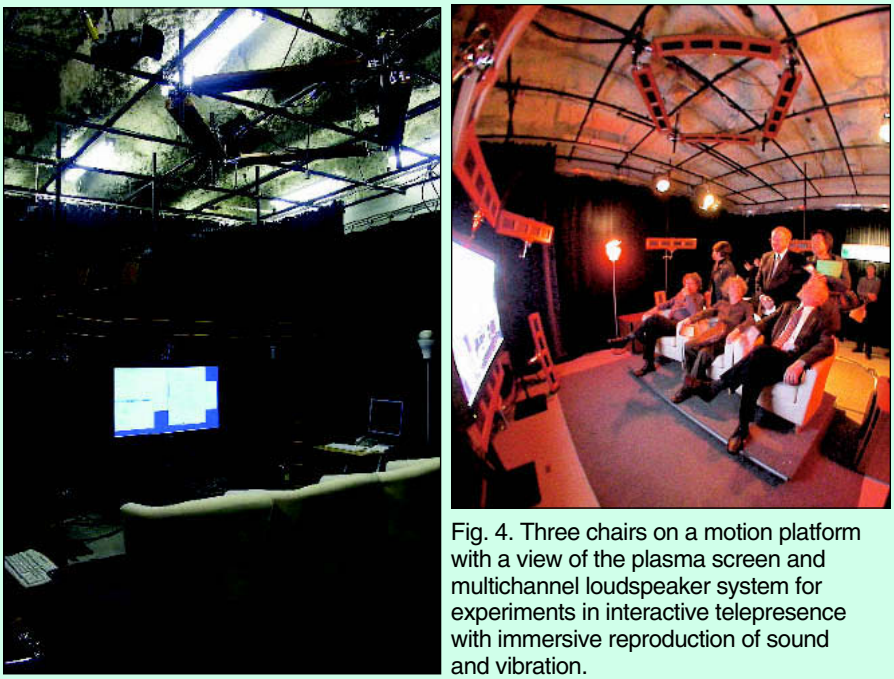

Fig. 4. Three chairs on a motion platform with a view of the plasma screen and multichannel loudspeaker system for experiments in interactive telepresence with immersive reproduction of sound and vibration.

bodily motion can be corrected. Adding touch (haptic) sensations and motion (vestibular) sensations consistent with what is seen and heard heightens the sense of presence in the simulation. We have reached the conclusion that a comprehensive model of human spatial hearing cannot be formulated without the inclusion of the human perception of self-motion, both in terms of angular and linear acceleration of observers within their immediate environment (Martens, 2004).

NETWORK CAPABILITIES AND TRANSMISSION SOFTWARE

The project uses the transmission software developed at McGill (Xu et al., 2000; Cooperstock and Spackman, 2001). The existing software was made available free for noncommercial purposes and has been downloaded by over 100 individuals and organizations around the world (see http://ultravideo.mcgill.edu). For example, York University in the UK in partnership with British Telecom used it for a project entitled BT Music On-Line (http://music.york.ac.uk/musiconline) that is intended to bring remote professional coaching to orchestras around the UK. The software is capable of transmitting standard-definition SDI video over IP networks; it will be developed further to support high-definition video. The existing software supports unicast transmission of standard-definition video, high-resolution audio, and haptic data between only two sites. We will further develop the software to support multicast transmission of video and audio among multiple sites, a capability frequently requested by current users.

METHODS OF SOUND CAPTURE FOR NETWORKED TRANSMISSION

When recording the audio representation of an event we typically capture both the direct sound of the sources and the acoustic ambiance of the surrounding space. This can be done either through a common microphone system capturing and integrating direct and ambient sound or through separate microphones dedicated to each of these components. In capturing sound for real-time networked transmission of acoustic events, ambient sound is not desirable since it carries loudspeaker signals used to monitor the remote venue and the common ambiance enhancement. Especially when these signals are delayed in transmission, they can cause disturbing echoes that degrade intelligibility and impede communication.

The solution to avoiding the echoes is using independent microphones placed at a very close distance to each source; for example, miniature omnidirectional microphones placed near the mouths of speakers. If possible, cardioid or ambiance-canceling differential microphones can also be used but their placement is critical to achieving optimum results. These solutions provide a clear direct sound that is almost completely lacking any ambiance information. Therefore, in this case, shared acoustic space must be synthesized at each site using room simulators to produce the effect of the distant space being added to the local space. Either algorithmic synthesis or sampling-convolution room ambiance, or both, can be used; each has specific advantages and disadvantages. In all cases some feedback suppression is needed if ambiance is amplified locally using a live microphone feed.

