To appear in *Interacting with Computers*

Evaluating Audio and Video Quality in Low-Cost Multimedia Conferencing Systems

Anna Watson

Martina Angela Sasse

Department of Computer Science

University College London

Gower Street

London WC1E 6BT

tel: 0171-380 7212

fax: 0171-387 1387

email: a.sasse@cs.ucl.ac.uk

Evaluating Audio and Video Quality in Low-Cost Multimedia Conferencing Systems

Anna Watson & Martina Angela Sasse

Abstract

Real-time audio and video transmission over shared packet networks, such as the Internet, has become possible thanks to efficient data compression schemes and the provision of high-speed networks. Low cost multimedia conferencing technology could benefit many users in different areas, such as remote collaboration, distance education and healthcare. It is likely that diverse tasks performed by users in different application domains will require different levels of audio and video quality. Established methods of rating audio and video quality in the broadcast and telephony world cannot be applied to digital, lower quality images and sound. The providers of networks and services are looking to HCI to provide means of assessing audio and video quality. This paper describes two different approaches to assessing audio and video of desktop conferencing systems - a controlled experimental study and an informal field trial. We discuss the advantages and disadvantages of both approaches for providing task-specific quality assessment, and outline future work to integrate lab-based and field trials into a valid and reliable assessment approach.

Keywords: Multimedia Conferencing, Internet Conferencing, Audio Quality, Video Quality, Quality Assessment Methods

1 Introduction

Multimedia conferencing has become increasingly popular over the last 5 years. From expensive videoconferencing suites used by executives for special occasions, the technology has moved onto users' desktops for regular use. Conferencing over the Internet offers a cheap alternative to videoconferencing services offered by phone companies. Due to improved compression technology and higher speed networks, the three media (audio, video and shared workspace) can be transmitted in real time over shared packet networks, such as the Internet. With the introduction of multicast conferencing, multiway conferences can be held between very large numbers of participants all over the world, at comparatively low cost. The feasibility of using this technology for remote research collaboration has been demonstrated by the MICE project (Handley et al., 1993; Kirstein et al., 1995). Low-cost conferencing for large numbers of participants could benefit users in many areas, such as distance education, health care and small businesses. The question to be answered is whether the quality of media offered through low-cost conferencing, especially audio and video, will be sufficient to support the tasks to be performed in this domain.

We know from Gestalt psychology that "*the whole is greater than the sum of its parts*" (Kohler, 1930). Applied to multimedia conferencing, we observe that the different component media, especially audio and video, interact and influence the perception of each other. As Bjorkman et al. (1992) state, "assessing picture and audio quality simultaneously when two sensory channels are interacting represents a ... severe and complex problem". Conversely, it is likely that assessing the quality of a component medium in isolation will not produce accurate predictions as to perceived quality in a full multimedia conference, since the possible perceptual benefits afforded by interaction will not be present. There is a need to develop co-existing and complementary methods of assessment for these components, such as controlled experiments in addition to field trials of prototype systems.

The interaction between the senses in multimedia conferencing deserves further research. Studies have indicated potential trade-offs that could be exploited. For example, it may be possible to influence user perception of quality. In experimental studies with High Definition Television (HDTV), Neuman (cited by Negroponte, 1995) improved the perceived quality of video by increasing the audio quality only. An investigation of videophone systems indicated that any increase in visual representation of the speaker increases the viewer's tolerance to audio noise (Ostberg et al., 1989). On the other hand, Anderson et al. (1996a, in press) report that with visual contact with a listener, the clarity of a speaker's speech will drop significantly, a finding which could have severe consequences for multimedia conferences where audio and video are generally not synchronised. Obviously any means of increasing perceived quality of one component by raising the quality of another could be extremely important in getting the best performance possible out of a multimedia conferencing system.

The advantages to be gained from a thorough investigation of the perceived quality of the audio and video components are believed to be many. It is generally agreed that the critical element in multimedia conferencing at present is sufficient quality audio (Sasse et al., 1994). However, it can also be assumed that the quality required, both for audio and video, is likely to vary between conference participants and their tasks. For example, where the participants in a conference all speak English as their first language, and are using the conference as a weekly progress meeting, the required quality from the components is likely to be less than the quality that would be required were the purpose of the conference teaching/learning a foreign language (e.g. the ReLaTe Project, which will be discussed later). The perceived and required quality for different applications of multimedia conferencing need to be addressed, with a view to ultimately producing guidelines for quality thresholds required for a desired level of task performance.

This paper discusses two different approaches to evaluating quality of the multimedia components of audio and video for specific users and their tasks. The first approach is that of a controlled experimental study, and the second is that of a more informal field study. The pros and cons of both approaches will be discussed, and a proposal for a closer integration of the two techniques will be put forward. The paper will focus mainly on the assessment of audio, since this has been identified as the critical component of multimedia conferencing, but we believe that the issues raised and solutions proposed can be transferred and adapted to video. We will begin with a brief introduction to multicast conferencing.

2 Background to multicast conferencing

Traditionally, shared packet networks, which allow low-cost transmission of computer data, have not been considered suitable for real-time multimedia applications. Unlike circuit-switched networks, where users are allocated a specific amount of bandwidth, shared networks provide a “best-effort” service. When there is more traffic than bandwidth, data for the three separate streams of audio, video and shared workspace are held in queues and delivered when possible. Until recently, it was assumed that because of these delays, and the size of files required to represent digitised audio and video, shared networks could not be used to support real-time audio and video applications.

