Interpolation of Packet Loss and Lip Sync Error on IP Media

Licha MUED, Benn LINES and Steven FURNELL
Network Research Group, University of Plymouth
Plymouth, Devon, United Kingdom

ABSTRACT

The work presented in this paper, outlines the test conducted to investigate the important factors that define the percieved multimedia quality in desktop videoconferencing, such as packet loss, delays and lip synchronisation (lip sync). The work focuses upon investigating the effects of lip sync as well as packet loss, on the perceived quality of audio only, video only and audiovideo overall, using the subjective test method, known as Mean Opinion Score (MOS). The test has been design based upon five (5) different categories as explained in the Experimental Design and Method section. The results obtained from the experiments are presented in the Result section, followed by the discussion of the findings, in the subsequent heading. The study has suggested that, the subjects were less susceptible to poor video and, hence lip sync while engaged in the interactive communication, as opposed to the passive communication. Therefore, different task performed by end user required different level of multimedia quality. It is also concluded that the perceived quality of one media (e.g. audio or video), interacts and influences the perception of the other.

Keywords: Packet Loss, Lip Sync, Delay, MOS, Audio and Video Quality, Interactive and Passive Communications.

INTRODUCTION

Desktop Videoconferencing (DVC) offers the opportunity to develop a global multimedia communication system and will become mainstream both professionally and personally. Despite its increased popularity, the current low cost DVC is facing a challenge as it is often questioned whether the quality of the audio and video provided is adequate to perform the required task performance. This is because, the IP networks are not designed to support real-time applications and factor such as network constraints and lips synchronisation error lead to unpredictable deterioration in the perceived Quality of Service (QoS).

Packet loss i.e. the number of lost packets, reported at the total traffic could cause interrupted speech that leads to 'bubbly' sound. It has been claimed that, a packet loss of 2% is acceptable to obtain tool quality speech. Delay is defined as the time passed between the sending of a packet and its arrival at the destination. For delay more than 450ms, the nature of interaction is clearly awkward and generally considered less than satisfactory [1]. Like audio, video is also sensitive to delay, although, there is no distinctive figure to justify the accepted delay of video in multimedia conferencing. 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 to determine the quality of service in multimedia applications [2]. Current desktop videoconferencing systems transmit between 2 and 8 frames of video per second [3] (Quarter Common Interchange Format, QCIF–176x144 pixels/ Common Interchange Format, CIF–352x288 pixels), with poor resolution and unsynchronized audio and video. It is claimed that, the frame rate should exceed 8 frames per-sec to achieve substantial lip sync. To date, a lot of work has been focused on implementing new techniques and approaches to minimise lip sync error [4][5].

There are numerous factors that can influent user's perception of audio quality, such as loudness, intelligibility, naturalness, pleasantness of tone and listening effort [6]. While for video, dress/background, lighting, frame rate, packet loss, field of view, size of image, 'blockiness', and degree of lip sync are the important factor s to determine its quality [7].

The work presented in this paper, outlines the test conducted to investigate the important factors that define the percieved multimedia quality in desktop videoconferencing, such as packet loss, delays and lip synchronisation (lip sync). The work focuses upon investigating the effects of lip sync as well as packet loss, on the perceived quality of audio only, video only and audiovideo overall, using subjective test method. The test has been design based upon five (5) different categories as explained in the Experimental Design and Method section. The subjective rating method, known as the Mean Opinion Score (MOS) has been employed for the test [8].

EXPERIMENTAL DESIGN AND METHOD

The experimental design can be described into five (5) sections, as follows:

Section 1: Passive Test, i.e. listening and viewing to 'talking head'
Section 2: Interactive Test, i.e. informal interactive conversation (one-to-one person)
Section 3: Interactive Test, with the introduction of packet loss
Section 4: Lip Sync Test (4 category rating method)
Section 5: Controlled Experiment, i.e. test under ideal network condition

Prior to transmission, for each test section, except for Section 5, a delay within the range of 40-520 ms was randomly introduced, separately to the audio and video streams. A step of 40 ms interval was selected due to the fact that multimedia software and hardware are capable to refresh motion video data every 33/44 ms. Each test step lasted for approximately one minute and one test section would be completed in 30-40 minutes.