ECHO-CANCELLATION

Ultimately, echo cancellation is required in bidirectional transmission to remove the crosstalk between the local microphones and the loudspeakers used to monitor the remote site (see Fig. 6). Because acoustic feedback is delayed due to latency in transmission and processing, it becomes highly audible when longer delay times are ➥

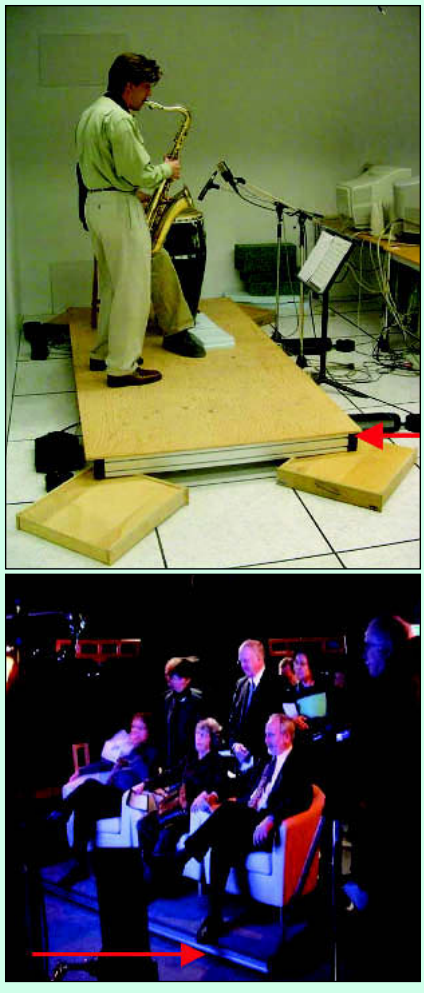

Fig. 5. Vibrational signal transmission using motion platforms. The red arrows point to the platforms. The audience can feel the vibration generated by musicians in the studio.

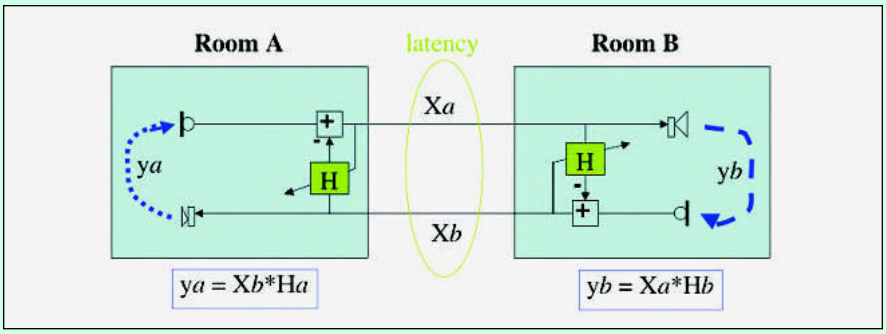

Fig. 6. Single-channel acoustic echo cancellation. Filter H models the crosstalk transfer function y and tries to remove it from the signal X at each location (a and b).

encountered. Adaptive digital filtering can remove the crosstalk, and this is moderately successful for speech signals in mono and stereo (Usher et al., 2004). However, multichannel echo cancellation for music signals is not yet ready for prime time.

A number of adaptive filters estimate the acoustic contribution of the loudspeaker output and subtract this from the microphone signal. Each such filter runs with different time and learning parameters and is maintained independently, such that the best performing one can be applied at any stage. This is important in order to provide both stable performance and rapid adaptation, for example, during a period of double-talk. However, the filtering introduces additional delay due to block processing and is imperfect: adaptation is noninstantaneous and transitions between the filters may introduce additional audible artifacts. Since the echo-cancellation process affects the primary signal path, it should, in the ideal, be totally free of distortion and artifacts for use in recording, archiving, and transmission. At present, one is faced with a difficult choice between the artifacts that do arise from echo-cancellation or the often more serious effects of echo itself.

