Multimedia Conferencing via Multicast: Determining the Quality of Service Required by the End User

Anna Watson & Martina Angela Sasse
Department of Computer Science
University College London
Gower Street, London, WC1E 6BT, UK

ABSTRACT: *Network congestion affects all users, but real-time multimedia communication is particularly vulnerable. It has been proposed that bandwidth reservation should be used to guarantee a certain quality of service (QoS) for such applications. In order to employ bandwidth reservation effectively, we need to know how much bandwidth a particular application requires. This depends largely on the quality at which different media have to be delivered. Since end-users' perception of quality will determine whether applications are successful, it is essential to establish minimum and maximum media quality required by a particular user group for a particular task. This paper discusses suitable methods and measurements to determine perceived media quality. The design of, and results from, an audio quality experiment illustrate the complexity involved.*

1. INTRODUCTION

The Internet community has grown considerably over the past few years, and there is a danger that it could become a victim of its own success. As network congestion is increasing, so are calls for *quality of service* (QoS) guarantees, which could be achieved by mechanisms such as *bandwidth reservation*. Quality of Service is a term that can mean different things to different bodies. According to Fluckiger [1], QoS in networking terms is a *"concept by which applications may indicate their specific requirements to the network, before they actually start transmitting information data"*. This raises the question *'What requirements does the application have?'* Ultimately, an application's requirements are the application user's requirements, i.e. *"a residential person, a professional individual, a subscriber organisation, a computer program, a computer system"* [1]. In the context of this research, we focus on the end-user of a system, the person who interacts with a network application to achieve a particular task.

Our long-term aim is to define the level of quality required to complete a particular task successfully, and the point of diminishing return, i.e. the threshold beyond which increasing objective quality (and hence bandwidth) does not increase user performance or satisfaction.

Quality of media delivery can be expressed in objective measurements (e.g. packet delay and loss) which can be translated into bandwidth required. The quality of media as perceived by the end-user, however, can rarely be predicted from these objective measurements. Subjective perception of quality is influenced by a number of factors, and varies according to the task being undertaken. Determining the quality required by a particular user for a particular application is a non-trivial problem.

- There are numerous variables which can affect the end user's perception of a system, and complex interactions occur between the component media (and their qualities) of multimedia conferences.
- Even when the variables involved can be separated and controlled, the methodology for measuring user opinion/performance is not established.
- Generalising to the real world from results gained when interacting variables have been controlled will always be problematic.

Our short term research aim is therefore to address these issues, by establishing methods of measuring user perception of multicast audio and video subject to packet loss at given bandwidths. Preliminary steps towards this aim are discussed in this paper, with results from a recent audio experiment.

2. WHAT AFFECTS USER PERCEPTION?

In investigating user perception of a multimedia conferencing system, the influence of many variables needs to be considered. A representation of some, though by no means all, of the factors that can influence user perception is shown in Figure 1.

Of these many different factors, it is of course possible to identify some that play a larger role than others. In the case of audio perception, for example, the single most important factor has been identified as packet loss [2]. With respect to video, it is frame rate that is most likely to affect user perception. The exact impact of these factors on perceived quality, however, is likely to depend on the task that is being undertaken in the conference. For example, it is obvious that an application such as a remote interview would require both a high quality of audio and a high quality of video such that the participants could acquire an accurate impression of each other.

At the other end of the range, a research project meeting where participants already know each other and are using the system only to exchange routine information is likely to place less stringent quality demands.

