Multicast Audio: The Next Generation

Colin Perkins <c.perkins@cs.ucl.ac.uk>
Vicky Hardman <v.hardman@cs.ucl.ac.uk>
Isidor Kouvelas <i.kouvelas@cs.ucl.ac.uk>
Martina Angela Sasse <a.sasse@cs.ucl.ac.uk>
University College London
United Kingdom

Abstract

The paper summarizes recent technical developments in second-generation multicast audio tools (packet loss protection and repair, improved scheduling, and cross-media synchronization) and the resulting improvements in terms of performance and usability. More sophisticated applications of Internet multimedia conferencing will, however, require further improvements in both reliability and quality. We outline a number of technical solutions (improved redundancy and repair, higher quality audio, heterogeneous distribution, and improved acoustic effects) required to achieve these improvements, which should therefore be incorporated into the next generation of multicast audio tools.

Contents

- Introduction
- The state of the art
- Limitations of current audio tools
- Future developments
  - Improved control of redundancy
  - Improved waveform substitution/packet repetition
  - High-quality audio
  - Improved acoustic feedback
  - Sound localization
- Conclusions
- Acknowledgments
- References

Introduction

In the past few years, multimedia conferencing systems have become increasingly popular: from the expensive video-conferencing suites used by executives for special occasions, the technology has moved onto user desktops for regular use. With the growth in bandwidth and improvement in compression techniques, conferencing over the Internet is emerging as an inexpensive alternative to ISDN-based solutions. With the introduction of IP multicast technology, it is possible to conduct ad hoc conferences among large numbers of participants at comparatively low cost. As a result of this increased experience, a number of problems that were apparent in older versions of multicast tools have been solved. At the same time, a number of limitations have emerged that reduce the applicability of multicast tools.

In this paper, we provide a brief overview of the state of the art in multicast audio-conferencing, together with a discussion of the recent work undertaken at UCL leading to the release of our second-generation audio tool, RAT. This is followed by a discussion of the limitations of these tools, as applied to a number of demanding application areas. Finally, we highlight those areas where we believe further development work is required.

The state of the art

In the past few years, many multicast audio-conferencing tools have become available. These tools typically provide an 8-kHz audio sampling rate, with the option of using several compression algorithms, to transmit data at bit rates of between 66.8 kilobits/second (PCM u-law) and 9.8 kilobits/second (LPC), including IP/UDP/RTP headers, although silence suppression may reduce this further. The choice of audio compression algorithm affects the sound quality, but many algorithms are available that provide quality comparable to that of the telephone system, provided the underlying network is not congested.

Audio quality is impaired by packet loss on the network, caused by congestion, and by the lack of real-time support in general-purpose operating systems. Acoustic aspects of packet network audio systems also produce problems, because one channel of restricted bandwidth is not the natural means of human audio communication. A number of enhancements that attempt to solve these limitations have been made to current audio tools, compared to the first-generation tools. These enhancements are discussed in turn:

Packet Loss Protection
: The most noticeable problem with multicast audio-conferencing tools is that of packet loss, caused by congestion in the networks. Hardman et al. [1] suggested that low-bit-rate redundancy could be used to improve the quality of multicast audio in the face of packet loss. This proposal was first implemented in a multicast audio tool called RAT (Robust-Audio Tool), developed at UCL [2] and has been shown to significantly improve perceived audio quality. It has since been incorporated into both the UCL/Exeter ReLaTe and INRIA FreePhone systems and is currently the subject of an IETF standardization effort [7].

Packet Repetition
: In addition to redundancy, the use of packet repetition and waveform substitution can also improve perceived audio quality in the presence of packet loss. Packet repetition is implemented in RAT to patch single-packet loss (if redundancy is not used or if the redundant copy of a packet is also lost) and relies on the short-term self-similarity of speech waveforms to provide improved audio quality compared to silence substitution.

