Broadband Networked Audio: Entering the Era of Multisensory Data Distribution

Jeremy R. Cooperstock, John Roston, Wieslaw Woszczyk

Centre for Interdisciplinary Research in Music Media and Technology
McGill University
Montreal, Canada
jer@cim.mcgill.ca

Abstract

As high bandwidth networks have become widespread, Internet audio, once relegated to the second class status of AM radio, is now recognized as an effective means of audio distribution for both pre-recorded and live transmissions, including support for distributed, real-time interaction. Recent advances, including transmission of surround sound (e.g. Dolby Digital AC-3), uncompressed multichannel 96kHz/24 bit streaming, and low-latency networked audio, have each inspired a range of novel technologies and applications, ushering in a new era of audio in which distance no longer dictates limitations for high quality musical experiences and interaction.

In the non-networked world, audio tends to be experienced in conjunction with other modalities, and thus, efforts to advance the state of the art must not neglect the challenges of integrating high-fidelity audio with rich video and vibro-sensory data. In this context, this paper explores some of the issues involved in developing a scalable architecture and suitable protocols for broadband multicast network distribution of multimodal data. These issues include coding for heterogeneous clients, both with respect to their capabilities and needs, balancing the conflicting demands of optimal immersive quality, minimal latency and maximal reliability, and suppressing echo in such a manner that interaction is improved without impairing perception of the original signal.

As music is one of the most demanding of applications, the research described here relates to our goal of supporting the activities of both performer and audience, or student and teacher, in the process of creating and experiencing music. Achieving success in these domains requires contributions ranging from psychology, signal processing, to network engineering.

1. Introduction

Despite the claims of various marketing departments, anyone familiar with the technology will readily admit that the experience of videoconferencing falls short of physical presence. Participants behave differently in videoconferencing situations from how they do when together in person and typically prefer the latter. In our own experiences with the development of Ultra-Videoconferencing, we have identified three largely unexplored axes that are major determinants for such preferences and of user perception of quality and immersion: multimodality, spatial extent, and latency.

In its evolution over traditional voice-only communication, videoconferencing technology has tended to emphasize the role of video. Unfortunately, this focus has often come at the expense of other modalities, notably audio. Recent research demonstrates the tremendous impact that one modality may have on others, for example, how high quality audio can influence one's perception of video. [1] Despite the allocation of videoconferencing bandwidth two or more orders of magnitude greater than that available to conventional telephony, high resolution, multichannel audio, long since understood by the audio engineer as a rich and powerful means of conveying *presence*, seems to have been largely neglected, even in supposed "high-end" videoconferencing systems and facilities. Similarly, vibro-sensory data such as floor vibration and low-frequency subsonic effects, remains virtually untapped in distance communication. This modality is arguably of greater importance in entertainment or distributed musical applications than for conventional videoconferencing, but we expect to see it play an increasingly important role in immersive environments.

Spatial extent applies both to video and audio, the former in terms of image size and the latter in terms of multichannel capture and playback. The display of one or more videoconference participants on a small video monitor simply cannot convey the subtle visual cues of gaze awareness, facial tension, and other gestures that we take for granted as a part of human communication. Furthermore, as we rely, to a certain extent, on object size to gage distance, a less-than-life-size display of a remote participant immediately violates the intended illusion of *virtual* proximity. With respect to audio, we subconsciously exploit binaural audio cues to localize sound sources (i.e. a human speaker) and discriminate between multiple simultaneous conversations (the "cocktail party effect"). When sound capture and reproduction in a videoconference are stereophonic at best, we are deprived of these important communication cues.

Finally, delay, previously an annoyance only for pre-fibre-optic transatlantic telephony, has returned with a vengeance to videoconferencing. While various telephony standards have long held that end-to-end latency should be less than 150ms and that latencies in excess of 200ms are unacceptable for human communications, few videoconferencing systems in use today satisfy this requirement. The effects of latency are dramatic: normal interaction is inhibited and conversation by necessity degenerates to a formal turn-taking interaction.

