THE EFFECTS OF LIP SYNCHRONIZATION IN IP CONFERENCING

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

University of Plymouth, United Kingdom
E-mail: lmued@jack.see.plymouth.ac.uk

INTRODUCTION

In multimedia IP conferencing, audio and video are separate streams of data, routed separately through the network. Packets that are transmitted simultaneously are not guaranteed to arrive at the same time at their destination, and hence, cause lip synchronization (lip sync) error.

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. However, it is difficult to obtain lip sync in IP conferencing systems as the frame rates obtained are generally very low, i.e. 2-5 frame per-sec, Rudkin et al (1). The frame rate should exceed 8 frames per-sec to make lip sync a meaningful term.

In addition to frame rate, the various factors that can affect lip sync are, network traffic (packet loss, delay jitter and delay), CPU activity (like launching and closing other applications while running the videoconference), and other task operations (e.g. T.120 data operation). Typically, data packets are sent at higher priority than video packets and consume some of the communication bandwidth, and hence, cause some reduced frame rate and loss of synchronisation.

It is claimed that audio may be played up to 120 ms ahead of video, whilst video can be played up to 240ms ahead of audio, Steinmetz (2). This is due to the fact that, people are more tolerant to audio lagging video, rather than vice-versa, because they are more used to perceiving an event before they hear it i.e. light travel faster than sound. Ideally, before the lip sync error becomes apparent, audio should be synchronized within +/- 90 ms of the video, with a maximum range of +/- 160 ms, Steinmetz (2). Also, it is indicated that out of sync is perceived when the mismatch time between audio and video exceeds 80 to 100 ms, Jardetzky et al (3). Audio delay above 400 ms, would compromise the quality of two-way communication in IP conferencing.

To date, many new techniques and approaches have been implemented to minimize lip sync problems, Ravindran (4)

This paper focuses upon investigating the effects of lip sync on the perceived quality of audio and video in desktop videoconferencing, over two different task performances namely passive communications and interactive communications. This is because, it has been stated that different tasks performed by the end user will require different levels of audio and video quality, Finholt et al (5), Mued et al (6).

The study shows a comprehensive subjective evaluation of achievable multimedia quality undertaken based upon different set of impairments i.e. packet delay between audio and video, prior to transmission. The test has been design to investigate the impact of lip sync error on the perceived quality of audio only, video only and audiovideo overall, using subjective test method. Previous research stated that, different component media, especially audio and video, interact and influence the perception of each other, Mued et al (6). Therefore, it is suggested that the combined audio and video quality needs to be considered.

OUTLINE OF EXPERIMENTS

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 (7).

Two identical processors, Pentium 200 MHz (64.0MB RAM), were used. The Quarter Common Information Format (QCIF-176x144) frame size was use as 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 video CODEC was used. For the audio CODEC, we used G723.1, 6400bit/s.

Microsoft NetMeeting (Version 3) was selected over the other existing IP telephony tools due to its readily available software and its popularity in the current market. Figure 1 below depicts the VoIP (Voice over IP) test bed configuration used for the experiments.

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

At the receiving end, the subjects were asked to evaluate individual the quality of audio and video components and the combined audiovisual quality, in terms of MOS. The method of assessment being used is the subjective test method, called Mean Opinion Score (MOS) which is the standard recommended by the International Telecommunications Union, ITU-T P800 (8). It is a 5-point rating scale, covering the options Excellent (5), Good (4), Fair (3), Poor (2) and Bad (1).

The test candidates were also required to classify a perceived synchronization error based upon 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 results, based upon the percentage of students responding in each category, are shown in Figure 4 and 5.

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. 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.

For each test, a delay within the range of 40-440 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 lasted for approximately one minute and one test section would be completed in 30-40 minutes.

As previously stated, our experiments were based upon investigating the effects of lip sync on the perceived quality of multimedia components (audio, video and audiovideo overall), in two different task performances i.e., Passive Test (listening and viewing 'talking head') in Section 1 and Interactive Test (two communicative parties, casually chatting), in Section 2.

The test scenarios can be clearly described in Table 1.

