Wieslaw Woszczyk, Jeremy Cooperstock, John Roston, William Martens

Centre for Interdisciplinary Research in Music Media and Technology (CIRMMT), Faculty of Music, McGill University, Montreal, Québec, Canada H3A 1E3

Environment for immersive multi-sensory communication of music using broadband networks

Eine Umgebung für immersive multisensorische Musikkommunikation in Breitbandnetzwerken

Abstract

Broadband Internet enables bidirectional real-time transmission of multiple streams of audio, video, and vibro-sensory data with latency dependent on distance plus network and processing delays. In this paper we describe a new immersive multi-sensory 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 twenty-four 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 post-production for remotely-captured live events. The technical and operational challenges of this undertaking are described, as well as potential future applications.

Breitband-Internetverbindungen ermöglichen eine bidirektionale Übertragung von mehrkanaligen Audio-, Video- und vibrosensorischen Daten in Echtzeit, mit Latenzzeiten die von der Distanz sowie von Netzwerk- und signalverarbeitungsbedingten Verzögerungen abhängen. In diesem Artikel wird eine neue immersive multimodale Umgebung vorgestellt, die kürzlich an der McGill Universität entwickelt wurde, um eine netzwerkbasierte Kommunikation über große Distanzen zu ermöglichen, bei der Musikdarbietungen zwischen verschiedenen Orten koordiniert werden. Durch die Verwendung von mehreren Sensoren, Effektoren, und hochauflösenden Übertragungskanälen erlaubt die Systemarchitektur den Teilnehmern eine verstärkte Musikpräsenz zu erfahren. Bis zu 24 Audiokanäle, digitales Video und vier Vibrationskanäle können gleichzeitig über das Netzwerk gesendet und empfangen werden, was eine Reihe von verschiedenen Anwendungen wie Musikunterricht, Probespiele von Studenten, Jam-Sessions, Konzerte, Aufnahmesitzungen und Post-Produktionen von entfernt stattfindenden Live-Präsentationen ermöglicht. In dem Vortrag werden sowohl die technischen und betriebsbedingten Herausforderungen des Projektes als auch mögliche Anwendungen für die Zukunft beschrieben.

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 multi-sensory 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 an immersive multi-sensory experience laboratory. Important non-technical members of the development team include musicians, classical and jazz, 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 multi-sensory, multi-channel music with subtlety and detail afforded by high-resolution.

During the last two years, system design team has faced many technical and performance considerations in their effort to match the limitations of current technology with high musical sensitivity and expectations of the users. In this paper we review historical progress on this project and present critical issues that were addressed, as well as those that still need to be addressed. 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 communication requires sharing of expertise and techniques of music studios, teaching labs, television, live concert, telephony to achieve fruitful synergy between technologists and artists who challenge each other to raise the bar of networked quality higher and higher. Challenges of quality networked communication were discussed in the AES Technical Council White Paper (AES White Paper, 1999; Bargar et al, 1999) and addressed by the Technical Committee on Networked Audio Systems, and others (Xu et al, 2000).

Telepresence. Shared reality. Immersive reality.

High-fidelity networked communication brings forward the concept of 'telepresence'. The goal of 'telepresence' is to connect physically separated spaces using multiple sensory links so that they appear to participants as constituting a single larger space integrating the two remote spaces (Figure 1). A sophisticated multimodal display system should recreate in each space a comprehensive "portal" (view) into the remote site as though by building a virtual multimodal extension to the local site, or at least a large open window through which light, sound, and structural vibrations can pass (De Bruin, 2004).

The virtual representation of a distant space and the activity within it must be made as close as possible to what any participant would expect in physically connected spaces, but this will often not be enough. A musician practicing with a colleague in the same room relies on temporal cues, subtle tonal and spatial details of the sound, with occasional quick look at body movement, or instrument handling, bowing techniques; and musicians may even rely subconsciously on the vibration of the floor they can feel under their feet or in a chair. A musician may choose to review any of these elements in any modality at any time, and expects them always to be present and accurate in time, location, and magnitude. Only when these multimodal events become perceptually fused through sensory integration into one construct in consciousness, can we consider that the technology facilitating the communication is transparent.