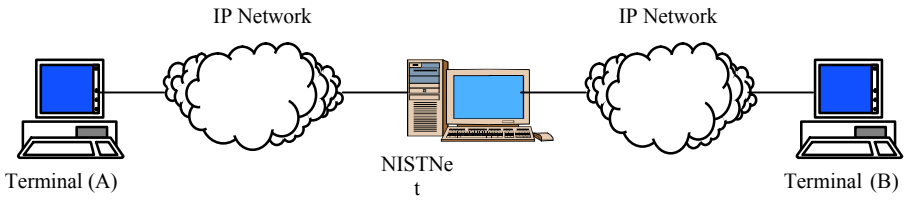


Figure 1: Test Bed Configuration

In Section 3, apart from the delay, the packet loss was also interpolated to the separate audio and video streams, randomly. For audio, packet loss of 5%, 10%, 15% and 20% were selected and 1%, 1.5% and 3%, for video.

In the experiments, a network emulation tool (NISTNet) [9] is used to introduce the different sets of impairments, i.e. packets delay and loss, on each audio and video stream, randomly. Hence, different levels of lips sync and packet loss impairments were produced.

In Section 4, the test candidates were required to classify the perceived synchronization error based upon four (4) different categories, i.e. (a) audio is ahead of video, (b) audio is behind video, (c) not sure, whether audio is ahead or lagging video, and (d) no synchronization error. The result, based upon the percentage of students responding in each category is shown in Graph 7.

In Section 5, as a common reference, the subjects were introduced to the perceived quality of audio and video where the media data were sent in the ideal network condition, i.e. without loss, delay jitter, delay and no lip sync error.

At the receiving end, the subjects were asked to evaluate the perceived quality of (a) audio, (b) video and (c) combined audiovisual components. The method of assessment being used is the subjective test method, called the Mean Opinion Score (MOS), which is the standard recommended by the International Telecommunications Union, ITU-T P800. It is a 5-point rating scale, covering the options EXCELLENT (5), GOOD (4), FAIR (3), POOR (2) and BAD (1).

The 38 subjects were mostly students (of multiple nationalities) of the University of Plymouth, aged between 18-35 years old. The two communicative parties selected were already acquainted (and thus fully at ease with one another) to maximise the task being performed. This is vital to ensure the validity of the results. For the same reason, in the case of the interactive test, the subjects were allowed to select their own issue for discussion. The tests were undertaken based upon the terms and condition stated in International Telecommunications Union, ITU-R P500 [10].

Two identical processors, Pentium 200 MHz (64.0MB RAM), were used. The Quarter Common Information Format (QCIF-176x144) frame size was use as the Common Information Format (CIF-325x288) provided an almost still-like picture. The video setting was unchanged throughout the test, i.e. 'better quality' video and the H.263 (p x 64Kbit/s, p = 1 to 30), video CODEC was used [11]. For the audio CODEC, we used G723.1, 6400bit/s [12]. Microsoft NetMeeting (Version 3) [13] was selected over the other existing IP telephony tools due to its readily available software and its popularity in the current market. Figure 1 above depicts the AVoIP (Audiovideo over IP) test bed configuration used for the experiments.

Variables that would cause inconsistence in the subjective test result, such as different room lighting levels, background noise and task performance were kept to minimum [14]. The test candidates were also trained to maintain their movements throughout the test to minimise dynamic variation in frame rates that could lead to inconsistent in image degradation.