Adaptation to Workstation Scheduling Jitter
: The problem of network timing jitter and the subsequent requirement for an adaptive playout buffer to be included in conferencing tools has been known for many years [10]. Experience has shown that most modern operating systems provide poor support for real-time processes, and this manifests itself as disrupted audio when a workstation is under heavy load. Recent work [5] has illustrated a technique whereby a workstation's audio device driver buffers can be used to adapt to scheduling jitter in a manner analogous to that required for adapting to network jitter, resulting in improved performance under conditions of heavy load.

Cross-Media Synchronization
: The use of the IETF standard Real-Time Transport Protocol (RTP) [8] allows for the association of multiple media streams and for the mapping of media clocks between streams. This allows for lip synchronization between audio and video streams, and has been demonstrated at UCL [4] as part of the ReLaTe project, using RAT and a specially modified version of the vic video tool [9].

The techniques described above allow second-generation audio-conferencing tools, such as RAT, to perform significantly better than first-generation multicast audio tools. Experience within the MICE [21] and ReLaTe [3] projects has shown that with the application of these techniques, it is possible to conduct conferences over low-quality network links (it has been demonstrated that audio is intelligible at up to 30-percent packet loss) and on heavily loaded workstations (for example, when decoding multiple video streams). In addition, the use of cross-media synchronization provides a first step toward more sophisticated, integrated real-time conferencing applications on the Internet. It is, however, clear that the quality of even these second-generation tools is insufficient for many applications, and there are a number of areas in which these tools may be further enhanced This is discussed further in the remainder of this paper.

Limitations of current audio tools

The enhancements described above are indicative of the progress that can be made in the development of multimedia conferencing applications that operate over "best-effort" packet networks, such as the Internet Mbone. The study of network characteristics, exploitation of knowledge about human perception, and application of improved computational techniques can lead to considerable improvement in existing tools and the perceived quality of audio.

Although these second-generation audio tools are well suited to many application areas, it is clear that a number of applications exist that demand still higher performance and additional features. Current work at UCL is focusing on two such applications: distance learning and the integration of multicast audio with networked virtual reality.

One of the more demanding environments for the use of multicast audio is distance learning, in particular, remote language teaching. Since 1995, the ReLaTe project [3] has studied the problems inherent in this area, fueling the development of features such as packet-loss robustness (through the application of redundancy) and cross-media synchronization. Extensive user trials have been conducted as part of this project [18,19,20], and these illustrate a number of points that, despite progress made to date, still cause considerable annoyance to users:

Manual Control
: In current audio tools, features such as redundant audio and packet repetition are manually controlled: a user can configure the audio tool to match the characteristics of a particular network. Often, however, the user does not have the necessary knowledge to perform this task well, resulting in these features being underused.

Narrow-Band and Poor-Quality Audio
: Although audio sampled at 8 kHz is sufficient for communication, such narrow-band speech does not sound natural, leading to user frustration after prolonged exposure and making language learning difficult.

Lack of Distance Cues and Acoustic Feedback
: Current multicast audio tools provide little feedback to the user regarding distance, loudness, etc., and provide only manual gain control. Experience has shown that most speakers have difficulty in judging how loud they are speaking, because the channel sounds acoustically "dead." This results in some speakers shouting, while others speak too quietly.

Monaural Sound
: In a natural environment, listeners use sound localization as an aid to identify the speaker and to focus on one sound in the presence of other sound sources. When combined with narrow bandwidth, monaural sound results in reduced speech intelligibility and restricts user ability to identify and focus on individual speakers [23].

The ability of multicast conferencing tools to scale to large numbers of participants and to cope easily with groups comprising many senders has resulted in interest from the virtual-reality and distributed interactive simulation communities. Not only is multicast data transfer well suited to distribution of world data for use in these simulations, but audio-conferencing tools have immediate applicability in this environment. The major limitation of current multicast audio tools when applied to this environment is the use of narrow-band, monaural sound. When used in an immersive virtual environment, sound localization is of fundamental importance, and current multicast audio tools do not support this.

To summarize, it is clear that future development must focus primarily on high-quality audio: wideband, stereo, sound localization, and acoustic feedback techniques are all important. Further, automatic control of features such as packet repetition, waveform substitution, and redundancy is a must. The next section of this paper will describe several impending developments in the field that attempt to solve these problems.