This view has changed due to three major developments: an increase in available bandwidth, more efficient compression schemes, and the implementation of multicast routing. In many countries, high-speed networks have been put in place over the last few years. SuperJANET in the UK is an example of a high-speed academic network. More efficient video compression schemes, such as H.261 (Turletti, 1993) have reduced the amount of data that needs to be transmitted, and have led to a proliferation of software codecs which can be run on computer workstations. They require a certain amount of processing power, but allow further processing of digitised audio and video data, and therefore more flexibility. With the implementation of multicast routing (Deering, 1988), the amount of bandwidth required for multimedia traffic has been reduced substantially.

Apart from cost, multicast conferencing has another significant advantage. Telephony (e.g. ISDN)-based conferencing can support point-to-point and 3-way conferences; sessions with more than 3 participants require advance booking and incur charges for the high number of circuits required. In contrast, multicast conferencing over the Internet scales from small conferences to audiences of several hundred. It has become increasingly popular: from a small number of experimental sites in late 1992, the multicast backbone (MBone) network has grown to more than 2000 sites worldwide, and is offered as a service by many academic network providers. Large-scale events transmitted worldwide range from technical conferences to shuttle launches (Macedonia & Brutzman, 1994).

In Europe, the MICE project (Handley et al., 1993; Kirstein et al., 1995) has successfully piloted the use of multimedia conferencing over the Internet for international research cooperation between Europe, US and Australia. Many collaborative research projects - including industrial ones - now use Internet conferencing for weekly project meetings, to improve communication and coordination between team members, while at the same time reducing the amount of travel required. As the use of Internet conferencing for different applications grows, so will the need for accurate information on quality requirements.

2.1 Multicast audio

Digitised audio information is sent in packets of duration 20, 40 or 80 ms. MBone audio is subject to various problems (e.g. high background noise, poor quality headphones resulting in echo etc.), but the most significant of these is packet loss. Packet loss can occur for the following reasons: congestion of routers and gateways, leading to packets being discarded; packets being delayed too long to be played out; and overloading of the local workstation. Audio that is subject to packet loss can be very disruptive for the listener, especially as larger packet sizes approach the size of meaningful units of speech, phonemes. When packet sizes are extremely small, losses can barely be detected. Jayant and Christensen (1981) found that when the packet size was 16-32 ms, perceptible glitches were heard when packet losses occurred. However, when the packet size was greater than 64 ms, the speech sounded ‘bubbly’ due to phoneme losses. It is this last point that is critical with respect to speech transmission over the MBone: larger packet sizes or consecutive losses of the smaller sizes can result in important, meaningful parts of speech being lost. Most existing audio tools, such as vat (Jacobson, 1992) and nevot (Schulzrinne, 1992) substitute lost packets with silence to maintain the playout order.

2.2 Multicast Video

The transmission of video information requires far more bandwidth than audio. The processing power required by the workstation at either end to encode and decode real-time video information is high, and this increases incrementally according to the number of video streams being received and transmitted. For these reasons, video information is generally updated relatively infrequently: 5-7 frames per second is the maximum most workstations can manage with four simultaneous video streams, compared to television quality at 25 frames per second. At the very low frame rates used in the early days, it was neither possible nor sensible to synchronise the separate audio and video streams, and so the facility of video has been viewed as being of secondary importance to audio in communicating in multimedia conferences over the Internet. Users describe video that is undergoing packet loss as ‘blocky’, a result of partial upgrading of parts of the video image.

3 The assessment of audio and video quality

The state of the art in audio/video assessment has generally focused on finding the point at which degradtion is not discernible. Telecommunication companies try to determine the best quality possible and charge accordingly. Real-time audio and video over the Internet is a low-cost, low-bandwidth alternative to these expensive forms of communication, and must look for the minumum quality (i.e. requiring a low amount of bandwidth) necessary to support a multimedia conferencing application. Thus, HCI knowledge is needed to determine the best methods of revealing the quality (and thus also the bandwidth) required.

3.1 Assessing speech quality

Speech quality can be measured either subjectively or objectively. Subjective methods (where ‘subjective’ refers to opinion rating and/or measurement of task performance) are agreed to be more reliable than objective methods involving instrumental assessment (Flanagan, 1965). The speech that is assessed in these tests is generally specific material, recorded or spoken under defined conditions.

Subjective speech assessments can be divided into two main groups, Mean Opinion Scores (MOS) and intelligibility tests. MOS have traditionally been used in speech quality assessment in the telecommunications world where the speech that is assessed is at, or approaching, telephone or ‘toll’ quality speech. The MOS is typically a 5-point rating scale, covering the options Excellent, Good, Fair, Poor and Bad, and is the standard recommended by the CCITT (1984) (Figure 1). Listening subjects rate the quality of the speech they hear according to this scale. Other rating scales exist, for example scales which attempt to address the impairment of quality rather than the quality itself (Bjorkman et al., 1992), but the MOS remains the most well-known and commonly used one.

Intelligibility tests have commonly been employed for the quality assessment of synthetic speech (Kryter, 1972). These types of test usually entail the subject listening to parts of words (e.g. consonant-vowel tests), words (e.g. rhyme tests) or sentences (e.g. Harvard or Haskins sentences), and writing down what is heard, or answering comprehension questions on a passage of speech.

The application of either of these methods to assessing the quality of speech transmitted over the Internet has certain difficulties. These methods of assessing speech quality and intelligibility have been developed for assessing the quality of telecommunications systems and synthetic speech. Their suitability for assessing speech quality over the MBone needs to be explored since packet loss of the sizes involved presents a novel type of degradation, the effect of which is hitherto unknown. In addition, the MBone is a highly unpredictable network in terms of quality of received information. Network load cannot be predicted beyond broad generalisations about peak loading times of day, and loss rates can alter drastically within the space of a few minutes or even seconds. Studies that have attempted to measure the loss characteristics of the Internet have found that packets tend to be lost individually rather than in clumps (Bolot et al., 1995), but beyond this the Internet remains far more unpredictable than any other type of communication network. It is therefore difficult to apply traditional intelligibility tests to Internet speech, since there is no means of predicting whether or where the packet loss will occur. It is this constantly changing and unpredictable level of quality that presents complex problems for the evaluation of audio quality over the Internet.

