The Effects of Audio and Video Correlation and Lip Synchronization

L.Mued, B.Lines, S.Furnell and P.Reynolds

Network Research Group, Department of Communication and
Electronic Engineering, University of Plymouth, United Kingdom
E-mail: lmued@jack.see.plymouth.ac.uk

Abstract

This paper is based upon (a) investigate the interaction effect of audio and video and, (b) study on lip synchronization (lip sync). The study shows a comprehensive evaluation of achievable audio and video quality undertaken based upon different sets of impairments between audio and video, prior to transmission. The tests have been conducted on two different task scenarios, i.e. passive communication and interactive communication (person to person). The research concentrates on quantifying the effects of network impairments (packet loss) on perceived audio and video quality, as well as finding the correlations between audio and video in multimedia applications. The results presented in this paper show the strong interaction dependency between audio and video. It was justified that video has a unique benefit on multimedia quality for its psychological effects. The findings also concluded that the sensory interactions, and the attention given to a particular aspect of performance, are clearly content-dependent.

***Keywords* – Audio, video, lip sync, multimedia, MOS, task performance.**

Introduction

The aim of the paper is to investigate the interaction effect between the perceived audio and video quality in multimedia services. The study on lip sync is also described in this paper. Lip sync refers to the synchronization between the movements of the speaker's lips and the spoken voice. Lip sync is one of the important issues in multimedia applications.

Previous research has claimed that a user's assessment of audio quality is influenced by the presence of video in multimedia applications (Watson and Sasse, 1996). For this reason, the experiments were based on investigating and quantifying the potential interaction effect between audio and video when transport mechanism carrying the two medias is subject to packet loss.

The importance of good quality audio in a conference cannot be overstated (Kawalek, 1995), (Kitawaki and Nagabuchi, 1998). Since true lip reading is impossible for most people, effective communication cannot be achieved without intelligible audio. Likewise, audio delay can make interactive communication difficult. Also, audio that is not synchronized with video can be distracting due to loss of lip synchronisation.

Current desktop videoconferencing systems transmit between 2 and 8 frames of video per second (Quarter Common Interchange Format, QCIF–176x144 pixels/ Common Interchange Format, CIF–352x288 pixels), with poor resolution and unsynchronized audio and video. The presence of video which enables interpersonal face-to-face communication is prevalent and much preferred over all human means of interactions (Tang and Issacs, 1993). Studies show that, in workplace settings, even when people are given a choice between different means of communication, such as email, phone and face-to-face, they still choose face-to-face meetings for planning and definitional tasks (Finholt *et al*, 1990). This is evidence that videoconfencing has unique benefits over audio only commmunication for most class of task.

Many studies have investigated the influence that video mediation has on the process of communication. Some research findings claim that the presence of a video channel does not directly improve the task performance in the context of desktop videoconferencing (DVC) (Wilson and Sasse, 2000a). However, it has been suggested that the main use of the video link in DVC is psychological (Hardman *et al*, 1998) such as to clarify meaning, to provide a means of common reference, to check whether anyone was speaking during an unusually long silence, to give psychological reassurance that the other participants were actually there by creating a sense of presence etc. Thus, it is stated that, in general, video is better than audio for interruptions, naturalness, interactivity, feedback and attention (Sellen, 1992).

In summary, whilst good quality video is beneficial to enhance many interactive tasks, sufficient audio quality is an essential for real-time interaction. The question is, what quality is good enough to meet end user's requirements?

To date, there is no standard consensus to clarify multimedia quality of service (QoS). In conjunction, effective evaluation methods are vital to determine the quality the users need to successfully perform tasks in videoconferences. However, it is stated that assessing the quality of audio and video over IP network offers a great challenge due to its constantly changing and unpredictable nature (Wilson and Sasse, 2000a). On the other hand, to determine multimedia conferencing quality has certain difficulties, as there is no recognized industry consensus of what really determines audio and video quality. At present, it is often questioned whether the quality of the audio and video in multimedia conferencing is adequate to carry its task performance (Wilson and Sasse, 2000b). Many researchers claim that different tasks performed by the end user will require different levels of audio and video quality. In some cases it may be necessary to prioritise video over audio, or vice versa, depending on the type of session. For example, language teaching in a distance learning application will require better audio, as opposed to a remote interview that demands a good quality of video as well. Therefore, it is essential to investigate what quality is necessary for each specific application. The aim of this research is to establish taxonomy of real-time multimedia task and applications, and to determine the maximum and minimum audio and video quality boundaries for the given tasks.