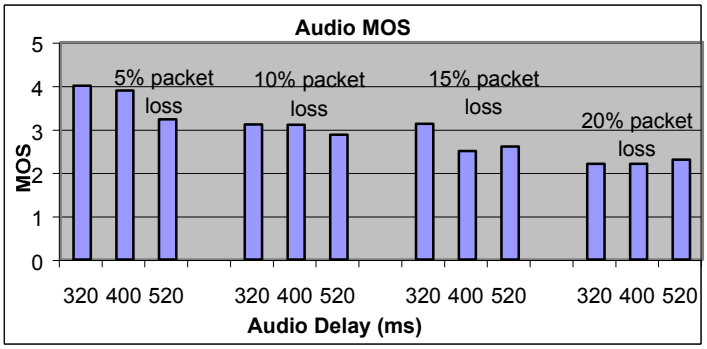
RESULT

Graph 1 shows the perceived audio MOS for the Interactive Test, for audio packet loss of 5%, 10%, 15% and 20%. The MOS are ranging from 4 (GOOD) and 3 (FAIR) when audio loss 5% and drops to around 3 MOS, for audio loss of 10%. At 15% audio loss, the MOS for audio are between 3 and 2.5. However, at 20% audio loss, the scores are around 2.2, which are approaching the POOR threshold i.e. 2 MOS. Notice that, the audio delay has no significant effect on the MOS as the audio loss is reaching 20%. The conclusion is that, at 20% audio loss the audio quality was so poor that it was difficult to evaluate the perceived quality, precisely. The MOS at this stage is claimed to be around 2.5 and below.

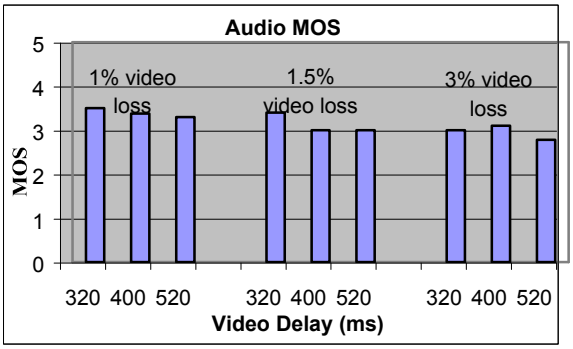
Graph 2 shows the MOS of the perceived audio for the Interactive Test, for video packet loss of 1%, 1.5% and 3%. It can be seen that the degradation of video quality, due packet loss and delay, has a significant impact on the perceived Audio quality. At 1% video loss, the MOS drop from 3.5 to 3.3, i.e. above FAIR quality. The MOS drops to around 3 (FAIR) for video loss of 1.5% and above.

The test candidates claimed that evaluation of audio quality is very straightforward and the distortions could be easily detected as opposed to video. It is observed that, the assessment of video quality is very difficult and complicated since the degrees of deteriorations are constantly changing.

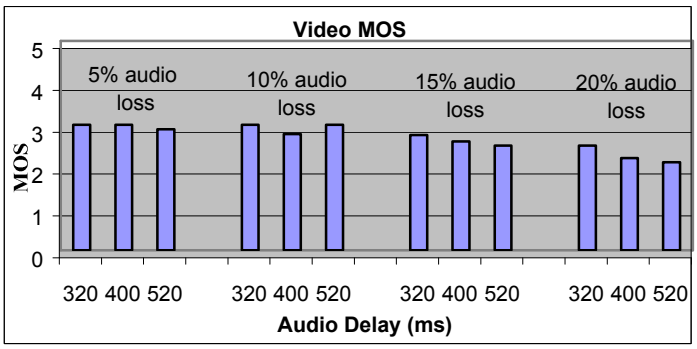
Graph 3 shows the perceived Video MOS for the interactive test for audio packet loss of 5%, 10%, 15% and 20%. The MOS are around 3 (FAIR) for audio loss of 5% and 10%. It was observed that the video MOS decrease as the audio packet loss increases, i.e. from 15% to 20%, while the quality settings of the video stream was unchanged, prior to transmission. Hence, it is concluded that the perception of video MOS is affected by the quality of audio, i.e. the subject opinion of perceived quality of video is degraded in relative to the increased deterioration of the audio quality.