While it is beyond the scope of this paper to explore these issues in great detail, we endeavour to provide a general overview of some key research areas in which the problems identified above can be addressed as we head into the era of broadband multisensory data distribution.

2. Network Transport Architecture

Assuming input and output devices of reasonable quality across the various sensory modalities of interest, one of the primary determinants in the experience of high-fidelity interaction rests on the network transport layer. Conventional IP transports, TCP and UDP, are both known to be inadequate for real-time interaction [2]; the former due to its lack of timeliness guarantees and the latter to its inherent unreliability. The real-time protocol (RTP) [3][4] and the associated real-time streaming protocol (RTSP) [5], were developed in the late 1990s to meet the needs of applications that operate on continuous media. While these now find widespread use in popular streaming media applications (e.g. RealPlayer, Windows Media, QuickTime), they do not provide the flexibility to support the specific requirements of different media types in isolation. For example, although buffering of audio data to ensure lossless delivery is generally desirable, this is not necessarily the case for video, where the loss of a single frame may be preferable to additional delay.

To this end, we have invested considerable attention in developing *bronto*, a transport that can operate in a variety of modes as required by the individual application, ranging from minimum-latency best-effort to low-latency semi-reliable to unbounded fully reliable [6]. This transport, currently at the heart of our Ultra-videoconferencing software, has facilitated the performance of a real-time violin duet, a cross-continental jazz jam, remote sign-language interpreting, and distance Masters classes with Maestro Pinchas Zuckerman.

A key differentiating factor of Ultra-videoconferencing is that unlike most conventional videoconferencing systems, which employ frequency transforms and discretization algorithms to reduce bandwidth requirements, our system transmits uncompressed data, i.e. PCM audio and SDI or raw analog video frames.[1] The motivation for uncompressed data transmission is twofold: first, to ensure the maximum possible quality of the reproduced signal, and second, to avoid the encoding cost and thereby minimize end-to-end latency, critical for the demands of real-time interaction.

3. Data Coding Issues

To date, our public demonstrations of high-bandwidth data communication have been confined to point-to-point transfers in which zero loss of audio data and only minimal loss of video data could be tolerated, as the entire content of both audio and video streams was reproduced with uniform quality. This approach follows from our conditioning to television broadcasts, in which all clients receive exactly the same audio and video, subject to the quality of receiver and reproduction equipment, and indeed, this typifies the requirements of any high-quality videoconferencing system. As such, we were constrained to carry out our experiments on carefully controlled research networks where we were generally the only serious users of the available bandwidth.

3.1. Region of interest coding

Transmission of the *entire* data content does not take into account the potentially diverse interests or capabilities of heterogeneous clients nor the relative importance of different components of the scene. Attempts to date to address the former concern appear limited to the sole factor of varying bandwidth capability, for example, the scalable bitstreams of H.26x and MPEG audio [7] and video codecs [8]. A base layer provides a minimal quality representation of the entire signal, and clients may optionally receive supplemental layers that provide added quality, uniformly distributed over the image. This seems to be a reasonable starting point, but it offers the clients only passive control over the quality of the reconstructed signal without any ability to specify regions or content of greater interest or importance.

This point has been addressed in part by the content description scheme of MPEG-7 [9], but this is aimed primarily at *describing* content, rather than permitting control over the quality at which different portions of the scene are transmitted or rendered. A more apt example of client-based control over the allocation of data to such content is the User-Centered Video of Yammaashi et al [10]. In Yamaashi's case, the system allowed the client to allocate bandwidth as desired over multiple parallel streams or within a particular *region* of a single stream, as appropriate to the interests of the client. Assuming

[1] Ultra-videoconferencing also supports JPEG and DV codecs, as well as a pseudo-run-length-encoding compression, in order to satisfy the requirements of full-frame video over 100 Mbps links. The need for (possibly compressed) data encoding is discussed in further detail in the next section.

operation on a multicast network, the challenge here is to ensure that individual client *requests* are balanced against overall system constraints, such as total available server bandwidth and limit of multicast channels. Our long-term goal is for such region selection to be automated with the assistance of intelligent agents, possibly given some *hints* from the user, for example, "I'm interested in this person's face" or "follow that object."