The Experiments

The two main experiments described in this paper are (a) Experiment A: Investigate interaction effects between perceived quality of audio and video and, (b) Experiment B: Study the effects of lip sync on multimedia quality.

Experiment A: Investigate interaction effects between perceived quality of audio and video

As previously stated, the experiments were based upon investigating a potential interaction effect between audio and video media in DVC systems in the presence of packet loss. The approach is to send the audio and video component with respect to the assigned quality for each media, in two different task performances (i.e. interactive and passive interactions). The proposed method will be to degrade the quality of audio and to upgrade the quality of video, or vice-versa, before sending it through a “connectionless” network. At the receiving end, the subjects will evaluate individual quality of audio, video and combined audiovisual of low bit rate videoconferencing.

Figure 1 below depicts the VoIP (Voice over IP) test bed configuration used for the experiments, and the various elements illustrated are described below.

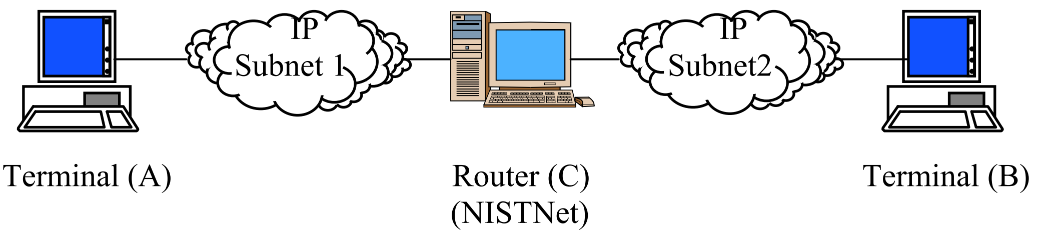


Figure 1: Test bed configuration

Terminal A & B: Two identical videoconferencing systems (hardware and software), running Microsoft NetMeeting, placed in two separate rooms, to be used by the subjects to rate Mean Opinion Scores of the perceived audio and video quality. CPUs: 200 MHz Pentium processors, 64 MB RAM, were used. The QCIF - 176x144 pixels frame size is used. The video setting was unchanged throughout the test, which was ‘better quality’ video. For the audio CODEC, a G723.1, 6400bit/sec was employed.

Router (NISTNet): A network emulation package (Carson, 2000) that runs on Linux. By operating at the IP (Internet Protocol) level, it allows a PC-based router to emulate numerous complex IP networks performance scenarios. In our experiment, it was used to introduce different sets of packet loss for audio and video streams.

Subnet 1& 2: IP networks

The test activity of the project is organised in a number of steps. First, tests are carried out under error free network environment. Second, different set of network impairments (packet loss) are introduced to the separate audio and video stream in order to evaluate their impact on the perceived quality. The conditions under considerations are shown in Table I, below:

video (v)/audio (a) (%)	v/a	v/a	v/a	v/a	v/a	v/a	v/a
v degraded/a (0% loss)	0/0	1/0	1.5/0	2/0	2.5/0	3/0	4/0
v (0% loss)/a degraded	0/0	0/9	0/10	0/15	0/25	0/30	0/35
v (%) degraded/a (%)	0/0	1/9	1/10	1.5/15	2/25	2.5/30	3/35
v poor (4%)/a degraded	0/0	4/9	4/10	4/15	4/25	4/30	
v degraded / a poor (35%)	0/0	1.5/35	2/35	2.5/35	3/35		

Table I: Packet Loss of Video (v) and Audio (a) Under Test, in Percentage

The test was conducted on two different task scenarios i.e. (a) Interactive test and (b) Passive test

(a) Interactive test

There were 20 adult individuals involved in the test. The subjects were allowed to select their own issue for discussion, with which they were comfortable, so as to enable the interactions. It is stated that informal communication tends to be representative of individuals who are familiar with each other (Issacs and Tang, 1994). Hence, to maximise task motivation and to ensure subjects are fully at ease with each other, individuals (subjects) who were acquainted with one another were selected for the tests. This is vital so as to ensure the validity of the results.

