Vicky Hardman,
Martina Angela Sasse,
and Isidor Kouvelas

# Successful Multiparty Audio Communication over the Internet

The Internet was once perceived as a computer network used by researchers to transfer files and send text messages. Today, more users are becoming aware of its potential as a general communication network.

**Multicast conferencing over the Internet has the potential to offer low-cost real-time multimedia solutions to a wide range of user groups, provided that sufficient audio quality can be sustained.**

Commercial interest in Internet audio has focused primarily on point-to-point applications such as Internet telephony, which provides roughly the same functionality as Public Switched Telephone Networks (PSTNs) over a computer network. The second focus of Internet audio developers has been downloading audio files—typically from a World-Wide Web server—for playout on a remote user's workstation [9]. Multicast conferencing [1], on the other hand, allows real-time multiway audio and video communication over the Internet and is now moving from the pilot stage [7] to a usable service in countries like the U.K. and the U.S.

Multicast audio allows groups of users to participate in real-time, simultaneous audio conferences, supporting communication that goes beyond the possibilities of telephony or broadcast technology. Since the multicast backbone (Mbone—an overlay over the Internet [7]) can also support video and shared workspace, collaboration environments can be tailored to support the requirements of many distributed user groups. Another important benefit, particularly for applications such as distance education, is that multicast conferencing costs a fraction of the cost of other solutions. While video and shared data are essential to many distributed tasks, audio of sufficient quality is a necessary condition for almost any successful real-time interaction. Therefore, ensuring sufficient audio quality is a major stepping stone for realizing the potential of multicast conferencing.

Audio quality is impaired by packet loss on the network and lack of real-time support in general-purpose operating systems. Acoustic aspects of packet network audio systems also produce problems, since one channel of restricted bandwidth is not the natural means of human audio communication. In this article, we describe state-of-the-art Internet audio communication and its applications, and outline impending technical developments and applications that will be available in the short- and medium-term. We describe first-generation multicast audio tools and how problems observed in their use have led to the conception and development of a Robust-Audio Tool (RAT). Subsequent user trials demonstrated a notable improvement in perceived audio quality.

The origins of speech transmission over long-haul packet networks stem from the ARPANET and SATNET experiments, which spawned packet-based multimedia conferencing research on both sides of the Atlantic. Indeed, the Mbone is used by a number of research and teaching collaborations. The existence of an international infrastructure, and the use of general-purpose computer workstations to access it, has created a critical mass of users who are piloting the technology. Thus, first-generation multimedia conferencing tools and the effect of real-time multiway conferences on the network have been studied with a range of diverse user groups:

- Researchers have attended project meetings, seminars, and conferences from their desktop workstation and conference rooms [4].
- School children in the U.S. and Europe have been able to participate in submarine excursions from their classrooms [7].
- Students have participated remotely in lectures and tutorials [11].

The Mbone currently supports multicast traffic distribution to around 3,000 sites worldwide. Senders submit packets to a group destination address, and receivers express an interest in receiving traffic destined for this group.

The network conspires to deliver the packets to interested receivers. The best-effort service model of the Internet causes problems for real-time traffic.

Many audio tools are available for most general-purpose workstations and multimedia PCs. Apart from the software, all the user needs is either a headset or a microphone and speakers. Currently available audio tools, such as Visual Audio Tool (VAT) [3], provide multiway telephone-quality speech, using the proposed Internet Engineering Task Force (IETF) standard called Real-time Transport Protocol (RTP) over the network [10].