Audio/Video Delay Set-up	**Section 1**	**Section 2**
Audio (no delay) Video (no delay)	Passive Test	Interactive Test
Audio (delay 40-440 ms) Video (no delay)	Passive Test	Interactive Test
Audio (no delay) Video (delay 40-440 ms)	Passive Test	Interactive Test

TABLE 1- Test Scenario

As a common reference, the subjects were introduced to the perceived quality of audio and video where the data were sent in the ideal network condition i.e. without loss, delay jitter and delay.

RESULTS AND OBSERVATIONS

Figure 2 shows the audiovideo overall MOS, obtained from the interactive test when audio or video streams were delayed from 40 ms up to 440 ms. The MOS were in the range of 2.4 to 3.1, with video delaying less negative effects than audio.

Figure 3 displays the MOS for audio, in both interactive and passive tests. Audio MOS obtained were generally higher, followed by audiovideo overall, while video scored the lowest MOS (see Table 2).

The passive test gives higher MOS values than the interactive test, e.g. by referring to Figure 3 and Table 2, the average MOS for audio in the passive test are 3.5 for audio delay and 3.4 for video delay, whereas in the interactive test the scores are 2.9 for audio delay and 3.13 for video delay. Therefore, passive test was less affected by either audio or video delay. For both passive and interactive tests, video delay has less significant effect on the perceived multimedia quality, i.e. the average MOS obtained from video delay test is much higher, as compared to that of audio delay. This is clearly indicated in Table 2 and Figure 3.