For each new set of impairments of audio and video, after every discussion, the subjects were asked to rate the perceived quality of (a) audio, (b) video and (c) combined audiovideo. The discussions were limited two minutes. For control purposes, initially, tests were carried out under error-free condition, i.e. 0% packet loss.

(b) Passive test

A number of 20 adult individuals volunteered for the test. They were asked to view and to listen to a 'talking head', reading a short sentence to them. First, for control purposes, tests were carried out under conditions that used no packet loss and each medium (i.e. audio, video and combined audiovideo) were evaluated. Second, packet loss was introduced in order to evaluate its impact on the perceived quality. For each set of impairments, the subjects were asked to rate the perceived quality of (a) audio, (b) video and (c) combined audiovideo, which took approximately two minutes for each setting.

Experiment B: Study the effects of lip sync on multimedia quality

For this experiment, the same test-bed configuration as shown in Figure 1 were used. The test candidates were asked to qualify a detected synchronization error (while viewing and listening to a 'talking head') in terms of four different categories, i.e. (a) audio is ahead of video, (b) audio is behind video, (c) cannot tell, if audio is ahead or lagging and, (d) no synchronization error. The subjects were also required to give the MOS for the perceived quality of audio, video and audiovideo overall.

For each test, a delay (ranging from 40msc to 440msc) is introduced separately to the audio and video streams, in random order. A step of 40 minutes interval was selected due to the fact that multimedia software and hardware are capable to refresh motion video data every 33ms/40ms. Each test lasted for approximately one minute and the whole section took not more than 30–40 minutes to complete.

The method of assessment being used in both experiments (i.e. Experiment A and B) is the subjective test method, called Mean Opinion Score (MOS) which is the standard recommended by the International Telecommunications Union (ITU-T, 1984). The MOS is typically a 5-point rating scale, covering the options Excellent (5), Good (4), Fair (3), Poor (2) and Bad (1).

The perceived quality of audio and video over one conference is affected by different network factors (e.g. packet loss), hardware (e.g. headset), CPU power, CODEC, task performance, background noise and lighting, and loading on the individual's workstation. Therefore, in the experiments, maintaining the above variables constant (for both end users), except packet loss and delay for Experiment A and Experiment B, respectively, is vital to ensure the validity of the results.

Current Internet-based solutions for multimedia conferencing involve the use of separate TCP/RTP sessions for the audio and video signals (Schulzrinne *et al*, 1996). In the experiments, a network emulation tool (NISTNet) is used to introduce different sets of impairments (packet loss or delay) on each audio and video stream (for example, audio is degraded by 5% packet loss while video quality is unimpaired or vice versa).

Results and Observations

All the figures below show the results obtained from the test and the observations made are described in this section. Figure 2-11 show the result obtained from Experiment A, while Figure 12 is deduced from Experiment B.

