2004

A MODEL FOR PREDICTING THE PERFORMANCE OF IP VIDEOCONFERENCING

LICHA MUED

PHD

A MODEL FOR PREDICTING THE PERFORMANCE OF IP VIDEOCONFERENCING

MUED, LICHA

Award date:
2004

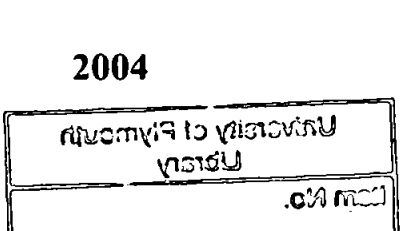
Awarding institution:
University of Plymouth

A MODEL FOR PREDICTING THE PERFROMANCE OF IP VIDEOCONFERENCING

MUED L.

DOCTOR OF PHILOSOPHY

2004



Dedicated to Sulu, Manjut Anak Mingging

COPYRIGHT STATEMENT

A Model for Predicting the Performance of IP Videoconferencing

Licha Mued

With the incorporation of free desktop videoconferencing (DVC) software on the majority of the world's PCs, over the recent years, there has, inevitably, been considerable interest in using DVC over the Internet. The growing popularity of DVC increases the need for multimedia quality assessment. However, the task of predicting the perceived multimedia quality over the Internet Protocol (IP) networks is complicated by the fact that the audio and video streams are susceptible to unique impairments due to the unpredictable nature of IP networks, different types of task scenarios, different levels of complexity, and other related factors. To date, a standard consensus to define the IP media Quality of Service (QoS) has yet to be implemented. The thesis addresses this problem by investigating a new approach to assess the quality of audio, video, and audiovisual overall as perceived in low cost DVC systems.

The main aim of the thesis is to investigate current methods used to assess the perceived IP media quality, and then propose a model which will predict the quality of audiovisual experience from prevailing network parameters.

This thesis investigates the effects of various traffic conditions, such as, packet loss, jitter, and delay and other factors that may influence end user acceptance, when low cost DVC is used over the Internet. It also investigates the interaction effects between the audio and video media, and the issues involving the lip sychronisation error. The thesis provides the empirical evidence that the subjective mean opinion score (MOS) of the perceived multimedia quality is unaffected by lip synchronisation error in low cost DVC systems.

The data-gathering approach that is advocated in this thesis involves both field and laboratory trials to enable the comparisons of results between classroom-based experiments and real-world environments to be made, and to provide actual real-world confirmation of the bench tests. The subjective test method was employed since it has been proven to be more robust and suitable for the research studies, as compared to objective testing techniques.

The MOS results, and the number of observations obtained, have enabled a set of criteria to be established that can be used to determine the acceptable QoS for given network conditions and task scenarios. Based upon these comprehensive findings, the final contribution of the thesis is the proposal of a new adaptive architecture method that is intended to enable the performance of IP based DVC of a particular session to be predicted for a given network condition.

A MODEL FOR PREDICTING THE PERFORMANCE OF IP VIDEOCONFERENCING

by

LICHA MUED

A thesis submitted to University of Plymouth

in partial fulfilment for the degree

of

DOCTOR OF PHILOSOPHY

School of Computing, Communications and Electronics

24 February 2004

Contents

List of Figures

List of Tables

Glossary

ACELP	Algebraic Codebook Excitation Linear Prediction
ACR	Absolute Category Rating
ADPCM	Adaptive Differential Pulse Code Modulation
AMR	Adaptive Multi-Rate
ARQ	Automatic Repeat Request
ATM	Asynchronous Transfer Mode
AV	Audiovideo
CCITT	International Telegraph and Telephone Consultative Committee
CELP	Code-Excited Linear Prediction
CIF	Common Interchange Format
CODEC	COder and DECoder
CPU	Central Processing Unit
CSRC	Contribution Source Identifiers
DCR	Degradation Category Rating
DCT	Discrete Cosine Transform
DMOS	Degradation Mean Opinion Score
DSCQS	Double Stimulus Continuous Quality Scale
DVC	Desktop Videoconferencing
DWT	Discrete Wavelet Transform

DIFFSERV	Differentiated Service
EMBSD	Enhanced Modified Bark Spectral Distortion
ETSI	European Telecommunications Standards Institute
FEC	Forward Error Control
GSM	Groupe Spécial Mobile
GSTN	General Switched Telephone Network
HDTV	High Definition Television
IETF	Internet Engineering Task Force
IP	Internet Protocol
ISDN	Integrated Services Digital Network
ITS	Institute for Telecommunication Sciences
ITU-R	International Telecommunication Union, Radiocommunication Section
ITU-T	International Telecommunications Union-Telecommunication
JPEG	Joint Photographic Experts Group
jit	jitter
LPC	Linear Predicative Coding
Lip Sync	Lip Synchronisation
LAN	Local Area Network
MAVT	Mobile Audio Visual Terminal

MNB	Measuring Normalizing Blocks
MOS	Mean Opinion Score
MP-MLQ	Multipulse, Multilevel Quantization
MPEG	Moving Picture Expert Group
MPLS	MultiProtocol Label Switching
MRLE	Microsoft Run Length Encoding
NACK	Negative Acknowledgement
NTIA	National Telecommunications and Information Administration
PAMS	Perceptual Assessment of Speech Quality
PCM	Pulse Code Modulation
pck	packet loss
PESQ	Perceptual Evaluation of Speech Quality
POTS	Plain Old Telephone System
PSQM	Perceptual Speech Quality Measurements
PWFQ	Priority Weighted Fair Queuing
QCIF	Quarter Common Interchange Format
QUASS	Quality Assessment Slider
QoS	Quality of Service
RPE-LPC	Regular Pulse Excited - Linear Predictive Coder with a Long Term Predictor Loop

RSVP	Reservation Protocol
RTCP	Real-time Transport Control Protocol
RTP	Real-time Transport Protocol
rtt	round trip time
ReLaTe	Remote Language Teaching
SNR	Signal to Noise Ratio
SQCIF	Sub-Quarter CIF
SSCQE	Single Stimulus Continuous Quality Environment
SSRC	Synchronisation Source Identifier
stdev	standard deviation
3G Mobile	Third Generation Mobile
ToS	Type of Service
TCP	Transport Control Protocol
UDP	User Data Interface
UNIMAS	University Malaysia Sarawak
UoP	University of Plymouth
VDO	Video-on-Demand
VLC	Variable Length Code
VQ	Vector Quantization
VoIP	Voice over IP

Acknowledgements

I would like to express my sincere appreciations to my Director of Studies, Dr Benn Lines, who has been providing valuable help, knowledge, and suggestion, which made indispensable contributions to the completion of this thesis.

I would also like to give my boundless gratitude to my Supervisor, Dr Steven Furnell, who has always been enthusiastic, encouraging and patient. I have constantly benefited from his deep insights and constructive criticisms.

I would like to acknowledge the assistance of: Prof. Paul Reynolds, my third Supervisor, who provided further valuable input to the research programme; Bogdan Ghita and Paul Dowland, for their significant technical assistance; and Dr Mohammed Zaki Ahmed, my LaTeX advisor. I owe my endless gratitude to them.

My special thanks are also due to the various friends, colleagues and all the master students (MSc Comm. Eng & Signal Processing and MSc Network Systems Eng., Year 2000–2003) who spared their time to participate as test subject in the subjective assessment. I THANK YOU ALL for the beautiful life in Plymouth!

I am always indebted to my family for their constant love, understanding, and support. Without their love, I would have never been able to finish this work.

Lastly, a special tribute to a very dear friend of mine, the late Dr Harjit Singh;

..... *"A Dedicated Life That Illuminated Many Hearts And Minds!"* .....

Author's Declaration

At no time during the registration for the degree of Doctor of Philosophy has the author been registered for any other University award. This study was financed by the University Malaysia Sarawak (UNIMAS) with support from the Network Research Group at the University of Plymouth. Relevant scientific seminars and conferences were regularly attended at which work was often presented. Several papers were published in the course of this research project. The work presented in this thesis is solely that of the author.

Signed .....

Date 5/05/04

Chapter 1

Introduction

Attempts to achieve real-time delivery of multimedia information over the Internet have increased rapidly over the past few years. It has enabled the opportunity to develop a global multimedia communication system. Many new services, such as, audio and videoconferencing, audio and video-broadcasting, or audio and video on demand may become prevalent and world-wide spread within the near future. The increasing use of the technology has led to the difficulties of having a best-effort service that is the key issue in real-time multimedia applications over the Internet.

This thesis investigates the effects of various network parameter variations on end user acceptability, when low cost desktop videoconferencing is used over the Internet. As visual communication is part of the human existence, there is little doubt that videoconferencing will become popular both professionally and personally. The increasing demands of desktop videoconferencing stems from its many benefits, namely, improved person-to-person communication, reduction in the need to travel to meet colleagues and clients. However, despite its increased popularity, the current low cost Internet Protocol (IP) conferencing is in its infancy, with substantial

improvement needed to achieve higher performance. The various drawbacks faced by desktop videoconferencing (DVC) today, stem from a number of factors which will be described in the following.

For current Internet-based solutions for multimedia, concerning real-time audio and video applications, the UDP-IP protocol is used which has no end-to-end delivery guarantees. As such, it only provides 'best effort' service, offering no timeliness or bandwidth assurance, as the techniques to guarantee the delivery of IP packets (with sufficient level of Quality of Service (QoS)), have yet to be implemented successfully. Many research efforts are now being directed toward some potential methods to upgrade perceptual media quality. These methods are designed to improve QoS through bandwidth conservation, develop new effective technique to evaluate audio and video quality and implement more advanced coding techniques. The study presented in this thesis, aimed to establish a taxonomy of real-time multimedia task and applications, and to determine the maximum and minimum audio and video quality boundaries for the given tasks. System developer and network designer will be able to employ the model developed to enhance DVC system design. Hence, better bandwidth utilization and improved media quality can be achieved.

The characteristic of IP network presents a great challenge, in that, due to its constantly changing and unpredictable nature, the inter-observer reliability for the perceived multimedia quality over one conference has becomes a critical issue. Hence, this in effect, presents a complex problem for the evaluation of multimedia quality over the Internet. Today, there are a lack of standards to define low cost multime-

dia QoS over the Internet. Those that do exist are rather more inclined towards the audio and video quality over the more sophisticated and higher bandwidth communications and entertainment (broadcasting) systems. Thus, it is this dearth of the standard methodology to evaluate the perceived quality of IP media that has became the catalyst for the research work presented in this thesis. Therefore, the research aims to address this problem by investigating a novel approach to establish a model for assessing the perceived quality of audio, video, and audiovisual overall, required for a specific task performance which can be used as a basis of a control mechanism to predict the QoS of audio and video in IP multimedia applications.

1.1 The Research Objectives

The main aim of the thesis is to investigate and analyze the influencing factors that could affect the end user's perception of IP media quality and later, to establish quality threshold for each media required for specific task performance. To achieve this objective, the research has focussed specially upon desktop videoconferencing, which is considered to be an representative application in terms of its demands for real-time multimedia communications.

A number of considerations and factors involved, need to be thoroughly investigated and analyzed, as follows:

- to benchmark the current state of the art of the existing DVC systems;
- to generate in depth knowledge of the methodologies and techniques of the existing objective and subjective test for assessing the end-user QoS, both theoretically and practically, and to justify which method that should be implemented for the research work;
- to investigate the fundamental factors that affect the perceived media quality in DVC (i.e network constraints, CODECs, task performances, background noise, hardwares, and etc.), to gain an understanding of the audio, video and combined audiovideo quality issue;
- to investigate the correlation effects between the IP medias (especially audio and video);
- to investigate the impacts of lip synchronisation (sync) error on the perceived multimedia quality;
- to make the comparison of results between controlled classroom-based experiment and real-world environment test and to investigate other involving factors and issues;
- to establish quality threshold for audio and video required for specific task performance, for reasonable user satisfaction at low cost.

1.2 Outline of Thesis

This thesis describes research, leading to the formulation of a novel model for the optimization of the perceived quality audio and video in IP communications. The material is structured into ten chapters as outlined below.

In Chapter 2, an introduction of videoconferencing systems configuration and the protocols involved are provided. The ever increasing range of applications and services using IP conferencing, in the real world are also outlined, suggesting the growing interest in the designated study. The overview of RTP/RTCP protocol, videoconference technology (coding techniques) and audio and video optimization techniques are given in this chapter, providing the foundations for the research context.

Chapter 3 focuses upon the definition of the perceived audio and video quality in IP multimedia conferencing. The fundamental factors that may affect the perceived quality are described and explained in detail. The existing evaluation methodologies of audio and video quality, subjective and objective methods, are theoretically investigated in this chapter. The subjective test method is then practically applied in the subsequent chapters, i.e Chapter 4, 5, and 6. The critical issues and drawbacks of the current QoS in IP multinedia are extensively identified and addressed, which then form the core focus for the research, i.e. aims to significantly improve the perceived quality of audio and video media in IP communications.

In Chapter 4, the concept and methodology of the research approach are described, and a description as to why they are considered as being relevant focus for the designed experiments is thoroughly explained.

Chapter 5 outlines the initial work that have been undertaken, i.e. to investigate the current state of the art in desktop videoconferencing system. This chapter focuses upon benchmarking the performance of the popular Microsoft NetMeeting (which represents over 90% of the current market) with respect to the related issues that affect the perceived audio and video quality, such as network congestions, computing resources, tasks performance, CODECs, and conferencing hardware. The test experiments were based upon two different tasks performance i.e. passive test and interactive test. The designed task scenarios are continued throughout the remaining chapters.

The study then proceeds to investigate and analyze the interaction effect between audio, video and audiovideo overall, as perceived in low cost videoconferencing systems.

Chapter 6 provides substantive evidence of the strong interaction between the perceptual quality of these media (especially audio and video), in that it is clearly content dependent. While, in Chapter 5, the same quantity of network impairments are simultaneously introduced to both audio and video streams, the experiments in Chapter 6, however, were based upon interpolating the different levels of packet loss, into the two media, separately.

Chapter 7 outlines the test conducted to investigate the effects of lip sync due to delay, jitter as well as packet loss, on the perceived quality of audio, video and audiovideo overall. The chapter reveals that the attention given to the assessment of lip sync error is dependent on the different tasks being performed by the end users. An exhaustive research has negated the previous finding which stated that, the nature of two-way interaction is claimed to be awkward and annoying when audio delay reaches 400ms [1], [2], and [3]. However, the study presented in this chapter provides the substantive evidence that the subjective MOS of the perceived multimedia quality is unaffected by lip sync error for low cost desktop videoconferencing operating over low bandwidth systems.

The tests experiments conducted in Chapter 5, 6, and 7 are classroom-based experiments, with contrived tasks assigned to the subjects. In Chapter 8 the study was conducted between the University of Plymouth and University Malaysia Sarawak (UNIMAS), which aims to make the comparisons between controlled classroom-based experiment and real-world environments, to provide actual real-world confirmation of the bench tests.

Chapter 9 presents a new adaptive architecture method that is intended to enable the performance of IP based DVC of a particular session to be predicted for a given network condition. By inferring the multimedia quality scores (MOS) and the congestion control information from a network monitoring system, the proposed technique (that can be implemented within the videoconferencing architecture) can automatically adapt to the network change by negotiating for the best configuration

within the adaptive architecture tools to give the best quality and improved bandwidth utilization.

Finally, Chapter 10 concludes all the findings obtained throughout the entire study, highlighting the achievements and contributions, as well as stressing the limitations and problems of the research work. Here, the potential further developments suggested for the future work are discussed.

A number of appendices are included in the thesis, which provide a range of supporting materials. Copies of a number of the published papers are listed in Appendix D.

Chapter 2

Desktop Videoconferencing Systems

This chapter presents an overview of videoconferencing systems, technologies, standards, and applications. The chapter begins with the introduction of the basic concept of videoconferencing systems and RTP/RTCP protocols, and then proceeds with the fundamentals of audio and video compression technologies. Later, it reviews the network environments for which videoconferencing systems have been defined, and provides a list of the ITU-T Recommendations that specify these systems, as well as their components. The number of existing loss recovery techniques and bandwidth conservation methods are also discussed. The chapter concludes by presenting the various videoconferencing applications and services that are available today.

2.1 Videoconferencing Systems Configurations

This section provides an introduction to audiovideo systems configuration and the protocols involved. Typical desktop videoconferencing (DVC) systems are

equipped with video and audio capture and compression subsystems, and decompression and display subsystems. A simplified block diagram of these components is shown in Figure **2.1**. In general, it can be divided into two major parts, i.e. Sender terminal and Receiver terminal.

The sending terminal is consists of Coder and Packetizer, meanwhile, the receiving terminal is consists of depacketizer and Decoder. The CODEC (COder and DECoder), the most important device in any videoconferencing system, performs the function of coding, decoding, compressing, and decompressing the video and audio to conserve bandwidth on a transmission path. CODECs can be found in either hardware or software form. The Packetizer is used to packetize the media frames. Packetisation is a simple process of placing audio or video frames into RTP packets [4]. The bigger the packet, the longer the delay for sampling and encoding, packetisation and transmission. The smaller the packet, the larger the relative overhead of the packet header, and therefore the poorer the bandwidth utilisation.