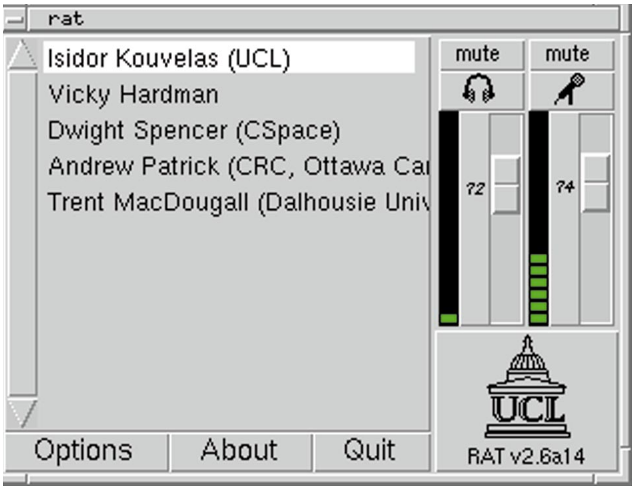
Isidor Kouvelas (UCL)
Vicky Hardman
Dwight Spencer (CSpace)
Andrew Patrick (CRC, Ottawa Ca
Trent MacDougall (Dalhousie Univ

**Figure 1.** Typical audio tool interface

The audio tools provide a user interface (Figure 1), which includes power meters to give the user an indication of send and receive volumes. Since it is difficult to distinguish among multiple speakers using monaural sound (one channel copied to both ears), a list of participants' names appears in the interface window, with the current speaker highlighted.

Speech samples are collected from the audio device, compressed, assembled into packets, and transmitted to the receiver. At the receiver, the speech samples are retrieved, decompressed, and fed out to the audio device for the user to hear. (The structure of a typical packet audio tool is depicted in Figure 2.)

*Silence detection* is commonly used in packet speech systems, since it can provide bandwidth savings of up to 50%. Input speech samples are categorized as either "speech" (a talkspurt) or "silence," and during periods of silence, packets are not transmitted. Silence detection algorithms in currently available audio tools use a simple average energy measure; they work well in acoustically quiet environments. The push-to-talk facility is often used in conjunction with a silence detection mechanism to restrict the number of channels open at any one instant to approximately half of one PSTN voice channel per multiway conference.

*Speech compression* reduces the network bandwidth consumed. At the sending end, speech samples are compressed before being transmitted across the network. At the receiving end, the inverse coding algorithm decodes the samples. Most speech compression algorithms originate from the PSTN world, where the goal is to increase the number of channels carried by one bearer circuit.

*Receive buffer* software enables audio tools to transfer a stream of speech samples in packets. Each packet contains a segment of speech from a time interval; packet sizes of 20, 40, and 80ms are common. Increasing the packet size reduces the rate at which packets are sent into the network, and this reduces the overhead associated with the protocol header. Increasing the packet size, however, also increases the end-to-end delay, because the samples for one packet all have to be collected before the packet can be transmitted.

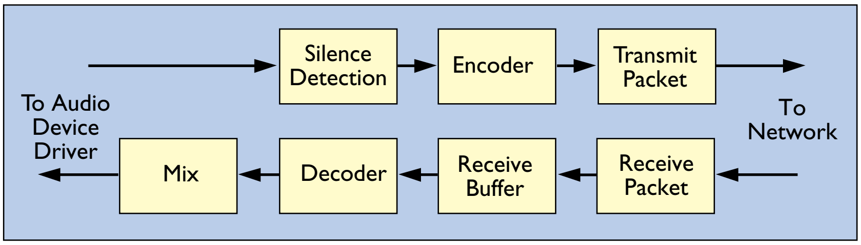

**Figure 2.** Structure of a typical packet audio tool

Since the network distorts the original timing relationship of the transmitted audio packets, a buffer is needed at the receiver to smooth out the packet flow. The receive buffer artificially adds a small extra delay to the mean network delay, so that packets stand a good chance of being received before their playback time. A particular receiver will experience different and variable delay and jitter from different senders. Consequently, delay adaptation must be on a per-source basis. Adaptive playout algorithms represent a compromise between low delay requirements and user perception of loss, and a delay budget sufficient to receive 99% of the transmitted packets is commonly used.

Some Internet audio tools use the Transmission Control Protocol (TCP) instead of User Datagram Protocol (UDP). TCP is designed for reliable data transmission and automatically instigates flow control mechanisms when packets are lost. The use of TCP results in unbounded delay, which is detrimental to audio conferencing interactivity. This effect was identified early on in the ARPANET and SATNET projects, which led to the development of more lightweight protocols for delay-sensitive traffic.