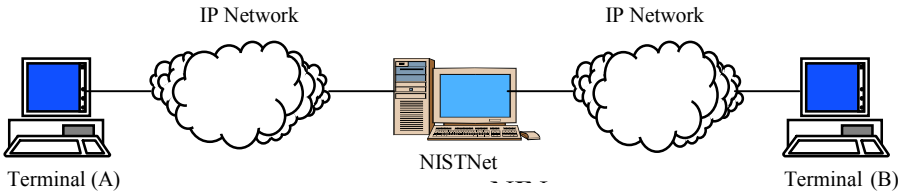

Figure 1: Test Bed Configuration

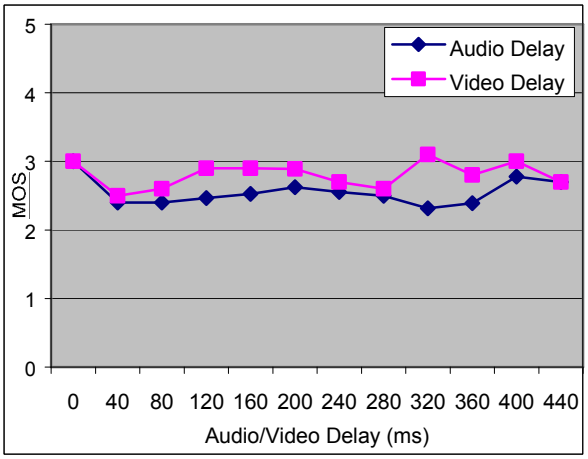

Figure 2: Interactive Test-Audiovideo Overall MOS

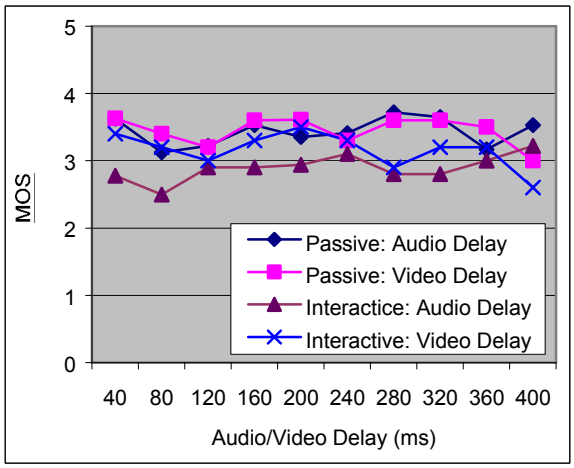

Figure 3: Audio MOS-Interactive Vs Passive Test

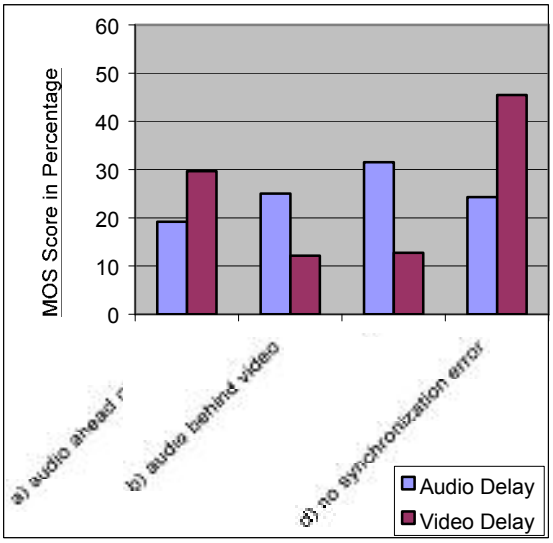

Figure 4: Passive Test – Audio Vs Video Delay

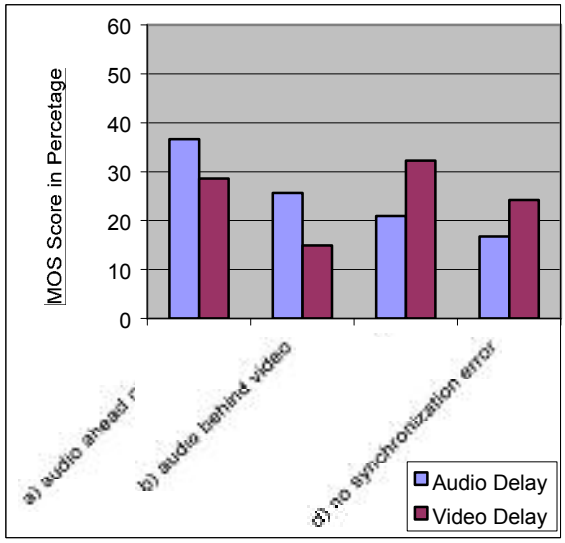

Figure 5: Interactive Test – Audio Vs Video Delay

Media type	Test scenarios	Audio delay MOS	Video delay MOS
Audio	Interactive	2.9	3.13
	Passive	3.5	3.4
Video	Interactive	2.4	2.63
	Passive	2.6	2.89
Audiovideo Overall	Interactive	2.5	2.79
	Passive	2.9	3

TABLE 2 - Average MOS

Figure 4 and 5 show the number of scores of the test candidates, based on the 4 categories rating in passive and interactive test, respectively.

The passive test (see Figure 4), 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 12.12% noticed that audio is lagging video. When video was sent ahead of audio, 25% candidates scored correctly, but 19.18% of them claimed that audio is ahead of video. However, the majority of the subjects, i.e. 45.45% indicated that there was no synchronisation error for the test when video was delayed, in the passive test.

Likewise, in the interactive test (see Figure 5), a higher percentage of participants noticed the synchronisation error, i.e. 32.29% for video delay and 20.94% for audio delay. However, majority of them were giving the wrong answer or not sure if audio was ahead or vice-versa. For example, in the case where audio was sent behind video, a number of 36.65% of the subjects indicated otherwise, i.e. audio ahead of video.

It has been observed that, when audio and video data were delayed separately, in the range of 40-440 ms, the MOS ratings were generally between POOR (2) and FAIR (3). While, GOOD (4) and EXCELLENT (5) ratings were hardly indicated. Moreover, by comparing these results with those when both audio and video were sent simultaneously using the same amount of delay, the latter has shown a higher MOS, as depicted in Figure 6.