Future developments

In this section we highlight a number of areas where current Internet multicast audio-conferencing applications can profitably be enhanced. In some cases, these enhancements will involve extensive research (for example, efficient, scalable, multiparty control of redundancy); in other cases, integration of existing work in other fields is sufficient.

Improved control of redundancy

The use of low-bit-rate redundancy as a means of overcoming the problem of packet loss is now well understood and has been implemented in a number of audio-conferencing tools. However, the majority of those tools require the user to manually enable redundancy and select the audio compression scheme to be used. If the use of multiple redundant copies of a packet, each compressed using one of a range of possible compression algorithms, is allowed, it becomes difficult to manually select the appropriate combination of redundancy and compression schemes to give optimal quality for a given network condition. A clear improvement that may be made in this area is the provision of automatic control mechanisms, which sense and adapt to network conditions. Although a limited amount of work has been conducted in this area [11], it is clear that further work is required before a general-purpose scalable solution is found.

In addition, layered-coding algorithms provide for a greater range of transmission rates, allowing such a scheme to function more effectively. Some work has been conducted with the combined use of layered encoding and layered compression schemes, applied to video-conferencing [12], and it seems clear that some variation on this technique may also be profitably applied to audio-conferencing.

Improved waveform substitution/packet repetition

The use of waveform substitution and packet repetition schemes provides a receiver-based solution to the problem of packet loss, which complements the use of packet redundancy. The aim of such techniques is to generate, at the receiver, a replacement for a lost packet, based on the contents of the surrounding packets. This process is based on the observation that speech waveforms often exhibit a degree of short-term self-similarity, and hence generation of a replacement packet with similar spectral qualities to the lost packet is possible, provided that packets are relatively short (at above 30-ms packet duration, this technique breaks down).

Previous work [1,17] has shown that the use of simple packet-repetition schemes can result in a significant perceived improvement in audio quality at relatively low loss rates, and it is clear that simple techniques such as these have considerable potential for improving the perceived quality of audio in the presence of packet loss. In addition, a number of waveform substitution algorithms have been presented in the literature (see [15], for example), which have been shown to perform better than simple packet repetition: the incorporation of these techniques into multicast audio-conferencing tools would be a worthwhile future development.

High-quality audio

Assuming the existence of a good-quality network, existing multicast audio tools provide toll-quality audio and are typically optimized for speech compression, performing poorly with other classes of sound (music, for example). It is clear that such quality is not sufficient for certain classes of application: for example, remote language-teaching applications would benefit from high-quality audio, as would music transmissions.

The provision of high-quality audio is a two-stage process: wideband sampling (16 kHz, 32 kHz) is required, and improved audio compression schemes must be integrated into the multicast audio tools. The

first of these is a simple matter, and the RTP audio/video profile [13] provides the facilities necessary for the transport of such wideband audio.

The provision of improved audio compression schemes is more complex. A number of compression schemes for high-quality audio have been defined--for example, MPEG audio [14]--but these are typically optimized for broadcast media and operate on large data packets. This results in increased coder delay, when compared to toll-quality compression algorithms, and this delay limits interactivity. Further, in a lossy packet network, loss of a single large packet will have a more noticeable effect on audio quality than the loss of a smaller packet, since a larger gap will occur. The development of high-quality audio compression schemes that have low delay and are insensitive to packet loss is an area of active research.

Improved acoustic feedback

Improved acoustic feedback can be provided by reverberation, which simulates normal echoes that follow closely after the original sound and which is a result of reflections off walls. More simply, however, improved acoustic feedback can be provided by a single echo after the original sound, which is similar to side-tone [26] provided in normal telephone handsets. Reverberation or side-tone should also be combined with automatic gain control [24], to cope with impedance mismatch caused by different headsets, in order to maximize use of the input dynamic range.