At the same time, just the re-creation of perceptually veridical reality is not good enough to satisfy the requirements of advanced communication. To justify the use of such technology, one may wish to hear and see better than in reality, with greater detail or larger perspective, or simply 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 large canvas, the immersive enveloping display in each of the modalities, auditory, visual, and haptic, because the information provided for the participants must be laid out fully and openly, and must be easily accessible without constraints. To achieve this requires a wide-angle video screen extending into peripheral vision, presenting life-size objects, immersive multichannel sound with height, width and depth in high-resolution digital audio, and full-body motion with mutliple degrees of freedom.

There is also a challenge of blending two often quite different environments. The two connected yet independent spaces should appear to be present at each site. However, depending on the character of each, these spaces could be quite distinct 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 characteristics 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.

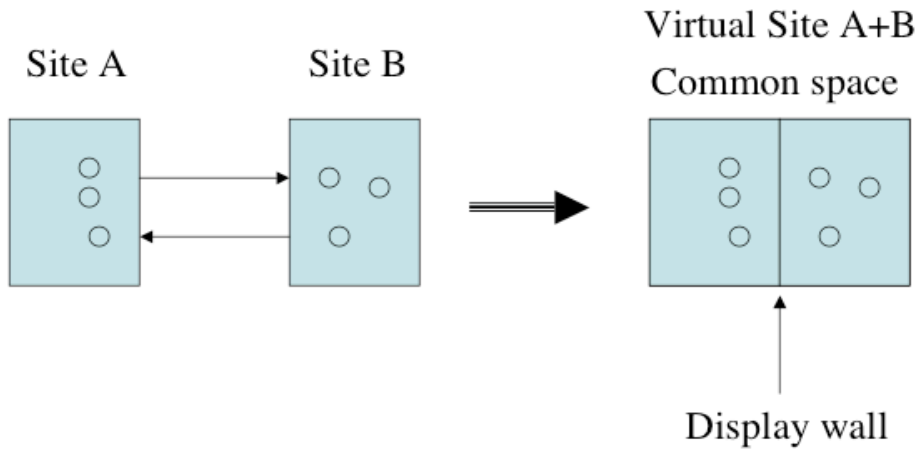

Figure 1. The concept of Telepresence. Two physically remote sites become connected electronically via a common display wall that recreates for each the presence of the other site using audio, video and haptic channels transmitted via broadband networks.

Synthesis of shared ambience

The above discussion reveals that each connected site must have 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 consonance 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 much one another.

Audio Environment

To achieve the perceptual goals of unimpeded (transparent) bi-directional 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 Figure 2) as 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 electro-dynamic ribbon transducers featuring "Planar Focus Technology", and these 96 loudspeakers are placed 4 units wide in 24 panels (two loudspeaker 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 extreme high elevation, the spatial organization of the upper-frequency drivers is defined by 3 planes at elevation angles of –15°, +25°, and +45° degrees relative to the horizontal plane. 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; rather, a more comprehensive simulation of a sound field is attempted 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 by 6 low-frequency drivers, even with 90^{o} head turning, there can always be a reproduction of low-frequency incoherence naturally associated with room acoustics for binaural listeners (Martens et al, 2004).