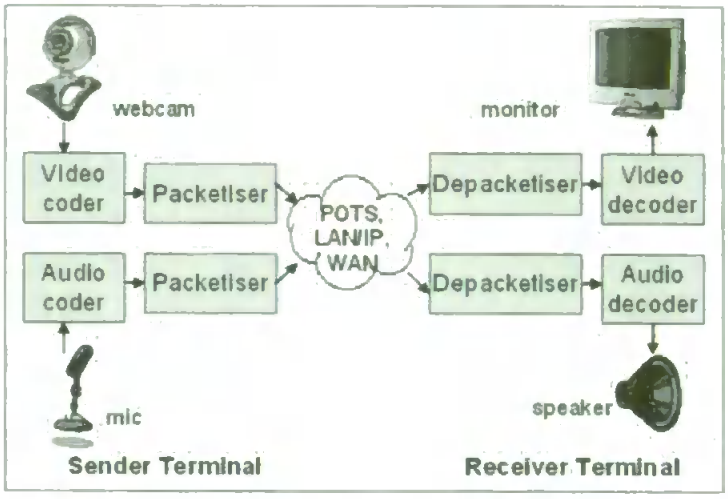


Figure 2.1: Components of Videoconferencing Systems

The functions of each terminal is described as follows:

- Sender terminal: digitizing, encoding, and packetising - the voice and image/video of the sender is first digitised, then encoded, to reduce the amount of bandwidth required to transport it, and then packetised into IP datagrams, which are to be sent over the Internet;
- Receiver terminal: depacketising, decoding, dequantisation - the reverse operations, as performed at the sending point, in order to deliver the voice and image/video to the receiver

2.2 RTP/RTCP Protocol

An overview of audiovideo over IP protocol architecture is shown in Table 2.1 below [5].

Table 2.1: Audiovideo Over IP Protocol Architecture

Application Layer	Audio/video			
Transport Layer	RTP	RTCP	SIP	H.323
Transport Layer	UDP		TCP	
Network Layer	IP			
Physical Layer	e.g Ethernet/SDH			

Note:

RTP - Real Time Protocol

RTCP - Real Time Control Protocol

SIP - Session Internet Protocol

UDP - User Data Protocol

TCP - Transfer Control Protocol

IP - Internet Protocol

SDH - Synchronous Digital Hierarchy

The Real-time Transport (RTP) Protocol [5], normally runs on top of UDP, provides end-to-end network transport functions suitable for application transmitting real-time data, such as, audio, video, and simulation data, over multicast or unicast network services. RTP provides a standard packet format, with a header containing media-specific timestamp data, as well as identifiers, payload type information,

number of sequence, etc. Table 2.2 shows the RTP header architecture [5]. This enables RTP to provide end-to-end transport functionality suitable for Real-time applications, over multicast or unicast networks.

Real-time Transport Control Protocol (RTCP), a control protocol that enables the participants in a multicast group to be able to exchange control information, supplements RTP by providing data, such as, sender identifier, quality of service of the received data, loses etc.

Different RTP sessions are set up for different data types in the same communication session. For example, audio and video of the same communication are sent in different RTP sessions. Sources are identified by the SSRC (Synchronisation Source Identifier) field in the RTP packet headers, which is guaranteed to be unique in each session.

Timestamps (refer to Table 2.2) are useful to smooth out the effect of network jitter. Timestamping is media specific and different reference clock are used for each media. For example, audio uses its own device interface as a clock and for video stream, it depends on the type of compression employed [6].

Table 2.2: RTP Packet Header

V=2	P	X	CC	M	PT	Sequence number
Timestamp						
Synchronisation Source (SSRC) Identifiers						
Contribution Source (CSRC) Identifiers						

Note:

V - Version

P - Padding

X - Extension

PT - Payload Type

CC - number of CSRC identifiers that follow

2.3 Videoconference Technology

This section describes the basic theory and technology needed to capture and compress audio and video in videoconferencing applications over packet networks. Audio and video must be captured in their analog form and stored digitally by a CODEC before the computer can manipulate them. Uncompressed data, especially video, would require massive amount of bandwidth and CPU cycles to transmit, therefore it is necessary to perform compression prior to transmission over communication channel. The compression must occur in as near real time as possible to satisfy the strict timing constraints of the videoconferencing session.

2.3.1 Audio Coding and Compression

An analog audio signal has amplitude values that vary continuously with time. Analogue speech is digitised by first limiting the top frequency to less than 4kHz (see below for explanation). The next step in the process is sampling. This is followed by quantisation where each sample is given a binary code ready for serial transmission. The number of quantisation levels depends on how many bits are used to store the sample value. The Nyquist theorem states that if you sample the analog signal at least twice the rate of the highest frequency of interest, the original analogue form can be accurately reconstructed [7]. A human voice can only produce frequency between 30Hz to 17kHz, whereby a human ear can perceive between 20Hz and 20kHz. Since most of the speech energy is below 4000Hz, the sampling rate needed is 8000 times per second [8].

Some of the audio compression methods used in videoconferencing systems are described below. In general, audio encoding format can be categorised as waveform CODECs or Source CODECs. PCM and ADPCM are known as waveform CODECs in that they exploit redundant characteristics of the waveform itself. Source CODECs, compress speech by sending only simplified parametric information about voice transmission (as opposed to a compressed version of the voice transmission), hence, requiring less transmission bandwidth. Examples of source codecs include linear predicative coding (LPC), code-excited linear prediction (CELP), and multipulse, multilevel quantization (MP-MLQ).

These techniques can achieve real-time compression and decompression in software or using inexpensive hardware.

2.3.1.1 Linear PCM

Linear PCM (Pulse Code Modulation) is considered as an uncompressed audio encoding format, where the quantizer values are uniformly spaced, with the top frequency of 3.4kHz maximum. The telephony form of PCM uses 8 bits for each sample. Thus, the transmission rate is obtained by multiplying 8000 samples per second times 8 bits per sample, giving 64,000 bits per second. Therefore, the standard transmission rate for one channel of digital telephone communications is one 8bit byte transmitted every 125 microseconds.

2.3.1.2 Non-linear PCM: μ-law and A-law

Non-linear PCM has two basic formats, i.e. μ-law and A-law, operating at 64-kbps. The fact that the ear is more sensitive to variations at low amplitude, enables non-linear conversion scales (logarithmic) to be configured which allows larger quanta to be used for larger level signals (louder), and smaller quanta sizes to be employed for low level signals. Hence, a compression ratio of 1.77:1 (Original data: compressed amount) can be effectively achieved. μ-law is used in the US and Japan and A-law is used in European countries. μ-law has a slight advantage in low-level signal-to-noise ratio performance.

2.3.1.3 ADPCM

Adaptive Differential Pulse Code Modulation (ADPCM) is a compressed version of PCM, whereby, previous PCM sample are used to indicate the value of the current sample. ADPCM encodes using 4-bit samples, giving the transmission rate of 32 kbps. The technique employs linear prediction coding methods by encoding only the difference (hence the term 'differential') in speech samples as well as the rate of the change of that amplitude instead of the complete sample value. Thus, fewer bits need to be decoded. The following equation is used to calculate the amplitude:

$$\text{Amplitude} = (\text{Amplitude of last sample}) + (\text{difference} \times \text{step})$$

The encoder can adapt to signal by changing quantization or prediction parameters, the term 'Adaptive'. ADPCM can achieves compression ratio of 2:1, as compared to μ-law and A-law.

ADPCM technique is employed in ITU-T Recommendation G.721, G.722, G.723, G.726, and G.727 audio CODECs, as used in desktop videoconferencing systems. These CODECs specify sample size ranging from 4 bits to 14 bits.

2.3.1.4 LPC-Linear Predictive Coding

Linear Predictive Coding (LPC) is one of the most powerful speech analysis techniques. LPC is used to compressed audio data to 16 Kbps and below. In this tech-

nique, the LPC encoder fits speech signals to a simple analytical model of the vocal track. The best-fit parameters are transmitted and used by the decoder to generate synthetic speech that is similar to the original. GSM (Groupe Speciale Mobile) encoding uses a variation of LPC called RPE-LPC (Regular Pulse Excited - Linear Predictive Coder with a Long Term Predictor Loop) [9] [10]. GSM compresses 160 13-bit samples (2080 bits) to 260 bits, which is an 8:1 compression ratio. For 8 KHz sampling, this means GSM encoded speech requires a bandwidth of 13 kbps. ITU-T Recommendation, G.723.1 [11], G.728, and G.729 [12] describe this technique.

2.3.1.5 CELP (Code Excited Linear Prediction)

CELP employs the same vocal track modelling as LPC encoder. The added technique is that it computes the error between the input speech data and the synthetic model and transmits both the model parameters and representation of the errors. Thus, the error signal is very much compressed. The errors represented as indices are encoded into a common codebook, shared between encoder and decoder. The computational complexity and speech quality of the coder depend upon the size of the code books, which can be reduced at the expense of sound quality. ITU-T Recommendation G.728 (16 kbps) and U.S. Federal Standard 1016 (4.8 kbps) use CELP [13].

Several overview of audio CODECS most commonly used for videoconferencing, IP telephony and packet voice are presented in Table 2.3 [14].

Table 2.3: ITU-T Audio Compression Standard (G Family)

ITU Standard	Year Approved	Algorithm Used	Bit Rate	Bandwidth (kHz)	End-to-end delay(ms)	Application vc=videoconferencing tel=telephony
G.711	1977	PCM	48, 54 and 64	3	1	GSTN tel H.323, H320 vc
G.723	1955	MPE-ACELP	5.3, 6.3	3	67-97	GSTN videotel H.323 vc
G.728	1992	LD-CELP	16	3	2	GSTN, H.323 vc
G.729	1995	ACELP	8	3	25-35	GSTN tel, videophone modem h.324
G.722	1988	subband ADPCM	48, 56 and 64	7	2	ISDN vc

2.3.2 Video Coding and Compression

Video is a sequence of still images, that when presented at a high enough rate, gives the illusion of fluid motion. TV image is presented at 25 frames per second (fps) in Europe, 30 fps in USA, while video over desktop videoconferencing are normally between 2-8 fps [15].

The uncompressed digital video signal is impractical, in that the bandwidth is far too large to deal with, either for storage or for transmission. Desktop PCs or small workstations suffer from lack of computing power for compression of these large video streams, and hence, desktop video conferencing (DVC) is considered the most

suitable method to offer 'best-effort' real-time audiovideo communications. This section will focuses upon the various video compression techniques, normally used in multimedia services over packet network.

The term *video coding* comprises the stages of applications, such as, video/image capture, pre-processing, and compression techniques to obtain compressed digital video images. The important metrics of video coding are the compression ratio, the data rate, and the bits per pixel, i.e. the number of bits required to represent one pixel in the image.

2.3.2.1 Overview of Video Coding

Basic video compression techniques are namely:

- Run Length Encoding;
- Vector Quantization;
- Discrete Cosine Transform;
- Discrete Wavelet Transform;
- Motion Compensation.

2.3.2.2 Run Length Encoding

Run length encoding encodes a sequence or run of consecutive pixels of the same color (such as black or white) as a single codeword. For example, the sequence of pixels;

77 77 77 77 77 77 77 could be coded as 7 77 (for seven 77's)

Microsoft RLE (MRLE) is an example of a video CODEC that uses run length encoding method. It is also used to encode the DCT coefficients in the block Discrete Cosine Transform (DCT) based on international standards MPEG [16] [17] [7], H.261 [18], H.263 [19], and JPEG [20] [7].

2.3.2.3 Vector Quantization

In the vector quantization (VQ) compression technique, a frame image is divided up into blocks of 4x4 pixels. Some blocks are generally similar to another although not necessarily identical. The similar blocks are then identified and replaced to form a "generic" block which represent the class of similar blocks. A lookup table (codebook) that maps short binary code to the "generic" blocks is encoded by the encoder, where the shortest codewords represent the most common classes of blocks in the image. This codebook is used by vector quantization decoder to assemble an approximate image comprised of the "generic" blocks in the lookup table. Since the sent generic block is just a 'good enough' approximation to the original block thus,

vector quantization is inherently a lossy compression process, i.e. distortion is introduced such that the original sample value can no longer be exactly recovered. The smaller the lookup table, the higher compression ratio can be achieved. However, the reproduced approximation of image quality degrades as the lookup table becomes smaller. Vector quantization is used in Indeo 3.2 (developed by Intel in the 1980's and it is well-suited to CD-ROM) and Cinepak. Cinepak was originally developed to play small movies on low CPU - '386 and '030 systems, from a single-speed CD-ROM drive.

The encoding and decoding scheme of vector quantization is shown in Figure 2.2 [21].

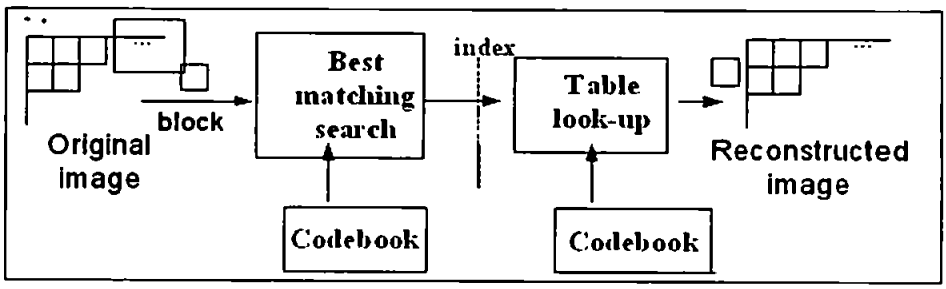


Figure 2.2: VQ encoder and decoder

2.3.2.4 Discrete Cosine Transform (DCT)

In this technique, a two-dimensional (2D) Discrete Cosine Transform (DCT) is applied to the compressed image of 8 by 8 blocks of pixels. The 64 (8x8) coefficients produced by the DCT are then quantized, and later, entropy coded using VLC (Variable Length Code) to obtain substantial image compressions. The disadvantage of

this technique is due to the small values of the DCT coefficients, which then become zero after quantization. This problem is eliminated by implementing a large quantization factor to the higher frequency component (as the human eyes are less sensitive to the high frequency components of the image represented by the higher DCT coefficients). However, the higher the quantization factor, the lower the signal to noise ratio (SNR) of the quantization, and hence, the compression quality decreases. DCT is used in the JPEG still image compression standard, the H.261 and H.263 videoconferencing standard and the MPEG (MPEG-1, MPEG-2, and MPEG-4) digital video standards.

2.3.2.5 Discrete Wavelet Transform (DWT)

Discrete Wavelet Transform (DWT) coding technique is based on passing a signal through a pair of filters, i.e. low pass and high pass filter. A low resolution version of the signal is produced by the low pass filter and an added detail or difference signal is generated by the high pass filter. These outputs are then downsampled by two, to obtain the same number of bits as the input (original) signal. DWT is used in VDONet's, VDOWave, VxTreme, and Intel Indeo 5.x.

2.3.2.6 Motion Compensation

In Motion Compensation coding technique, the video frame is divided into a number of neighboring macroblocks of N x M pixel. Each block in the macroblock is

assigned a vector from a common origin, say zero origin. Later, every macroblock is then compared with every N x M block in the previous frame, called the reference frame. After all possible N x M block from the previous frame are compared with the macroblock, the block that represents the best match is used for motion compensation. By sending or storing only the motion vector instead of the pixel values for the entire block, high compression ratio is achieved. Interframe prediction error can be improved by estimating the motion of image pixels between frame blocks. The motion vector are restricted at the picture edges to ensure that all pixel reference lie within the coded area. Generally, motion compensation technique is best applied in coding moving object across a background and unsuitable for spinning objects, resizing objects, or camera zooms. This technique is used in MPEG-1, 2, and 4, H.261, and H.263 CODECs.

2.4 Videoconferencing Standards and Networks

Traditionally, the widespread standardized recommendation by the ITU for desktop videoconferencing was H.320 [22], [23], which defines a methodology for transporting videoconferencing data over ISDN. In 1996, the emergence of H.321 [23], H.323 [24] and H.324 [25] standards took place, allowing Internet telephony and IP based real-time multimedia communications across various networks, such as, ATM, IP networks (LANs and Internet), and Plain Old Telephone System (POTS) networks, respectively. Each of these standards has advantages and disadvantages in video-

conferencing transmission and offers different capabilities and certain quality levels.

The Ethernet LANs of at least 10Mbps are commonly employed in most companies today [7], and these minimum available bandwidth are enough to support desktop conferences. Larger organizations may require larger bandwidth (100Mbps or more) to support multiple simultaneous calls. However, as IP videoconferencing becomes more popular, existing networks may become bogged down with its traffic which leads to the slowing down of other network application, such as web browsing, as well as degrading the audio and video quality. Hence, the management of network resources has become the key element of IP videoconfencing standards, to ensure other network applications still function while launching videoconferences.

Table **2.4** shows the various characteristics and constituent elements of videoconferencing standards and the network environments for which these standards have been defined [23]. Table **2.5** shows the standards for video resolution format, supported by H261 and H263. Note: H261 supports CIF and QCIF; H263 supports CIF, QCIF, SQCIF, 4CIF, 16CIF.