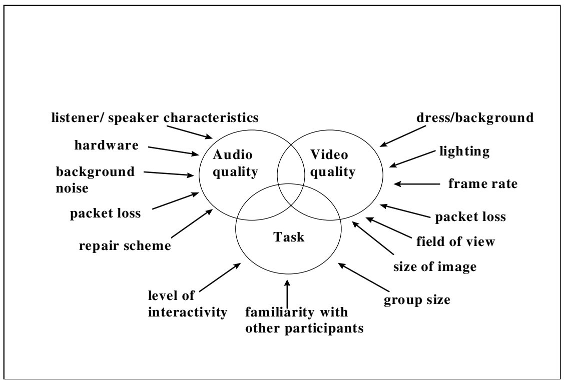

Figure 1: The interaction between audio quality, video quality and task, and factors that affect subjective opinion

Before considering the role of task, however, it is necessary to determine how best to measure perceived quality of multicast packet audio and video.

3. INVESTIGATING USER PERCEPTION

There are two approaches to measuring what users think of systems - in real field trials, and in controlled experimental conditions which mimic aspects of the real world situation. Although field trials are more desirable in that they actually *are* what a user would be experiencing, they are costly and time-consuming, can be frustrating for the user, and do not always provide the means of acquiring the information that is required by the human factors investigator. Controlled experiments, on the other hand, can be criticised in that generalising from them to the real world situation is never easy - their ecological validity is seldom high. A discussion of these issues can be found in [3].

Whether user perception is measured in controlled experiments or in field trials, the most suitable means of measurement needs to be determined. There are standard methods of measuring the perception of speech and video, methods that have been designed and used in the telecommunications industry. However, the large size of multicast packets (and the size of the consequent loss), in addition to novel encoding and repair methods used in the multicast community, mean that many of these standard methods cannot be easily applied to Mbone (multicast backbone of the Internet) audio and video.

Before turning to these methods, it is necessary to define more clearly what we mean by ‘quality’. We have already stated that quality must not be equated to bandwidth, so what exactly it is that we mean, and what do we need to measure?

3.1 What is Quality?

When we talk about investigating perceived quality, what exactly do we mean by the term ‘quality’? With respect to speech, ‘quality’ can be conceived of as an umbrella term. There are many variables that contribute to forming a perception of speech quality, such as listening effort, loudness, pleasantness of tone, and intelligibility [4]. It is therefore wrong to equate intelligibility with quality, since many factors other than intelligibility play a part in the forming of quality perception. However, it is true that intelligibility is the single biggest determinant in perceived speech quality, and it is therefore vital to investigate the intelligibility of Mbone speech.

With respect to video and picture quality, we can again identify a number of factors that affect overall perceived quality, such as frame rate, lighting, image size, ‘blockiness’, packet loss and degree of synchronisation with the audio. Again, it would be an error to assume that frame rate determines video quality perception, but it is the largest determining factor, and therefore deserves to be the focus of investigation.

3.2 Assessing Speech Intelligibility and Quality

Mbone speech is subject to packet loss: at present some degree of packet loss is inevitable and unavoidable. There is a need to investigate the effects of the loss (and attempts to compensate for the loss) on speech intelligibility and quality. What methods for assessing speech intelligibility and quality are there, and are they suitable for assessing Mbone speech?

The most common established methods for assessing speech intelligibility are syllable tests, word tests and sentence tests. These tests all involve listening to samples of speech at that level of description, and recording what was heard. It has been determined that syllable tests and some word tests are not suitable for Mbone speech intelligibility assessment since Mbone packet sizes (20ms, 40ms or 80ms) often exceed the duration of a syllable or the critical phoneme of a word test [5]. Sentence tests involve scoring how many key words in a sentence were intelligible to the listener, and as such provide a better test for the characteristics of Mbone speech. The suitability of these test methods with respect to Mbone audio is discussed at greater length in [3].

Speech quality tests have been developed and widely used throughout the telecommunications industry. For ease of comparison between different studies and countries, standard scales are recommended by bodies such as the ITU (see, for example, [6]). A widely used scale is the 5-point listening quality scale (Figure 2). However, there are some concerns with this scale. Firstly, it has been observed that the labels on the scale do not conform to interval data points, but are ordinal [7]. Virtanen et al. [8] have observed that the labels on the scales do not always translate well across different languages. In addition to these points, Mbone speech at the moment is rarely experienced with no packet loss, and so the term 'Excellent' is unlikely to be selected very often without anchoring the speech in controlled experiments (which can risk jeopardising the relationship between results in experimental conditions and field trials). Without extensive training, the vocabulary of the scale is not suitable, and relabelling the points would be a complex undertaking [9].