More adaptive filters are needed when many loudspeakers are used because each filter must be dedicated to a given transfer function between the microphone and a specific loudspeaker. Using many loudspeakers and microphones in both directions is very difficult to process in real time, especially when microphones and objects move and change the transfer functions that filters are attempting to approximate. The cancellation of multiple far-end signals reproduced by multiple near-end loudspeakers captured by multiple near-end microphones has not been developed yet to a quality level acceptable to musicians. In our environment we found more success with echo-suppression methods than with echo cancellation, mostly by using microphones close to the source and by synthesizing the early reflection patterns to be inserted between the signal and its echo to make it less apparent perceptually. Additional benefit may result from automatic gain adjustment that dynamically reduces gain in the feedback loop.

LATENCY

Our continuing work with the current system is corroborating the results of other research groups with whom network-based interactive musical events have been demonstrated (see: http://ultravideo.mcgill.edu/overview/). For example, the group at Stanford University (Chafe et al., 2004) has recently completed a study of deviations in musical rhythm that can be expected from remote performers as a function of network latency. Our results and theirs point to the conclusion that latencies roughly on the order of 10 to 40 ms are easily tolerated, as would be predicted from the fact that such latencies are normally experienced when musicians in a common physical space are separated by 10 to 40 feet. Learning and practice can develop greater tolerance; however, longer delays than these begin to produce increasing degradation in rhythmic accuracy.

Because latency inhibits interactivity and is highly undesirable in ensemble music performance, our goal is to use technology that minimizes latency as much as possible. We only use uncompressed signals because lossy compression requires frequency domain processing that adds to the latency. Aside from the interface and signal-processing hardware, which is steadily improving, the limiting factor in latency is the distance between the connected sites and the delay due to the speed of light. This delay is substantial for intercontinental transmissions, therefore some standard musical activities may not be feasible. This creates the opportunity to develop new types of musical activities that embrace the delay in a creative way while reserving fully interactive real-time modes for distances less than 1,000 miles.

EXAMPLES OF UNIDIRECTIONAL AND BIDIRECTIONAL APPLICATIONS

What follows is a brief description of several applications of networked mulitsensory communication between remote sites and users. These applications are carried out using ultravideoconferencing technology developed at McGill University.

Medicine

McGill uses this new immersive ultravideoconferencing environment for teaching and remote communication applications where high quality is necessary, such as in medicine (Fig. 7) and music. The team relies on the collaboration of the Instructional Multimedia Services (IMS), which is responsible for the development of the video components, the Centre for Intelligent Machines, which is responsible for development of the transmission software, and the Centre for Interdisciplinary Research in Music Media and Technology of the Faculty of Music

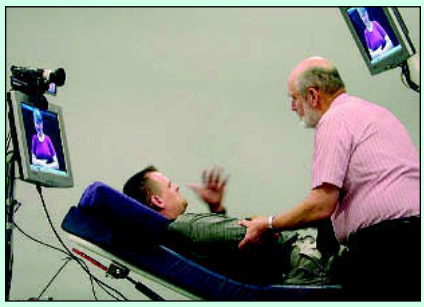

Fig. 7. Sign-language interpreter in a remote location assists in physician's communication with a deaf patient. Video cameras and monitors in the hospital are connected by network to a remote interpreter.

(CIRMMT), which is responsible for development of high-resolution multichannel audio capture and projection.

Music Master Classes

Pinchas Zukerman, Donny Degan, Pace Sturdevant, and Douglas Burden of the National Arts Centre Orchestra conducted music teaching classes with McGill University students using broadband CA*net3 (Canadian fiber-optic network, now CA*net4 capable of transmission rates of up to 40-gigabits per second) connectivity from the National Research Council in Ottawa to the IMS studio at McGill in Montreal (Fig. 8). The technology—McGill Advanced Learnware Network—was developed using a Canarie ANAST (Advanced Networks Applications, Services and Technologies Program) grant.

Fig. 8. Violin Master Class between Maestro Zukerman at NRC and a student at McGill University.

Jam Sessions

On June 13, 2002, the first crosscontinental jam session took place using our UltraVideoconferencing system in a minimal latency configuration. Musicians at McGill University and Stanford University jammed together over the Canarie CA*net4 and Internet2's Abilene research networks. The event featured full-screen bidirectional video and multichannel audio in what was heralded as the first demonstration of its kind over IP networks (see Fig. 9).