Sound localization

Sound localization [25] is the artificial manipulation of a single input channel to produce 3D spatial sound. The sound can be presented over headphones or via loudspeakers, but the effect is better than stereo, since the sound appears to emanate from a particular location (outside the listener's head if using headphones). A low-cost approach to sound localization is being developed for RAT [6], and the mechanism can be used to separate conference participants out in space. This feature substantially eases speaker identification and improves speech intelligibility.

Conclusions

To summarize, it is clear that although first-generation audio tools are sufficient for many needs, there are a number of application areas that require more advanced tools. For example, in the realm of distance education and language teaching, it is clear that wideband audio and automatic hands-free control of features such as redundancy and waveform substitution provide a much enhanced learning environment. When coupled with 3D audio to distinguish sound sources, these features provide a better environment for large audio-conferences by facilitating enhanced participant identification. The use of 3D audio also fits naturally into shared visualization systems, where users manipulate objects in 3D space.

The developments we envisage have applicability in three broad areas:

Networks
: With the current use of manual configuration, users tend to select tool control parameters at the start of a session and continue using these throughout, even if network conditions are such that these are inappropriate. This may result in unnecessary network traffic for no gain. The use of automatic control of redundancy, packet repetition, and waveform substitution will allow applications to adapt faster to network conditions and reduce the amount of resource wastage. When coupled with enhancements from other fields, such as, for example, RTP header compression [22], it will be possible to conduct conferences over more heterogeneous networks, including low-bandwidth modem links.

Applications
: The ReLaTe project has demonstrated that many users find integrated user interfaces easier to use than separate systems for audio, video, and shared workspace applications. In addition, virtual-

reality and distributed interactive simulation applications require a greater degree of integration between audio and other media tools. The use of wideband, stereo, sound localization, and acoustic feedback techniques will also improve the user experience.

Usability

The role of usability assessment should not be underestimated. Extensive user trials have been conducted as part of the ReLaTe remote language teaching project and these have had a major influence on the development of the UCL RAT. The improvements suggested in this paper should go a long way toward improving the usability of multicast audio-conferencing tools, but future assessment will be required to evaluate the scale of these improvements.

We believe that the improvements we propose will provide a valuable addition to multicast audio-conferencing tools and extend the realm in which such tools are applicable into new areas.

Acknowledgments

The authors wish to thank Orion Hodson for his comments on an early draft of this paper. In addition, this work has benefited from many discussions with our colleagues at UCL: Mark Handley (now at ISI), Jon Crowcroft, Peter Kirstein, and the numerous other members of the multimedia research group. Work on the RAT and ReLaTe systems has been funded by the BT/JISC SuperJANET applications project ReLaTe, EPSRC projects RAT (GR/K72780) and MEDAL (GR/L06614), and the Commission of the European Communities Telematics for Research project (MERCI). Audio redundancy was developed in consultation with partners of project MICE (ESPRIT 7602/6), in particular Jean Bolot and colleagues at INRIA Sophia Antipolis, France.

References

[1] V. Hardman, M. A. Sasse, M. Handley & A. Watson, *Reliable Audio for Use over the Internet*, Proceedings, INET'95, Honolulu, Hawaii, June 1995.

[2] V. Hardman, M. A. Sasse & I. Kouvelas, *Successful Multi-party Audio Communication over the Internet*, To appear in Communications of the ACM.

[3] J. Buckett, I. Campbell, T. J. Watson, M. A. Sasse, V. Hardman & A. Watson, *ReLaTe: Remote Language Teaching over SuperJANET*, Proceedings of UKERNA Networkshop 1995.

[4] I. Kouvelas, V. Hardman & A. Watson, *Lip Synchronisation for use over the Internet: Analysis and Implementation*, Proceedings, IEEE GLOBECOM'96, November 1996.

[5] I. Kouvelas & V. Hardman, *Overcoming Workstation Scheduling Problems in a Real-Time Audio Tool*, USENIX Annual Technical Conference, January 1997.

[6] V. Hardman & M. Iken, *Enhanced Reality Audio in Interactive Networked Environments*, Proceedings of the Framework for Interactive Virtual Environments Conference, Pisa, Italy, December 1996.