Quality of the speech	*Score*
Excellent	5
Good	4
Fair	3
Poor	2
Bad	1

Figure 2: ITU 5-point listening quality scale

The listening effort scale [6] is another 5-point scale, which requires the listener to rate the ease with which the meaning of the speech was understood. Assessment of Mbone speech is yet to be undertaken using this method.

In addition to listening quality tests, it is common practice to carry out conversation tests to determine whether a system can be used efficiently and effectively in two (or more) way communication. These are again assessed via a 5-point scale, with participants also being asked to report perceived reasons for any problems with the connection.

3.3 Assessing Video Quality

The quality at which video is assessed in the telecommunications industry far exceeds that which is currently common in the Mbone community. TV quality video is transmitted at 25 or 30 frames per second, whereas Mbone video is regularly transmitted somewhere in the range of 2-8 frames per second, since video is bandwidth and processor intensive. In the broadcast world the degradations are so small as to be almost negligible, and a test that is commonly applied is that of the forced-choice comparison [9]. This kind of approach makes little sense for Mbone video - what we should be asking is whether a certain quality (in terms of frame rate) is adequate for the task that is being engaged in. It is commonly asserted that low frame rate video is of little communicative benefit in videoconferencing (see [10] for a fuller discussion), however, in common with other researchers we have found that the psychological value of the video component is high [11]. In addition, we have demonstrated that synchronisation between the audio and video streams begins to be perceived at frame rates as low as 8-10 frames per second [12]. The communicative benefit of this has yet to be evaluated in experimental studies.

The remainder of this paper discusses a recent Mbone audio study carried out at UCL, illustrating the difficulty in assessing subjective quality, and presents ideas for future research on determining quality requirements for multimedia conferences over multicast networks.

4. MEASURING SPEECH INTELLIGIBILITY AND PERCEIVED QUALITY

As discussed in section 3.1, intelligibility cannot be equated with perceived quality, and therefore investigation of only one or the other may not be sufficient in generating user requirements for a system. Established quality rating scales may not be suitable for assessing packet loss speech, and alternative methods should be investigated. It is also important to bear in mind that the perceived quality rating that is awarded when a participant is passively listening to speech may not be the same rating awarded when that same person is engaged in an interactive task. These issues need to be addressed, and as a first step towards this end, an experiment with three main aims was designed and carried out:

1. To compare the intelligibility of Mbone speech subjected to packet loss and repaired by two different methods.
2. To compare the perceived quality of passages of speech repaired with the two methods, using both a common 5-point quality scale, and a novel unlabelled one.
3. To compare the perceived quality of the two repair methods when they were being used in a real conversational task in a multicast multimedia conferencing environment.

24 subjects participated in the experiment. All had normal hearing and were native English speakers. The experiment was divided into three parts, speech intelligibility assessment, listening quality rating, and conversational quality rating. The material for the intelligibility and quality sections was recorded by a male speaker with Received Pronunciation in 16-bit linear, 8kHz format and encoded using the DVI4 codec. A Sennheiser HMD25-1 headset was used for both the recording and the playback. The trials were conducted using Sun SPARC 10 workstations.

The experimental details for each part are presented below.

4.1 Intelligibility Assessment

The material used in assessing speech intelligibility was phonetically balanced (PB) sentences published by the IEEE [13]. These take the form of short syntactically varied sentences with five key words, for example "*Nine rows* of *soldiers stood* in *line*" where the key words are italicised. Different percentages of packet loss were generated on the sentences (no loss, 10%, 20%, 30%, 40% and 50%) using a packet reflector mechanism in the audio tool RAT [14]. The loss was either repaired by packet repetition or by LPC (a summary of these techniques can be found in [15]). 15 sentences were generated in total, 4 practice (reference) sentences, and 11 test sentences.