Table 2.4: Videoconferencing Systems in Various Network Environments

Network	POT	Narrowband ISDN	Guaranteed QoS LANs	Non-guaranteed QoS LANs	ATM
Channel Capacity	up to 28.8 kbit/s	up to 1536 or 1920 kbit/s	6/16 Mbit/s	up to 10/100 Mbit/s	up to 600 Mbit/s
Standard	H.324	H.320	H.322	H.323	H.310, H.321
Aproval date	96/03	90/12	96/03	96/11	96/11, 96/03
Audio	G.723.1	G.711 G.722 G.728	G.711 G.722 G.728	G.711 G.722 G.723.1 G.728 G.729	G.711 G.722 G.728
Video	H.261 H.263	H.261	H.261	H.261 H.263	H.261 H.262
Data	T.120	T.120	T.120	T.120	T.120
Control	H.245	H.242	H.242	H.245	H.245, H.242
Multiplexing	H.223	H.221	H.221	H.225.0	H.222.0
Call setup signaling	National standards	Q.931	Q.931	Q.931 H.225.0	Q.2931

Table 2.5: Standard Video Resolution Formats

Acronym	Meaning	Resolution (pixels)
CIF	Common Interchange Format	352x288
QCIF	Quarter CIF	176x144
SQCIF	Sub-QCIF	128x96
4CIF	Four CIF	704x576
16CIF	16CIF	1408x1152

2.5 Audio and Video Optimization Techniques

Previous studies involving multimedia application over the Internet [26], [24], [27] have stated that the audio quality greatly suffers from a range of network imperfections, namely, packet loss, jitter, loss distribution, end-to-end delay, and packetisation interval. As the QoS is still an issue, it has become extremely important to minimize these factors. A number of play-out adaptation and loss recovery techniques have been developed to address the problems, which will be discussed in the following subsections. These methods can be divided into two categories, namely, error control mechanisms and bandwidth conservation techniques.

2.5.1 Error Control Mechanisms

The various mechanisms available to recover from the loss or corruption of IP packets are, namely, Forward Error Control (FEC), Silence Suppression, Silence Substitution, Packet Repetition, and transmission control technique. These techniques are presented below.

2.5.1.1 Forward Error Correction

Forward error correction (FEC) algorithms have been developed to minimize the effect of packet loss, by sending additional information (called parities) to allow reconstruction of lost packets at the receiver [28]. The redundant data is derived

from the original data using coding theory techniques. The recovery of lost data at the receiver requires very little time, but considerable local processing power and hence, FEC is employed for applications with real-time requirements as in packet networks videoconferencing and mobile telephony.

2.5.1.2 Silence Suppression

Silence Suppression is a technology implemented in voice transmission to free up bandwidth on the voice channel, by removing the pauses in speech before transporting voice traffic over a network, thus allowing the extra bandwidth to be used by speech or data from another channel.

2.5.1.3 Silence Substitution

Silence substitution refers to a technique of error recovery by the insertion of silence whenever there is a gap left by a lost packet. As such, the timing relationship between the surrounding packets is maintained. The disadvantage of this method is that, its performance degrades rapidly as packet sizes increase (perceptible glitches or gaps) and becomes unusable for packet size of 40ms (which is commonly used in network conferencing tools). Silence substitution is only effective with short packet lengths (4ms) and low loss rates (2%). At low lost rates, the sound becomes 'bubbly'. Despite this, the use of silence substitution is widespread, primarily because it is simple to implement.

2.5.1.4 Packet Repetition

This technique replaces the lost packet with copies of the unit that arrived immediately before the loss. Packet repetition has the advantages of being simple to implement, provides low computational complexity and performs quite well. Clicks may be heard where frequency and amplitude change and produces 'stutter' effect with large packet sizes. The performance of packet repetition can be improved by the use of fading technique. For example, in the GSM system, the first 20ms of the repeated signals are of the same amplitude as the original sample, while the next 320ms are faded to zero amplitude.

Further information on the related error correction methods described above can be found in [29], [30], [27], and [31].

2.5.1.5 Transmission Control Mechanism

Adopting a scalable bit-rate CODEC, and a prioritized transmission algorithms, at the network layer, to stabilize the level audio degradation during burst congestion periods [32]. This transmission control mechanism employs the method of combining the three mechanisms, namely, layered encoding scheme, priority dropping of packet in the network, and fair weighted queuing technique. The scalable audio encoders, i.e. Mobile Audio Visual Terminal (MAVT) is used to transmit audio information in differentiated flows, and hence, allowing straightforward prioritized packetization of audio data. The highly increased multimedia traffic cannot sup-

port congestion control, and will overloading the network, that could lead to network collapse. The Priority Weighted Fair Queuing (PWFQ), a queuing algorithm is introduced to implement the priority mechanism and guarantees fair network utilization, when network is congested. Therefore, a smooth degradation of quality can be achieved during network congestion periods.

2.5.2 Bandwidth Conservation Techniques

RVSP, Diffserv, and MPLS are the most common separate standards which purport to help solve the problem in IP networks. These network-control protocols are intended to enhance the current Internet architecture with support for Quality of Service (QoS) flows. Further descriptions of each standard are presented in the following.

2.5.2.1 RVSP

The RSVP is an acronym for Reservation Protocol. It is an IEEE standard protocol used to provide QoS on current Internet applications, such as, videoconferencing, IP telephony, and other forms of multimedia communications by reserving bandwidth before packet transfers to guarantee its availability. The RSVP enables Internet services to obtain differing QoS for their data flows as it is recognized that different applications have different network performance requirements.

The RSVP protocol is used by a host to request specific resources from the network for particular application data streams. It is also used by routers to deliver QoS requests to all nodes along the path(s) of the flows and to establish and maintain state to provide the requested service. The RVSP is most commonly used as bandwidth control mechanism.

2.5.2.2 DIFFSERV

DIFFSERV is a short term for Differentiated Service. It is a new standard defined by the Internet Engineering Task Force (IETF) working group to provide a way to indicate a packet's priority level using the IPv4 packet header's Type of Service (ToS) field [33]. ToS is an Internet header field which indicates the type (or quality) of service for the Internet datagram. The proposed project will redefine part of the existing ToS byte in every IP packet header to mark the priority or service level that packet requires. DiffServ is designed to achieve a new bandwidth-management scheme for IP networks, thereby provide a framework for delivering better QoS.

2.5.2.3 MPLS

MPLS is a short form of MultiProtocol Label Switching. It is standards-approved protocol, defined by IETF that is used in IP traffic management. Basically, it is a technology for speeding up network traffic flow and making it easier to manage. It works by providing a means for one router to pass on its routing priorities to

another router by means of a label and without having to examine the packet and its header, thus allows core network routers to operate at higher speeds, and enable more complex services to be developed, and hence, facilitates the Quality of Service issue. Therefore, MPLS can ensures that voice or video always receives the bandwidth it needs and its quality does not suffer.

2.6 Videoconferencing Applications

The popularity of videoconfencing applications will continue to grow as the technology advances. As the Internet becomes more and more ubiquitous, videoconferencing has the tendency to follow the success of Internet telephony, and it has shown great promise to be beneficial in a variety of areas in human activity, both professionally and personally.

The increasing demands of videoconferencing stems from its many benefits:

- videoconferncing saves the time, cost, and trouble of travelling to a collaborator;
- visual communication is part of the human existence and it is needed because people can communicate best when they can use voice inflection and body gesture [34] to express themselves rather than just type text;

- Internet has becomes more ubiquitous and most existing workstations are connected by packet-switched networks, and hence, it is readily accessible;
- it has enhance global collaboration as distance is reduced;
- tools and facilities of videocoferencing systems are easy to use and manipulate (mostly plug and play) and the prices have become lower as the market demand increases.

Since visual communication is part of the human existence, videoconferencing would satisfy the many demands of the different enterprisers and societies, such as, electronic commerce, distance learning and training, banking and financial services, interactive government services, telejustice (judicial application), telemedicine, telecommuting, administrative application, as well as personal 'informal' videoconferencing. Detail descriptions of these services can be obtained from [14]. A number of commonly used videoconferencing services, today are presented in the following.

2.6.1 Distance Learning and Training

Videoconferencing is widely used in distance learning applications because unlike old distance learning methods such as correspondence classes and videotapes, the modern multimedia distance learning technologies are live and interactive. These two attributes of the new distance learning process enable real-time interaction between the teacher and the students.

2.6.2 Telemedicine

Videoconferencing is also used in telemedicine services. Telemedicine is defined as the delivery of care to patients at any location by using communications technology and the hardware and software tools used in medical practice with medical expertise [35]. It includes the transfer of medical information (voice, data, graphics, video, etc) between patients, primary physicians and medical institutions that are geographically dispersed. Telemedicine also links the healthcare specialist and his patients at a distance location for diagnosis, treatment, consultation, and continuing education [35].

The advantages of telemedicine [35] include:

- improve access to healthcare in rural area;
- better service due to acceleration of treatment and diagnosis - improved efficiency;
- reduced travel cost for both the patients and healthcare professional.

2.6.3 Telejustice

Telejustice has been introduced mainly because the technology saves court money and reduce logistical difficulties. Consequently, there are increasing number of courts and lawyers who claim that telejustice increases the opportunity for partici-

pation in the legal process [36]. By linking together previously disparate agencies, departments, corrections institution, and law enforcement branches, productivity can be increased, interdepartmental can be improved. Moreover, public safety can be improved, operating costs can be better attained, and fee and revenue collection can be increased [36].

2.6.4 Telecommuting

Videoconferencing without doubt, will become an integral part of emerging telecommuting environment. Telecommuting is defined as periodic work out of the principal office (one or more days per-week), either at home, a client site, or in a telework centre. It emphasis on the reduction of the daily commute to and from the workplace. Telecommuting has several advantages and disadvantages. The detailed advantages and disadvantages of telecommuting can be found in [37].

2.6.5 Personal Videoconferencing

Videoconferencing for personal use has attracted increasing attention since the use of packet networks, especially of the Internet, has increased rapidly over the past few years. Moreover, the ready accessibility and low cost facilities of desktop-to-desktop videoconferencing have moved the market in this direction.

In the context of the thesis, the research focuses at low cost/bandwidth DVC applications, such as, personal or more informal business uses. Despite the disadvantage of poor video quality (at least for the time being), before long, as the technology improves, many users will commonly launch videoconferencing tools with their friends and loved-ones over the Internet connections.

2.7 Summary

This chapter has presented the overview of study background within which the thesis research is undertaken. Various videoconferencing applications and why they become popular have been briefly explained. Despite the growing interest in the use of videoconferencing applications over the Internet, unfortunately, it is inevitably afflicted with problems that affect its quality, such as, network constraints (i.e. loss, jitter, and delay) and other 'external' factors (e.g. background noise). Accordingly, significant research is taking place towards defining multimedia quality over IP and ways to improve it. The next chapter deals with the issues of network factors and other various aspects that affect IP media QoS. It also presents a critical review of the existing assessment method, latterly, clarifies the technique used in this thesis.

Chapter 3

Audio and Video Quality

3.1 Introduction

This chapter presents the various issues that determine the perceived quality of audio and video in the context of IP videoconferencing. First, some factors, such as network impairments, lip synchronisation, external 'noise', and combined effect of audio and video (audiovideo overall) quality on the perceived multimedia quality are discussed. Later, the chapter focuses upon the various existing techniques and methodologies to assess its quality. The advantages and disadvantages of the two major techniques (i.e. subjective and objective evaluation methods) are compared and discussed. The key issues and difficulties involved in assessing the delivered quality are also discussed in detail. Finally, the chapter concludes by stating that the subjective method is considered to be the most suitable assessment technique to realize the objectives of the study in the thesis.

3.2 Factors that Affect Multimedia Quality

Before overall quality requirement can be assessed, it is necessary to investigate individual factors that can influence perceived quality of audio and video. For example, audio and video quality is susceptible to variables such as network congestions (i.e. packet loss, delay, and delay jitter), background environment, bandwidth, processing power, CODECs, conferencing hardware, and task performance can affect the quality of both audio and video. Even when the variables involved can be separated and controlled, unfortunately, there is yet no recognised industry standard for determining audio and video quality. All these factors should be carefully addressed before evaluation procedures can be proceeded.

The various factors that affect the perceived audio quality in multimedia services are described in the following sections.

3.2.1 Packet Loss

At this point in time, Internet video does not support the quality of service that most people would find acceptable. Packet loss occurs when the network becomes congested where router buffers fill-up and start to drop packets. A similar effect to packet loss happens when a packet experiences a large delay in the network and arrives too late to be used in reconstructing the voice signal.

Packet lost is defined as the number of lost packets, reported in the total traffic and is usually quoted as a percentage (%). For non-real-time applications, such as file transfers, packet loss is not critical. However, for real-time IP conferencing services, perceived audio (and video) quality is affected primarily by the presence and degree of packet loss. It could cause interrupted speech and lost of intelligibility that lead to 'bubbly', 'glitches', and 'robotic' sound occurrence. Viewers describe video that is undergoing packet loss as 'blocky' or 'jerky', as a result of partially upgrading parts of the video image [38] [39]. The level of degradation caused depends on the type of error correction method implemented in the system (see Section 2.5.1), CODECs and packet size. The larger the packet size, the more severe the packet loss impact.

The minimum acceptable audio packet loss, required by users in most IP conferencing application is below 30% [40]. As for video, packet loss of 3% is the maximum amount of loss (in most CODECs) before the image degradations become discernable [41].

3.2.2 Delay and Variation in Delay (Jitter)

Delay is another network factor that can affect multimedia quality. Delay is defined as the time passed between the sending of a packet and its arrival at the destination. The major problem arises from end-to-end delay caused by latency and jitter that leads to loss of packet.

On the other hand, a long delay can lead to lack of patience between listeners. Audio end-to-end delays should be kept to less than 200ms, recommended in ITU G.114 [1], for effective interaction. It is stated that, a maximum delay of 450ms is tolerable before the disruption in two-way-communications become annoying. However, other findings [42] have concluded that a round trip delay of 500ms was the maximum limit that could be tolerated without serious degradation in conversational effectiveness. Also, it is reported in [34] that delay of 320ms to 420ms are acceptable in interactive communication. However, delay tolerance varies with respect to the task scenario [42]. For example, in a lecture mode, long delay is more tolerated especially when there is a need for efficient data transfer along the channel.

Ideally, video should be synchronised with audio so that it has the same amount of delay. Generally, people are more tolerant to audio lagging video in an audiovisual clip [43], [44], rather than vice-versa, because of the nature of human behavior i.e. people are more used to perceiving an event before they hear it (i.e. light travels faster than sound). As of today, however, there is yet no definitive figure specifying the accepted delay of video in IP multimedia conferencing.

Several sources of delay are outlined below:

- ***algorithmic delay*** - the time to accumulate audio sample before coding begins;
- ***processing delay*** - the time to execute the coding and decoding algorithms;
- ***multiplex delay*** - the time coded audio must wait before transmission begins;

- ***transmission delay*** - the time needed to transmit the bits representing audio;
- ***modulation delay*** - time use to modulate or demodulate the signal;
- ***propagation delay*** - time taken for the signal to reach its destination;
- ***buffering delay*** - the time data is passively stored, including time for smoothing out jitter in signal arrival time.

Jitter is defined as the maximum difference between end-to-end delays experienced by any two consecutive packets [45]. Jitter in packetised audio transmission produces glitches sounds, cause gaps in the play-out of audio streams, choppy appearance on a video display, and in severe conditions, can cause similar effect to that of packet loss.

Many methods are being employed to reduce delay and delay jitter [46], [47]. Resizing of the jitter buffer under varying network conditions is one of the techniques to eliminate variation in arrival delay. Another approach is to hold the collected packet long enough so that the slowest packet to arrive is still in time to be played in the correct sequence. This approach however causes additional delay at the receiving end. As previously explained, in the case of real-time voice information, packets must arrive within a relatively narrow time window to be useful to reconstruct the voice signal. Therefore, packets that are too late are ignored, i.e equivalent to them being lost. Hence, jitter should be kept to a minimum to ensure good quality of audio.

Two technologies have been proposed by the Internet Engineering Task Force (IETF) [48], i.e. Header Compression and Segmentation.

- **Header Compression**: a compressed voice packet header reduces the overhead of the packets that leads to the efficient use of network resources, such as intermediate routers and links, and as a result, decreases delay.
- **Segmentation**: it is applied to TCP packets sharing network resources with voice packets, which makes TCP packet transmission time smaller and hence, reduces the waiting time of voice packets.

3.2.3 Lip Synchronisation

Lip Synchronization, defined as the synchronization of the audio and video signals received during a videoconference, is another factor to determine audio and video quality. It is suggested that audio should be synchronized within +/- 90ms of the video, with a maximum range of +/- 160ms [2]. Normally, a 100ms gap between received audio and video signal is considered very acceptable for most videoconferencing applications. Subjective studies [49] have indicated that the out of sync time between audio and video streams can be in the region of 80–100ms before a lack of synchronisation is perceived.

Frame rate is considered as one of the major factors that can affect lip synchronisation. Frame rate is defined as the amount of frames displayed in one second.