This project forms a component of the McGill Advanced Learnware Network project, funded by Canarie Inc. and Cisco Systems. It is also part of our four-year research project, Real-Time Communication of High-Resolution Multisensory Content via Broadband Networks, funded by the Valorisation-Recherche Québec.

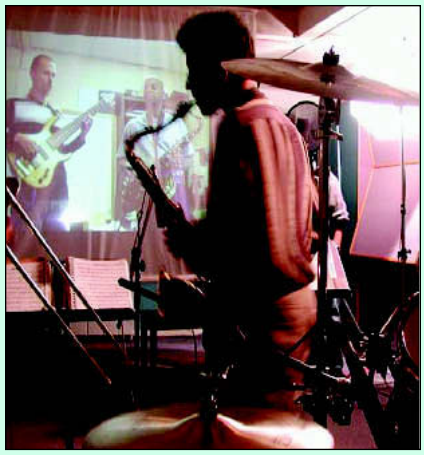

Fig. 9. Musicians at McGill jam with colleagues (on screen) at Stanford (photo courtesy Peter Marshall).

Remote concerts

The world's first transcontinental studio was demonstrated on Saturday, September 23, 2000, as part of the 109th Audio Engineering Society Convention in Los Angeles. The McGill Jazz Orchestra performed in a university concert hall in Montreal while the recording engineers mixing the 12 channels of audio during the performance were not in a control room at the back of the hall, but rather across the continent in a theater at the University of Southern California in Los Angeles, mixing for a live audience. This was the first time that live audio of this quality was transmitted over the CA*net3 and Internet2 networks with the resolution of 96kHz/24bits linear-PCM. Once the 12 channels of audio were mixed into six 96/24 outputs in a digital console in the USC theater, the six signals were converted to analog by 96/24 D/A converters before being sent to the theater's 6.1-monitoring system (see Fig. 10) for the immediate enjoyment of the audience.

Fig. 10. Audience at Norris Theater of the University of Southern California in Los Angeles.

On October 31, 2004, an AES audience at Genentech Hall of the University of California in San Francisco enjoyed the first transcontinental networked multichannel transmission of Direct Stream Digital audio (1-bit, 64Fs, Fs=44.1kHz) with SDI video. The transmission was a quadraphonic concert of the Haydn Quartet and the McGill Jazz Orchestra from McGill's historic Redpath Hall in Montreal. Two audio streams at 5.6 Mbps each and one video SDI stream at 270 Mbps were transmitted using CA*net4 and Internet2 networks as part of a special event prepared by the Technical Council and the Technical Committee on Networked Audio Systems at the AES 117th Convention (see Fig. 11).

MULTIMODAL INTERACTIONS: VIDEO–AUDIO AND VIBRO–AUDIO

A great deal of asynchrony between video and audio events can typically be tolerated, perhaps due to the wide range of intermodal delay values that can be observed for increasingly distant events (due to differences between the speed of sound and the speed of light). Nonetheless, our system allows us to delay audio signals to bring them back into synchrony with video signals that always require greater processing and transmission times. In our experience it is not always necessary to match the latency of the video signal by delaying the audio. In the jam sessions between Stanford and McGill we chose to run audio with minimum latency (approximately 50 ms) and video with its minimum latency (approximately 80 ms). The musicians were guided mainly by the audio, using the video occasionally to read gestures and emotions of the remote musicians just as they do when following the conductor (who could be located 30 feet away, matching an intermodal delay of 30 ms).

In contrast, vibro–audio synchrony was found to be much more critical (Martens and Woszczyk, 2004a). Though adjusting intermodal delay values can enhance perceptual results (Woszczyk and Martens, 2004), the ➥

Fig. 11. Quadraphonic DSD (Direct Stream Digital) transcontinental transmission of a concert using SDI video. Jeremy Cooperstock at UCSF control room (left). UCSF audience enjoying the view of the concert hall in Montreal, and the high-resolution quadraphonic sound from 3000 miles away in real-time.

natural multimodal experience occurs when multisensory components are properly timed across different modalities, increasing the chances that the musicians at the separate locations will be able to perceive them and integrate them into a harmonious performance (Martens and Woszczyk, 2004b). Psychophysical tests allow these multimodal components to be perceptually synchronized for our immersive display, and this improves perception of music and, in particular, the sense of rhythm and timing provided by music transmitted from a remote site. Furthermore, the relative intensity of the vibration and sound can be adjusted to keep results within an acceptable region as the loudness of the reproduced sound is varied, assuring perceptual fusion of sound with the user's feeling of bodily motion (Martens, 2004).