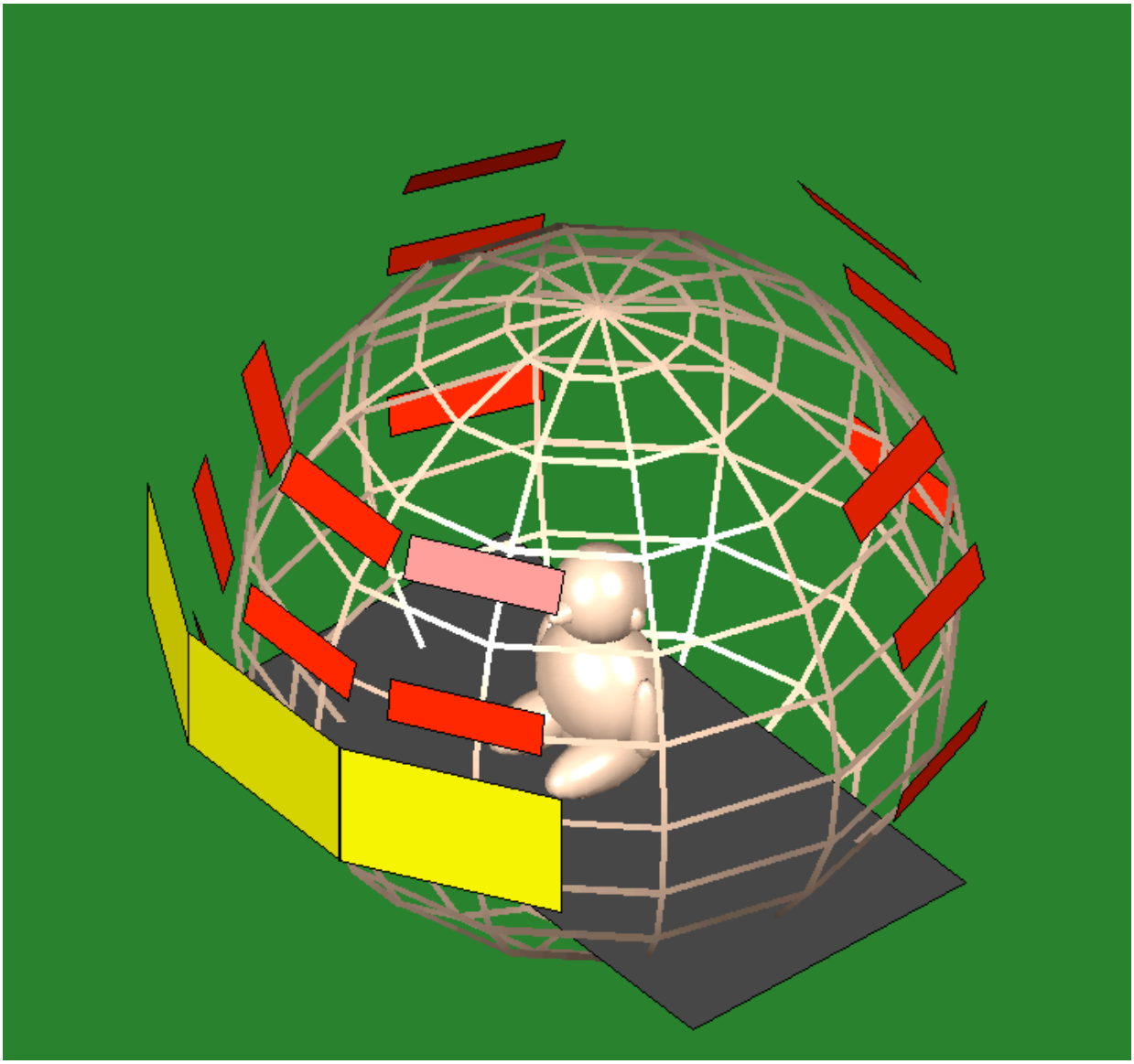

Figure 2. Graphic depicting the 3-D configuration of 24 loudspeakers (shown here as red panels, some hidden from view) on the surface of an imaginary sphere of 2 meter radius. The listener is situated on a 4' by 8' motion platform (show here in grey) facing three video screens (shown here in yellow)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

Video Environment

The goal of this research project 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.

Figure 3. Pinchas Zukerman in Ottawa giving master class to violin students in Montreal (Photo: Owen Egan)

In the initial experiments in distance music teaching, broadcast standard digital video (SDI at 270 Mbps) was used and dispalyed on a 50“ (127cm) diagonal measurement plasma display. The project is now moving to broadcast standard high definition video (HD-SDI at 1.5 Gbps) on a 65“ (165cm) 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 and minimize video processing by eliminating the need for de-interlacing which would be necessary if the competing 1080i60 standard were used. Some recent model plasma displays have direct SDI / HD-SDI inputs which again minimizes video processing and reduces latency.