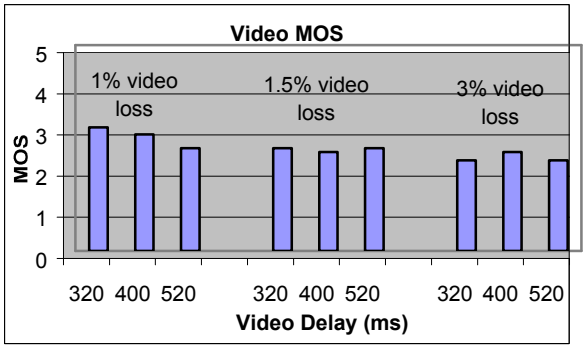
Graph 1: Audio MOS Vs Audio Delay and Loss



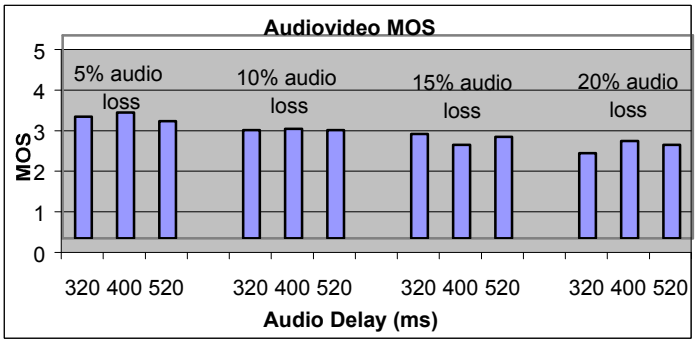
Graph 2: Audio MOS Vs Video Delay and Loss



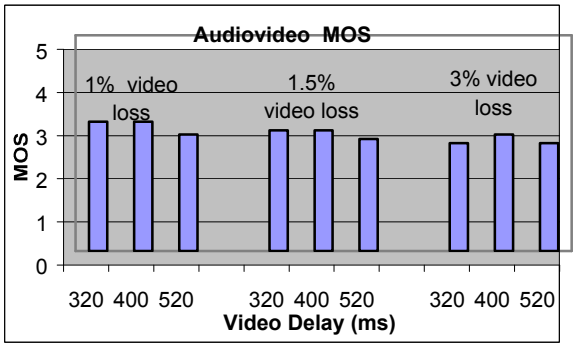
Graph 3: Video MOS Vs Audio Delay and Loss



Graph 4: Video MOS Vs Video Delay and Loss



Graph 5: Audiovideo MOS Vs Audio Delay and Loss



Graph 6: Audiovideo MOS Vs Video Delay and Loss

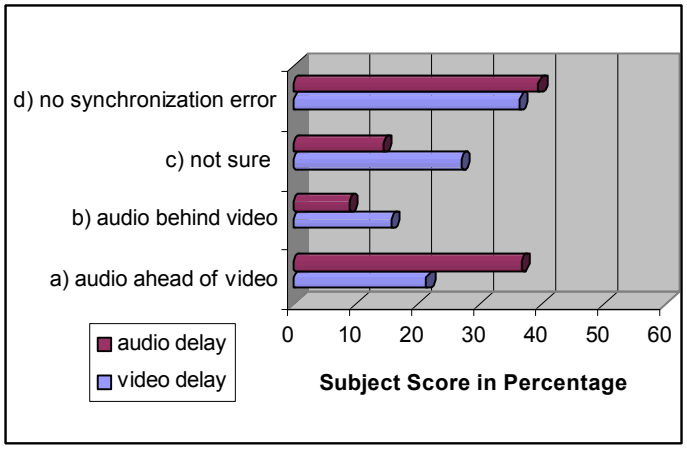
Graph 4 shows the MOS of the perceived quality of video for video loss of 1%, 1.5% and 3%. It has been noticed that, for 1% video loss there is a gradual degradation of the perceived video score as the video delay increases from 320ms to 520ms. It is al so suggested that the result becomes less meaningful when the video loss increases i.e. from 1.5% to 3%, where the MOS of 2.5 has been reached.

Graph 5 shows the perceived combined audiovideo MOS for the interactive test, for audio loss of 5%, 10%, 15% and 20%. There has been no significant effect of audio delay (i.e. between 320ms, 400ms and 520ms) on the perceived quality of audiovideo overall. For 5% audio loss, the average MOS is around 3 and it drops to around 2.6 and below, when the audio loss exceeds10%.