4. Quality vs. Latency

An important consideration in networked interaction is the tradeoff between quality and latency. At one extreme, sophisticated compression algorithms can transform a complex stream of audio or video into a (perceptually) nearly identical equivalent, requiring a small fraction of the original data size. However, such processing entails a cost, either due to the computational complexity or the algorithm's requirement of a non-trivial buffer for analysis. While this is not an issue for asynchronous playback operations, latency requirements become critical when considering interactive applications, particularly when there is little tolerance for synchronization errors, such as that for distributed musical performance or telesurgery. In such demanding cases, it may be necessary to forgo the tremendous bandwidth savings of data compression in favour of the reduced latency benefits of uncompressed data. Naturally, this approach is not for most users: for example, a reasonable quality MPEG-4 audiovisual stream can be obtained at 1 Mbps, whereas the uncompressed serial digital interface (SDI) equivalent requires approximately 270 Mbps.

In addition to compression algorithms, other sources of latency, easily ignored, include acquisition and display technology, as well as the interface hardware between these and the computer. As but one example, we were surprised to discover that our SDI interface cards were adding two video frames of delay[2] as was the video processing circuitry of our plasma displays.[3]

Two related matters are those of lost data recovery (retransmission policy) and error concealment. It is worth noting that appropriate policies are highly dependent on the modality and the application. For interactive applications, it is typically more important to be viewing the most recently available video data, even if this entails discarding one or more previous frames that have not yet arrived successfully. However, the unpleasant artifact resulting from a similar *gap* in an audio stream often motivates more aggressive recovery attempts, and hence, the use of a buffer for audio data. In the event that the necessary audio data has not arrived before its scheduled playback, a client has two options: either it may defer playback further, thereby increasing observed latency, or it may proceed with playback, ideally adding some intelligent signal shaping, such as detecting and extending a period of silence in the preceding samples.

4.1. Lossy coding

As alluded to above, it is unrealistic to assume that unbounded network resources are available to the majority of potential users. But more importantly, the reliance on an essentially lossless transmission mechanism is unnecessary for successful reproduction at the destination, as is evident by the perceptual indistinguishability of the results of certain *lossy* data compression techniques in both the auditory and visual domains. Furthermore, at least for faithful video reproduction, it is clear that different portions of a scene inherently require a greater encoding resources than others, for example, a moving foreground object versus a static background.

This observation forms the motivation for all perceptual codecs (e.g. MP3, AC-3, AAC for audio, JPEG for image, and MPEG for video data). These employ a potentially lossy encoding algorithm to reduce bandwidth requirements, in which data bits are allocated to components of the signal in proportion to their significance, or, in effect, to their need of resources. While not the case for most perceptual codecs, in the ideal, the encoding also permits lossless reconstruction, assuming all necessary encoded data is received, at no greater a cost than transmission of the original, unencoded data.

Furthermore, for low-latency networked interaction, a client may be forced to truncate the incoming data stream prematurely, either due to time or bandwidth constraints, in order to reproduce the next audio segment or video frame within allowable time constraints. This suggests that any encoding for real-time applications satisfy the *embedded coding* property, which requires the data to be transmitted in decreasing order of significance rather than spatial position. Truncation of such a data stream at an arbitrary point permits the reconstruction of the entire source, although possibly with degraded quality.

Following this principle, we have developed a hierarchical data representation, based on wavelet coding, suitable for the encoding of spatial data and are integrating this work into the *bronto* transport. Although our experiments have so far been limited to image data, we find that this new representation is amenable to a computationally efficient encoding and decoding process and provides perceptually superior results to competitive approaches (e.g. SPIHT [11]).

[2] This delay resulted from a decision by the hardware manufacturer to provide access to the video data through a double buffer, as required to ensure clean transitions between multiple video sources. These interfaces are often designed to be used with computer-based video editing systems, where latency is less of an issue.