3.2 Assessing video quality

Picture quality can likewise be assessed by either subjective (related to task performance) or objective means. The assessment of video quality has traditionally followed the same route as that of audio quality i.e. the MOS. Other rating scales have been applied successfully (Allnatt, 1983) in conjunction with a certain task. The applicability of these types of measures to video transmitted over the Internet is called into doubt by the simple fact that the frame rate is generally so low (i.e. 2-3 frames per second) that it makes little sense in trying to determine how good the quality was or how severe the impairment is when the normal standard of comparison is TV quality. It has been suggested that the main use of the video link in these types of conference is psychological, and evaluation of video has often concentrated on the kinds of interaction it supports (Whittaker, 1995). However, it is of importance to gain an understanding of how the quality of the image interacts with and affects speech comprehension, especially in view of the fact that distance education is likely to become one of the major uses of multimedia conferencing over the Internet. Forced-choice comparisons, whereby a viewer must determine which of two video clips is of better quality may be a better option than asking subjects explicitly to rate quality.

3.3 Applying traditional methods to multicast audio and video

The MOS was developed as a method of assessing already high quality sound, and is typically used where the type of degradation that will be encountered is noise and encoding distortion, which is not as disruptive as packet loss. The MOS has proven to be an accurate measure for very fine-grained quality detriments in audio and video, however the traditional labelling of the 5 points of the scale (*Excellent, Good, Fair, Poor* and *Bad*) is unlikely to be used to best effect in assessing speech or picture quality over the Internet, where one suspects that the quality would rarely warrant the description “excellent”, at least not when compared to telephone quality speech and TV quality video. The labelling on a 5-point scale therefore warrants close attention. Re-labelling the points to suit Internet speech and video would be a non-trivial matter since much research went into the selection of these terms, believed to represent discrete intervals (Allnatt, 1983). The training of subjects in rating quality in Internet transmissions would therefore appear to be critical, and the selection of an ‘average’ example to be used as a control is highly advisable.

One final issue that deserves to be pointed out is that in a typical multimedia conference, with three or more participants, the quality of audio and video that each participant receives is likely to be different. This can arise for many different reasons: different local network loads, different hardware (e.g. headsets), different background noise and lighting, and loading on the individual’s workstation. These differences mean that it is not easy to get inter-observer reliability for perceived quality over one conference.

4 Experimental approaches to evaluating quality of audio and video

The previous section decribed the existing techniques for assessing audio and video quality, and discussed why these techniques may not be suitable for Internet audio and video. Section 4 reports studies that have been carried out to investigate audio and video quality, utilising and adapting some of the techniques discussed above. We will present results from experimental trials, and present findings from an extended field trial, before comparing and contrasting the results obtained from the two types of study.

4.1 Compensating for packet loss in Internet audio

An experimental study was carried out in order to compare the efficacy of some different techniques of repairing packet loss from a perceptual point of view. The experiment used speech material on which random packet loss was artificially generated in order to mimic the behaviour of the Internet. Three different reconstruction methods were compared: silence substitution, waveform substitution and LPC redundancy.

Silence substitution fills the space where a missing packet should be with silence in order to maintain the playout order of the packets. It is the cheapest and easiest method of audio repair, and therefore the most common, but its effect on perception is severe. It has been demonstrated that listening performance is worst when unexpected silence is encountered in speech (e.g. Miller & Licklider, 1950; Warren, 1982). Performance can be greatly improved by the insertion of noise of any type at the same frequency as the missing speech. This is approximated by the technique of *waveform substitution*, which fills the missing packet by repeating the last correctly received packet. This technique is more likely to work well where the packet sizes are small: when the packet is large, the speech signal is likely to have changed dramatically within the missing packet, and repeating the previous packet may be more detrimental than helpful to perception. *Linear Predictive Coding* (LPC) is a synthetic quality speech coding algorithm which preserves about 60% of the information content of the speech signal. It can be used a means of reconstructing the lost packets of speech at the receiving end if it is transmitted with the original signal, whereby the LPC speech for a missing packet is carried in the following packet(s). A more detailed discussion of these techniques can be found in Hardman et al. (1995).

The experimental hypotheses were that speech quality and intelligibility would be worst where silence was used, and that quality and intelligibility would be enhanced where the loss was repaired using the techniques of waveform substitution and LPC redundancy. It was also predicted that waveform substitution would begin to fail as a method when loss rates were high and packet sizes large, since the speech characteristics would change within the lost packets of sound. A more detailed description of the hypotheses and experimental design can be found in Watson (1994).

4.2 Experimental design and method

It was decided to use both an intelligibility test and MOS to collect data. Of interest was whether there would be agreement between the results of the two types of measures. A comprehensive survey of available intelligibility tests was carried out. Ultimately it was decided to use Egan's (1948) Phonetically Balanced (PB) word lists as the test material since the words are all the same length (monosyllabic) and proportionally represent the sounds found in everyday English, thus eliminating the chance that loss would be generated on words that are contextually easier than others to identify. (Most other intelligibility tests e.g. rhyme tests and sentences tests have key parts of the word/sentence that provide the focus for scoring results. In generating random packet loss, it would not be possible to use these tests since hitting the relevant part of the material could not be guaranteed.) The word lists were recorded by a speaker with Received Pronunciation[1], and loss was generated on these lists randomly. The lists were stored, manipulated and played out on a Sun SPARC Station 10, using the OGI speech tools software (OGI, 1993). The independent variables were loss rate (from 0 - 40%), packet size (20, 40 or 80 ms) and reconstruction scheme (silence substitution, waveform substitution or LPC redundancy).