Graph 6 shows the MOS of the perceived quality of audiovideo for the Interactive Test, for video loss of 1%, 1.5% and 3%. The score for the perceived quality for audiovideo are slightly higher than that for video. At 1.5% video loss, the MOS of the perceived audiovideo quality are degraded gradually, with respect to video delay (i.e. from 320ms to 520ms). At 3% video loss, the average MOS is around 2.5, which is POOR. On the other hand, the overall score is higher that that of the MOS of audiovideo where the audio losses and delays were introduced.

* Int. Test (A) = no packet loss Int. Test (B) = with packet loss	Passive Test		Int. Test (A)		Int. Test (B)	
	Video delay	Audio delay	Video delay	Audio delay	Video delay	Audio delay
a) audio ahead of video	19.2	29.7	36.6	28.6	21.5	36.8
b) audio behind video	25.0	12.1	25.6	14.9	15.9	9.2
c) not sure	31.5	12.7	20.9	32.3	27.1	14.5
d) no synchronization error	24.3	45.4	16.8	24.2	36.4	39.5

Table 1: Lip Sync Test - Scores in Percentage



Graph 7: Lip Sync (Interactive Test) Vs Packet Loss and Delay

Table 1 below shows the number of scores of the test candidates, based on the four (4) categories rating in the passive and the interactive test (with and without packet loss), versus variable delays.

The passive test, gives more accurate result, i.e. when audio was sent ahead of video, 29.7% of the subjects stated that audio is ahead of video, while only 12.1% noticed that audio is lagging video. When video was sent ahead of audio, 25% candidates scored correctly, but 19.2% of them claimed that audio is ahead of video. However, the majority of the subjects, i.e. 45.4% indicated that there was no synchronisation error for the test when video was delayed, in the passive test.
Likewise, in the interactive test as shown in the Int. Test (A) column (Table 1), a higher percentage of participants noticed the synchronisation error, i.e. 32.3% for video delay and 20.9% for audio delay. However, majority of them were giving the wrong answer or not sure if audio is ahead or vice-versa. For example, in the case where audio was sent behind video, a number of 36.4% of the subjects indicated otherwise, i.e. audio ahead of video.

Graph 7 shows the results for the lip sync test of the interactive test, with respect to packet loss and delay, as indicated in Table 1, in Int. Test (B) column.

CONCLUSION

A number of subjects claimed to notice the lip sync error but having the difficulties to distinguish between the perceived audio and video delay. The majority of them were not sure whether audio was played ahead of video or vice versa, especially in the interactive test. The subjective test has shown more correct outcomes. It has been observed that, audio that is not synchronized with video can be distracting or appeared strange due to loss of lip synchronisation. However, despite experiencing varying lip sync error (without the introduction of packet loss), the MOS of the subjects remain almost constant throughout the test, i.e. between FAIR and GOOD quality. On the other hand, a large number of students (approx. 40%) rated the same range of scores although they stressed that there is no lip sync error. The finding also suggested that, the perceived multimedia quality was not affected even though the delay goes as high as 520 ms, when there is no packet loss occurred. Hence, it is concluded that, in application scenario where the subjects are having an informal conversation and that they are well acquaintance with one another, lip sync error is not a critical issue. This finding is contradicted with the ITUG.114 Recommendation [1], which stated that audio delays should be kept less than 200 ms, for effective interaction.

In the experiments where both packet loss and delay were introduced, the multimedia perceptual scores decreased as the packet loss increased. At 3% video loss, the viewer described

that the video quality suffered from severe impairments, such as 'blocky' and blurring, as a result of partially upgrading parts of the video image. While for audio, at 20% packet loss, the perceived quality suffered from glitches, feedback and became less intelligent. It is agreed that at 2.5 MOS and below, the result has no meaningful term.