Video applications are sensitive to irregularities in bandwidth that is one of the main problems in IP networks. Hence, frame rate also varies immensely. Frame rates in videoconferencing signal can vary from 2–15 frame per second (fps), depending on network impairments i.e. available bandwidth, packet loss, latency/delay, and jitter. In low cost DVC, it is difficult to accurately measure lip sync as the frame rates obtained are generally very low i.e. 2–5 fps [15], [50], [51]. It is claimed that the frame rate should exceed 8 fps to achieve substantial lip sync. The rate must be higher than 16 fps for humans to perceive a smooth moving picture [52] and above 8 fps to obtain lip synchronization [6], [15]. It is also claimed that audio and video is not perceived being synchronised for frame rates between 5 and 6 frames per second [53]. Thus, it is generally agreed that below 8 fps, the video quality is too low to make lip synchronisation a meaningful parameter to measure.

Other factors that can affect lip synchronisation are namely,

- other task operation – data operations as in T.120 (i.e. ITU-T standard that describes data conferencing), consumes some of the communication bandwidth, and hence, can affect lip sync. Typically, data packets are sent at higher priority than video packets and cause some reduced frame rate and loss of synchronisation;
- CPU activity – like launching and closing other application while running videoconference;
- network traffic (packet loss, delay jitter and delay).

The method to minimise lip synchronisation problem in multicast videoconferencing applications has been developed by [6].

3.2.4 Multimedia CODECs

Multimedia compression technique has direct effect on the perceived audio and video quality. Hence, to improve the QoS in DVC application, a number of more advanced coding techniques have been implemented. These CODECs are known as G.723.1 (MP-MLQ/ACELP: Multipulse Excitation with a maximum likelihood-quantizer/Algebraic Codebook Excitation Linear Prediction - Dual rate: 5.3/6.3 Kbps) [11], ITU-T H.263+ (or H.263 Version 2) [54] [51], and Adaptive Multi-Rate (AMR).

Perceptual Speech Quality Measurements (PSQM) [55] at KPN Research of different IP telephony gateways showed that a good audio quality with videoconferencing products could be reached with a G.711 64 kbit/s CODEC [55]. The speech quality drops to a moderate level with speech compression CODECs G.723.1, G.726 or G.729.

The audio CODECs, such as, G.711, G.726, G. 729 and G.723.1 are being used in Microsoft NetMeeting. Both G.729 and G.723.1 CODECs have built-in concealment algorithms and adopt Silence Suppression method for quality improvements and better bandwidth utilisations. However, G.723.1 has a better loss concealment scheme than G.729. For example, G.723.1 interpolates a loss frame by simulating the vocal

characteristics of the previous frame and slowly damping the signal.

The video CODEC, H.263+ [56] is an improved version of H.263. The source formats are similar to that of H.263, i.e. CIF, QCIF, 4CIF, 16CIF, and SQCIF. It offers many improvements over H.263 such that it has a broader range of applications, for example, wide format picture, resizable computer window, higher refresh rate etc. The drawback of this CODECs is that it has greater coding delay, and hence, increases end-to-end delay (i.e by 0.65 seconds).

AMR speech CODEC was developed by ETSI , and has been designed for use in digital cellular telecommunications system applications [10]. It is a multi-mode CODEC with 8 narrow band modes with bit rates of 4.75, 5.15, 5.9, 6.7, 7.4, 7.95, 10.2, and 12.2 Kb/s. The scalable bandwidth adaptation and the tolerance in bit errors of AMR CODEC are not only beneficial for wireless links, but are also desirable for enhanced videoconferencing applications. However, since the scope of this thesis is dedicated to the existing low DVC systems, AMR CODEC is not being covered in the study.

3.2.5 Task Performance

Different task performance has different effect on the perceived multimedia quality. A study concluded that the impact of end-to-end delay on different conversational tasks is depended on the test scenario [42]. It is reported in [57] that the impact of frame rate on task performance depends on the task scenario being performed.

Also, some findings stated that user perception of audio and video quality is directly influenced by the task difficulty [58], [59], and [60].

3.2.6 Other Factors

All noise, regardless of its source, has the potential to reduce the clarity of audio and video signals.

Echo speech that is echoed back to the speaker such that it is perceived during conversations can have a significant effect on perceived speech quality. For example, on hearing your own voice echoed back to you as you are talking can be annoying and perhaps disruptive. The degree of annoyance of talker echo depends on the one-way delay and the level of difference between echo and speech signals [61]. Echoes are usually caused by a mismatch in impedance, normally found in bad headset and poor line connection. Echo cancellation, a process implemented to remove echo [62], is needed to improve the performance of DVC application. An echo canceller, operates by keeping a sample of the previously sent speech signal and upon hearing the inverted speech sample transverses back in the opposite direction, it subtracts the original speech from the inversed signal.

Below, are some factors that affect the perceived quality of video, as stated in [63]:

- video artefacts – these might appear as blocks (macrocells), color splotches, image distortions;
- sharpness – should be able to see individual hairs on the person's head, the line of the shoulder should be sharp and smooth (not jagged or fuzzy), eyes should be crisp and clear;
- contrast, brightness, color saturation, and color depth;
- stability – the image should be perfectly stable with no motion in the background due to video artefacts, TV interlace jitter, or video noise ('snow');
- background clarity – the background should be still rich in color and texture although slightly blur.

Distortion due to hardware problems and high levels of background noise can be very annoying to the end users. In addition, end-user mood, expectations, and other intangible factors, are all potential issues that can cause the multimedia quality to be perceived as unacceptable. The group size (i.e. one-to-one or one-to-many person group and etc.), familiarity with other participant, duration of interactivity, and listener/speaker characteristic (e.g. talker voice and language) also affect the subjective opinion on audio and video quality assessment.

These variables affect testing methods and make true subjective testing with human subjects more difficult.

3.3 The Impact of Multimedia Quality

Audio is generally considered as the most important element of videoconferencing systems. It is well stated that good audio quality is important in DVC [64], while the video channel is often provided for its psychological effect [65].

Generally, 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 face-to-face communication is prevalent and much preferred over all human means of interactions, as observed in [34]. Previous findings also show that, in workplace settings, face-to-face meetings has been chosen above the other means communications being offered, such as email, and phone, for planning and definitional tasks [66]. This implies that videoconfencing has unique benefits over audio only commmunication, in general.

On the other hand, some studies claim that the presence of a video channel does not directly improve the perceived multimedia quality in respect to task performance [67]. It is claimed that a poor video channel results is a cognitive load that interrupts the conversational process [57]. It is also suggested in [68] that access to audio information is more important in determining when and whether a caller interrupts. Video media however, is mainly used for its psychological effect, such as, to clarify meaning, to check reaction, to provide a means of common reference, and to give

psychological reassurance that the other participants are actually there and so on. In addition, the video can authenticate the users of one system which is very important in secure communication systems.

The following describes how video influences personal conferencing, as well as collaborative works in office environments.

3.3.1 Cognitive Cue

One of the advantages of adding video channel is that it ameliorates cognitive cue. Visual cues such as head nods and gaze help speakers to evaluate listener's understanding and attention. It also provides visual access to facial expression, posture and gesture that allows talkers to make interference about other participants' affective or emotional state. Many studies assert that users perceive the video channel to add value by offering a sense of presence to the remote participants [68], [65].

3.3.2 Turn-taking

Subjective test, as in [69] [65], claim that a high quality video channel in interactive task (e.g. collaborative meeting) is preferred to audio only because it supports turn-taking, and hence, improves the degree of interactions. Thus, more natural and smooth conversation is achieved. It is also suggested that video seemed to allow participants to manage pauses better than in speech only communication [34].

3.3.3 Social Cue

There is clear evidence that video supports the transmission of social cues and affective information [68]. It is reported that, communications using audio and video are more personalized, less argumentative, more polite, broader in focus and the conversational group tend to like each other more. Participants believe that face-to-face interactions are better than audio only for task requiring effect, such as getting to know other participants or 'ice-breaking' section. Moreover, by adding a video channel, the interaction is less likely to end in a deadlock than speech only communication.

Other video functions are to increase the ability to listen selectively to particular speakers and allow one to determine whether one is attended to or otherwise.

In summary, for applications that involve high interactivity between two communicative parties, there is ample evidence that video is preferred than audio for interruptions, naturalness, interactivity, feedback and attention [65], [34]. Users make more use of the video channel than they care to appreciate.

3.4 Combined Audio and Video Quality

Combined audio and video communication, is inarguably perceived to be better and much preferred than speech only channel in a numbers of ways.

Studies conducted in this thesis focus upon the quality of audio, video and audiovideo overall. It is agreed that, assessing the quality of a component medium (audio and video) individually will not produce informative predictions compared to evaluating the overall multimedia performance, and hence, combined audio and video quality has been considered.

Some studies suggest that the perception of one component media, influences the perception of the other [70], [71], [72]. It has been observed that the different component media, especially audio and video, interact and influence the perception of each other, for example, the evaluation of video quality is influence by the presence audio [70]. An experimental study with High Definition Television (HDTV) , states that an improvement in the perceived quality of video can be achieved by increasing the audio quality only [72]. A study also indicate that when a talker has a visual contact with a listener in a conference, where audio and video are not synchronised, the clarity of a speaker's speech will drop significantly [57]. Another study indicates that any increase in visual representation of the speaker also increases the viewer's tolerance to audio noise [71]. Therefore, it can be concluded that upgrading the quality of one media could benefit the quality of the other and vice-versa. In addition, evaluating the quality of a component medium in isolation will be unlikely to give accurate scores as to perceived quality in overall or full multimedia services, since the subjective and affective perceptual benefits supported by video channel will not be present. Thus, the overall effect of audiovideo quality was investigated due to the above reasons.

3.5 Assessment of Audio and Video Quality

Before the test can be carried out, there are a number of factors that need to be considered, such as test methodology, test materials and conditions. These factors can be generalized, and known as data-collection methods. The method used is dependant on the goals and the objectives of the evaluation. The characteristics of these data-collection techniques should be carefully investigated before being used, mainly for several reasons [73], such as;

- *validity* – the technique must be (most) suitable for the purpose;
- *reliability* – tool used provides stable and repeatable results;
- *sensitivity* – able to measure even small variations in what it is intended to measure;
- *intrusion* – interference between method or systems used, users, and task performance;
- *acceptance* – test subject is willing to operate the technique;
- *ease of use* – test subject is comfortable with the tool;
- *cost* – method employed is affordable;
- *availability* – whether the measuring tool is free or commercially available.

There are two approaches to investigating user perception of DVC QoS, namely, controlled experimental condition (laboratory-based test) and real field trials. The evaluation methods of audio and video quality over one conference can be categorised as objective and subjective testing. In general term, an objective method is based on perceptual-based techniques, for example, psychoacoustic modelling [74], [58] for objectively estimating coded audio or video quality. These techniques suffer from the practicality when applied to the real world situation [70]. The objective methods often result in highly reliable information but limited in the scope for interpretation. A subjective test method refers to a procedure of obtaining any information (written or oral data) that originates from users/observers, which can be in the form of participative observation or interview. Subjective methods can often be time consuming but have the advantage of providing a wealth of information, and are user centric.

Both the objective and subjective test methods can be employed in laboratory-based studies. The subjective measures are normally used in field trials where genuine tasks, under real world environments (i.e. work place or home), are assessed.

To date, apart from the subjective traditional ITU-T Recommendations rating scheme (i.e. Mean Opinion Score, MOS), there are many new methods and techniques have been implemented to effectively evaluate the perceived quality of audio and video over the Internet, such as, QUASS - Quality Assessment Slider [75], Physiological Responses to Stress [76] and [67], Double stimulus Continuous Quality Scale (DSCQS) [75], Single Stimulus Continuous Quality Environment (SSCQE)

[77], Perceptual Speech Quality Measurement (PSQM) [55], and PESQ-Perceptual Evaluation of Speech Quality (ITU-T P.862) [78].

The methodologies and techniques of the most widespread approaches are discussed further in the following subsections.

3.5.1 Assessment Methods of Audio Quality

As previously mentioned, methods used for assessing perceived audio in DVC can be categorized as subjective or objective assessment methods.

3.5.1.1 Subjective Assessment Method

Currently, the traditional ITU Recommendations for subjective test are most commonly used to measure multimedia quality [70], [79] in IP networks. A numbers of ITU-T recommendations that can be found in P series, are originally designed to assess speech transmission quality over the telephone networks. However, since the emergence of the Internet these methods are also widely applied to the IP multimedia applications. Several methods and procedures used for conducting subjective evaluation of transmission quality can be found in ITU P.800 [80].

There are two commonly used methods for subjective assessment test, i.e. Absolute Category Rating (ACR) and Degradation Category Rating (DCR) [81], [80]. The recommended rating scale for both is based up on a 5-point category scale, also

known as quality scale which covering the options of Excellent (5), Good (4), Fair (3), Poor (2), and Bad (1). These scales are shown in Table **3.1** and **3.2**. Results from the quality scales are averaged across a number of subjects (typically 20/22 candidates) in order to provide a Mean Opinion Score.

In ACR, the untrained subjects are asked to rate the perceived audio quality without listening to the original (reference) sample first, as a comparison. Table **3.1** depicts the opinion scale used in ACR method [80].

Table 3.1: Absolute Category Rating (ACR)

Category	Speech Quality
5	Excellent
4	Good
3	Fair
2	Poor
1	Bad

Table 3.2: Degradation Category Rating (DCR)

Category	Degradation Level
5	Inaudible
4	Audible but annoying
3	Slightly annoying
2	Annoying
1	Very annoying

DCR is an alternative to the Absolute Category Rating. In DCR, subjects are asked to rate annoyance or degradation level by comparing the audio sample under test to the high quality fixed reference (i.e. original sample). This method has been found to be most effective when the impairments to be measured are relatively small. There-

fore, DCR is used when ACR becomes insensitive to the small degradation differences while evaluating good samples of audio. The rating sample as used in DCR method is shown in Table **3.2** [80].

Listening-only test can be assessed via the listening effort scale, which is illustrated in Table **3.3** [80].

Table 3.3: Listening Effort Scale

Effort required to understand the meaning of sentences	Score
Complete relaxation possible; no effort required	5
Attention necessary; no appreciable effort required	4
Moderate effort required	3
Considerable effort required	2
No meaning understood with any feasible effort	1

For a conversation test, the conversation difficulty scale is used. The subjects are normally asked this question [80]:

'Did you or your partner have any difficulty in talking or hearing over the connection'?

- **Yes 1**
- **No 2**

3.5.1.2 Objective Assessment Method

A review of existing objective methods for assessing perceived audio quality in multimedia are now described. Traditionally, the objective assessment methods are perception-based techniques for objectively estimating speech quality in narrow-band telephony networks (300-3,400 Hertz) with low bit-rate CODECs.

These techniques transform speech signals into a perceptually relevant domain such as bark spectrum or loudness domain, and incorporate human auditory models.

A block diagram of the basic perceptual-based approach is illustrated in Figure 3.1 [82].

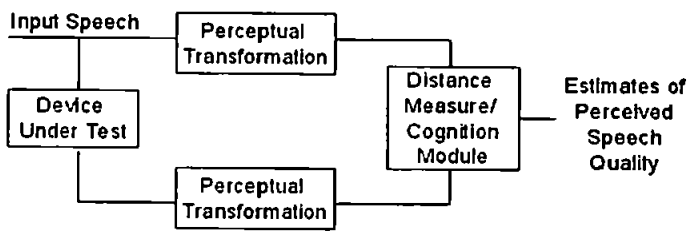


Figure 3.1: Perceptual-based Approach to Quality Estimation

Initially, the input signals are transformed into a appropriate perceptual domain, and only perceptually-relevant speech information is retained (to ensure the accurate quality assessment). The perceptually-transformed speech data are then compared by a distance measure or cognition module that estimates the perceived quality of the coded speech.

There are numerous objective assessment methods currently being used to estimate speech quality, such as, Perceptual Speech Quality Measurement (PSQM) [55] - limited to the assessment of telephone-band speech codecs only [74], Perceptual Assessment of Speech Quality (PAMS) [83], [77], Measuring Normalizing Blocks (MNB), Enhanced Modified Bark Spectral Distortion (EMBSD), and Perceptual Evaluation of Speech Quality (PESQ) [78], [84], [74]. These methods are more accurate for measuring end-to-end perceived speech quality and are unsuitable for monitoring live traffic.

PESQ is widely used to access speech quality in Voice over IP (VoIP) systems. PESQ is the new ITU standard, developed by KPN Research, the Netherlands and British Telecomunications (BT), by combining the two advanced speech quality measures PSQM+ and PAMS. PESQ uses a unique psycho-acoustic hearing model to objectively estimating the perceived quality of coded speech. It predicts the quality scores similar to those that would be given in a subjective test, and both methods correlate quite well. The PESQ scores are calibrated using a large database of subjective tests. This method has proved to be the most successful objective speech quality measure so far. Unlike the other methods, PESQ also addresses the effect of filters, variable delay and coding distortion and is thus suitable for real-time VoIP applications.

The objective method offers the benefits of being more practical, convenient, and inexpensive than that of the subjective test. The major drawback of this technique is due to the limited scope of interpretations.

3.5.2 Assessment Methods of Video Quality

The quality assessment of interactive video in real-time communication has proved to be critical as the quality varies tremendously and unpredictably due to the nature of IP networks and other factors as previously described in Section 3.2.

Since the main use of the video quality in the context of multimedia conferencing is psychological, thus the evaluation of video has often concentrated on the kinds of subjective effects it supports [85].