Method

All subjects were presented all 15 sentences, but the 11 test sentences were randomised for each subject to prevent any order effects. The subjects were asked to write down as much as possible of each sentence they heard.

Results

Scoring of the sentences was carried out by marking the correctly perceived key words and working out a percentage intelligibility for each sentence from this. The results are presented in Figure 3.

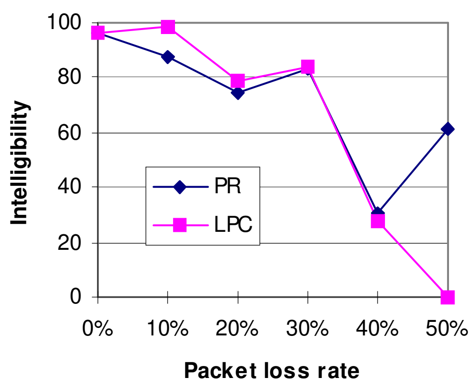

Figure 3: The intelligibility of sentences with packet loss repaired by either packet repetition (PR) or LPC redundancy.

Discussion

The results indicate that the LPC repair method offers superior intelligibility over packet repetition, at least at lower packet loss rates (5-20%). The surprising result for packet repetition at 50% loss may be a function of using only one sentence and loss pattern - it is possible that the loss did not hit the critical key words in this sentence. The study needs to be repeated with a larger number of test sentences and loss patterns to confirm this (see section 5).

4.2 Listening Quality Assessment

As discussed in section 3.2, there are some potentially serious drawbacks with the 5-point quality scale, namely the vocabulary and the fact that is does not appear to be an interval scale. We decided to investigate whether an unlabelled continuous scale might allow listeners to indicate their perceived quality rating in a less restrictive, and hopefully more indicative manner than the 5-point scale. We wished to make comparisons between the two scales. For example, would it be possible to discern whether subjects actually had a preference for one scheme over the other when both were rated as 'Fair' on the 5-point scale? In addition, we hoped to make comparisons between the listening quality ratings and the intelligibility results.

Materials

15 passages of approximately 30 seconds duration were recorded from a student newspaper, with loss and repair generated on them in the same manner as before. Again, 4 passages were used as test/reference conditions, and there were 11 test passages.

Method

Half of the subjects rated the passages on a 5-point scale first, and then on an unlabelled scale, and the other subjects undertook this in the reverse order. The unlabelled scale was a vertical line 20cm long, with no labels other than a '+' sign at the top and a '-' at the bottom to indicate the polarity. The subjects were instructed to place a cross on this line at a point they felt represented the quality of the speech to which they were listening.

Results

Scoring of the 5-point scale gave rise to the Mean Opinion Scores (Figure 4). The unlabelled scale results were translated into a 100-point % scale, and are shown in Figure 5.

Discussion

It is clear that the two scales follow the same trend, which is encouraging in the case of the unlabelled scale where subjects control their own rating criteria. The unlabelled scale was popular with many of the subjects, who commented that they preferred it since it allowed them more flexibility in making quality judgements.

It can be seen from these results that there appears to be a slight preference from the listeners for speech that has been repaired with packet repetition over speech that has been repaired with LPC redundancy, especially at lower loss rates. This does not agree with the intelligibility assessment results at these loss levels (Figure 3).