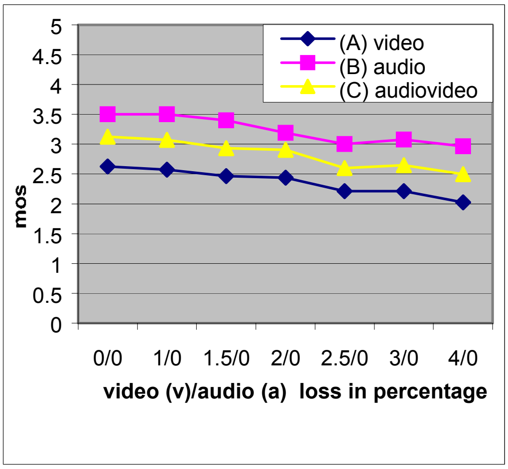


Figure 2: Interactive test – Video Degraded; Audio Constant

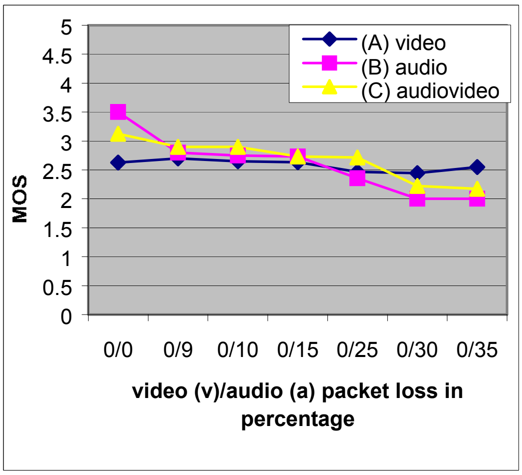
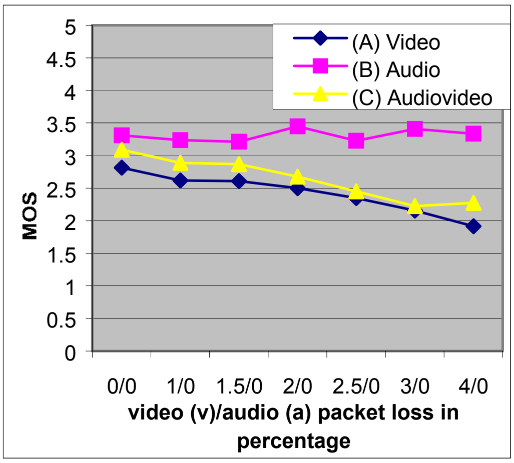


Figure 3: Interactive test – Video Constant; Audio Degraded

Figures 2 and 3 show MOS of packet loss impact on the perceived quality of (A) video, (B) audio and (C) combined audiovideo as obtained in interactive test. It can be seen that when video is degraded, audio scores also decreased by 0.5 (MOS), for video packet loss in the range of 1%-4%, even though the audio quality was kept constant. However, the MOS for video, while its quality being held at constant (i.e.0 % loss), is not affected by the change in audio quality. The rating for video stays at ± 2.6 (MOS) for audio loss raging from 9%-35%. However, the MOS for the perceived quality of combined audiovideo for both test scenarios is approximately the same, i.e. ± 0.1 (MOS) difference, when audio loss is below 30% loss. The score for combined audiovideo drops by 0.4 (MOS) upon reaching 30% audio loss and above. This implies that good audio is critical in interactive test.



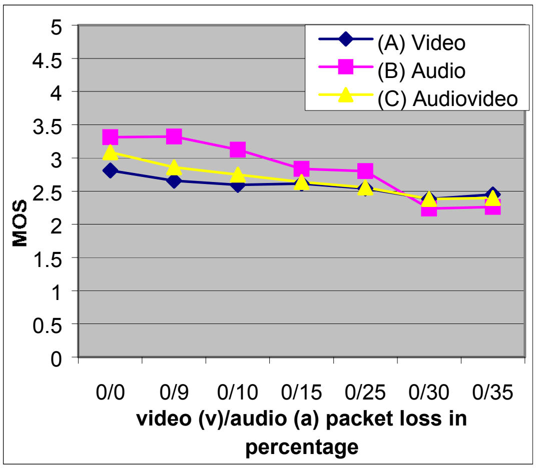
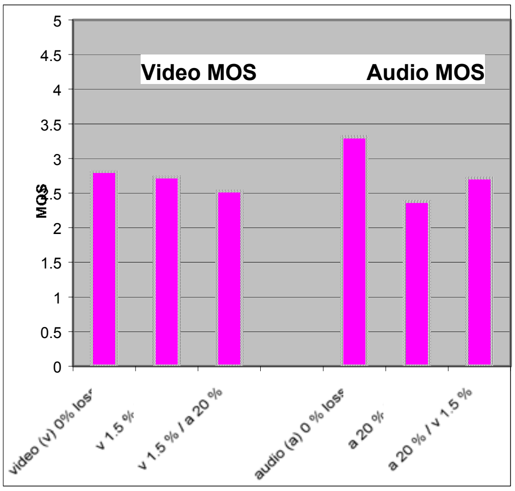


Figure 4: Passive test – Video Degraded; Audio Constant

Figure 5: Passive test – Video Constant; Audio Degraded

Figures 4 and 5 show MOS of packet loss impact on the perceived quality of (A) video, (B) audio and (C) combined audiovideo as obtained in passive test. Unlike the interactive test, the MOS for audio is not affected by the degradation in video (see figures 2 and 4). Also, by referring to Figure 5, there is slight drop in video score, i.e. 0.36 (MOS), when audio loss ranging from 0%-35%. The MOS for combined audiovideo is affected severely by the change in video loss as compared to audio loss.