## Problems with Existing Audio Tools

Despite problems with them, existing audio tools are widely used, especially VAT. Use of VAT in piloting activities helped to identify a number of key problems that affect the perception of Mbone audio: shared network effects (packet loss) [2], scheduling in multitasking operating systems [5], and acoustic problems.

Generally speaking, delay is not a problem in multicast audio since the maximum round-trip delay is well below that at which normal conversational patterns cannot be maintained (400–600ms). With network congestion, however, packets may be excessively delayed or dropped. When packet loss occurs, a fill-in section of audio is provided by the receiver to maintain timing synchronization. Silence was used as the replacement in first-generation audio tools, such as VAT, but this results in degradation of speech quality, which increases rapidly with loss rate and packet size.

Mbone packet size is in the range 20ms–80ms, with 40ms packets being the default. Especially with packet sizes of 40ms and 80ms, a degradation in speech intelligibility, as well as in perceived quality, can be expected during packet loss [6]. Our experience shows that a 10% loss rate for 40ms packets is the highest that can be tolerated with silence substitution before a significant drop in speech intelligibility occurs. Packet loss on international links is frequently about 20–25%. This means that most existing audio tools cannot provide the minimum quality required to conduct an audio conversation over international—and sometimes national—links.

The performance of audio tools suffers from the lack of real-time scheduling support on general-purpose computers. The operating system may not schedule the audio tool at regular intervals since scheduling depends upon the other tasks being run on the host.

Erratic scheduling means the audio device may run out of buffered samples before the audio tool process is scheduled to provide more audio. This results in gaps in the output audio stream and an increased delay due to excessive stored audio in the device output buffers [12].

Acoustic problems that affect audio tools comprise:

- *Echoes and feedback.* The audio tool provides a closed audio loop between participants; if a participant uses a microphone and loudspeaker, then the system will have feedback. This problem is addressed in current audio tools by using the push-to-talk mechanism. Hands-free operation requires the use of headsets or expensive echo cancellers.
- *Silence detection.* The simple average-energy silence detection mechanism used in most audio tools is inaccurate and adversely affected by background noise. The beginnings and endings of words are lost, or the mechanism cuts in and out inappropriately. Better silence detection algorithms exist, but they require significantly more processing power.
- *Lack of distance cues*. Audio tools aim to support a wide range of audio headsets, microphones, and speakers, with a wide range of gains and impedance levels. The user has manual gain control which can be used to compensate for impedance mismatches to some degree. In practice, however, most speakers cannot judge how loudly they are speaking, because the communication sounds acoustically dead. In the real world, we judge the distance of the listener, and the size of the room, by using visual information and the amount of reflections (reverberation) caused by audio.

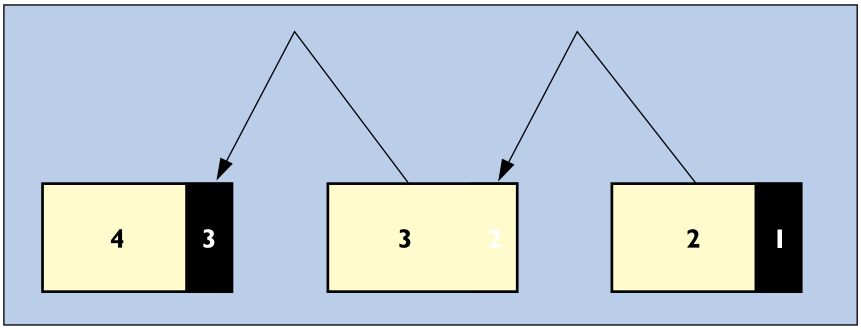

**Figure 3.** Redundant packet loss protection

- *Restricted intelligibility*. Toll-quality speech compression algorithms are suitable for most voice communication over the Internet, but some specialist applications require better-quality audio. Higher-quality audio can be produced only by increasing the audio frequency range transmitted. The use of toll-quality audio both restricts the intelligibility of speech and makes speaker identification more difficult (individual voice characteristics partially depend on frequencies higher than those preserved in toll-quality speech).
- *Monaural sound*. Audio transmitted over the Mbone uses one channel. In a natural environment, listeners use sound localization to identify a particular speaker through the position of the audio in space (in conjunction with visual information) and to focus on one sound source in the presence of other sound sources. Lack of sound localization both reduces speech intelligibility and restricts the ability of users to identify and focus on individual speakers.