High resolution video enables the viewer to be in much closer proximity to the display (5’ or 1.5m) without seeing video scan lines. This more closely approximates the distance customary for one-on-one human interaction. Since there is a camera above the display sending an image of the viewer to the far end, moving the viewer any closer to the display than this makes it obvious that the viewer is looking below the camera, not directly into it. This is disconcerting at the far end since the lack of eye contact detracts from effective human interaction. Live-size plus live-distance image of remote participants contributes to the increased sense of presence with them in a shared virtual space.

Figure 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.

Whole-body haptic stimulation

Whole-body haptic stimulation from motion and vibrational signals (captured from live performance using accelerometers or synthesized via automatic analyses (Martens, 2004; Woszczyk, 2004) are presented via a commercially available motion platform (the Odyssée™ system from the Quebec based company D-BOX Technology (http://www.d-box.com) that uses four coordinated actuators to vertically displace a wooden platform of 4 by 8 feet on which the users' chairs are fixed. When all four actuators move together, users can be displaced linearly upwards of downwards, with a very quick response and with considerable force (the feedback-corrected linear system frequency response is flat from DC to 50 Hz). The controller also enables two types of angular motion of the platform, so the potential mismatch between audiovisual stimulation and bodily motion can be corrected. When simulation includes touch (haptic) sensations and motion (vestibular) sensations that are consistent with what is seen and heard, a heightened "sense of presence" is to be expected. We have proposed that a comprehensive model of human spatial hearing cannot be formulated without the inclusion of human perception of self-motion, both in terms of angular and linear acceleration of observers within their immediate environment (Martens, 2004).

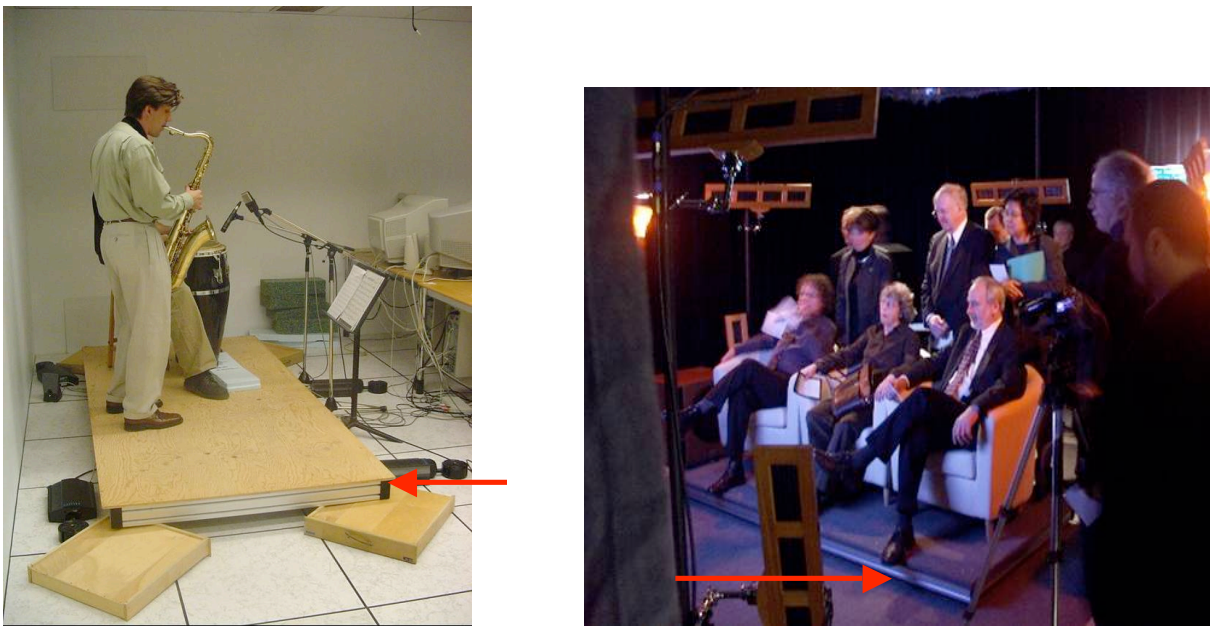

Figure 5. Vibrational signal transmission using motion platforms. The arrows point to the platforms. The audience can feel the vibration generated by musicians in the studio.