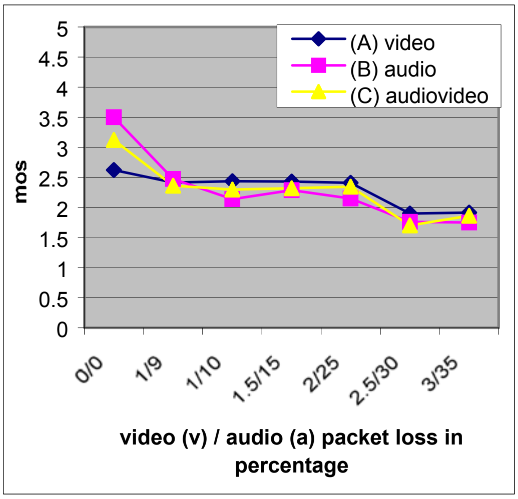
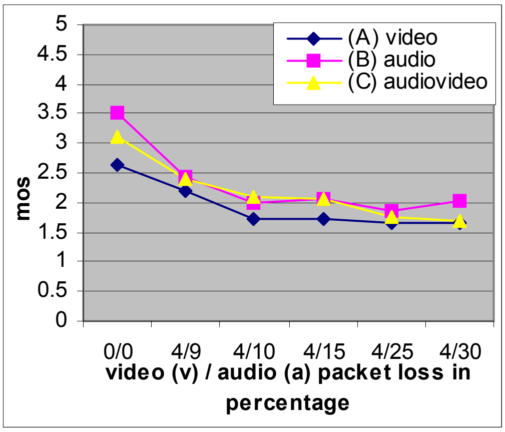


Figure 6: Passive test – Packet Loss Impact on Audio and Video

Figure 7: Interactive test – Audio and Video Degraded

Figure 6 above shows the MOS results of the perceived audio and video quality, indicating the impact of having audio only or video only and comparing these results that when audio and video are both present during the test. The result indicates the strong interaction dependency between audio and video. It is revealed that the perceived quality of audio increases with the presence of video. For example, for 20% audio loss (the 5th column in Figure 6), the MOS is 2.3 without the presence of video. However, with the presence of video, the same audio sample gives MOS rating of 2.7 (final column in Figure 6). This indicates that video information enhances speech only communication. On the other hand, perceived video quality degrades when poor quality of audio was present. Another example, Figure 2 shows how perceived audio quality (for a specific audio condition) changes as the video quality deteriorates. When the video quality is high (0% loss), the audio MOS is 3.5, and when the video quality is poor the audio MOS is 2.9, even though the actual audio quality used is unchanged. This shows that video is an important determinant to justify multimedia quality.

Figure 7 shows the effect of packet loss on the perceived multimedia quality as observed in the interactive test. By comparing this result with that in Figure 2 (video constant; audio degraded), it is evident that the audio score gives higher rating with good video (i.e. 0% loss), even though the audio was degraded by the same amount of loss through out the test.



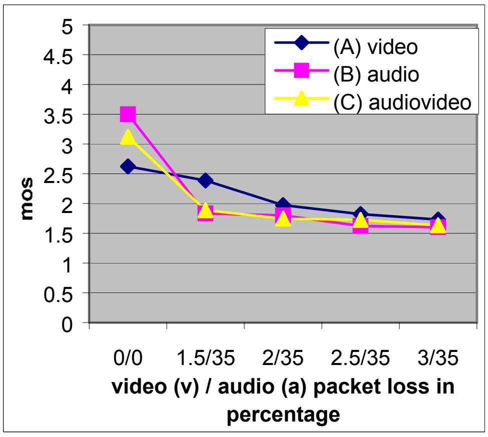


Figure 8: Interactive test – Video Poor (4% loss); Audio Degraded

Figure 9: Interactive test – Video Degraded; Audio Poor (35% loss)

Figures 8 and 9 show the MOS rating of the perceived quality of video, audio and combined audiovideo with respect to high video loss (4%) and high audio loss (35%). Figure 9 shows that, when audio is very poor, interactive test scores very low MOS for the perceived multimedia quality. Hence, interactive test severely depends on sufficient audio quality.