## The Robust-Audio Tool

RAT attempts to address the problems experienced with other currently available Mbone audio tools. Packet loss is perceived as the most significant problem facing audio over the Internet, so effort has been focused mostly in this area. Redundancy [2] was developed to counteract packet loss. Problems caused by the multitasking operating system have been addressed through the development of an adaptive control mechanism [5]. Some acoustic problems, such as silence detection, have also been addressed. The first major RAT application and trial was used as part of a multimedia conference system for remote-language teaching [11]. Packet loss robustness, workstation adaptation techniques, and hands-free operation were improved in response to ReLaTe project requirements. Further requirements led to the development of an integrated interface for novice computer users and the provision of lip synchronization of audio and video components.

Given the packet size used and the range of packet loss rates that can occur on the Mbone, receiver-only mechanisms—such as packet repetition—cannot repair speech successfully [2]. Consequently, packet loss repair must be achieved by sending extra information from the transmitter to the receiver. This can easily be accomplished over the Mbone, since IP allows variable-length packets. Protection sent from the transmitter to the receiver could be in the form of packet-level forward error correction (FEC) tech-

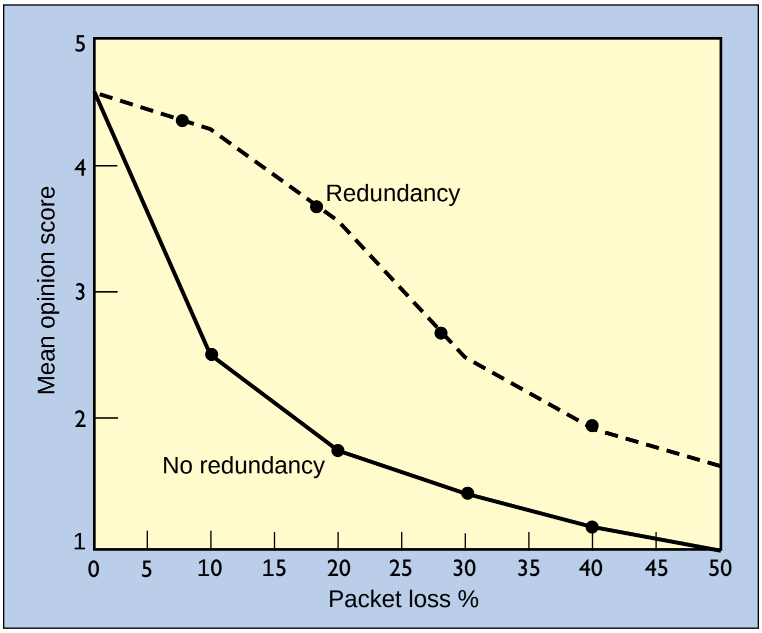

**Figure 4.** User perceptions of audio repaired with redundancy

niques. However, it is not necessary to repair speech with bit-level accuracy, since the brain is capable of restoring phonemes that have been replaced with noisy, but frequency-correct, information.

RAT uses redundancy [2], which involves transmitting extra information along with the main audio packets. Redundant information is piggybacked onto a later scheduled packet, which is preferable to the transmission of extra packets since the extra header is an overhead. Linear predictive coding (LPC) is a lossy low-bit-rate coding algorithm that is perceptually acceptable as the source of the redundant information for speech communication (see Figure 3). The use of LPC adds only a small overhead to each packet (a large increase in packet size is likely to increase the loss rate).