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 non-commercial 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 United Kingdom in partnership with British Telecom uses it for a project entitled *BT Music On-line* 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 and it will be developed further to support high definition video. The existing software only supports unicast transmission of standard definition video, high-resolution audio, and haptic data between two sites. It will be expanded to support multicast transmission of standard definition video (SDI) and audio among multiple sites, a capability frequently requested by current users.

Methods of sound capture for networked transmission

In 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 as it carries loudspeaker signals used to monitor the remote venue and the common ambience enhancement. Especially when these signals are delayed in transmission, they can cause disturbing echoes that degrade intelligibility and impede communication.

To avoid the echoes, the solution is to use 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 ambience canceling differential microphones could 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, the ambiance information must be synthesized locally either from room simulators or reverberation generators to create the sense of shared acoustic space that is different from the local room as it contains the effect of the other space being added to the local space. Algorithmic synthesis or sampling convolution room ambience can be used, or both, each having 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 Figure 6). Because acoustic feedback is delayed due to latency in transmission and processing, it becomes highly audible when longer delay times are 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 still not ready for prime-time just yet.

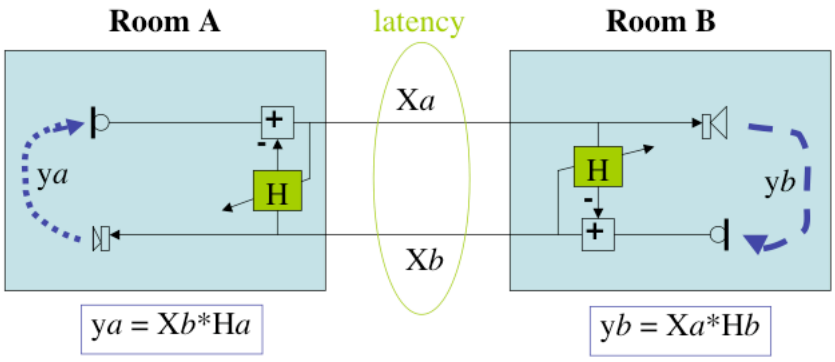

Figure 6. Single-channel acoustic echo cancellation. Filter H models the crosstalk transfer function y and tries to remove it from the signal X at each location (a and b).

Adaptive filter estimates the contribution of the far-site signal to the microphone signal and subtracts it from the microphone signal. However, filtering takes time and causes additional delay due to block processing. Also, adaptive process is time-varying and creates artifacts that are highly audible because filters require finite ramp up time to adapt and reach the most effective cancellation setting. Since adaptive filtering affects primary signal path, it must be totally distortion and artifacts free. Unfortunately, adaptive filtering can delay and distort the microphone signal making it unsuitable for recording, archiving, transmission, and for low-latency interactive applications.

Larger quantity of 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 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.

Latency

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 degredation in rhythmic accuracy.

Because latency kills 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. 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 of 1000 miles or so.

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 ultra-videoconferencing technology developed at McGill University.

Medicine

McGill uses this new immersive ultra-videoconferencing environment for teaching and remote communication applications where high quality is necessary such as in Medicine (Figure 7) and Music. The team involves collaboration of the Instructional Communications Centre which is responsible for the development of the video components, the Centre for Intelligent Machines of the Faculty of Engineering 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 (CIRMMT) which is responsible for development of high resolution multichannel audio capture and projection.