There were three groups of seven subjects, each group hearing a different method of reconstruction. Each subject heard 10 lists of 25 words each, the first list being a no loss control condition. The subject was required to write down the word he/she had heard, or make a best guess. After each list the subject was asked to indicate the quality of the speech they had just heard on the 5-point MOS scale.

4.3 Experimental results

The results from the intelligibility tests broadly confirmed the hypotheses outlined above. Silence substitution did produce the lowest intelligibility for all packet sizes, failing at around 15% loss for 80 ms packets (more than the average length of a phoneme), and between 15 and 20% loss when the packet sizes were smaller at 20 and 40 ms (see Figures 2 & 3).

When waveform substitution is the reconstruction method and packet sizes are small, the drop in intelligibility is found at significantly higher loss rates compared to silence substitution. The results showed that waveform substitution is better than silence substitution for packet sizes of 20 and 40 ms, but the advantage is not present for packet sizes of 80 ms. The decrease in intelligibility for packet sizes of 40 ms does not become noticeable until between 30 and 40% loss. However, when the packet size is 80ms, intelligibility performance starts to drop noticeably at between 15 and 20%, indicating that the repetition of a packet where speech characteristics may have changed has a detrimental effect on perception.

Intelligibility using LPC redundancy is significantly better than using silence substitution at all loss rates and packet sizes. A comparison between waveform substitution and LPC redundancy revealed that for 20 and 40 ms packets, there is little advantage to using LPC rather than waveform substitution, but for 80 ms packets, LPC provides better intelligibility than waveform substitution.

The MOS results were in broad agreement with the intelligibility results, although the responses rarely used the 'good' option, and the 'excellent' choice was even more seldom indicated. This reflects the fact that speech over the Internet is not toll quality at best - usually 'adequate' audio is what is striven towards. It also highlights a concern (as stated in section 3.3) that the vocabulary on the MOS scale may not be suitable for use in quality judgements for speech over the Internet.

4.4 Conclusions from the experimental approach

4.4.1 Audio

The experiment provided some important results as to the advisability of applying different reconstruction methods to compensate for packet loss other than the silence substitution method which is in most common use at the moment. However, two main criticisms can be raised. Firstly, doubt can be cast on the ecological validity of the approach - writing down monosyllabic words is not the same as a conversational exchange. It is not possible to generalise from highly controlled experimental results such as these to the everyday fluctuating audio conditions and contextual speech that are commonly experienced on the Internet. (This issue of generalisation also applies to video studies.) Secondly, no account was taken of the interaction that often happens in multimedia conferences between audio and video. The experiment provided no visual information about the speaker, not even a still photograph. Multimedia conferences do not usually provide audio only - even if the video image is updated very infrequently (2 or 3 frames per second), its input to communication and perceived quality of the conference should not be underestimated, until comprehensive investigations of its use have been undertaken.

Additional experimental studies have been carried out more recently, and a preliminary attempt to address these problems has been made. Firstly, a study was carried out in order to determine whether using LPC redundancy in a new Robust-Audio Tool (RAT, Hardman et al., 1996) provided a better speech quality when loss was occurring than not using the redundancy option. Speech material used in this study was passages of text taken from a magazine lasting about 30 seconds rather than context-free monosyllabic words as were used in the first

1. Received Pronunciation is defined as a non-regional British accent (Ainsworth, 1976).

experiment discussed. The material was read and recorded by a male speaker with Received Pronunciation. Subjects were required to sit at a SPARC 10 workstation wearing a headset and listen to the passages, which had had loss generated on them and were either repaired with redundancy or not, and then rate the quality of the speech on the 5-point MOS scale. Results are shown in Figure 4. In agreement with the first experiment, a clear subjective preference for LPC repaired speech was demonstrated.

4.4.2 Video

Another experiment, this time involving both audio and video, was carried out in order to assess whether there was any benefit to synchronising the audio and video streams at the low frame rates that are commonly found on the MBone. The first method of synchronising multicast audio and video steams was recently developed (Kouvelas et al., 1996). Subjective studies have indicated that the mismatch in time between these audio and video streams can be in the region of 80-100ms before a lack of synchronisation is perceived (Jardetzky et al.,1995). A fuller discussion of this work can be found in Kouvelas et al. (1996), and subjective performance results can be seen in Figure 5.

The perceived level of synchronisation for different frame rates, and for both synchronised and unsynchronised audio and video was investigated.The material consisted of audio and video transmission of a speaker counting from 1 to 10. Eight subjects assessed this material at rates of 2, 5, 6, 8 and 12 frames per second. They were asked to indicate perceived quality on a 5-point rating scale, where 5 is rated as synchronised, and 1 as unsynchronised. The experiments were performed on a pair of Sun SPARC 10 workstations, using H.261 video coding, and CIF sized frames. The results shown in Figure 5 indicate that audio and video is not perceived as being synchronised for rates of less that 5 frames per second, a result which appears to support the findings of Frowein et al. (1991), who found a significant difference for speech reception between 5 and 6 frames per second. The results also support the finding (Jardetzky et al., 1995) that the jitter associated with video frame presentation times is below the limit of human perception.

4.4.3 Summary

These studies have provided useful data for making software design decisions, but the problem remains that the material that has been used is highly artificial. The rating scales that have been used to gather the data use vocabulary that may in some cases be unsuitable, and also provide a limited set of results. Some subjects, especially in the tests involving passages of speech, felt dissatisfied with having to circle only one quality rating, since within that relatively short audio experience, the quality may have fluctuated quite noticeably.