Video quality can be measured either subjectively or objectively. The International Telecommunication Union, Radiocommunication Section (ITU-R) P.910 or ITU-R Recommendation BT.500, provides guidelines for the subjective video quality assessment methods for multimedia applications [55], which suggest viewing conditions, criteria for the selection of observers and test material, assessment procedures, and data analysis methods.

3.5.2.1 Subjective Assessment Method

Subjective assessment of video quality has traditionally followed the same route as that of audio quality i.e. the MOS and Degradation Mean Opinion Score (DMOS), another recommendation of the ITU-R. The standards are originally concerned with establishing the subjective performance of television systems. Table 3.4 illustrates the image impairment scale as used in DMOS.

Table 3.4: Image Impairment Scale

Image impairment	Score
Imperceptible	5
Perceptible but not annoying	4
Slightly annoying	3
Annoying	2
Very annoying	1

Double Stimulus Continuous Quality Scale (DSCQS) is another example of subjective test method. In these methods, the video clips are presented in pairs, i.e. source and processed video clips. The sequence duration is usually 10 seconds and a maximum of 30 mins of constant viewing is suggested in order to having the effect of fatigue and boredom introduced to the results. The clips are presented in random order, and viewers are instructed to grade each clip's quality. Generally, non-expert viewers are used to evaluate the quality. For rating purposes, a 10 cm graphical scale is divided into five equal intervals. In the middle of each interval the quality terms namely, Excellent 100–80, Good 79–60, Fair 59–40, Poor 39–20, and Bad 10–0 are associated from top to bottom. These scale are shown in Figure 3.2. Unlike MOS and

DMOS methods, DSCQS enables the subject to score between the categories, where the subject places a mark anywhere on the rating line, which is then translated into a score. The data is gathered in pairs.

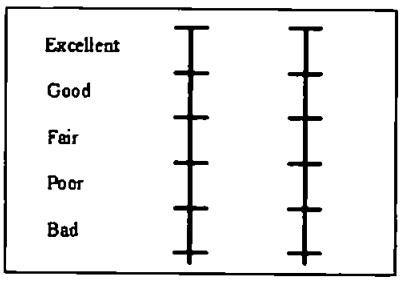


Figure 3.2: Double Stimulus Continuous Quality Scale

As previously mentioned, the DSCQS method uses 10 second sequence segments. However, there is a concern that digital impairments that are short-lived and temporally spaced may not be captured within the 10 second segments. This problem could be overcome by increasing the sequence duration. Hence, a method called Single Stimulus Continuous Quality Evaluation (SSCQE) is introduced. The technique uses the similar quality scale as in DSCQS 3.2, except that only one video clip is being assessed at a time. A slider device with a continuous grading scale composed of the adjectives Excellent, Good, Fair, Poor and Bad is used to evaluate the image quality in real time. The SSCQE method presents a digital video sequence once to the subjective assessment viewer that last about 20–30 mins. The video sequences may or may not contain impairments. Thus, the subject is not aware that he/she is evaluating the reference or processed sequence. The votes are collected continuously i.e. twice per second during the entire test session. This allows the number of sampled points gathered increased by a factor of more than 20 compared

to that of DSCQS method. The disadvantage of this measure stems from the increased data manipulation and analysis. However, it allows for the interpretation of the entire program stream and for a better evaluation of time-varying impairments.

Another subjective assessment approach, developed by University of Central London (UCL) is called Quality Assessment Slider (QUASS). The QUASS is a dynamic software version of DSCQS [75]. The tool consists of a slider bar that allows users to rate the quality of video during the test session. This allows us to see users' perception of quality at any one point in time or continuously during the session, which means that we can see the effect of fluctuations of quality during a session.

The QUASS, DSCQS, and SSCQS approaches are consistent with real-time video broadcasting or in applications that involve passive viewer but are unsuitable for real-time interactive communication, such as videoconferencing. The drawback of these methods is that the continuous rating schemes can result in task interference. Therefore, they have not being selected for the purposes of the tests conducted in the thesis.

3.5.2.2 Objective Assessment Method

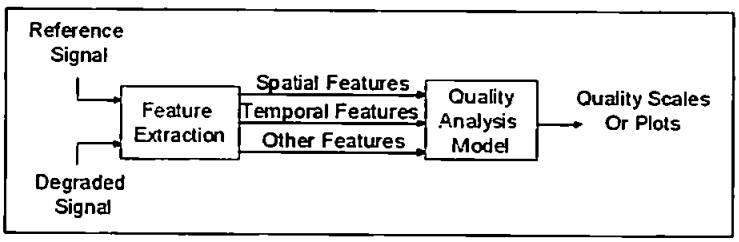


Figure 3.3: Perception-based Objective Picture Quality Measurement System

Figure 3.3 [86] illustrates an example of the perception-based objective measurement systems developed by the National Telecommunications and Information Administration (NTIA)/ Institute for Telecommunication Sciences (ITS), a part of the US Department of Commerce. It has been used for the assessment of video CODECs operated at 56 kbps to 1.5 Mbps for videoconferencing and television signal [87] [88]. The proposed objective test methods do not use human subjects, but rather measure and analyze the video signal using an automatic perceptual measurement metrics, as shown in Figure 3.3. These methods implement an algorithm that measures image quality usually based on the comparison of the source i.e. reference signal and the degraded sequences. These algorithms, referred to as quality analysis models, incorporating the workings of the human visual system while trying systematically to measure the perceptible degradation occurring in the video imagery. Another example of these techniques is known as Multi-modal Perceptual Model, developed by BT Laboratories that combines both audio and visual perceptual model to produce the multi-sensory model [58], [89].

In some situations these objective methods may successfully replace the use of subjective measures. For example, the objective measure is much preferred in applications to assess the quality of multimedia CODECs, still image, video-on-demand or web-based video services, and etc., simply because it is less tedious and cheaper, as well as more practical (i.e. easy to implement automatically). Unfortunately, the objective method is inconsistent with the application involving real-time interactions such as videoconferencing since the physiological data, task performance, and user

behavior become unrealistic. Hence, the subjective MOS is considered the most reliable multimedia assessment method thus far [86], [90]. However, it has several drawbacks that are presented in the next subsection.

3.5.3 Disadvantages of Subjective Methods

As previously mentioned, currently, the subjective methods are widely used to evaluate multimedia quality. This is simply due to the fact that, since it is the end-user who will determine whether a service is satisfactory; it is vital to carry out subjective assessment of the multimedia quality over IP networks. However there is a rising concern regarding the validity of MOS test results, which mainly stems from the inadequacy of the subjective method used for assessing multimedia audio and video. The drawbacks suffered by subjective test, MOS, are namely:

- time consuming, tedious, and stressful - subjective test in a prolong field trial can be quite frustrating and requires a lot of energy and efforts. It requires special viewing room and equipment. It also needs a big group of people and large amount of post processing of data.
- inadequate MOS scale - the MOS scales were originally designed to access high-quality television picture of 24fps and toll-quality audio. Low cost DVC, however, usually provides 2-8fps where lip synchronization is extremely difficult to achieve. Thus, there is a major concern that the vocabulary on the MOS scales cannot be applied to a much lower quality of audio or video obtained

from videoconferencing. Also, the rating scales, thus far, are generally focused on finding the point at which degradation is not discernable;

- cognitively mediated - subjective assessment is cognitively mediated, e.g. a recent study found that users accepted significantly lower media quality when it had a notion of financial cost attached [59]. Also, subjective opinion varies considerably between subjects, possibly due to different level of users expectations and experiences of the technology involved.

- inter-observer issues - the degree of inter-observer interaction is a complex issue and problematic as there are numerous variables which can affect the end user's perception of video quality, such as, network loads, hardware components (e.g. headset), background environment, system configurations (CODECs), and loading on the individual's workstation

3.5.4 Disadvantages of Objective Methods

The subjective method is still preferred to evaluate image quality despite the above mentioned drawbacks [75]. This is because the objective measure possess much greater inadequacies compared to the subjective test. The following, outlines some of the disadvantages of the objective measure.

- the objective methods, which is based on psychoacoustic modelling [91], [92], can only be validated through correlation with subjective test. Today, how-

ever, the subjective testing procedures itself are yet to suffer from different perceptual weights due to a large number of factors (i.e. network constraints, environment noise, systems configuration, etc.) In addition, the more complex issues such as physiological data, task performance, and user behavior become impossible to implement. Thus, there is a dearth of subjective results against which to measure any new model;

- the subjective methods have the advantage of providing a wealth of information as compared to that of objective measures which are often limited in the scope for interpretations. For example, some important complex issues, such as, creating a sense of presence and users interactivity are clearly impossible in the objective test;

- the objective test suffers from the practicality when applied to the real world situation. It is mandatory to create the hypothesis of the imitation of how the particular task would be judged in the real environment, for the validity of results. For example, the best measure of subjective quality will be gained from people engaging in a conversation over a connection. In laboratory settings these conversation can be quite artificial. Furthermore, in the context of assessing videoconferencing systems, it is the end-users who interact with a network application (in real-time) to achieve a particular task, that determines whether a media QoS is satisfaction;

- it is stated that the objective test methods, namely, PSQM, PSQM+, MNB have poor correlations with subjective MOS in some commonly-occurring condi-

tions in IP networks, such as, packet loss, speech clipping, background noise, variable delay and filtering [84].

Hence, for the above reasons the subjective test method, MOS has been employed for the thesis.

3.5.5 Problems in Assessing IP Media

To date, multimedia conferencing is facing a challenge as to whether the current methods used in assessing the multimedia quality provided is feasible or not. As previously mentioned, there is lack of the standard methodology to determine the IP media quality. The major concern is that the continuous assessment scale can occur when the test is conducted in isolation and carried out without any reference to task or unsupervised. Recently, there are numerous techniques being introduced, either subjective or objective, as previously explained. Both these techniques, inevitably suffer from the variability in the parameter imposed by the nature of IP networks, such as:

- the quality varies unpredictably with time – the range of impairments of audio, video, and combined audiovideo quality in real-time IP networks is highly time varying, due to low bit rates, error prone environments, different network load etc.;

- unique impairments – audio and video streams delivered over IP packet networks are susceptible to unique impairment due to the unpredictable nature of IP networks. In addition, different levels of task scenarios and difficulties, introduce unique impairments to the subjective perceptual of multimedia quality [93].

- subject opinion factor – with respect to the subjective assessment method, the test candidates opinions can be very subjective in that they vary largely between one person and another. For example, one person's 'Good' rating scale might be another person's 'Fair' or even 'Poor'. Hence, to address this issue, a large number of subjects is needed (typically between 20–22 participants as recommended by the ITU-T P.800 [80]) for test, so that the average results can be obtained. Also, the subjects were introduced to the ideal quality of the videoconferencing system (Microsoft NetMeeting Version 3.01), i.e. without network constraint that can be used as a common reference before the tests commence.

These variables prove that the task to predict the perceived quality of audio and video media over IP videoconferencing systems is very challenging and needs to be thoroughly addressed.

3.6 Summary

This chapter has discussed some aspects that influence the perceived quality of IP multimedia applications, in terms of end user's QoS. It has also discussed the importance of developing methods of evaluating audio and video quality for videoconferencing over the Internet. A critical review of the most commonly used assessment methods has been presented, providing the advantages and disadvantages of each technique, in order to establish which methods give superior performance or are most suitable for the research objectives.

Based on the findings described in this chapter, the subjective evaluation method (MOS) has been employed for the research work. Both the passive and interactive task performance have been considered, that involved two parties communicated in real time, using the Microsoft NetMeeting, in order to closely imitate the real world scenarios.

As stated in Chapter 1, the work in this thesis concentrated in benchmarking the state of the art in low cost IP videoconferencing systems. All the research findings that addressing the multimedia quality thus far, are based on experiments involving higher system bandwidth or more sophisticated multimedia CODECs, such as, TV system, video on demand, MPEG CODECs etc. [2], [3], and [47]. To date, little research has been carried out on the assessment of multimedia quality over low cost DVC system. Therefore, the study has been focussing upon the investigation of the perceived multimedia quality (with respect to network constraints, CODECs, task

performances etc.), in low cost desktop videoconferencing systems.

The next chapter presents the research approach employed in the thesis. The concept and methodology of the research approach will be clearly described, and a description as to why they are considered as being relevant focus for the designed experiments will be thoroughly explained.

Chapter 4

Research Methodology

4.1 Introduction

The previous chapters have described the DVC technology and discussed the issues of the assessment of the current multimedia quality over the packet network. Given the facts and differences that have been discussed thus far, the concept and methodology of the research approach is described in this section and why they are considered as being relevant for the designed experiments is thoroughly explained.

The basic aim of the thesis is to investigate a novel method used to optimise the perceived quality of audio and video in videoconferencing systems. The approach is to establish a quality threshold for audio and video required for specific task performance and application, and to use it as the basis for a control mechanism to predict the transmission quality of audio and video in multimedia applications. The work involved in benchmarking the state of the art in Internet-based multimedia conferencing. The problems, inherent in clarifying QoS as required by the end user, were

investigated. The work concentrates on the evaluation of perceived audio and video quality in the existing videoconferencing system with respect to different types of network impairments (packet loss, jitter and delay), CODECS, system configurations, and task performances. Microsoft NetMeeting was selected over other existing IP telephony tools due to its popularity and readily available software.

4.2 Testbed

4.2.1 General Overview of Test Bed

The test bed configuration used is shown in Figure 4.1.

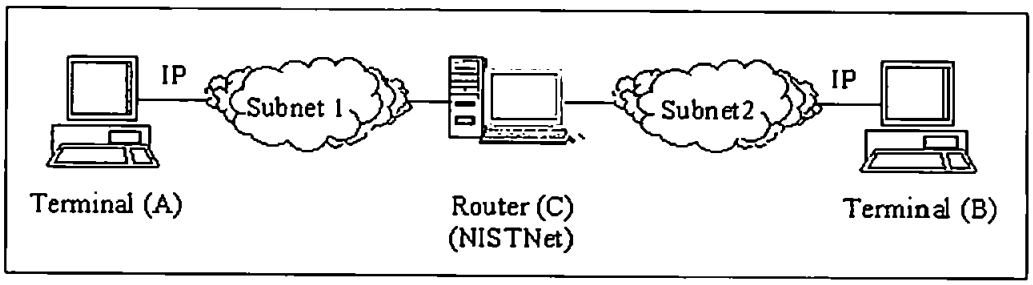


Figure 4.1: Test Bed Configuration

The test bed consists of three machines: two machines (Terminal A and Terminal B) running audio-video clients (videoconferencing) using Microsoft NetMeeting, and one Linux machine (Terminal C) with NIST Net emulator [94] in between (see Figure 4.1).

NIST Net is a software package that can be used to emulate the performance of a variety of TCP/IP networks and network paths. NIST Net (version 2.0.12) is installed on a Linux box with two Ethernet Cards (10/100Mbit) which is routing between the two subnets, (i.e. Subnet 1 and Subnet 2) as shown in Figure **4.1**. This allows it to increase or decrease all incoming traffics flowing between the two networks. Detailed information on NIST Net is described in the following section.

4.3 NIST Net

NIST Net is a simple Linux kernel-based network emulator that operates at the IP level [94]. It allows a single Linux box set up as a router which provides the ability to emulate common network effect such as packet loss, duplication or delay, router congestion and bandwidth limitations. The studies described in the thesis were based on packet loss, jitter, and delay. It is a requirement that the value be reliable, controllable, and repeatable throughout the studies for the validity of results.

NIST Net has been used for emulation up to line rate over 100Mbps Ethernet with system specification; 200MHz Pentium-class processors and PCI-based 10/100 Ethernet cards.

NIST Net consists of two main parts: (1) a loadable *kernel module*, which hooks into the normal Linux networking and real-time clock code, implements the run-time emulator proper, and exports a set of control Application Programming Interfaces

(APIs) that define how to access software-based services: and (2) a set of *user interfaces* which allows controlling and monitoring a large number of emulator entries simultaneously. Figure 4.2 show the NIST Net emulator architecture [94].

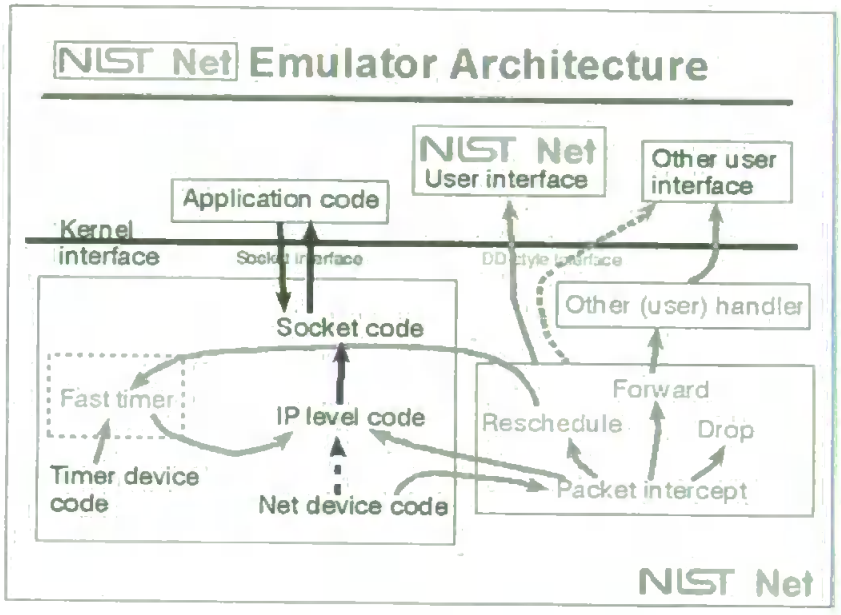


Figure 4.2: NIST Net Architecture

4.3.1 Running NIST Net

Before running NIST Net, the kernel emulator module must be installed through *insmod nistnet* or *Load.Nistnet* (see Appendix C). There are two tools for controlling the emulator, Hitbox which is a command line interface and Nist Net which is a graphical user interface (GUI). By using its command line interface or GUI, different parameters like delay, packet loss, jitter, bandwidth etc. can be set for all incoming IP packets.