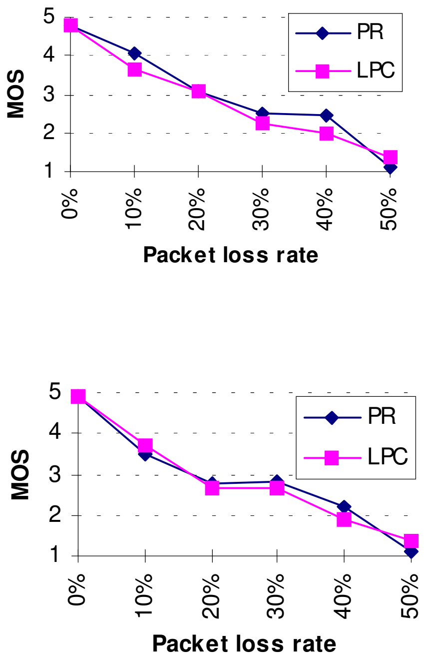

Figure 4: Quality ratings for speech passages using 5-point scale first (upper graph) and second (lower graph)

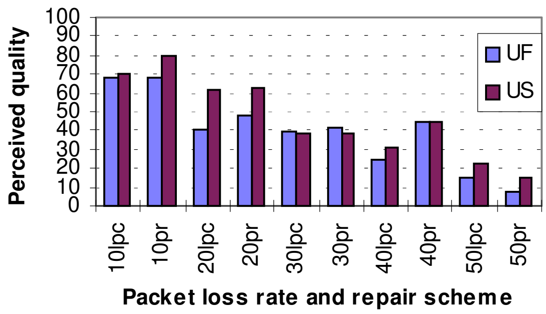

Figure 5: Quality ratings from the unlabelled scale used first (UF) and second (US)

There is no clear trend to suggest that the unlabelled scale allows subjects to indicate 'degrees of fairdom' between PR and LPC.

It is also interesting to compare the results between those who experienced the unlabelled scale first and those who experienced it second. We were interested to see if some kind of shaping behaviour might have occurred for those subjects who experienced the 5-point scale first - would they mentally conceive of the unlabelled scale falling into 5 basic divisions? It is difficult to draw conclusions about this, but it is interesting to note the significantly higher ratings awarded at 20% loss by subjects who rated with the unlabelled scale second. Further research would be required to determine whether this is due to shaping by the 5-point scale.

4.3 Conversation Task

In an attempt to investigate how subjects perceived the quality of an audio connection when they were actively engaged in a multimedia conferencing task, rather than passively listening to Mbone speech, we paired the subjects and asked them to play ten 2 minute games of the word game Taboo across a simulated network.

Materials

The multimedia conferencing tools RAT, vic [16] and wb [17] were used in order to simulate a multimedia conference desktop set-up. Again, loss level and repair scheme were variables, but we excluded the 50% loss level on the basis of pre-trials which indicated that this level of loss would make a conversation too difficult and frustrating for participants.

Method

The subject pairs were seated at SPARC 10 workstations in different rooms, and given a brief introduction to the tools. The rules of Taboo were explained to them, although it was made clear to them that they should just take turns in describing a word: the purpose was to have a conversational exchange rather than to win the game. The order in which the different conditions were experienced was randomised across pairs. After each game the participants were asked to rate the quality of the connection on a 5-point scale.

Results

The MOS results can be seen in Figure 6.

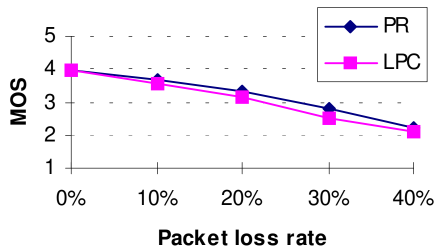

Figure 6: Quality rating for conversational task

Discussion

There is a clear depreciation in perceived quality as the packet loss rate increases, much as is seen in the listening quality results. The relative depression of the 'no loss' rating in the conversation test compared to that of the listening test can be attributed to the differences expected between listening passively to one male speaker and taking an active part in communicating with a different speaker/voice across a 'lossy' network.

5. GENERAL DISCUSSION & CONCLUSIONS

We recognise that the studies reported here are limited in their scope and generalisability, but nevertheless we feel that there are valuable conclusions to be drawn from this initial attempt at investigation, and many issues worthy of future research have arisen.