When the opportunity to evaluate audio and video quality in a real application context arrived, however, different problems arose.

5 The ReLaTe Project - an informal field trial

ReLaTe (Remote Language Teaching over SuperJANET) is a joint project between UCL and Exeter University, with the aims of providing a working demonstrator of a multicast-based conferencing system for remote language tuition, and to assess the feasibility of using this to provide remote tutoring in a field trial with teachers and students. It was hoped that the results of the trial would provide guidance for the development of networks and workstation technology fit for distance learning applications, and help us to identify corresponding changes required in distance education pedagogy.

Language teaching places high demands on the quality of the audio (Foot, 1994). Learners of a foreign language do not possess the native speaker's facility for compensating for poor audio quality, and lip synchronisation is required for at least some tasks. If multicast audio can support language teaching, it should be good enough for most distance education tasks.

We chose to investigate the technology in the context of small-group tutoring sessions. This was partly for practical reasons (number of workstations and students required), but also because groups with fewer students can be expected to yield a higher degree of student involvement and interaction, which is more demanding for the technology than a one-to-many lecturing scenario.

The field trial took place from October to December 1995, with tutors and students from the Language Centres at UCL and Exeter University, over the SuperJanet SMDS service between the two sites. Four weekly sessions (Advanced French, French for Business, Portuguese for Beginners and Latin) ran over 10 weeks. The sessions were observed and assessed from both a usability and pedagogical point of view.

5.1 The ReLaTe System

The ReLaTe system is based on multicast conferencing (audio, video and shared workspace) tools piloted and developed by the MICE project. It incorporates three new developments:

1. The Robust-Audio Tool (RAT), which can repair packet loss (see above).
2. Stream synchronisation between multicast audio and video.

3. A user interface in which the conferencing tools (audio, video and shared workspace) are integrated into a single conferencing window[1] (see Figure 6).

Each student or teacher sat at a workstation with a camera and headset. All tools are started up automatically with the conference system. Initially participants had to 'push-to-talk' by placing the mouse in the audio tool area of the interface and depressing the left hand button. However, this was rapidly replaced so that audio could be activated solely by a person speaking - full-duplex audio (see section 7.1). The packet size in the sessions was 40 ms long, and average loss rates were 10-15%.

6 Evaluating the ReLaTe system

Established evaluation methods for field trials developed in human factors and psychology include questionnaires, interviews, participant observation, and content analysis. However, multiway multimedia conferencing over the Internet is novel and more complex than the point-to-point interactions that have typically been the focus of work into producing usability and evaluation guidelines (e.g. RACE ISSUE 1992; LUSI, 1993) in this field. Controlled experimental studies of videoconferencing systems have often measured task performance while aspects of the system quality (typically video) are varied e.g. size of image, presence or absence of video (Frowein et al., 1991). These types of studies investigate users' performance and perception of the system. However, the tasks that subjects are asked to perform in these studies tend to be limited and artificial, and measures are restricted to time taken to perform problem solving tasks etc. To our knowledge, there has been very little quantitative or qualitative analysis carried out using subjects in a field study of a system in extended use over a real network such as the MBone, although occasional observational studies such as the MICE seminars (Sasse et al., 1994) have been reported.

Evaluation of the ReLaTe system needed to address two different aspects of the system: the effect of the technology (including the support it can/cannot provide for the application), and the pedagogical effectiveness of teaching and learning a language in this manner. The technological evaluation, comprising the three main areas of audio, video and shared workspace, is discussed below. Details of the pedagogical evaluation can be found in Watson & Sasse (1996).

6.1 Technological evaluation

6.1.1 Assessing the audio component

Since the ReLaTe system was to be tested in a field trial, many of the methods mentioned above were not suitable to use in ascertaining subjective audio quality and intelligibility. It was not possible to use a large number of subjects to ascertain interobserver reliability. Furthermore, not all participants in a language lesson would experience identical audio quality since different local networks and workstations will be subject to different loads. As already discussed, audio quality over a wide-area network can change rapidly and unpredictably, often meaning that audio will be good one minute and terrible the next, so a request to describe the quality for a two-hour lesson may be rendered meaningless. However, it was decided to apply the 5-point MOS scale after a series of one-hour trials, to see whether an indication of the subjective opinion of the overall audio quality during each particular lesson could be gained.

6.1.2 Assessing the video component

The video component of a conferencing system is often assessed in terms of comparing a task outcome using audio-only, audio and video, and face-to-face scenarios. There is much debate as to what, precisely, a visual channel adds to a conference (Whittaker, 1995). Given that the frame rate in a system such as ReLaTe would be slow, this question was of interest to the project: how was the video facility used? However, conclusions could only be drawn through observation and interviews, since controlled study would not be possible.

In addition it was of interest to ascertain whether the size of the video images was large enough for use in a teaching application.

---

1. In standard multicast conferences, users have to position and juggle windows for audio, video and shared workspace tools, plus a separate window for every video stream displayed. This allows experienced users to set up their own conferencing screens, but is a cumbersome and often confusing task for less experienced users. Asking teachers and students to arrange their own windows for their own lessons seemed unreasonable, when it is possible to present a unified conferencing ReLaTe interface. An integrated user interface was developed for use in the project trials, incorporating the video tool (*vic*) , the audio tool (*RAT*) and the shared workspace tool (*wb*). The front-end of vic was modified to display the video streams as one large and three smaller images on the left hand side of the screen. Individual participants can select the image to be displayed as the larger picture by clicking on the name bar above it. This feature was designed to give the teacher (and students) the choice of who to see "at large".