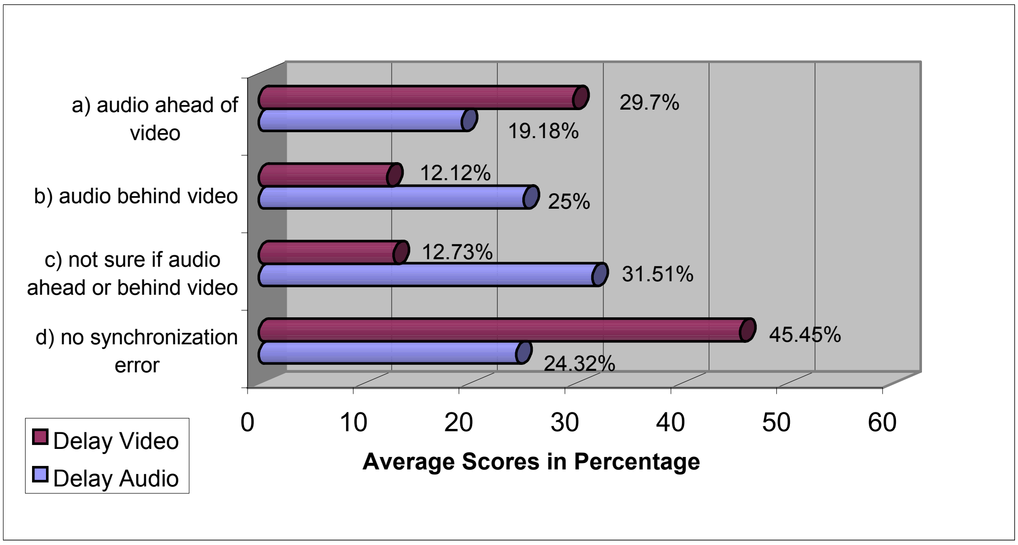


Figure 10: Effects of Delay Video or Audio on Lip Sync

Figure 10 shows the subjective average scores of lip sync effect when audio or video was delayed, obtained in Experiment B. From the observation, majority of the subjects indicated no synchronization error occurred (i.e. 45.45%), when video stream was delayed up to 440msc. However, 29.7% of the subjects noticed that audio was played ahead of video, while 12.12% stated otherwise. A number of 12.73% claimed that they are not sure if audio is ahead of video or behind video.

For the test where Audio was delayed up to 440msec, 25% of the subjects noticed that audio was played behind video and only 19.18% of them claimed that audio was ahead of video. Whereby, a number of 31.51% of the subjects were not sure if audio was ahead of video or behind video and 24.31% stated that there was no synchronisation error.

Conclusions and Future Work

The results concluded that there is strong interaction independency between audio and video media. For example, it can be seen that the MOS of audio increases with the presence of video. It is also observed that, video adds value to a conference and enhances interactivity. Thus, it is evident that video is an important determinant to justify multimedia quality. As in the case of interactive test, video scores are not affected by the audio quality, whilst audio scores deteriorated as video is degraded. Therefore, it is justified that, the importance of video at the expense of audio cannot be underestimated, as video has a psychological effect on interactive communications, such as for interruptions, naturalness, interactivity, feedback and attention.

From the observation, the sensory interactions, and the attention given to a particular aspect of performance, are clearly content-dependent, i.e. if a person is reading text from a screen, the quality of the audio has little significance; likewise, if a person is casually chatting (interactive communication), the quality of the video is of less important of than that of the audio. This finding also confirmed with the previous research result which that states subjects are less susceptible to poor video in interactive communication, i.e. users did not report the difference between 12 and 25 frames per second (fps) when involved in an engaging task (Anderson *et al*, 2000).

The results also suggested that, increase in task difficulties have the effect of decreasing the subjective video and audio quality. For example, in passive test, where user are required to understand the read material, the overall scores for the combined audio and video quality in passive test are much lower than that in interactive test.