The loss characteristics over the Mbone vary; a reasonable working assumption is that packet loss follows a random pattern. Using this model, a single copy of redundancy is good enough to provide packet loss protection for loss rates of up to 20–30% (see Figure 4). RAT uses waveform substitution [2] to repair bursts of up to two packets—one packet is repaired by using redundancy, the other by using waveform substitution. Results show that a single copy of redundancy used with waveform substitution at packet sizes of 40ms repairs packet loss of up to 30% and provides perceptual improvement over this range by using waveform substitution.

RAT was also used in the ReLaTe trials, in which small groups of students participated in remote foreign language tutorials [11]. The results support the improvement in perception shown in Figure 4.

Voice reconstruction techniques cannot completely protect audio from the effects of packet loss and thus should be considered only as part of the solution. Of all the causes of packet loss, transient and long-term congestion are the most prevalent and must be addressed. Transient loss may result in only a small burst of loss, which can be repaired using the voice reconstruction scheme. However, some transient bursts are known to last up to a second, and a loss of this length cannot be repaired.

Loss that is a result of long-term congestion requires the network load be reduced. This can often be accomplished by increasing the level of speech compression. Increasing compression is often associated with an increase in the time length of the packet in order to restrict the overhead of the header to reasonable values. This results in an increase in the end-to-end delay and in the perception of loss. With increased packet sizes, redundancy must be used instead of receiver-only mechanisms [2].

The lack of real-time support on most general-purpose multitasking operating systems means that audio tools frequently suffer from gaps in the audio output and from an unnecessary increase in delay. RAT uses an adaptive scheme that analyzes the current real-time performance of the workstation and adjusts the amount of samples held in the device driver buffers. The scheme monitors the number of samples maintained in the audio device buffers (cushion), resulting in greatly reduced gaps in audio output due to output buffer starvation. Results also show a great reduction in the average end-of-talkspurt delay [5].

Silence detection is often inaccurate and difficult to implement. Many relatively successful algorithms exist, such as the voice activity detector developed for Global System for Mobile (GSM), but adequate silence detection can be obtained for much less processing power. RAT currently uses a simple average-energy silence detection mechanism with an adaptive silence threshold but supplements the operation with a rule-based approach. This adds only a small amount of extra processing but can increase the performance of the algorithm. The use of this mechanism within RAT has improved the performance of the silence detection mechanism to the point that hands-free operation can be used instead of push-to-talk for small-group interaction, if users wear headsets.

The application and user group of the ReLaTe project [11] presented an additional set of requirements: Communication must be intuitive and natural. In most applications of multicast conferencing to date, audio, video, and shared workspace tools are presented to the user as separate entities, each with associated tool and activity information (to indicate who is currently speaking, drawing, and so on). Managing multiple windows and scanning separate tools for activity information was an unacceptable level of overhead for language teachers and students, who had used computers for word processing only. Therefore, we developed an integrated user interface tailored to small-group teaching (Figure 5).

Since audio is used in hands-free mode, users see only two sliders to adjust the volume of the microphone and headphones. The interface adds speech power bars with each video window to indicate active speakers.

Since multicast audio and video are sent in separate streams, users of separate tools may hear a person speak before the received video image shows that person speaking. Applications such as ReLaTe require lip-synchronization, so RAT has been synchronized with a suitably modified version of a video tool using RTP time-stamping [6].