6.1.3 Shared workspace

Evaluation concerns for the shared workspace (in this case, the tool wb) were: whether its features were easy to use and useful, what kind of teaching tasks it would be used to support, and whether advantage would be taken of the facilities it offers for importing text and postscript files.

6.2 Evaluation methods used in the ReLaTe trials

Taking into account the evaluation constraints discussed above, an evaluation plan was devised, whereby students (at Exeter or UCL) would be taught by tutors at remote locations (UCL or Exeter) over a course of ten lessons. Four different classes at different levels were involved: Portuguese for Beginners, Advanced French, French for Business and Latin. All the lessons were taught in two hour sessions except for Latin (one hour). Each participant sat at a UNIX workstation (a Silicon Graphics Indy) equipped with a headset and a camera. Two or three students participated per tutorial. Evaluation of the technological and pedagogical issues of the lessons took place through the techniques of

- observation (by HCI experts sitting with the teacher/student and by language teachers acting as unobtrusive 'expert observers', from a separate workstation)
- questionnaires and rating scales administered after the lessons
- comparison with face-to-face classes
- informal interviews with all participants
- a group discussion workshop with all the participants

7 Results/Observations

The main result was that the feasibility of teaching language over a system such as ReLaTe was clearly demonstrated. The teachers and students reported that the system was easy and fun to use, and in the opinion of the expert observers, produced learning at least "as good" as might be achieved in a conventional small group class. A detailed discussion of the results can be found in Watson & Sasse (1996).

Teachers and expert observers agreed that the system in its present state not only supports the teaching of the four main language skills (reading, writing, speaking and listening) but encourages the *integrated* development of these, something that is not always easily achieved in conventional teaching situations. The group discussion workshops showed a high level of agreement between students, teachers and expert observers as to the technological and pedagogical strengths and weaknesses of the ReLaTe system. In particular, all participants stressed the high level of concentration it promoted, and the fact that it encouraged students to participate actively in the sessions. The enthusiasm of all participants may, however, have led to some extravagant claims about the system's superiority over a conventional classroom environment.

7.1 Audio

From the earliest trials it became clear that the 'push-to-talk' mode used in existing multicast audio tools was hindering communication between the participants. No sound was communicated unless the mouse button was held down. Speaking interfered with use of the whiteboard, preventing, for example, the kind of 'writing with commentary' which is central to many lessons. A move from speaking to another activity requiring the mouse left other participants unsure as to the cause of the resulting silence, which might be interpreted as system error. Observations revealed also that visual contact (through the camera) was reduced by the need to keep locating the mouse and placing it on the audio tool. Perhaps most important of all in a teaching and learning context, being able to hear only one participant at once meant that the phatic function was lost and students were denied the paraverbal reassurances ("mmm", "uhuh" etc.) that encourage them over hesitations and give them the confidence to speak in the target language. Hands-free operation was therefore essential.

Full-duplex audio requires silence detection to avoid placing an unacceptable burden on the network through the constant transmission of background noise from all participants. An improved silence detection mechanism was incorporated into RAT (Hardman & Kouvelas, 1996). The workshop discussions endorsed the full-duplex audio.

Even after the switch to full-duplex mode, however, audio quality was the most frequently cited cause of dissatisfaction with the system. Network packet loss impaired quality. This was perceived to have a crucial impact on learning in a beginners' class (Portuguese), where the student was unfamiliar with the sound of the language and the teacher wanted to practise pronunciation of short sounds. The greatest frustration amongst users arose from the unpredictability of audio quality, which could vary dramatically within a single lesson. In spite of this, most of the participants believed they spoke at least as much of the target language as they would have done in a face-to-face class, and said that they felt able to rate the overall quality of the sound during a lesson. Users rated the majority of lessons as having "fair" or "good" quality audio. There was a noticeable trend within some subjects to rate the audio quality as better towards the end of their sessions, which did not correlate with the objective packet loss statistics available. This indicates that listeners may habituate to impaired audio quality over time.

7.2 Video

The main result concerning video was that lip movement could not be used as a comprehension aid because the frame rate used in the trials (2-3 frames per second) was not fast enough to allow lip synchronisation. Surprisingly, not everyone found this a disadvantage, one student commenting: “Tutor’s lip movement was delayed and didn’t match speech. This actually encouraged me to listen harder to the French sounds - very useful”. The lack of synchronisation meant users did not perceive an advantage in selecting the larger image for the current speaker; it was therefore not possible to draw conclusions about the optimum size of video image for language learning.

All the same, video was felt to be a valuable component, and participants made use of the images in many other ways, including: to check whether anyone was speaking during an unusually long silence which might suggest an audio problem; to find out whether other participants understood what had been said (expressed through smiling, laughing, nodding); as a way of gauging the teacher’s reaction to the student’s work; to provide a means of common reference (as when the tutor holds a worksheet in front of the camera to identify it to students); to learn some of the nonverbal gestures pertinent to the target language; and, since lack of sidetone in the microphone contributes to the feeling that the system is ‘dead’ when no sound is being received, to give psychological reassurance that the other participants are actually “there”. From the group discussions came the suggestion that relationships might be affected by the fact that, unlike in a conventional classroom, all participants’ faces were always visible to one another.

A common problem was with participants moving partly out of camera shot. This was often related to space management; when the teacher used a text book or students took notes on separate paper, they would lean sideways and out of the range of the camera. Most teachers found that the ‘computer desk’ allowed insufficient space for ease of work and sometimes were observed holding a textbook over the keyboard whilst attempting to type on the whiteboard. Tutors made a recommendation that bookrests should be provided as part of the physical workspace.