Figure 4.3 shows the NIST Net graphical user interface [94] where different sets of impairments can be introduced on each audio and video stream by simply typing in the desired values into the respective columns. To load a new set of traffic impairments, the *'Update'* button must be pressed to clear the old settings.

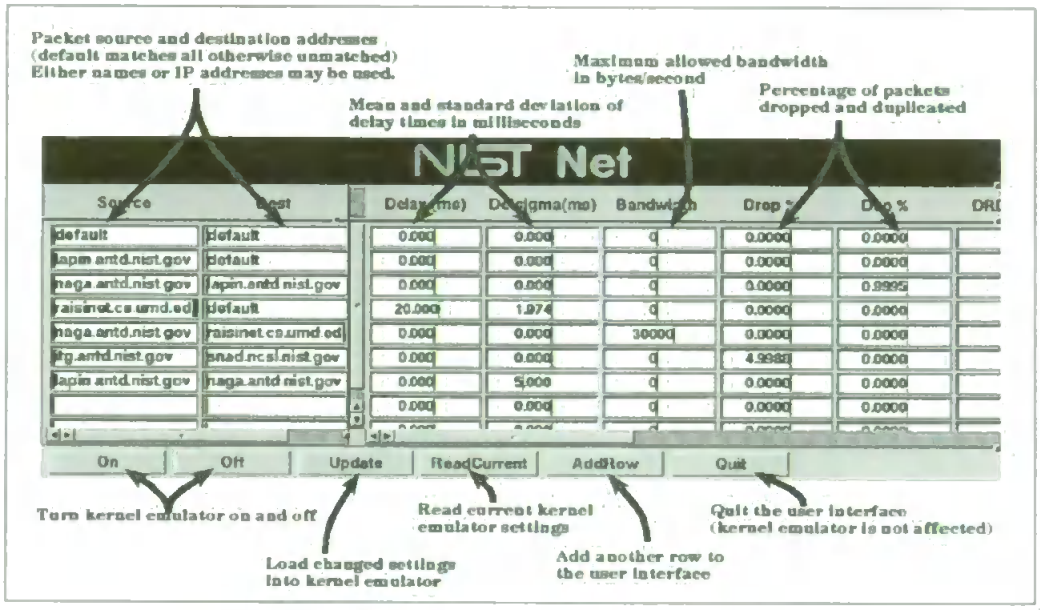


Figure 4.3: NIST Net Graphical User Interface

The Hitbox mode takes a lot of arguments some of which are explained further in Appendix C [94]. Table 4.1 and Table 4.2 show examples of data input for ideal network and 5% packet loss or drop respectively, using the command line.

Table 4.1: Command Line Data Input – Ideal Network

```
linux:~ #
cnistnet -a 141.163.49.182:49606.udp 141.163.50.2:49606.udp - -drop 0 - -delay 0 0 - -delay 0
cnistnet -a 141.163.50.2:49606.udp 141.163.49.182:49606.udp - -drop 0 - -delay 0 0 - -delay 0
cnistnet -a 141.163.49.182:49608.udp 141.163.50.2:49608.udp - -drop 0 - -delay 0 0 - -delay 0
cnistnet -a 141.163.50.2:49608.udp 141.163.49.182:49608.udp - -drop 0 - -delay 0 0 - -delay 0
```

Table 4.2: Command Line Data Input – 5% Packet Loss (Drop)

```
linux:~ #
cnistnet -a 141.163.49.182:49606.udp  141.163.50.2:49606.udp - -drop 5  - -delay 0  0  - -delay 0
cnistnet -a 141.163.50.2:49606.udp  141.163.49.182:49606.udp - -drop 5  - -delay 0  0  - -delay 0
cnistnet -a 141.163.49.182:49608.udp  141.163.50.2:49608.udp - -drop 5  - -delay 0  0  - -delay 0
cnistnet -a 141.163.50.2:49608.udp  141.163.49.182:49608.udp - -drop 5  - -delay 0  0  - -delay 0
```

4.4 Microsoft NetMeeting

NetMeeting is a software program developed by Microsoft Corporation. It is an end-user application that allows videoconferencing over any IP network connection, including LANs, the Internet or any Intranet. The software is a freely available conferencing solution that can be downloaded from the Internet at [95] and readily available for all desktop versions of Windows since Windows 95. The software supports four types of real-time collaboration, namely, audio, video, whiteboard window, and chart board window that allow users to enable any one, two, or three of the media simultaneously. However, only audio and video channels were being used in the study. In this thesis, NetMeeting Version 3.01 has been employed.

The audio and video CODECs used in Netmeeting are described in the following.

- **G.711**: this Microsoft ITU high-bit-rate CODEC is appropriate for audio over higher speed connections. It transmits audio at 48, 56, and 64 kilobits per second (Kbps). The frame size is 30ms. There are two type of PCM (G.711) encoding laws, i.e. the A-Law and the μ-Law;

- **G.723.1**: this compression technique can be used for compressing audio at a

very low bit rates, and transmits audio at two rates, i.e. at 5.3 and 6.3 Kbps which reduces bandwidth usage. For the high bit rate, Multi-pulse Maximum Likelihood Quantization (MP-MLQ) excitation is used, and for the low bit rate, an algebraic-code-excitation (ACELP) is used. The frame size is 30ms;

- **H.261**: this standard videoconferencing CODEC transmits video images at 64 Kbps. It is designed for low bit rates and relatively low motion applications, for example, videophone and videoconference over ISDN. The bit rate is represented as n x 64 kbps (n=1, 30). It supports two frame sizes, i.e. CIF (352x288), and QCIF (172x144). The coding schemes:

  - DCT based compression to reduce spatial redundancy

  - block based motion compensation to reduced temporal redundancy;

- **H.263**: this new scheme video CODEC is the advancement of the H.261. It is designed for low bit rate video application (10-384 kbps). The coding algorithm is similar to that used by H.261, however with some new developments and changes to improve performance and error recovery. Hence, it provides better quality even on higher bit rate than H.261. It supports Common Interchange Format (CIF), Quarter Common Interchange Format (QCIF), Subquarter Common Interchange Format (SQCIF, 128x96), and 16CIF (1408x1152) picture formats. H263 is designed with the goal of producing substantially better video quality below 64 kbps.

Figure **4.4** shows the main NetMeeting window showing a participant image taken from the field study i.e. between University of Plymouth and University Sarawak Malaysia (UNIMAS).

Figure 4.4: The Main NetMeeting Screen

The primary limitation of Netmeeting is in the performance of the video and audio over unicast connection and the lack of the effective multiparticipant conferencing support. Netmeeting would therefore only feasibly be useful for point to point videoconferencing or for application sharing within an Intranet between two users on Windows machines.

Figure 4.5 shows a typical desktop videoconverencing set-up, showing the webcam, headset, and Microsoft NetMeeting window.

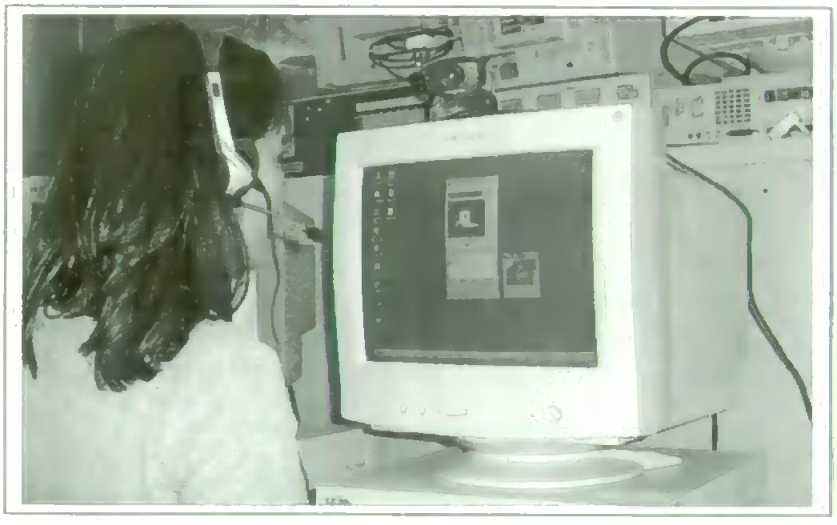


Figure 4.5: Typical Desktop Videoconferencing Layout

4.5 Different Network Conditions

The initial phase of the study was based on the static ratio of audio and video degradations, with respect to packet loss, jitter and delay. The objective was to assess the individual impact of the given network constraints on the perceived multimedia quality. Later, the study proceeded to investigate the appropriate audio and video deterioration ratios (to provide 'best-effort' QoS, with respect to task), interaction effects between audio and video media and lip sync problems. Hence, to achieve these objectives, the technique of introducing different sets of audio and video impairments ratio has been employed for each designated test, prior to transmission.

The network emulation tool (NIST Net) [94] was used to introduce different sets of impairments on each audio and video stream.

As perviously mentioned, the assessment of audio and video media over IP is a very complex issue. The range of audio, video and combined audiovideo quality is highly time varying, due to low bit rates (low bandwidth), 'noise' prone environments, varying background conditions and unreliable characteristic of IP networks. It is also suggested that, the same network impairments have different impact on perceived speech quality due to different languages and talker's voice (male or female) [96]. Moreover, audio and video delivered over IP network is susceptible to unique impairment, such as packet loss, that leads to novel type of audio and video degradations. Hence, to address these issues, the initial stage of the thesis was to investigate the individual effects of each network constraint (such as delay, jitter and loss), on the perceived quality of audio, video and audiovideo overall.

It is anticipated that the static weighting utilized in the initial phase will not be appropriate for all scenarios, as 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.

It is also suggested that the perception of one media can affect and interact with the perception of the other. For example, the perceived video quality can be increased by improving the audio quality and vice versa. Therefore, it is essential to first iden-

tify and understand the role that each of those parts plays, and how they interact with one another. Bearing this fact in mind, the research focuses on investigating the interaction between the perceived audio and video components in multimedia conferencing. One complicating factor has been observed, i.e. the requirements for audio and video will be task dependent. This also indicates that potential trade-offs between the audio and video quality could be exploited with respect to task. Indeed, a better understanding of the relationship between audio and video quality with related task is beneficial to maximise bandwidth utilisation.

Therefore, in the later phase of the experiment, different levels of network loads are employed in the audio and video sequences, individually. For example, audio is degraded by 5% packet loss while video quality is unimpaired or vice versa. The main objective is to investigate the correlation between audio and video media.

Different delay parameters were also introduced to the audio and video components, separately, to investigate the effect of lip sync error on the perceived multimedia quality in low cost DVC, in the context of passive and (informal) interactive communications.

4.6 Method of Assessment

As described in Section 3.5, there are two approaches to investigate user perception of DVC QoS, known as subjective and objective test methods. The subjective test,

in a prolonged field trial is costly, time consuming, prone to external variables (such as background noise) and can be frustrating for the subjects (due to technical problems), while the objective test methods suffer from the practicality when applied to the real world situation [70].

The test methods adopted in this thesis represent the state of the art in the subjective assessment of multimedia service. The methods and procedures for conducting subjective evaluations of transmission quality are based on the ITU-T Recommendation, P.800 (i.e. Methods For Subjective Determination Of Transmission Quality), namely, room setting/condition, selection of the test subjects, test set-up etc.

The subjective test method has been employed in the experiments conducted in the thesis, simply due to the fact that, since it is the end-user who will determine whether a multimedia service over an IP networks is a satisfactory or otherwise. Moreover, the subjective methods have the advantage of providing a wealth of information as compared to that of the collected data, obtained from objective test, which are often limited in the scope of interpretations. Detail discussions upon this matter have been elaborated in Subsection 3.5.4.

The Absolute 5-point Category Rating MOS, introduced by ITU-T (1996) (see Section 3.5.1.1) is used to assess the perceived quality of audio, video, and audiovideo overall. The 5-point rating scale has been selected since it has been widely used in evaluating audio and video quality and has been approved by the ITU organisation. The fact that the scale is easy to administer and score has become the main advan-

tage of the MOS 5-point rating method, since it is mandatory that the test candidates completely understand their tasks without much effort and not to be confused by a complicated evaluation technique. Furthermore, the scale labels are easy to define and translate into different languages across the world, and hence, the quality results can be globally generalised.

Once the test was completed, the results were statistically analysed, and graphically presented according to the procedures described in ITU-R BT.500 [97], for example, the mean and the standard deviation results were produced.

4.7 Audio, Video and Audiovideo Overall

Previous research stated that, different component media in videocoferencing applications, especially audio and video, interact and influence the perception of each other. The interaction effect between an IP media is a complex one and needs to be addressed in depth. Assessing the quality of a component medium (audio and video) individually will not produce accurate predictions as to perceive the quality in the overall multimedia conference performances. In addition, evaluating the quality of a component medium in isolation is unlikely to give accurate scores as to perceive quality in a full multimedia conference, since the possible perceptual benefits afforded by interaction will not be present. Therefore, alongside the audio and video components, it is suggested that the combined effect of audio and video quality needs to be considered in order to adequately assess the multimedia QoS.

4.8 Task Performance

The experimental procedures conducted in this thesis was designed for specific applications in controlled testing environments with defined viewing and listening to a 'talking-head', in passive test and informal 'person to person' conversation, in interactive test. It is important to establish a fixed benchmarking standard i.e. using a specific task with fixed goals throughout the stages of studies to validate the comparisons of results. For this reason, the assigned task performances were maintained throughout the study.

It is also mandatory that, the task scenario (interactive test) is carefully designed, in order to closely duplicate the real world situations. It is reported in the ITU-T Recommendation, P.920 that lively audiovisual conversations can be maximised if the communicative parties know each other [98]. Therefore, subjects who were acquainted with one another were selected and they were allowed to select their own issues for discussion.

For the passive test, the test candidates were required to listen and view to a female talker, in real-time, saying: "You just have to be quiet, there is nothing to be seen", as suggested in [80]. The ITU-T P.800 also stated that it is mandatory that the speech material should consist of simple, meaningful, short sentences, and easy to understand. The material should not be too long, i.e. should fill the slot of 2-5 seconds to avoid fatigue and loss of motivation [80].

4.9 Eligibility of subjects

One of the criteria of the test was that the subjects should have normal vision and hearing. Also, the subjects should be able to learn the task faster and they that are less susceptible to fatigue and loss of motivation. For these reasons, a group of subjects aged between 18 to 45 years old were selected, since it is believed that it will more likely to meet the subject's eligibility requirements.

According to ITU-T P.800, generally only non-expert viewers/listener are used to evaluate multimedia quality. Non-expert respondents are people who have no prior professional or extensive personal experience in dealing with multimedia display systems or devices. They should not have been directly involved in the related work such as assessing coding techniques or other multimedia systems. It is mandatory that they have not participated in any subjective test whatever for at least the previous six months, and not in a conversation test for at least one year [80].

For this reason, most of the participants have little experience (if any) with IP video-conferencing environment, but are familiar with Internet (e.g. listening radio or viewing video clips via Internet). Almost all of the candidates (95%) were categorised into the researcher, lecturer and student user groups.

Unfortunately, the majority of the tests participants were males, i.e. 80% (approx.) of the total number of the subjects. This is due to the larger population of male students in the Engineering department (UoP), and also the lack of response from the female

students, at the University during the time when the tests were conducted. It is essential to note that one of the main obstacles of the research programme is to find the test candidates.

4.10 Familiarization of Test Procedures

Before the tests commence, the candidates were introduced to use the videoconferenncing tools. Firstly, they were encouraged to get familiar with the systems and the demonstration of how the system works was given, when it was needed. They were also thoroughly informed about the tasks they were going to perform, both orally and in using written instruction sheets. The answer sheets, pen and pencil, and the sample of the ITU-T 5-point rating score (for the reference) were placed in front of the subjects, and it is important that the subjects were fully understood the whole procedures of the test before the commencement of the experiment. Only after they are totally at ease and fully understood the task that they are going to perform, then only the real test begins.

The subjects were asked to rate the quality of the video while audio was also being present (and vice versa) as would happen in the real life situation to minimise the gap between the experimental set-up and real world. For the same reason, the overall audiovisual quality was also considered.

4.11 Field Trial

The earlier stage of the studies are classroom-based experiments with contrived tasks assigned to the subjects. Although with a thorough research, control and well-trained subjects, the task can be made to look and feel real, it is considered that, only through a field study that the true sense of the impact of multimedia technologies on the users can be manifested. Thus, the later part of the study was conducted between the University of Plymouth and University Malaysia Sarawak in order to approximate real world situation more strictly. In addition, it is essential to make comparisons between controlled classroom-based experiment and real-world environments to enrich the results of the study.