A number of problems were encountered while conducting the tests. For example, task performance was dynamically varying. This could lead to varying in frame rates that could result in inconsistent in image degradation. Also, subjective quality evaluation in the prolonged field trial approach suffers from the problem of lack of control over a large variety of variables, such as different lighting levels, inconsistent task performance, the different of sensory and perceptual ability of subjects to identify errors in the perceived audio and video signal, and possibly the expected emotional state of a subject, etc.

At the time writing this paper, the results obtained from Experiment B were incomplete to deduce a more comprehensive observation and conclusion. Further analysis will be carried out, such as, to investigate the perceptions of the subjects in terms of MOS for the given fours different categories of answers, to find out the maximum threshold for delay tolerance for specific task performance, to justify if audio ahead of video is more tolerated or vice versa and so on. The continuing work in this area is to conduct similar test on varieties of task performances, i.e. interactive communication (person-to-person) and animation.

The future approach is also to investigate how audio and video degradation can affect subjective evaluations of audio/video quality with respect to different duration, intensity and frequency of error occurred in a single event. As we already justified that the quality requirements for audio and video will be task dependent, work is also needed to specify more precisely the set of tasks for which video information is useful and vide-versa.

References

Anderson, A.H., Smallwood, L., MacDonald, R., Mullin, J., Fleming, A. and O'Malley, C. (2000) Video data and video links in mediated communication: what do users value*? International Journal of Human-Computer Studies,* 52(1), 165-187.

Carson, M. (2000), NIST Net Home Page, <URL: http://snad.ncsl.nist.gov/itg/nistnet/ >

Finholt, T., Sproull, L and Kiesler, S. (1990), "Communication and Performance in Ad Hoc Task Groups". In J. Galegher, R.E. Kraut, and C. Egido, editors, *Intellectual Teamwork*, pages 291--325, Hillsdale, NJ, Lawrence Erlbaum Associates, Inc.

Hardman, V., Sasse M. A. and Kouvelas, I. (1998), "Successful Multiparty Audio Communication over the Internet", *Communications of the ACM*, Vol. 41(5), pp 74-80.

Isaacs, E. and Tang, J. (1994), 'What Video Can and Cannot Do for Collaboration: A Case Study', *Multimedia Systems 2*, pp 6-73.

International Telecommunications Union (ITU) (1996), Methods for Subjective Determination of Transmission Quality, *Recommendation P.800*, ITU-T.

Kawalek, J (1995), "A User Perspective for QoS Management", *Proceeding of 3rd International Conference on Intelligence in Broadband Services and Network*, IS & N 1995, Crete, Greece.

Kitawaki, N. & Nagabuchi, H. (1998), "Quality Assessment of Speech Coding and Speech Synthesis Systems", *IEEE Communications Magazine*, October, 1998, pp.36-44.

Schulzrinne, H., Casner, S., Frederick, R. and Jacobson, V. (1996), RFC 1889: RTP for Real Time Application, Audio-Video Transport Working Group.

Sellen, A. J (1992), "Speech Patterns in Video-mediated Conversations", *Conference proceedings on Human Factors in Computing Systems*, Monterey, California, US, pp 49-59.

Tang, J. C. & Isaacs, E. A. (1993), Why Do Users Like Video: Study of Multimedia Supported Collaboration", *Computer Supported Cooperative Work 1*, pp163-196.

Watson, A. and Sasse, M.A. (1996) " Evaluating Audio and Video quality in Low-Cost Multimedia Conferencing Systems," *Interacting with computers*, 8, pp. 255-275.

Wilson, G. and Sasse, M. A. (2000a), "Do Users Always Know What's Good For Them? Utilising Physiological Responses to Access Media Quality", In S. McDonald, Y. Waern & G. Cockton [Eds.]: People and Computers XIV - Usability or Else! *Proceedings of HCI 2000* (September 5th - 8th, Sunderland, UK), pp. 327-339. Springer.

Wilson, G. and Sasse, M.A. (2000b), "Investigating the Impact of Audio Degradations on Users: Subjective vs. Objective Assessment Methods". In C. Paris, N. Ozkan, S. Howard & S. Lu (eds.) Proceedings of OZCHI 2000: Interfacing Reality in the New Millennium, pp135-142, December 4th - 8th, Sydney, Australia