7.3 Conclusions from the ReLaTe field trial

Evaluating audio and video quality in the ReLaTe field trial was not straightforward. With respect to audio, a number of problems arose. Firstly, because the participants wore headsets during the lessons, it was not possible for an observer to hear the sound quality being experienced from that specific workstation. Secondly, the integrated interface had been designed so that only the windows that were strictly necessary for the purpose of the language lesson were visible. Therefore it was not possible to view a window with the objective packet loss statistics for the workstation in use. For these two reasons, the observer had to rely solely on comments the user made during or after the lesson with respect to the sound quality. The assessment of audio quality was also impacted by the fact that the participants in a conference were all sitting in different environments, with different levels of background noise, different headsets and different workstations.

Evaluating the video component revealed some interesting discrepancies between what the users thought they used the video for, and what they were observed using it for. The questionnaires that were completed suggested that the video use was mainly psychological, although the observers noted a great deal of varied use of it (see section 7.2 above). It is possible that the participants dismissed the video since the audio was not synchronised with it (confirmatory evidence of this was suggested by an additional lesson that was run with the new synchronisation method, discussed in section 4.4.2, in use at 6 frames per second, where subjects afterwards reported using the video far more than usual), and were unaware of how much use they made of the tool for indirect communication. This use of video falls under the nonverbal communication hypothesis (Whittaker, 1995). From our observations we strongly believe that there are communicative benefits to having a video link, regardless of frame rate. It has been stated elsewhere (e.g. Anderson et al, 1996b, in press) that video mediated communication is not as efficient as face-to-face communication, but it should not be dismissed on this account - multimedia communication over the Internet does not seek to replace face-to-face communication, but rather to offer a low-cost means of communication between geographically distributed groups.

Overall, the ReLaTe field trial suffered from what all field trials are subject to: lack of control over numerous environmental factors. It is not possible to prevent participants from behaving in a certain way e.g. individual preferences for volume and lighting often led to unsatisfactory sound and view for other participants. On the other hand, the ecological validity of the trials was very high, and over the course of the ten language lessons for each course, the students and teachers were able to learn and teach without being overly affected by the observations being carried out. The trials were judged to be a great success by all who participated. It is to be stressed that it is only by carrying out field work such as this that any technological developments can be assessed properly by the end-users who will ultimately have the final say as to whether the system is acceptable in its quality.

8 Discussion and Future work

This paper has discussed the importance of developing methods of evaluating audio and video quality for conferencing over the Internet. The achievable subjective quality of these two media is bandwidth dependent. Bandwidth cannot be guaranteed over the Internet at the moment, and this situation is not likely to change in the short or medium term. Therefore it is sensible to investigate what quality is necessary for individual applications. Better quality audio will be required for applications such as language teaching, while it is to be expected that remote interviews would require a good quality of video. Establishing the minimum quality required for adequate task performance will lead to HCI guidelines towards which applications are viable under certain bandwidth constraints. HCI researchers need to define methods of evaluating audio and video quality as related to task performance in this new area of communication. Existing methods of audio and video quality assessment need to be refined due to the unique conditions of packet network conditions, and new methods may need to be developed.

The use of forced-choice comparison tests will be explored in future work. Given the reservations voiced about using the MOS, or at least the labelling, it may be worthwhile asking subjects which of two samples they prefer, without asking them to make explicit quality ratings. In this manner it will be possible to play longer samples in which quality fluctuates and gain an understanding of the perceptual effect of differently changing quality. Forced-choice comparisons will be a particularly effective method of investigating the interaction effect of the two media of audio and video. To establish quality thresholds for specific tasks, such rating methods would have to be incorporated in task-specific benchmark tests.

We have presented and discussed two different types of study aimed at addressing the evaluation of audio and video quality. The controlled experimental approach produces clear results, but suffers from the constraint common to all experimental studies, that of difficulty in generalisation to the real-world situation. The prolonged field trial approach suffers from the problem of lack of control over a large variety of variables, and the inability for the evaluators to get 'too close' while the trials are ongoing. There will always be a trade-off between getting accurate measurements and interfering with the task at hand. One of the ways in which the gap between the experimental and field trial techniques can be bridged is by integrating 'real' material into controlled experiments. It is now possible to record and store the digital material of conferences being transmitted, and these streams of information can be played back after the event. Combining this facility with a record of the objective packet loss statistics, it will become possible to use 'real' conference material in experimental situations. For example, it will be possible to assess perceived audio quality with/without the presence of video, which will allow a closer examination of the interaction and trade-offs between these two media and senses. The recording technique also permits an unintrusive record of a conference to be made and studied after the actual conference for evaluation purposes. This method is therefore extremely valuable on two counts: as a source of genuine Internet audio and video for controlled studies, and as a means of keeping observer intrusiveness to a minimum in field trials. It is intended to begin using this technique in experimental studies at UCL shortly.

It is hoped that these new approaches will add significantly to the knowledge we have about required quality for different applications of multimedia over the Internet. The importance of increasing knowledge about this area cannot be exaggerated. HCI knowledge about the relationship between users, tasks and quality requirements should set the agenda for effective design and implementation of multimedia communication systems for a wide variety of applications. Network and service providers currently argue in terms of bandwidth and functionality offered at the network layer. Users, on the other hand, do not care about figures and features of the cable in the ground; they will pay for end-to-end quality and functionality which they perceive to be adequate for their chosen application area and its tasks. Fundamental decisions about network provision, management and charging to made over the next few years could be swayed by reliable information about what users actually require.

Acknowledgements

The authors would like to acknowledge the following contributions to the work reported in this paper: at UCL, Vicky Hardman, Isidor Kouvelas and Jane Hughes. At Exeter, John Buckett, Mark Pack and Gary Stringer. Special thanks are due to the teachers and students from the Foreign Language Centre at Exeter University (Director Elizabeth Matthews) and the UCL Language Centre (Director Dolores Ditner).