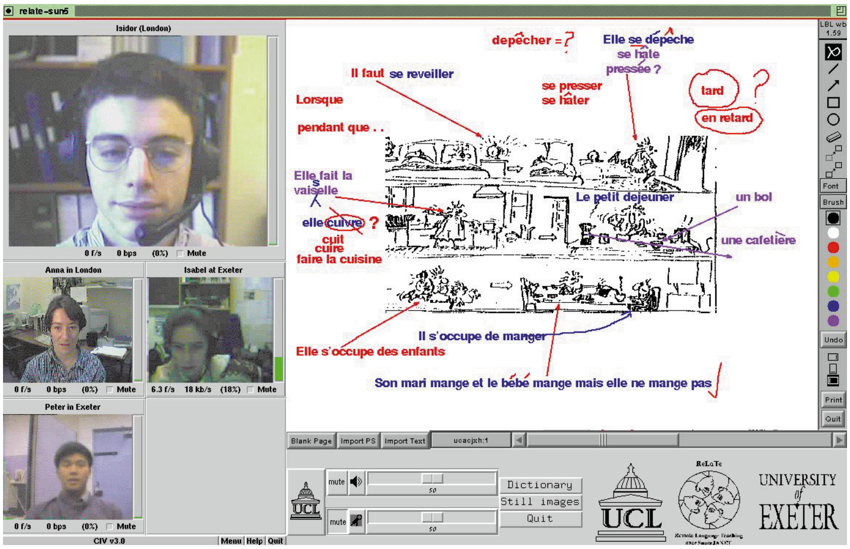

**Figure 5.** ReLaTe small-group distance-learning user interface

## Conclusions and Future Work

We stated at the outset that multicast has the potential to support multimedia conferencing solutions for many users and a wide range of tasks, provided sufficient audio quality can be sustained. Packet loss, multitasking in the workstation, and acoustic problems are the three major causes of impaired audio quality. We have presented methods of counteracting each of these problems and described how they have been implemented in a second-generation multicasting audio tool. Packet loss can be compensated by redundancy; results show a marked improvement in perceived audio quality with packet loss of up to 30%. The adaptive cushion algorithm is an effective and efficient mechanism that minimizes both delays and gaps caused by scheduling problems in typical operating systems. Improved silence detection allows hands-free use of the audio tool with headsets. Solutions to the other acoustic problems—wide-band speech, stereo effects, and spatial cues—have been devised and implemented. Informal trials have shown these further improve speech intelligibility in a way users feel is more natural. We have focused on multiway audio, but point-to-point applications, such as Internet telephony, will benefit from these improvements, too.

Ongoing research is addressing a number of current problems and future Mbone developments that have an impact on packet loss solutions. With the recent explosion in Web-based traffic [9], insufficient network capacity is likely to be a common problem. Audio communication becomes impossible when a user's available share of bandwidth falls below that required for even the most compressed audio. This could be avoided through bandwidth reservation. The IETF is standardizing a signaling protocol called RSVP [12] that includes hooks for classifying traffic and accounting policies to be implemented at routing nodes. Conference participants in an Mbone audio conference are likely to experience different loss rates. Rather than providing redundancy to all participants at the level required for the receiver at the end of the poorest link, multiple multicast addresses can be used [8].

The LPC algorithm used to provide redundancy forms the basis of many model-based coding algorithms. These algorithms can be made hierarchical, with LPC analysis forming the most important part of the information. Other layers successively add more quality to the received speech. A hierarchical scheme with LPC as its basis would be suitable for the heterogeneous nature of the Mbone. Currently, users may change compression schemes in response to long-term congestion. Audio applications must be made to adapt to congestion by automatically increasing compression of audio streams. This can be achieved by using RTP packet header that transfers audio session information on an end-to-end basis. If most of the network is well provisioned but heavily used, priority queuing for delay-sensitive traffic can be used to reduce network jitter without reducing the throughput of nonpriority traffic. Since queuing algorithms in routers can give priority to certain packets, marking delay-sensitive packets as high-priority would improve the quality of audio communication.

Acoustic problems associated with current audio tools reduce speech intelligibility. The acoustic environment provided by current audio tools is somewhat unnatural and may deter first-time users. Increased speech intelligibility can be provided only by coding and transmitting more of the input speech bandwidth. A less complex version of the wide-band speech coding algorithm is currently being developed for RAT. The wide-band speech coding algorithm transmits 7KHz audio and is known to preserve more of an individual's speech characteristics as well as increasing speech intelligibility.

Sound localization artificially manipulates 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 user's head with headphones). Sound localization is being developed for RAT and will be used to separate conference participants out in space. This substantially eases speaker identification and improves speech intelligibility, although discrimination of speech from the reverberation content of the original room is still marred. The lack of distance cues and problems with gain can be eased by artificial reverberation, telephone handset side-tone, and automatic gain control.