[7] C. Perkins, I. Kouvelas, V. Hardman, M. Handley, J.-C. Bolot, A. Vega-Garcia, S. Fosse-Parisis, *RTP Payload for Redundant Audio Data*, IETF Audio/Video Transport Working Group, Work in progress, January 1997.

[8] H. Schulzrinne, S. Casner, R. Frederick & V. Jacobson, *RTP: A Transport Protocol for Real-Time Applications*, IETF Audio/Video Transport Working Group, RFC 1889, February 1996.

[9] S. McCanne & V. Jacobson, *vic: A flexible framework for packet video*, Proceedings, ACM Multimedia'95, November 1995.

[10] R. Ramjee, J. Kurose, D. Towsley & H. Schulzrinne, *Adaptive Playout Mechanisms for Packetized Audio Applications in Wide-Area Networks*, July 1994.

[11] J.-C. Bolot & A. Vega-Garcia, *Control Mechanisms for Packet Audio in the Internet*, Proceedings, IEEE INFOCOM'96, March 1996.

[12] S. McCanne, V. Jacobson & M. Vetterli, *Receiver-driven Layered Multicast*, Proceedings, ACM SIGCOMM'96, August 1996.

[13] H. Schulzrinne, *RTP Profile for Audio and Video Conferences with Minimal Control*, IETF Audio/Video Transport Working Group, RFC 1890, February 1996.

[14] ISO MPEG 1991, *Coding of Moving Pictures and Associated Audio*, ISO/MPEG-90/176 Committee Draft Standard 1172, ISO Geneva, 1991.

[15] O. J. Wasem, D. J. Goodman, C. A. Dvorak & H. G. Page, *The Effect of Waveform Substitution on the Quality of PCM Packet Communications*, IEEE Transactions on Acoustics, Speech, and Signal Processing, vol. 26, no. 3, March 1988.

[16] J. D. Johnston, *Transform Coding of Audio Signals Using Perceptual Noise Criteria*, IEEE Journal on Selected Areas in Communications, vol. 6, no. 2, February 1988.

[17] A. Watson, *Loss of audio information in multimedia video-conferencing: an investigation into methods of assessing different means of compensating for this loss*, MSc Thesis, Department of Ergonomics, University of London, 1994.

[18] J. Hughes & M. A. Sasse, *Internet Multimedia Conferencing - Results from the ReLaTe Project*, To be presented at ICDE World Conference, 2-6 June 1997, Pennsylvania State University.

[19] A. Watson & M. A. Sasse, *Assessing the Usability and Effectiveness of a Remote Language Teaching System*, Proceedings of ED-MEDIA 96 - World Conference on Educational Multimedia and Hypermedia, 17-22 June 1996, Boston, MA, pp. 685-690.

[20] A. Watson & M. A. Sasse, *Evaluating Audio and Video Quality in Low-Cost Multimedia Conferencing Systems*, Interacting with Computers, vol. 8 (3), pp. 255-275, 1996.

[21] P. T. Kirstein, M. A. Sasse & M. J. Handley, *Recent Activities in the MICE Conferencing Project*, Proceedings, INET'95, Honolulu, Hawaii, June 1995.

[22] S. Casner & V. Jacobson, *Compressing IP/UDP/RTP Headers for Low-Speed Serial Links*, IETF Audio/Video Transport Working Group, Work in progress, December 1996.

[23] E. C. Cherry, *Some experiments on the recognition of speech with one and two ears*, JASA, 25, 975-9, 1953.

[24] H. Sharifi, *An Investigation of Acoustic Enhancing Features for Use in Audio Tools*, MSc Thesis, Department of Computer Science, University of College London, 1996.

[25] F. L. Wightman & D. J. Kistler, *Headphone Simulation of Free-Field Listening. 1: Stimulus Synthesis*, Journal of the Acoustical Society of America, vol. 85, no. 2, pp. 858-867, 1989.

[26] D. L. Richards, *Telecommunication by Speech*, Butterworth & Co., UK, 1973.