NEW PRACTICES IN MUSIC

Networked real-time communication over large distances using broadband networks introduces latency, echo, synchronization issues, and complex monitoring and mixing requirements that must be adequately resolved for each modality. The perceptual requirements of musicians, including cross-modal interactions, need to be addressed. Networked communication can be used in music studios to facilitate remote recording sessions, A&R reviews, remote overdubs and tracking, last-minute track changes and additions, long-distance session supervision, mix approvals, and auditions. As we discussed previously, networked communication can facilitate teaching when used in master classes, distance music teaching, rehearsals, and jam-sessions. It can also be used in film and sound design for remote team-composing, music scoring, and music mix approval when composer, producer, and orchestra are in separate locations.

Archiving telepresence

Networked musical interactions involving telepresence should be archived for future use. However, since a virtual representation of the remote site is combined with a multimodal representation of the local site, each performance can be experienced from at least two perspectives: local and distant. In an ideal case, when two connected venues are combined into a single virtual space, sources and environments need to be preserved in their elementary states as direct microphone and camara signals and captured impulse responses of the environments. They will be used for future reconstruction of virtual scenes with emphasis that depends on the intent and point of view of the presentation.

Quality levels

When the highest quality of audio is needed for recording, archiving, listening, and mixing, we should not use any echo cancellation and use only uncompressed high-resolution digital signals. For medium-quality applications—such as rehearsing, jamming, and auditioning—we can use some echo suppression but not echo cancellation, and should still use high-resolution digital audio having multichannel capabilities. For low-quality audio we can use echo cancellation as long as just voice and not music is being processed. Voice applications are similar to those for telephony: dialog, discussion, and commentary; in these applications echo cancellation is a must to ensure the appropriate intelligibility of speech. Increased latency due to compression may become a factor.

AUTOMATION

When each source has its own wireless microphone and moves freely within the remotely connected space, there is a need to dynamically adjust the direction and distance of the virtual source created in the reproduction space to correspond with its visual location. Having many sources captured with close wireless microphones can create confusion if they all appear virtually in the same acoustic location or in locations where they are not present visually. Thus, each source needs to be "followed" with a proper auditory design of perspective that includes dynamic adjustment of virtual reflections, reverberation, room boundary effects, and stimulated room modes. A system capable of such complex dynamic synthesis of spatial movement has been realized at McGill University. The system has shown excellent subjective results although only one source at a time can be dynamically positioned using a simple

controller, which is due to a large demand for processing power (Corey et al., 2001, 2001). We are developing a new controller of multisource perspective that will include distance, direction, and elevation cues and will allow additional simulation of source orientation for the 24-channel system described earlier.

At some point, a human mixer will be replaced with an automation system that tracks the position of each source and sends dynamic coordinates to the auditory positioner for composing a virtual presence of the source. For example, a software system that analyzes the video image to track objects, or a system tracking wireless microphones, or some other device identifying the position of each source will be used. Automation will have to be used in future communication systems to simplify the mixing of complex multisensory layers according to perceptual requirements for congruity and intelligibility. These requirements are currently being established.

CONCLUSIONS

A high degree of audio-video-haptic telepresence can meet the requirements of musicians performing in physically separated but electronically shared virtual spaces. Life-sized or larger video images help to provide detail and visual context supporting the auditory illusion that all participants are present in the same virtual location. Vibrosensory channels extend the low-frequency definition of audio channels and inform the participants of subsonic effects associated with the music and sound. To set up a proper multimodal environment with telepresence, sound designers and balance engineers need to collaborate with lighting and video-camera experts.

Typical sound design includes screen sound associated with video display and ambient sound associated with the created virtual space. Since echo may become audible in bidirectional transmission, good sound design may help to conceal the echo or arrange its redistribution into the ambient sound. The goal is to promote sensory awareness and involvement of all participants in musical interactions.