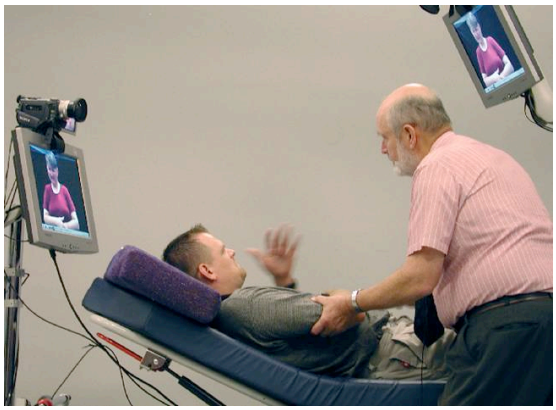

Figure 7. Sign language interpreter in a remote location assists in physician's communication with a deaf patient. Video cameras and monitors in the hospital are connected by network to a remote interpreter.

Music Master Classes

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

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 connectivity from the National Research Council in Ottawa to the ICC studio at McGill in Montreal. (Figure 8). The technology was developed using Canarie ANAST (Advanced Networks Applications, Services and Technologies Program) grant “McGill Advanced Learnware Network”.

Jam Sessions

World's First Low-Latency Videoconferencing System Enables First Cross-continent Jam Session

Figure 9. Musicians at McGill jam with Colleagues at Stanford (on screen). (Photo: Peter Marshall)

On June 13, 2002, musicians at McGill University jammed together with musicians at Stanford University in California, using McGill's low-latency ultra-videoconferencing system and next generation research & education networks: CA*net 3 and Internet 2. The event featured full-screen bi-directional video and multi-channel audio, in what was heralded as the first demonstration of its kind over IP networks. See Figure 9.

This project forms a component of the McGill Advanced Learnware Network project, funded by CANARIE Inc. and Cisco Systems as well as the Real-time Communication of High-Resolution Multi-sensory Content via Broadband Networks funded by Valorisation-Recherche Québec.

Remote concerts

World's First Studio that Spans a Continent

The studio was demonstrated on Saturday, September 23, 2000, as part of the 109th Audio Engineering Society Convention in Los Angeles. McGill Jazz Orchestra performed in a concert hall at McGill University in Montreal and 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 theatre at the University of Southern California in Los Angeles, mixing for a live audience. This is the first time that live audio of this quality has been streamed 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 theatre, the six signals were converted to analog by 96/24 D/A converters before being sent to the theatre's 6.1 monitoring system. See Figure 10.

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

First multichannel transcontinental transmission of Direct Stream Digital (DSD) audio with SDI video

First transcontinental networked transmission of Direct Stream Digital audio (1-bit, 64Fs, Fs=44.1kHz) provided a qudraphonic concert of Haydn Quartet and McGill Jazz Orchestra from McGill's historic Redpath Hall in Montreal to an AES audience at Genentech Hall of University of California in San Francisco. On October 31, 2004, two audio stream at 5.6Mbps each and one video SDI stream at 170Mbps were transmitted using CA*net4 and Internet2 networks as part of special event prepared by the Technical Council and Technical Committee on Networked Audio Systems at the 117th AES Convention. See Figure 11.

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

Multimodal interactions: video-audio, vibro-audio

Tolerance is typically fairly wide for asynchrony between video and audio events, 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, which always require greater processing and transmission times. In our experience, it is not always necessary to match the latency of video signal by delaying the audio. In the jam sessions between Stanford and McGill we chose to run audio with minimum latency (approx. 50ms) and video with its minimum latency (approx. 80ms) and the musicians played mostly guided by audio while using video occasionally to read gestures and emotions of partner musicians, as if looking at the conductor.

