Evaluating Real-Time Multimedia Audio and Video Quality

Anna Watson
Department of Computer Science
University College London
Gower Street
London, WC1E 6BT
+44 (0)171 419 3688
a.watson@cs.ucl.ac.uk

ABSTRACT

The aim of this research is to assess and establish quality thresholds for real-time Internet audio and video. Real-time multimedia conferencing over the Internet has huge potential, but there are limitations to the quality of audio and video that can be achieved, due to bandwidth limitations and the processing power of individual workstations. Assessing the effects of these limitations on the conference participant is not straightforward. The novel types of degradation found over the Internet means that existing speech and video quality assessment methods may not be applicable to multimedia conferencing experiences. This PhD will assess existing tests for measuring perceived quality from the psychology and telecommunications literature with respect to multimedia conferencing. The long term aim is to produce guidelines as to required bandwidth and quality for different multimedia conferencing tasks and applications.

Keywords

Multimedia conferencing, MBone, speech intelligibility, speech quality, video use, task.

BACKGROUND TO THE RESEARCH AREA

Multimedia conferencing involves three streams of real-time information available through an individual's workstation: audio, video and shared electronic workspace. Multimedia conferences are run over the multicast backbone of the Internet, known as the MBone, which makes multiway communication between large numbers of participants possible. In order for multimedia conferencing over the Internet to be used to its full potential, it is necessary to gain a complete understanding of the factors that affect the perceived quality of the component media, especially audio and video, so that recommendations can be made as to what bandwidth is necessary for the goal of a certain application to be achieved. Quality over packet networks such as the Internet is a function of bandwidth: as a general rule, the more bandwidth that is available, the better the quality will be. However, the major attraction of the Internet, that it is cheap and available to all, means that bandwidth is not, as yet, allocated to certain applications. It is therefore important to investigate what minimum bandwidth is *required* from a user point of view, for that application to be viable and successful. This is the long term goal of this research - to produce guidelines as to what bandwidth is required for a certain application or task to be possible.

Audio and video information is sent over the Internet in a digital fashion in small blocks known as 'packets'. These packets can be delayed or lost because of congestion on the MBone. Congestion occurs at the network routers through which the information must be passed in order to be sent to the correct destination. If information is held up for too long at the routers, the packets may arrive too late to be played out at the receiving end. Alternatively, if congestion at the router is very great, some packets may get dropped at that point. The end result of these two occurrences is the same: packet loss.

Packet Loss

Packet loss is the single-most disruptive factor in multimedia conferencing over the Internet. This is especially true with respect to the perception of speech. Video quality is also impaired by packet loss, but its effects are less harmful to the communicative process, since it is audio that is widely considered to be the most critical component [1]. When packets of speech information are lost, the effect on listener perception is dependent on the size of the packet and the loss rate [2]. Audio packets sizes are of 20 ms, 40 ms, or 80 ms duration. The smallest meaningful element of speech, the phoneme, has an average size of 80-100 ms [3], and so losses of this size can interfere with the intelligibility of the perceived speech. There are different methods of compensating for this packet loss, and an experiment was designed and carried out to compare the results of these schemes [2]. This study raised a number of interesting issues regarding the assessment of both audio and video quality over the Internet. The types of degradation that these media are subject to is novel and unique, and therefore existing methods for the assessment of speech and image quality and intelligibility are not always applicable in this area. This PhD seeks to identify which tests are suitable and which are not.

THE PROPOSED RESEARCH

There are 3 key aspects to this research:

An investigation of existing methods and tests for evaluating intelligibility and quality.

An investigation of the interaction with audio and value of

MBone video in its present state.

The relationship between and relative importance of these components with respect to task.

Speech Assessment

The standard methods for assessing speech are traditionally divided into methods for objectively assessing speech intelligibility and subjectively assessing speech quality. Speech intelligibility tests use different types of material, ranging from syllables and words to sentences and passages. Syllable tests have been identified as unsuitable for the assessment of packet speech since the size of the missing packet can be as large as the syllable itself. On the other hand, sentence tests rely on key words being degraded which is not likely with speech subject to packet loss, since loss is random. Phonetically balanced words [4] have been used in preliminary research [2], but generalising from word lists to genuine conference speech is problematic. Exploration of other test material will be carried out, with a view to maximising the ecological validity of experimental results.