The results from the experiment illustrate the complexity of the relationship between speech intelligibility and perceived quality. The results demonstrate that it is not safe to assume that more intelligible speech will receive a higher quality rating than less intelligible speech. Our results show LPC redundancy to produce significantly greater intelligibility than packet repetition at lower loss rates, yet in both passive listening and interactive conversations, packet repetition is awarded at least as good as, or a higher quality rating. Further research needs to be carried to validate and extend these findings across different speech coding schemes and packet loss repair methods. Experimental work of this nature is very time consuming. Having identified one of the study's drawbacks as being lack of variation in material and loss patterns, we propose to attempt a larger scale evaluation of speech intelligibility and quality by utilising the world wide web. In this study we will investigate also the utility of the listening effort scale [6].

It is important to explore results using different loss patterns and materials, and across a range of different speakers. Most importantly, perceived quality needs to be investigated under a range of different multimedia conferencing tasks. It is only when this has been undertaken that we can start making sensible recommendations for bandwidth requirements and reservations, based on loss tolerance results from experimental studies. Satisfactory task performance can be achieved at less than (objectively) ideal conditions - following this approach we will be able to make the best possible use of limited bandwidth resources.

6. ACKNOWLEDGEMENTS

The authors would like to acknowledge the following contributors to the work reported in this paper: Louise Clark, Vicky Hardman, Orion Hodson, Ismail Ismail and Daniel Pilon.

Anna Watson is funded through an EPSRC CASE studentship with BT Labs.

7. REFERENCES

[1] F. Fluckiger, "Understanding Networked Multimedia", Prentice Hall, 1995.

[2] M.A. Sasse, U. Biltung, C-D. Schulz, & T. Turletti, "Remote seminars through multimedia conferencing: experiences from the MICE project", Proceedings of INET '94/JENC5.

[3] A. Watson & M.A. Sasse, "Evaluating audio and video quality in low-cost multimedia conferencing systems", Interacting with Computers, 8(3), 1996, 255-275.

[4] J.E. Preminger & D.J. Van Tasell, "Quantifying the relationship between speech quality and speech intelligibility", Journal of Speech and Hearing Research, 38, 1995, 714-725.

[5] V.J. Hardman, M.A. Sasse, M.J. Handley, & A. Watson, "Reliable audio for use over the Internet", Proceedings of INET'95.

[6] ITU-T P.80 Methods for subjective determination of transmission quality.

[7] B.L. Jones & P.R. McManus, "Graphic scaling of qualitative terms", SMPTE Journal, November 1986, 1166-1171.

[8] M.T. Virtanen, N. Gleiss & M. Goldstein, "On the use of evaluative category scales in telecommunications", Proceedings of Human Factors in Telecommunications '95.

[9] J.W. Allnatt, "Transmitted-picture Assessment", John Wiley & Sons, 1983.

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

[11] A. Watson & M.A. Sasse, "Assessing the usability and effectiveness of a remote language teaching system", Proceedings of ED-MEDIA '96.

[12] I. Kouvelas, V.J. Hardman & A. Watson, "Lip synchronisation for use over the Internet: analysis and implementation", Proceedings of GLOBECOM '96.

[13] "IEEE Recommended Practice for Speech Quality Measurements", IEEE Transactions on Audio and Electroacoustics, Vol. AU-17(3), 1969, 225-246.

[14] V. Hardman, M.A. Sasse & I. Kouvelas, "Successful multi-party audio communication over the internet". To appear in Communications of the ACM.

[15] O. Hodson, C. Perkins, A. Watson & V. Hardman, "A comparison of channel coding techniques for internet audio", UCL work-in-progress, RN/97/63.

[16] V. Jacobson, Vic Manual Pages, 1994, LBL.

[17] V. Jacobson, Wb README file, 1993, LBL.