4.12 Summary

This chapter has described the concept and methodology of the research approach, and has explained why they are considered as being relevant for the designed experiments that proceed. The next chapter describes the early stages of the research work, i.e. to benchmark the current state of the art of desktop videoconferencing against the various factors that affect the perceived quality of audio, video and audiovideo overall.

Chapter 5

Investigating The Effects of Network Constraints

5.1 Introduction

The initial stage of the research has been designed to investigate the current state of the art in desktop videoconferencing systems. This chapter focuses upon benchmarking the performance of the popular Microsoft NetMeeting with respect to the related issues that affect the perceived audio and video quality, such as network congestions, computing resources, tasks performance, CODEC, and conferencing hardware. NetMeeting was selected over other existing IP telephony tools due to its readily available software and its popular usage in the current market.

At present, despite the increased popularity of low cost DVC, it is often questioned whether the quality of the audio and video provided is adequate for the tasks that users wish to perform. Previous research suggests that the perceived quality of au-

dio and video varies according to the task undertaken and user expectation also varies accordingly [59]. In the meantime, the issue of determining multimedia conferencing quality has certain difficulties, as there is no recognized industry standard of what really determines audio and video quality. In addition, assessing the quality of multimedia over the Internet is further complicated due to its constantly changing and unpredictable nature [67]. Many research efforts are now being directed toward developing new approaches in assessing audio and video quality in IP multimedia [76], [89], [75].

As stated previously, audio and video quality can be measured either subjectively or objectively. It is generally agreed that subjective methods are more reliable [93], but recent research findings suggest that the subjective method alone is inadequate to determine the audio and video quality in videoconferencing [67], [70], [76]. Throughout our research, the method of assessment being used was a subjective test method, called Mean Opinion Score (MOS) and is the standard recommended by the ITU-T (CCITT, 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 test experiments were divided into four different Tasks:

- *Phase 1*: General assessment of NetMeeting without network constraint;
- *Phase 2*: Evaluation of packet loss effects on the perceived quality of IP media;
- *Phase 3*: Evaluation of jitter effects on the perceived quality of IP media;
- *Phase 4*: Evaluation of delay effects on the perceived quality of IP media.

The NIST Network Emulation Tool (NIST Net), was used to introduce packet loss, jitter and delay (for Phase 2, 3, and 4) in the IP network between the two communicating parties (subjects), as shown in Figure **4.1**.

The associated study had three main aims:

1. To investigate the performance of NetMeeting. This involves assessment of audio quality, video quality and combined audio and video quality under a real network environment, and also under assigned network constraints, i.e. packet loss, delay and delay jitter;

2. To investigate the task performance effects on audio quality, video quality and the combined quality of audio and video under a real network environment, and also under assigned network constraints;

3. To investigate the impact of the two speech CODECs, i.e. PCM and G723.1, on perceived audio and video quality.

5.2 The Experimental Approaches

5.2.1 Test Bed Configurations

The test bed configuration is shown in Figure 4.1 (Section 4.2.1). The subjects were seated in the two separate rooms provided with 15 inch monitors. For the experiments, Pentium III 933 MHz systems with 128MB (CDRAM) were used and a USB video blaster Webcam Plus (capable of video capture up to 30 frames per second @ 352x288 pixels and 15 frame per second @ 640x480 pixels, was mounted on each monitor.

To send and receive audio, two identical Platronic PC headsets were used. The test viewing/listening conditions were designed to closely comply with those described in International Telecommunications Union Recommendation ITU-R BT.500-10.

The tests were conducted using two different audio CODECs, which are μ-Law (PCM) and G723.1 (6400bit/s) to give a comparison of results. The PCM CODEC was selected as it offers the best audio quality and can be used as a control mechanism.

The G723.1 CODEC was being used mainly for its growing popularity in the low bandwidth videoconferencing industries. It is a newly and more advanced coding techniques (using MP-MLQ/ACELP: Multipulse Excitation with a maximum likelihood-quantizer/Algebraic Codebook Excitation Linear prediction), purposely

designed to improve QoS in DVC. Apart from having low bandwidth, the G723.1 CODEC offers a good audio quality with dual rate, i.e. 5.3/6.3 Kbps.

The H.263 video CODEC, providing the Quarter Common Information Format (QCIF-176x144) frame size was used, as Common Information Format (CIF-325x288) displayed an almost still-like picture. Due to its better performance and low bandwidth, the H.263 has becomes popular and it is being used in the 3G mobile. The video setting was unchanged throughout the test, i.e. 'better quality' for control purposes.

Based on thorough laboratory experiments it was observed that 'better quality' setting offers the best option (compared to 'faster video') for the study since it is preferred to received a better video quality than a faster frame rate but with a much degraded image.

The audio and video were transmitted from one end to another using a testbed containing the NIST Network Emulation Tool (NIST Net) - a general-purpose tool for emulating performance dynamics in IP network [94] (see Section 4.2.1).

Each subject was provided with a self-view window, a remote view window and a talk window to send text to the remote partner.

5.2.2 Methodology

Ten pairs of volunteers from the university were involved in the interactive test. The same subjects were also participated in the passive test. These participants were aged between 20 and 40 years and have normal hearing and vision. It is essential to note that the subjects should learn the task faster and that they are less susceptible to fatigue and loss of motivation. It is believed that these subjects had met the criteria.

The subjects also had very little experience (if any) of using the software. However, they are all familiar with the multimedia quality over the Internet. According to ITU-T P.800 the test subjects should not been directly involved in the related work such as assessing coding techniques or other multimedia systems.

Previous research implies that informal communication tends to be representative of individuals who are familiar with each other [99]. Hence, to maximise task motivation and to ensure subjects were fully comfortable with each other, subjects who were acquainted with one another were selected and they were allowed to select their own issues for discussion.

Also, to ameliorate the task performance, and ensure that the participants were fully at ease during the conversation, they were encouraged to use their mother tongue while conducting the test.

For the experiment in Phase 1 to 4 the test scenarios, classified as below were introduced:

- ***Passive Communication*** : i.e. viewing and listening to a 'talking head', without interaction), was to benchmark the quality of: (a) Audio only, (b) Video only and (c) Audio and video overall;
- ***Interactive Communication***: where the subjects were involved in intensive informal communications and occasionally performed other tasks, such as marking the scores on the answer sheets and lift up an object (i.e. mugs, pen, etc.), was to benchmark the effects of content on the quality of: (a) Audio only, (b) Video only and (c) Audio and video overall.

Table **5.1** shows the categories of the subjects in terms of nationality and gender that involved in both interactive and passive test.

Table 5.1: The Categories of Subjects

Number of Subjects	Nationality	Gender
2	African	Males
2	Arabian	Males
4	Chinese	1 Male; 3 Females
3	English	Males
1	French	Male
5	Greek	3 Males; 2 Females
1	Indonesian	Female
2	Malaysian	Males
Total = 20		Total = 16 Males 4 Females

5.3 The Experiments and Results

This section explains the task of the experiments, the results and the observations derived from the studies.

5.3.1 General Assessment of NetMeeting Without Network Constraint (Phase 1)

5.3.1.1 Task Description

In Phase 1, the test was designed to evaluate the performance of NetMeeting in an ideal network environment (i.e no packet loss, delay or delay jitter). The experiment was conducted under two different task performance i.e passive and objective communications, as explained previously. The performance of two configurations was compared, operating firstly with no network constraints then subsequently with impediments introduced (as described in Phase 2, 3 and 4).

5.3.1.2 Results

Figure 5.1 shows the MOS for the perceived quality of audio, video and audiovideo overall for two different audio CODECs i.e., (μ-Law, PCM and G723.1, 6400bit/s), obtained from the experiment under ideal network condition. For the perceived quality of audio, PCM (μ-law) scored higher MOS, in both passive and interactive

test. The MOS were 3.9 (PCM) and 3.69 (G723.1) for passive test; 3.76 (PCM) and 3.56 (G723.1) for interactive test. Under error free network conditions, the perceived audio quality gave higher MOS as compared to that of video and audiovideo overall. In general, passive test scored higher ratings than interactive test.

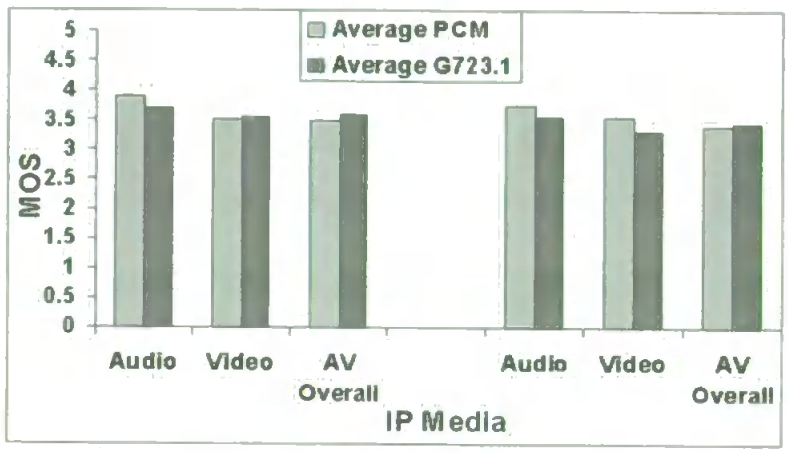


Figure 5.1: MOS Under Ideal Network Configuration

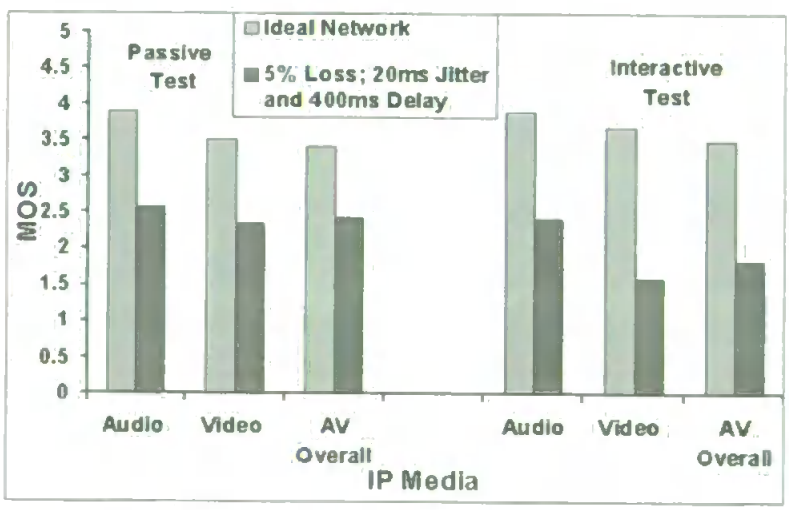


Figure 5.2: MOS Under Ideal Vs Congested Network (PCM CODEC)

Figure 5.2 shows the comparisons of the MOS between ideal and congested network obtained from test using PCM CODEC. The overall rating for general performance

of NetMeeting under ideal network was either Good (4) or Fair (3). On the other hand, the overall score for system under the congested network, was Poor (around 2.5 MOS for audio; 1.6-2.3 for video and 1.8-2.4 for audiovideo), in either passive or interactive test.

It can be seen that the MOS of the perceived quality of video in the interactive tests dropped significantly under the congested network as compared to that in passive test. The MOS dropped by 2.09 of that obtained from the ideal network test, for interactive and 1.6 for passive. This in effect, reduces the MOS for audiovideo in interactive test which is significantly lower than that in passive test.

It is observed that even under ideal network condition, the MOS is not a perfect, i.e. none of the subjects gave an Excellent (5) MOS rating. This indicates that Microsoft NetMeeting offers a much lower performance than the TV system, where the MOS rating scale was originally designed for.

These results were as expected with the MOS degrading for audio, video and audiovideo overall when network congestion was introduced.

5.3.2 Evaluation of Packet Loss, Jitter and Delay on the Perceived Quality of IP media (Phase 2, 3 and 4)

5.3.2.1 Task Description

In Phase 2, 3 and 4, the network emulator tool (NISTnet) was used to generate delays, jitter and packet loss between the IP networks. In Phase 2, packet loss of 3%, 4%, 8%, 10%, 15%, 20% and 30% were randomly introduced on both audio and video medias, prior to transmission. In Phase 3, the jitter values of 10ms, 20ms, 30ms, 40ms and 80ms were used for both audio and video streams. While for Phase 4, the delays of 300ms, 400ms, 600ms and 800ms were selected for the experiment. Task performances as described above were conducted and the subject were asked to rate the perceived quality of audio, video and audiovideo overall, at the end of each test section.

5.3.2.2 Results: Evaluation of Packet Loss on The Perceived Quality of IP media

Figures **5.3** and **5.4** show the effects of packet loss on video, for passive test and interactive test, respectively. Video was found to be error free for packet loss below 3%, (MOS is between 2.9 and 3.5). At 4%, MOS drops to 2.6 or 2.7 where the video quality deteriorated further and became apparent. On approaching 8% packet loss, the video channel is reported to become worse and unusable where the MOS is nominally 2.5. Therefore, it is concluded that the video media becomes insignificant

when the perceived MOS drops to 2.5.

Overall MOS for video, obtained by the system using G723.1, shows slightly lower results although video settings are not changed. This raises suspicions as to whether the change in audio quality would cause the change in perceived video quality and vice-versa.

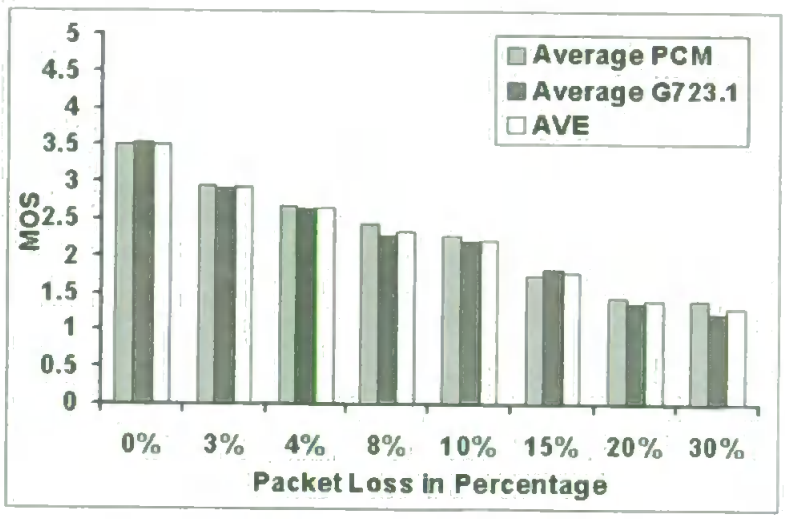


Figure 5.3: Passive–Loss Effects on Video

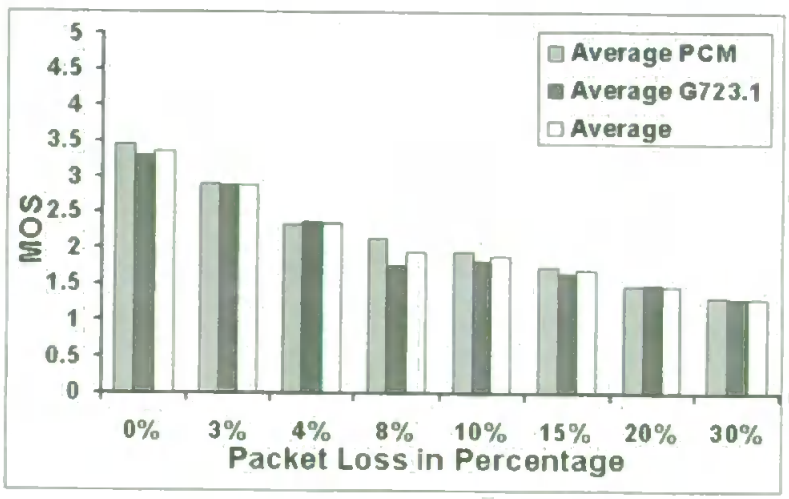


Figure 5.4: Interactive–Loss Effects on Video

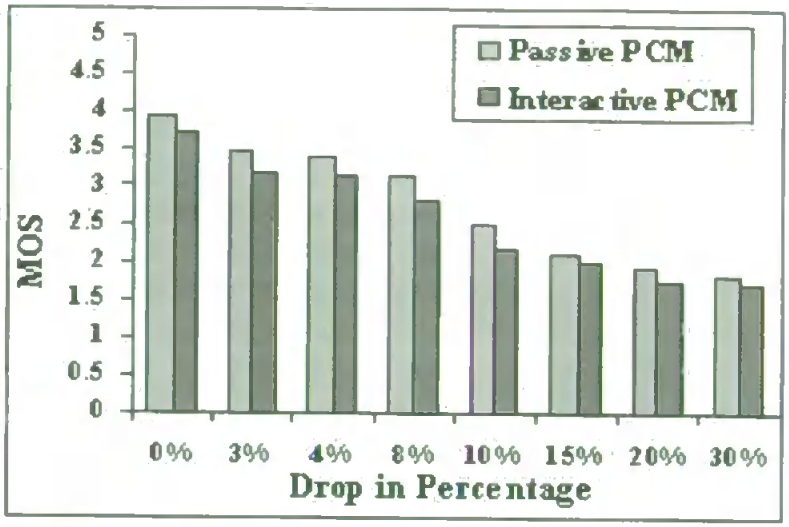


Figure 5.5: Packet Loss Effect on Audio - Passive Vs Interactive

The MOS rating for loss effects on audio, for system that using PCM (μ-Law), is Good (MOS 3-4) at 0% to 8%. Whereby, the system employing G723.1 produced lower MOS (2.5) at 8% loss. This indicates that systems employing PCM (μ-Law) are more tolerant to packet loss. Packet loss on audiovisual overall is generally the same as that on video.