Because networked music auditions, recordings, or jam sessions may last for many hours, technology should be transparent and not cause any delays in the communication. Two audio mixes will likely be needed at each site, a microphone mix and a loudspeaker mix; for this we need two audio-monitoring spaces. The loudspeaker mix can be done in the performance space where the musicians are. Microphone signals should ideally be mixed in a control room. Even if all microphone signals are transmitted independently, their gain and quality must be evaluated and adjusted before the transmission. Some microphones should also be used to excite the local ambiance to give the impression that ambient sound (the shared room) is responding to the participants. For this, a third "monitor" mix (ambient microphone mix) may have to be created from the local microphone signals to trigger room simulation and enhance the audio balance between local musicians. This mix will be combined with the loudspeaker signals used to monitor the distant site. Since tight synchronization and the co-location of audio and video displays enhances the localization of sources and improves the spatial awareness of participants, it is usually desirable to align the audio, video, and haptic displays using specific test signals. However, in networked communication, latencies are common and mismatch of synchronization can occur. In our experience musicians prefer to deal with these shortcomings by focusing on sound. Therefore, when sound quality is excellent and strong ambient sound accompanies rich screen sound, their ability to cope with timing delays is optimized. An auditory 3-D perspective is simply much more compelling to a musician than a visual perspective from a 2-D display. It is important, therefore, to ensure a large listening area allowing musicians to move or change their perspective for best personal audio balance. The recording of interactions for instant recall is always useful for participants, students, and teachers because it helps them to learn from reviewing their interactions.

We speculate that high-quality multimodal networked installations will be used regularly in the future for teaching music, concerts, and recording sessions. The installations are most likely to be fixed and permanent because of the complexity involved in creating a compelling telepresence. However, intelligent home environments will gradually be developed as network bandwidth to homes steadily increases. Will we have musicians making house calls or famous teachers giving music lessons in peoples' homes via networked telepresence? Such face-to-face personal interactions are not possible today by means of telephone, television, or film. Based on the positive reactions of our users to telepresence, we believe that, yes, we will see these developments in the not-too-distant future.

ACKNOWLEDGMENTS

The authors would like to acknowledge generous funding support from Valorisation-Recherche Québec of the Government of Québec and a grant from Canarie Inc. ANAST (Advanced Networks Applications, Services and Technologies Program) that have allowed us to carry out this research. We thank Alain Berry and Dan Levitin for their partnership in various aspects of the research and Stephen Spackman and Jonas Braasch for their participation as research associates. We also thank our musicians, system users, equipment manufacturers, and collaborators, most notably Chris Chafe, as well as graduate student engineers in Sound Recording at McGill for their help and enthusiasm.

REFERENCES

AES White Paper: "Networking Audio and Music Using Internet2 and Next-Generation Internet Capabilities." Technical Council Document available from:http://www.aes.org/technical/documents/i2.html.

Bargar, Church, Fukada, A., Grunke, Keislar, Moses, Novak, Pennycook, B., Settel, Z., Strawn, Wiser, and Woszczyk, W., "Networking Audio and Music Using Internet2 and Next-Generation Internet Capabilities," *JAES*, Volume 47, Number 4 (1999 April), pp. 300-310.

Chafe, C., and Gurevich, M., "Network Time Delay and Ensemble Accuracy: Effects of Latency, Asymmetry," Audio Engineering Society Convention Paper Presented at the 117th Convention in San Francisco, CA, October 28-31, ➥

2004, Convention Paper 6208.

Cooperstock, J., Roston, J., and Woszczyk, W., "Broadband Networked Audio: Entering the Era of Multisensory Data Distribution," Invited Paper, the Proceedings of the 18th International Congress on Acoustics, Kyoto, April 4-9, 2004, Japan.

Cooperstock, J., Roston, J., and Woszczyk, W., "Ultra-Videoconferencing Work at McGill," SURA/ViDe 6th Annual Digital Video Workshop, Indiana University-Purdue University Indianapolis, Indianapolis, Indiana, March 22-25, 2004.

Cooperstock, J. R. and Spackman, S., "The Recording Studio that Spanned a Continent." IEEE International Conference on Web Delivering of Music (WEDELMUSIC), Florence (2001).

Corey, J., Woszczyk, W., Martin, G., and Quesnel, R., "An Integrated Multidimensional Controller of Auditory Perspective in a Multichannel Soundfield," Audio Engineering Society Convention Paper, Presented at the 111th Convention, 2001 September 21-24, New York, USA (postponed until Nov. 30-Dec. 3, 2001), Convention Paper 5417.