The multicast audio work is indicative of the potential of multimedia conferencing tools and applications over the Internet. Studying network characteristics, exploiting knowledge about human perception, and applying computational techniques can lead to considerable improvement in existing tools and the perceived quality of audio and video. Such techniques also allow adaptation when bandwidth or workstation power is limited. The example of ReLaTe (see Figure 5) demonstrates how multicast software tools can be tailored and integrated into multimedia conferencing solutions for specific user groups and their requirements. The data provided by any media can be recorded and edited. Recorded audio and video materials can be brought into conferences, and conference recordings can be replayed. Digital audio and video can be processed further in the workstation so sophisticated adaptation to specific user or task requirements (for example, compensating for hearing problems) is possible. It does not require much imagination from a reader familiar with the power and flexibility of computers to realize they will be at the center of genuinely integrated communication services.

## References

1. Deering S. Host extensions for IP multicasting. Request for comments RTC 1112. Internet Engineering Task Force, 1989.
2. Hardman V.J., Sasse M.A., Watson A., Handley M. Reliable audio for use over the Internet. In *Proceedings of INET95.* (Honolulu, Oahu, Hawaii, Sept 1995); info.isoc.org/HMP/PAPER/070/html/paper.html.
3. Jacobson V. VAT manual pages. Lawrence Berkeley Laboratory, Feb. 1992; www.nrg.ee.lbl.gov/vat/
4. Kirstein P.T., Sasse M.A., Handley M.J. Recent activities in the MICE conferencing project. In *Proceedings of INET95* (Honolulu, Hawaii, Sept. 1995); info.isoc.org/HMP/PAPER/166/ps/paper.ps.
5. Kouvelas I., Hardman V.J. Overcoming workstation scheduling problems in a real-time audio tool. In *Proceedings of the USENIX Annual Technical Conference.* (Anaheim, Calif., Jan. 6–10, 1997.)
6. Kouvelas I., Hardman V., Watson A. Lip synchronisation for use over the Internet: Analysis and implementation. In *Proceedings of Globecom96.* (London, Nov. 96), pp. 893–898.
7. Macedonia M.R., Brutzman D.P. Mbone provides audio and video over the Internet. *IEEE Comput.* (Apr. 1994), 30–36.
8. McCanne S., Jacobsen V., Vetterli M. Receiver-driven layered multicast. In *Proceedings of ACM SIGCOMM 96* (Stanford, Calif., Aug. 28–30); www.acm.org/sigcomm/sigcomm96/papers/mccanne.ps
9. Press L. Net.Speech: Desktop audio comes to the Net. *Commun. ACM 38,* 10 (Oct. 1995), 25–31.
10. RTP: A transport protocol for real-time applications. *Audio-Video Transport WG.*
11. Watson A., Sasse M.A. Evaluating audio and video quality in multimedia conferencing systems. *Interacting with Computers 8*, 255–275.
12. Zhang L., Deering S., Estrin D., Shenker S., Zappala D. RSVP: A new resource reservation protocol. *IEEE Net* 7, 5 (Sept. 1993), 8–18.

**Vicky Hardman** (vhardman@cs.ucl.ac.uk) is a lecturer in the Computer Science Department at the University College, London, U.K., researching technical aspects of networked multimedia systems, especially audio over packet networks.

**Martina Angela Sasse** (a.sasse@cs.ucl.ac.uk) is a senior lecturer in the Computer Science Department at the University College, London, U.K., researching applications and usability of multimedia and communication systems.

**Isidor Kouvelas** (i.kouvelas@cs.ucl.ac.uk) is a Ph.D. student in the Department of Computer Science at the University College, London, U.K.

The redundancy standard (RFC2198) was developed in consultation with partners on project MICE (ESPRIT 7602/6), particularly colleagues at INRIA, Sophia Antipolis, France. Work on RAT was partly funded by the BT/JISC SuperJANET Applications project ReLaTe and is now funded by EPSRC grant GR/K72780.