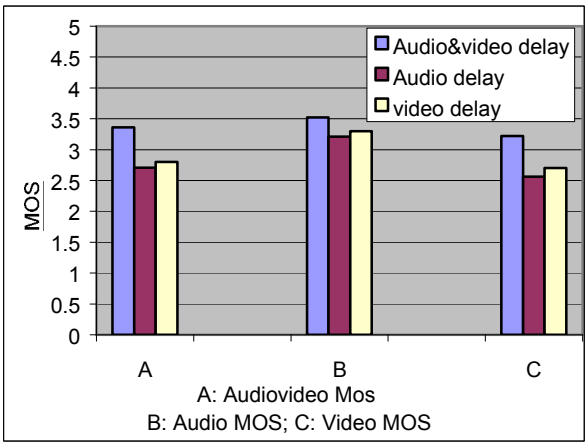

Figure 6: Combined Audio and Video Delay Vs Separate Audio and Video Delay

By referring to Figure 6, in the experiment where both audio and video were delayed by the same amount, the MOS ratings obtained were 3.36 (audiovideo overall), 3.52 (audio) and 3.22 (video); in audio (only) delay test, the ratings were 2.71 (audiovideo overall), 3.21 (audio), and 2.56 (video); and in video (only) delay test, the scores were 2.8 (audiovideo overall), 3.3 (audio) and 2.7 (video). Hence, it has been indicated that lip sync has more impact on the perceptual media quality. All the results in Figure 6 were deduced from the 400 ms delay test experiment.

DISCUSSION AND FUTURE WORK

The results suggested that video delay has less effect on the perceived quality of audio, video and audiovideo overall. This could be due to the fact that, since the facility of video has often been viewed as being of secondary importance to audio, little attention has been given to the changed in video data.

The subjective evaluation of lip sync effect on the perceptual multimedia components is depended on the task performances. Passive test has shown higher MOS throughout the test. It was observed that more attention was given to lip sync in passive test. In addition, a larger number of subjects scored the correct answer in passive test as compared to that of interactive test. More test candidates were not sure whether audio was played ahead of video or vice versa, in interactive test. Perhaps, when they were so involved in the conversations, the mind no longer perceived the lip sync error. Some subjects do not perceive every synchronization error to be annoying and some even go unnoticed. However, the interactive test, in general, scored lower average MOS. It is stated that, two-way communication is more susceptible to delay, and hence, lip sync error.

The impact of lip sync has been proven to be greater than that of delay (i.e. when both audio and video experience the same delay, see figure 6), and hence, has been considered to be a major problem in connectionless packet switched networks.

The MOS for audio, video and audiovideo overall for both audio delay and video delay tests increased when the delays reached 320 ms and decreased above 440 ms delays. Perhaps, at a certain point, media delays are advantages depending on the task performance and CODEC used. This observation will be investigated in the future work.

The results produced so far, were for the tests where the audio and video data were delayed up to 440 ms. Future work in this area is to increase the delays (i.e. >440 ms) until the lips sync can no longer make a 'meaningful term' or the subjective rating drops to POOR (2) MOS level.

REFERENCES

(1) Rudkin S, Grace A, Whybray M, 1997, "Real-time Application On The Internet", BT Journal, Vol 15 no. 2 209-224

(2) Steinmetz R, 1997, "Human Perception of Jitter and Media Synchronization", IEEE Journal on Selected Areas in Communications, Vol.14 No.1, 61-72

(3) Jardetzky P.W, Sreenan C.J. and Needham R.M, 1995, "Storage and Synchronisation for Distributed Continuous Media", Multimedia Systems, 3, 151-161

(4) Ravindran K, 1992, "Real-time Synchronisation of Multimedia Data Stream in High Speed Packet Switch Networks", Workshop on Multimedia Information Systems (MMIS 92), IEEE Communications Society, Tempe, AZ, 164-188

(5) Finholt T, Sproull L and Kiesler S, 1990, "Communication and Performance in Ad Hoc Task Groups". In Galegher J, Kraut R.E and Egido C, editors, Intellectual Teamwork, 291-325

(6) Mued L, Lines B, Furnell S and Reynolds P, 2002, "Investigating the Interaction Effect of Audio and Video as Perceived in Low Cost Videoconferencing", in the Proceedings of the Third International Network Conference (INC 2002), Plymouth, UK, 181-189

(7) ITU-R Recommendation BT. 500-7, 1997, "Method for the Subjective Assessment of the Quality of Television Pictures, RBT"

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