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

5. Echo-suppression

The use of a low-latency transport and avoidance of time-consuming data compression help reduce the problematic aspect of acoustic echo. However, router delay and the physical limit of the speed of light are nonetheless unavoidable. Thus, for cross-continental or cross-oceanic distances, audio signal processing for echo-suppression remains an important consideration. Most videoconference systems tend to employ fairly naive measures in this regard, such as speakerphone-inspired half-duplex transmission, although high-end systems often include a hardware "echo-cancellation" component.

The principle employed by such hardware is that knowledge of local room acoustics (i.e. the room's transfer function) and the far-end audio being delivered to the local speakers allows one to model, typically by convolution, the expected input at the microphones due to the far-end signal. In theory, this can be subtracted from the actual input at the microphones to isolate the near-end audio source. In practice, room dynamics are constantly changing due to the presence and movement of individuals, different frequency components exhibit dramatically dissimilar responses, and varying background noise complicates the tuning of adaptive filters. As a result, echo-cancellation hardware requires careful "tuning" and generally operates well only in the frequency range of human voice. Various attempts to employ such hardware for musical applications have resulted in highly disappointing results; careful sound engineering, involving judicious placement of speakers and acoustic baffling and the use of near field microphones, while certainly difficult and time consuming, is often preferable.

6. Conclusions

Multisensory data transmission over broadband networks promises to revolutionize distributed human interaction. However, we must be cognizant of several factors, often overlooked in conventional videoconferencing, in developing and deploying these systems, if they are to be acceptable to the user community. Notably, due attention must be paid to modalities other than video, such as high-quality audio and vibro-sensory content, the physical extent, both in terms of image size and audio spatialization, and end-to-end latency of the system. By ensuring that these factors are addressed in current research, there exists a very real potential to realize this revolution, especially as 'the underlying technology becomes increasingly affordable and "broadband connectivity" comes to mean bandwidths on the order of gigabits.

7. Acknowledgments

The authors wish to thank Canarie Inc. and Valorisation Recherche Québec, whose generous funding has supported the research described in this overview, as well as our numerous research partners who have offered valuable feedback throughout the development and evolution of our systems. The Ultra-Videoconferencing software, developed by Stephen Spackman, may be downloaded from `http://ultravideo.mcgill.edu`.

8. References

[1] Storms, R.L., "Auditory-Visual Cross-Modal Perception Phenomena," Ph.D. Dissertation, Naval Postgraduate School, September 1998.

[2] Xu, A., Woszczyk, W., Settel, Z., Pennycook, B., Rowe, R., Galanter, P., Bary, J., Martin, G., Corey, J., and Cooperstock, J.R., "Real-Time Streaming of Multichannel Audio Data over Internet," *Audio Engineering Society*, 48(7/8), 627–641, 2000.

[3] Schulzrinne, H., Casner, S., Frederick, R., and Jacobson, V. RTP: A Transport Protocol for Real-Time Applications, Network Working Group, Audio-Video Transport Working Group, Request for Comments (RFC) 1889, January 1996.

[4] Schulzrinne, H., RTP Profile for Audio and Video Conferences with Minimal Control, Network Working Group, Audio-Video Transport Working Group, Request for Comments (RFC) 1890, January 1996.

[5] Schulzrinne, H., Rao, A., and Lanphier, R., Real Time Streaming Protocol (RTSP), Network Working Group, RFC 2326, April 1998.

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

[7] Grill, B., "A Bit Rate Scalable Perceptual Coder for MPEG-4 Audio," Preprint #4620, 103, *AES Convention*, New York, 1997.

[8] Johanson, M., "Scalable Video Conferencing using Subband Transform Coding and Layered Multicast Transmission," *ICSPAT'99*, Orlando, FL, 1999.

[9] Manjunath, B.S., Salembier, P., and Sikora, T., Eds., Introduction to MPEG7: Multimedia Content Description Interface, John Wiley & Sons Ltd., 2002.

[10] Yamaashi, K., Kawamata, Y., Tani, M., and Matsumoto, H. "User-Centered Video: Transmitting Video Images Based on the User's Interest," Proc. *SIGCHI Conference on Human Factors in Computing Systems*, Denver, 325–330, 1995.

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