Perceived speech quality is ascertained by having listeners indicate on a scale how the sound quality seemed to them. The most common scale in use is the ITU Listening Quality Scale, which consists of 5 grades of quality (Excellent, Good, Fair, Poor and Bad). Although this scale has been used in early studies, we hold reservations about the vocabulary on the grades - getting listeners to indicate Internet audio as “excellent” is difficult even with negligible packet loss rates. Other researchers have also pointed out the accepted vocabulary on the scales may have shortcomings (see, for example,[5, 6]). For this reason it is proposed to explore the use of other techniques such as the continuous rating scale, unlabelled scales, and the forced-choice double stimulus method as a means of obtaining opinion data.

Video Assessment

Picture quality assessment is also carried out using rating scales [6]. It is likely that it will not be the video picture quality per se that will be of importance in evaluating the video component, but rather the frame rate that is required for the presumed benefits of video to be afforded. Packet video requires a much greater bandwidth than packet audio, and as a result it tends to be sent at a low frame rate (in the range of 2-8 frames per second, as opposed to the 25 or 30 frames per second that forms television quality). Applications and tasks that require a lower frame rate will use less bandwidth, thereby freeing up bandwidth for other multimedia conferences. This is in an important concept - as the popularity of the Internet continues to grow, any means of reducing the information being sent along it should be sought.

Synchronisation between the audio and video is not common, and lip synchronisation is not considered feasible at low frame rates. The utility of low frame rate video has been called into question (see [7] for a discussion of this issue), but observations from field trials indicate that even very low frame rate video has a communicative benefit [10]. However, assessing this observation in qualitative and quantitative terms is not straightforward. There is much evidence that the perception of audio can be improved by the presence of a video image [7], but this is likely to be linked to the task at hand [8]. For example, in multimedia conferences it is often the case that participants’ attention is focused on the shared electronic workspace, and this will obviously have an effect on the perceived utility and/or quality of the video. The potential interactions between variables in this area are many. It is planned to run a series of experiments on perceived audio quality and intelligibility with and without the presence of low frame rate video to address this issue further.

OVERALL FRAMEWORK AND SUMMARY

The research methodology for evaluating audio and video quality with respect to different multimedia conferencing applications needs to be defined. The general framework that is being followed in this PhD is that of grounded theory [9], whereby an area of study can be approached from a variety of different angles and methods, and the most useful approaches can then be used to refine and confirm the findings. In work carried out so far in this area, two main approaches have been used in trying to determine the perceived quality of multimedia conferencing components: controlled experimental studies and a large field trial [10]. Benefits and disadvantages of both approaches have been identified, and the PhD work will build on these findings and continue to try to refine an adequate approach to assessing audio and video quality in multimedia conferencing.

REFERENCES

1. Sasse, M.A., Bilting, U., Schulz, C-D. & Turletti, T. (1994). Remote seminars through multimedia conferencing. *Proceedings of INET ‘94/JENC5.*

2. Hardman, V., Sasse, M.A., Handley, M. & Watson, A. (1995). Reliable audio for use over the Internet. *Proceedings of INET ‘95.*

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

4. Egan, J.P. (1948). Articulation testing methods. *Laryngoscope*, 58(9), 955-991.

5. Virtanen, M.T., Gleiss, N. & Goldstein, M. (1995). On the use of evaluative category scales in telecommunications. *Proceedings of Human Factors in Telecommunications ‘95.*

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

7. Whittaker, S. (1995). Rethinking video as a technology for interpersonal communication. *International Journal of Human-Computer Studies*, 42, 501-529.

8. Anderson, A.H., Newlands, A., Mullin, J., Fleming, A., Doherty-Sneddon, G. & Van der Velden, J. (1996). Impact of videomediated communication on simulated service encounters. *Interacting with Computers*, 8(2), 193-206.

9. Strauss, A. & Corbin, J. (1990). Basics of Qualitative Research. Sage.

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