The results also concluded that the perceived video quality degraded, when poor quality audio was detected. Hence, it is concluded that the perceived quality of one media is affected by the perceive quality of the other. The result also justified that good audio quality is essential to determine the multimedia quality. The subjects were less susceptible to lip sync error while engaged in the interactive communication, as opposed to the passive communication. By comparing the effects of audio and video delay on the perceived multimedia quality, separately, in both passive test and interactive test, video delay has shown higher MOS throughout the test. This indicates that video delay has less significant effect on the viewers. It is considered that, the designated task performances have low video's temporal aspect and hence, the subjects may not notice the delayed or missing frames. From the observation, it has been suggested that the video media is mainly used to enhance psychological effects, such as for attention, naturalness, interactivity as well as a mean of assurance that the opposite party is actually presence.

The major drawback of the test experiment is that the subject may not be well trained to perform the task performance exactly as required, to obtain a dynamic result. Furthermore, subjective test result in the prolonged field trial method is susceptible to the lack of control over a large variety of variables, both internal and external. [15]

Future work will involve a comprehensive evaluation of achievable audio and video quality, following the experimental design described in this paper, to investigate the effect of jitter and the combination of delay, jitter and packet loss. The effect of these factors on the task performance with a high temporal aspect of video, such as animation will also be carried out. The work presented in this paper will eventually lead to the characterisation of the factors that define the perceived multimedia quality. The understanding of these effects is essential and beneficial for the network developer and provider to optimise the perceptual quality of audio and video in desktop videoconferencing systems.

REFERENCES

[1] **ITU-T Recommendation G114**, "One Way Transmission Line: Implementors' Guide No. 1 For Recommendation G.114", Study Group 12, February 2001.

[2] Raft Steinmetz, "Human Perception of Jitter and Media Synchronization", **IEEE Journal on Selected Areas in Communications**, Vol.14 No.1, January 1996.

[3] Steve Rudkin, Andrew Grace, and Mike Whybray, "Real-time Application On The Internet", **BT Journal**, Vol 15 no. 2 April, 1997.

[4] T. Ohmori, K. Maeno, S. Sakata and K. Watabe, "Cooperative Control for Sharing Application Based on Distributed Multiparty Desktop Conferencing System: MERMAID2, SuperCOMM/ICC'92", **Discovering a New World of Communications**, Vol.2, pp. 1069-1075, Jun.1992.

[5] K. Ravindran and V. Bansal," Delay Compensation Protocols for Synchronisation of Multimedia Data Streams", **IEEE Trans. On Knowledge and Data Engineering**, Vol. 5, No.4, pp 574-589, Aug. 1993.

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

[7] Gili Manzanaro, J., Janez Escalada, L., Hernandez Lioreda, and M.,Szymanski, "Subjective Image Quality Assessment and Prediction in Digital Videocommunications", **COST 212 HUFIS Report**, 1991.

[8] **ITU-T Recommendation P.800**, "Methods for Subjective Determination of Transmission Quality".

[9] Carson, M., **NIST Net Home Page**, <URL: http://snad.ncsl.nist.gov/itg/nistnet/ >

[10] **ITU-R Recommendation BT. 500-7**, "Method for the Subjective Assessment of the quality of Television Pictures, RBT".

[11] Guy Cote, Berna Erol, Michael Gallant and Faouzi Kossentini, "H.263+: Video Coding at Low Bit Rates, **IEEE Transactions on Circuits and Systems for Video Technology**", Vol.8.No.7, November 1998.

[12**] ITU-T Recommendation G.723.**1, "Dual Rate Speech coder for Multimedia Communication Transmitting at 5.3/6.3 Kbps", March 1996.

[13] **Microsoft NetMeeting Home Page**<URL: http://www.microsoft.com/windows/netmeeting/>

[14] L Mued, S Furnell, and B Lines, "Performance Evaluation of Desktop Videoconferencing", **the Proceedings of PGNET 2001**, Liverpool John Moores University, UK, 18th –19th June 2001.

[15] Mued, L., Lines, B., Furnell, S. and Reynolds, P, (2002) "Investigating the Interaction Effect of Audio and Video as Perceived in Low Cost Videoconferencing", **the Proceedings of the Third International Network Conference (INC 2002)**, Plymouth, UK, 16-18 July 2002.