ReLaTe (Remote Language Teaching over SuperJANET) was a BT/JISC funded SuperJANET Applications Project from Sept. 94 - Dec. 95. ReLaTe is a joint project between Exeter University and UCL. Anna Watson is funded through an EPSRC CASE studentship with BT.

References

Ainsworth, W.A. (1976): Mechanisms of Speech Recognition. Pergamon Press.

Allnatt, J. (1983): Transmitted Picture Assessment. Wiley.

Anderson., A.H., Bard, E.G., Sotillo, C., Newlands, A. & Doherty-Sneddon, G. (in press): Limited visual control of the intelligibility of the speech in face-to-face dialogue.To appear in *Perception & Psychophysics*.

Anderson, A.H., Newlands, A., Mullin, J., Fleming, A., Doherty-Sneddon, G. & Ven der Velden, J. (in press): The Impact of VideoMediated Communication on Simulated Service Encounters. To appear in *Interacting With Computers.*

Bjorkman, N., Goldstein, M., Hedman, L., Latour-Henner, A., Tholin, P. & Gil, L. (1992): Network performance and its relationship with quality of service in an experimental broadband network. In A. Casasc (Ed.) *Broadband Communications*. Elsevier Science Publishers B.V. (North-Holland)

Bolot, J., Crepin, H. & Vega Garcia, A. (1995): Analysis of audio packet loss on the Internet. *Proceedings of NOSSADV*, 163-174, Durham, NH.

CCITT (1984):Recommendations of the P Series, "Method for the evaluation of service from the standpoint of speech transmission quality". CCITT Red Book Volume V - VIIIth Plenary Assembly.

Deering, S.E. (1988): Multicast routing in internetworks and extended LANs. *SIGCOMM Symposium on Communications Architectures and Protocols*, (Stanford, California), 55-64, ACM, Aug. 1988.

Flanagan, J.L. (1965): Speech Analysis, Synthesis and Perception. Springer-Verlag, Berlin.

Foot, C. (1994): Approaches to multimedia audio in language learning. *ReCALL*, 6:2, 9-13.

Frowein, H.W, Smoorenburg, G.F., Pyters, L. & Schinkel, D. (1991): Improved speech recognition through videotelephony: experiments with the hard of hearing. *IEEE Journal on Selected Areas in Communicatio*n, 9, 611-616.

Handley, M.J., Kirstein, P.T. & Sasse, M.A. (1993): Multimedia Integrated Conferencing for European Researchers (MICE): Piloting Activities and the Conference Management and Multiplexing Centre. *Computer Networks and ISDN Systems*, 26, 275-290.

Hardman, V. & Kouvelas, I. (1996): RAT General Architecture. IN/96/2, Dept. of Computer Science, University College London, UK.

Hardman, V., Sasse, M.A., Handley, M.J., & Watson, A. (1995): Reliable Audio for Use over the Internet. *Proceedings of INET'95*.

Jardetzky, P.W, Sreenan, C.J. & Needham, R.M. (1995): Storage and synchronisation for distributed continuous media. *Multimedia Systems*, 3, 151-161.

Kirstein, P., Handley, M., Sasse, A. & Clayman, S.(1995): Recent activities on the MICE project. *Proceedings INET'95*.

Kohler, W. (1930): Gestalt Psychology. London:Bell.

Kouvelas, I., Hardman, V., & Watson, A. (1996): Lip Synchronisation for Use Over the Internet: Analysis and Implementation. RN/96/17, Dept. of Computer Science, University College London, UK.

Kryter, K.D. (1972): Speech communication. In Van, Cott & Kinkade [Eds.]: *Human Engineering Guide to Equipment Design*.

LUSI (1993)Specification of Usability Measures adopted for User Exposure Phase, HUSAT Research Institute, Loughborough University of Technology, UK.

Macedonia, M. & Brutzman, D. P. (1994): Mbone provides Audio and Video Across the Internet. *IEE Computer*, April 1994, pp.30-36.

Negroponte, N. (1995): Being Digital. Hodder & Stoughton.

OGI (1993): Speech Tools User Manual, Center for Spoken Language Understanding, Oregon Graduate Institute of Science and Technology. Available by anonymous ftp from speech.cse.ogi.edu:/pub/tools/

Ostberg, O., Lindstrom, B. & Renhall, P-O. (1989): Contribution of display size to speech intelligibility in videophone systems. International Journal of Human-Computer Interaction 1(1), 149-159.

RACE ISSUE Usability Guidelines (1992). Vol. 1-4, RACE ISSUE Project 1065. HUSAT Research Institute, Loughborough University of Technology, UK.

Sasse, M.A., Bilting, U., Schulz, C-D. & Turletti, T. (1994): Remote Seminars through Multimedia Conferencing: Experiences from the MICE project. *Proceedings of INET'94/JENC5*.

Schulzrinne, H. (1992): Voice communication across the Internet: a network voice terminal. University of Massachusetts, Technical Report, June 1992.

Turletti, T. (1993): H.261 Software Codec for Videoconfering over the Internet. Research Report No. 1834, INRIA, Sophia Antipolis, France.

Warren, R.M. (1982): Auditory Perception. Pergamon Press Inc.

Watson, A. (1994): Loss of audio information in multimedia videoconferencing - an investigation into methods of assessing different means of compensating for this loss. Unpublished MSc Thesis, Dept. of Ergonomics, University of London.

Watson, A. & Sasse, M.A. (1996): Assessing the usability and effectiveness of a remote language teaching system. *Proceedings of ED-MEDIA*, 1996.

Whittaker, S. (1995): Rethinking video as a technology for interpersonal communications: theory and design implications. *International Journal of Human-Computer Studies*, 42, 501-529.