For natural seeming experiences, we have found that vibro-audio asynchrony is much more critical (Martens and Woszczyk, 2004). Temporal coincidence of impact events in haptic and auditory modalities increases the chance for sensory integration effects and promotes perceptual fusion of remotely captured phenomena. Through psychophysical testing, synchronization of these multimodal components has been optimized in our immersive environment display system to improve perception of music, and in particular the sense of rhythm and timing

provided by music transmitted from a remote site. By adjusting intermodal delay values, enhanced perceptual results can be crafted, such as an increased sense of how powerful such bimodally-displayed impact events are (Woszczyk and Martens, 2004). 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, complex monitoring and mixing requirements that must be adequately resolved for each modality. Perceptual requirements of musicians need to be addressed and these include cross-modal interactions. 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, auditions, etc. **In teaching**, master classes, distance music teaching, rehearsals, jam-sessions can be facilitated as has been shown earlier. **In film and sound design**, team-composing and sound design, music scoring, music mix approval can be done remotely whenever composer, producer, orchestra, are distances apart.

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 state as direct microphone signals and as 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, Mixing one must not use any echo-cancellation, and use only uncompressed high-resolution digital signals. For **medium quality** applications including Rehearsing, Jamming, Auditioning, one may use some echo suppression but not echo cancellation, and use high-resolution digital audio having multichannel capabilities. **Low quality** audio may use echo cancellation as long as voice and not music are being processed. Here the applications are similar to telephony: Dialog, Discussion, Commentary were echo cancellation is a must to ensure 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 the visual location of the source. 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, stimulated room modes, etc. A system capable of such complex dynamic synthesis of spatial movement has been realized at McGill University and showed excellent subjective results although only one source at a time could be dynamically positioned using a simple controller due to a large demand for processing power (Corey et al., 2001, 2001). A new controller of multi-source perspective allowing additional simulation of source orientation is being developed for the 24-channel system described earlier including distance, direction, and elevation cues.

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 in order 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

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 bi-directional 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 be obstructing the efforts of communication. Two audio mixes will likely be needed at each site, microphone mix and loudspeaker mix, and for this to be achieved two audio monitoring spaces are required. Loudspeaker mix can be done in the performance space where musicians-listeners 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 distant site. Since tight synchronization and the

co-location of audio and video displays enhances the localization of sources and improves 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 best. 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. 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 become more common in the future allowing their regular use for music teaching and concert going, or recording. Fixed permanent installations are more likely because of the complexity involved in creating a compelling telepresence, however intelligent home environments will gradually develop with network bandwidth to homes steadily increasing. Will we have musicians making house calls, famous teachers giving music lessons in the neighborhood, or coaching teachers via networked telepresence? Most likely. Such face-to-face personal interactions are not possible today by using telephone, television, or film. The positive reactions of our users to telepresence are encouraging us to say yes.

Acknowledgments

The authors would like to acknowledge the generous funding support from Valorisation-Recherche Québec of the Government of Québec, and CANARIE Inc. ANAST (Advanced Networks Applications, Services and Technologies Program) grant allowing us to carry out this research. Alain Berry and Dan Levitin are thanked for their partership in various aspects of the research, and Stephen Spackman and Jonas Braasch are thanked for their participation as Research Associates. The authors also thank our musicians, system users and collaborators, most notably Chris Chafe, as well as student-engineers from the Graduate Program in Sound Recording at McGill University 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, Woszczyk, W., "Networking Audio and Music Using Internet2 and Next-Generation Internet Capabilities," Journal of the Audio Engineering Society, Volume 47, Number 4, 1999 April, pp.300-310.

Chafe, C., & 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. (2001) The Recording Studio that Spanned a Continent. IEEE International Conference on Web Delivering of Music (WEDELMUSIC), Florence.

Corey, J., Woszczyk, W., Martin, G., 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)

Corey, J., Woszczyk, W., Martin, G., 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 Bruin, Werner., „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. & 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).

Martens, W.L. & 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. Ottawa, Ontario, Canada. October 2-3, 2004. pp. 69-73. (2004).

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.

Martens, W.L., Braasch, J., & 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, October 28-31, 2004, Convention Paper 6208.

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, pp. 1-16.

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., Cooperstock, J., "Real-Time Streaming of Multichannel Audio Data over the Internet," Journal of the Audio Engineering Society, vol.48, No. 7/8, July/August 2000, p.627-639.