Corey, J., Woszczyk, W., Martin, G., and Quesnel, R., "Enhancements of Room Simulation with Dynamic Cues Related to Source Position," in *Surround Sound—Techniques, Technology and Perception*, Proceedings of the 19th International Conference of the Audio Engineering Society, Schloss Elmau, Germany, June 21-24, 2001.

de Bruijn, W., "Application of Wave Filed Synthesis in Videoconferencing," Ph.D. Dissertation, Delft University of Technology, Laboratory of Acoustical Imaging and Sound Control, October 4, 2004.

Martens, W., and Woszczyk, W., "Guidelines for Enhancing the Sense of Presence in Virtual Acoustic Environments," VSMM 2003 – Hybrid Reality: Art, Technology and the Human Factor. The Ninth International Conference on Virtual Systems and Multimedia, 15-17 October, 2003, Montreal, Montreal, Canada,. pp. 306-313.

Martens, W. L. and Woszczyk, W., "Perceived Synchrony in a Bimodal Display: Optimum Intermodal Delay Values for Coordinated Auditory and Haptic Reproduction," Proceedings of the 10th International Conference on Auditory Display, Sydney, Australia, July 7-9, 2004(a).

Martens, W. L. and Woszczyk, W., "Psychophysical Calibration of Whole-body Vibration in the Display of Impact Events in Auditory and Haptic Virtual Environments," Proceedings of the 3rd IEEE International Workshop on Haptic, Audio and Visual Environments and their Applications – HAVE 2004, pp. 69-73, Ottawa, Ontario, Canada, October 2-3, 2004(b).

Martens, W. L. and Woszczyk, W. R., "Subspace Projection of Multichannel Audio Data for Automatic Control of Motion-Platform-Based Multimedia Display Systems," ICASSP 2004, International Conference on Acoustics, Speech, and Signal Processing, Montreal, May 17-21, 2004(c).

Martens, W. L., Braasch, J., and Woszczyk, W., "Identification and Discrimination of Listener Envelopment Percepts Associated with Multiple Low-Frequency Signals in Multichannel Sound Reproduction," Audio Engineering Society Convention Paper, Presented at the 117th Convention in San Francisco, CA, USA, October 28-31, 2004, Convention Paper 6229.

Usher, J. and Woszczyk, W., "Visualizing Spatial Sound Imagery of Multichannel Audio," Audio Engineering Society Convention Paper, Presented at the 116th Convention in Berlin, Germany, May 8-11, 2004, Convention Paper 6054.

Usher, J., Cooperstock, J., and Woszczyk, W., "Multi-Filter Approach to Acoustic Echo Cancellation for Teleconferencing," 75th Anniversary 147th Meeting of the Acoustical Society of America, May 24-28, 2004.

Woszczyk, W. R., and Martens, W. L., "Intermodal Delay Required for Perceived Synchrony Between Acoustic and Structural Vibratory Events," Eleventh International Congress on Sound and Vibration, July 5-8, 2004, St. Petersburg, Russia.

Xu, A., Woszczyk, W., Settel, Z., Pennycook, B., Rowe, R., Galanter, P., Bary, J., Martin, G., Corey, J., and Cooperstock, J., "Real-Time Streaming of Multichannel Audio Data over the Internet," *JAES*, Volume 48, Number 7/8 (2000 July/August), pp. 627-641.

Upcoming Meetings

2005 May 16–20: *149th Meeting of the Acoustical Society of America,* Vancouver, British Columbia, Canada. ASA, Suite 1NO1, 2 Huntington Quadrangle, Melville, NY 11747, USA. Fax: +1 516 576 2377; Web: http:// www. asa.aip.org.

•

2005 May 28–31: *118th AES Convention*, Barcelona, Spain. See page 368 for details.

•

2005 July 7–9: *26th AES Conference, "Audio Forensics in the Digital Age,"* Denver, CO, USA. See page 368.

•

2005 July 12–14: *12th AES Regional Convention*, Tokyo, Japan, Science Museum, Chiyoda, Tokyo, Japan. See page 368 for details.

•

2005 July 18-22: *17th International Symposium on Nonlinear Acoustics*, State College, The Pennsylvania State University, University Park, PA, USA. Call 814-865-6364 or e-mail: ISNA17@outreach.psu.edu.

•

2005 September 2–4: *27th AES International Conference,* Copenhagen, Denmark, Hillerød, Copenhagen, Denmark. See page 368.

•

2005 October 7-10: *119th AES Convention*, Jacob K. Javits Convention Center, New York, NY, USA. See page 368 for details.