Figure **5.5** shows the packet loss effects on audio, comparing the results obtained from passive communication and interactive communication. It is evident that passive communication produces higher MOS than interactive communication. Therefore, it can be concluded that one-way communication is less affected by the packet loss as compared to the interactive test.

Results given by video quality (see Figures **5.3** and **5.4**) and audiovisual quality overall followed the same pattern of results given by loss effects on audio, but with lower MOS.

5.3.2.3 Results: Evaluation of Jitter Effects on The Perceived Quality of IP Media

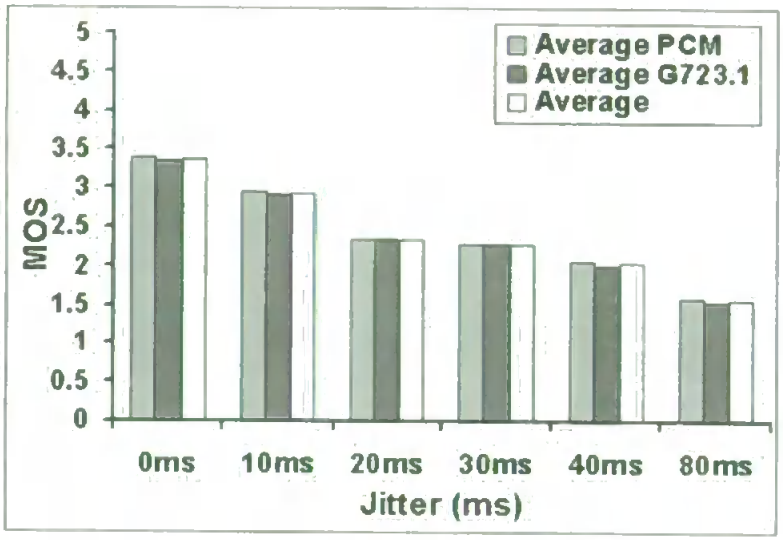


Figure 5.6: Passive–Jitter Effects on Audiovideo Overall

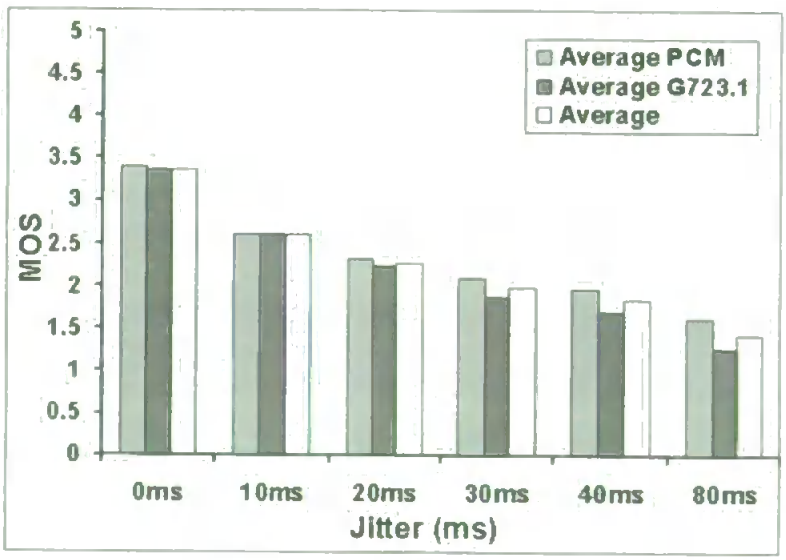


Figure 5.7: Interactive–Jitter Effects on Audiovideo Overall

Figures 5.6 and 5.7 show jitter effects on perceived audiovideo quality overall for passive and interactive, respectively. Between 0ms-10ms jitter, the perceived au-

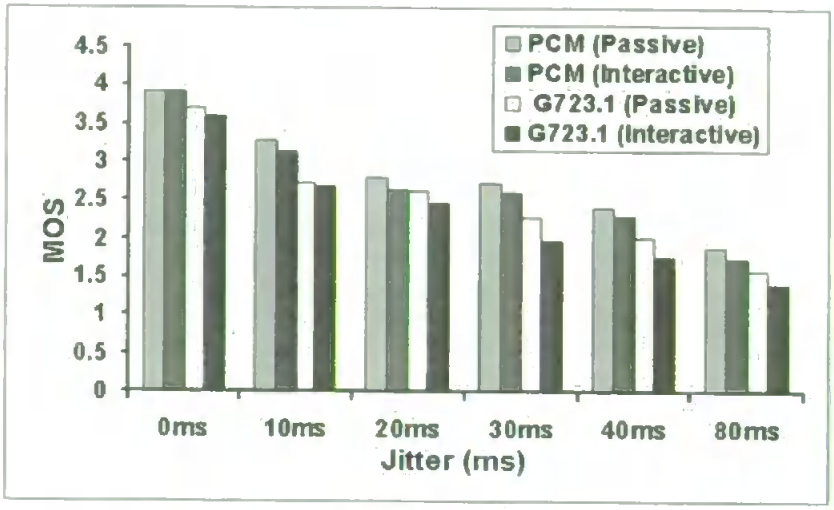


Figure 5.8: Jitter Effects on Audio - Passive Test Vs Interactive Test

diovideo quality is Good, i.e. MOS 3.4-2.9 (PCM) and 3.3-2.7 (G723.1) in passive test; 3.4-2.6 (PCM and G723.1) in interactive test. At jitter between 20ms - 30ms, both systems scored barely above Poor quality rating (i.e. 2.2-2.3 MOS, at the most). At 40ms jitter and above, the quality reached the Poor level (i.e. 2 and below MOS).

In passive test, for jitter effects on video, at 10ms, the scores are just below the Good (3) rating threshold, i.e. 2.81 (PCM) and 2.79 (G723.1), although the quality is still acceptable. At 20ms, MOS are 2.5 (PCM) and 2.4 (G723.1). At 30ms jitter, both systems produced Poor video. As expected, PCM gives better results, for jitter effects on audio (i.e. at 0ms to 30ms, MOS are 2.7–3.9). While for G723.1, the MOS are between 2.3-3.7, at 0ms to 30ms jitter. At 40ms, the perceived audio quality started to degrade and became poor at 80ms. Generally, the interactive test follows the same pattern as that in passive test, but with lower MOS.

Figure **5.8** shows the MOS of the jitter effect on audio, obtained from both passive and interactive test, using PCM (μ-law) and G732.1 audio CODECs. It can be seen

that, the PCM CODEC produced generally higher MOS than G723.1. The interactive test generated a slight effect on the perceived audio quality, where its MOS dropped to around 0.2-0.3 of that in passive test.

5.3.2.4 Results: Evaluation of Delay Effects on the Perceived Quality of IP Media

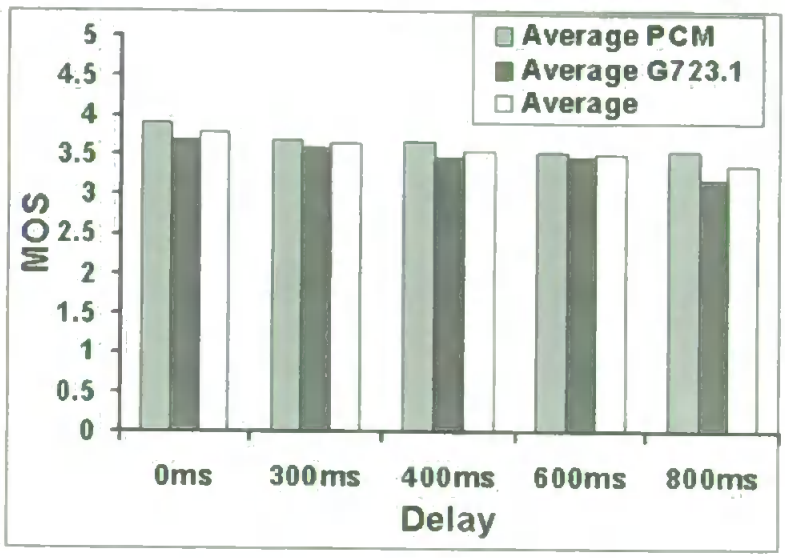


Figure 5.9: Passive–Delay Effects on Audio

Figures **5.9** and **5.10** show delay effect on audio in both passive and interactive test. As can be seen from the results, audio is less susceptible to delay as compared to packet loss and jitter. The MOS are around 3.2 to 3.7 (G723.1) and 3.5-3.9 (PCM), for the range of 0ms to 800ms delays in passive test. While in interactive test, the MOS are between 3.2-3.5 (G723.1) and 3.3-3.7 (PCM). By referring to Figures **5.11** and **5.12**, it is justified that delay also has little impact on video. For example, for the range

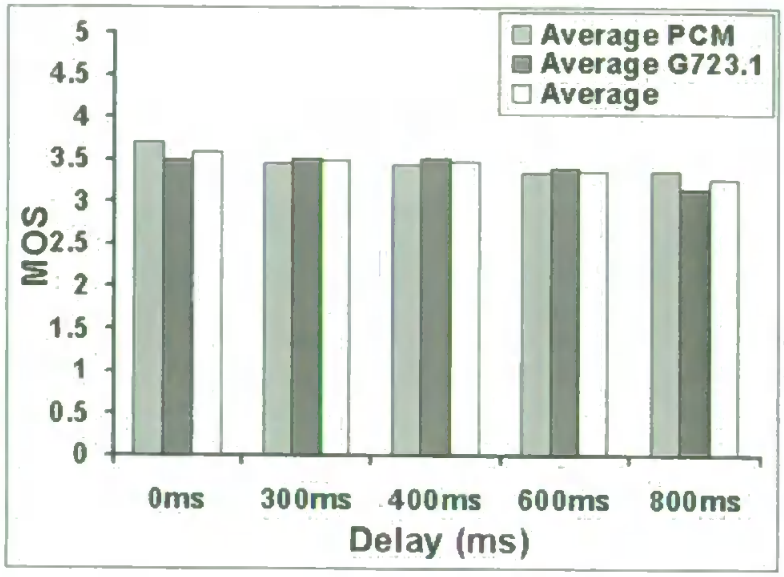


Figure 5.10: Interactive–Delay Effects on Audio

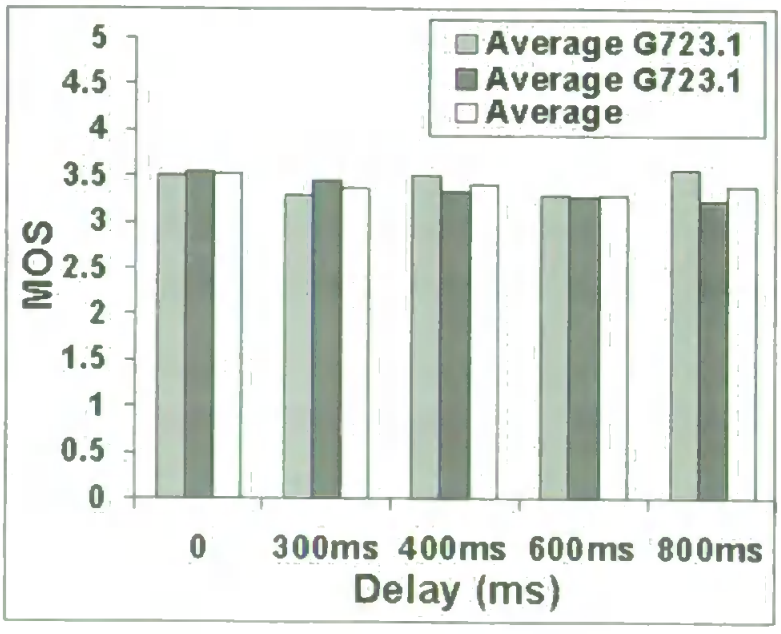


Figure 5.11: Passive–Delay Effects on Video

of 0ms to 800ms delay, the MOS for the perceived video quality in both passive and interactive test are around 3.2-3.6. The same pattern of results is repeated for audiovideo overall.

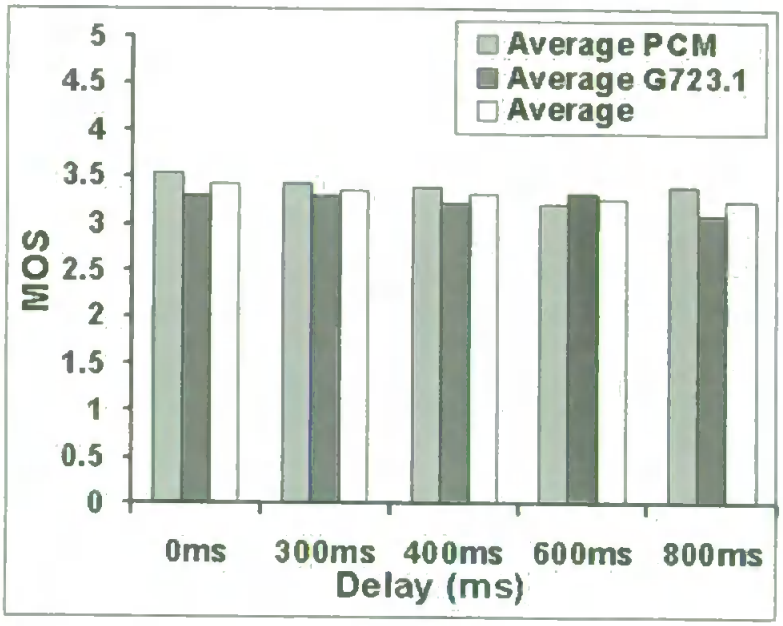


Figure 5.12: Interactive–Delay Effects on Video

5.3.3 Discussion

Our observation indicates that, to assess video quality in videoconferencing is a very complicated issue since the frame rates are constantly changing. Subjects found it hard to give the score for video in a limited period of time. Some subjects felt dissatisfied in giving only one score, as their mind fluctuated between two or more scores from one moment to another during the test. This is due to the fact that the frame rate is high if the subject is relatively still, but as the movement of the subject increases, the frame rates will vary inconsistently and/or the amount of video artifacts will increase. It was observed that frame rate varies with the motion content, the level of detail, and the percentage of image that changes from one frame to another. However, assessing audio quality is less complex since the audio degradations are relatively significant as the network constraint increases.

When evaluating audio quality only, the MOS rating is high but when evaluating the combined quality of audio and video, the rating drops to almost similar to video rating. This implies that video quality contributes an important element in benchmarking the overall performance of desktop conferencing system.

Task performance has small effects on both audio and video, with the difference is only below 0.5 MOS (i.e between 0.2-0.3 MOS). It is then suggested that, perhaps the task performance designed to carry out the assigned task was insufficient to obtain more outstanding results. Another possibility is that, the subjects were not fully familiar with the task experiments and were therefore, unable to perform the task exactly as required. Hence, it was decided that, for future experiments within the project, each specific task performance must be carefully designed and the subject must be well guided, so that the task will be conducted coherently in order to obtain more reliable results. Generally, the result shows that the task performance effects on both audio and video became apparent for packet loss between 8% - 15 for audio and between 4% - 8%, for video. Whereby, for jitter, it is between 10ms and 40ms.

5.4 Summary

Observations so far, indicate that audio, video, and overall audiovisual quality are susceptible to packet loss and jitter, but are less susceptible to delay. Over the range of 300ms to 800ms, the MOS results only drop by a maximum of 0.5 for all media.

Throughout the test, audio quality is rated higher when compared to video quality and overall audiovisual quality. The μ-law PCM CODEC has been proven to have a greater MOS than that of the G732.1 over the entire periods of the studies. The assessment of audio quality is very straightforward. The assessment of video, however, is very complex issue as its quality varied during the study, from very acceptable to almost useless.

Throughout the test, the best MOS rating for both audio and video is between Fair (3) and Good (4), although Good (4) MOS is seldom given. It is observed that none of the subjects gave an Excellent (5) MOS rating even under ideal network condition. Thus, it is evident that NetMeeting or IP videoconferencing in general, is in its infancy with substantial improvements needed to achieve higher performance. It has also been established that traffic related network factors (such as packet loss, jitter, and delay), CODEC, and task performance are all vital in maintaining the quality of video and audio service in NetMeeting.

The next chapter investigates the correlation between the perceived quality audio and video media. The individual effects of each audio and video media will also be investigated.

Chapter 6

Investigating the Interaction Effect between Audio and Video

6.1 Introduction

This chapter outlines the experiment undertaken to investigate the interaction effect between the perceived audio and video quality in low cost multimedia services. Previous research has claimed that a user's assessment of one multimedia quality is influenced by the perception of the other [70], [71]. As reported in [72], improving the quality of one medium can lead to the perception of improvements in another when there has been no actual quality upgrade. However. these findings are obtained from the higher performances multimedia systems, such as high definition television video (HDTV) and video-on-demand.

The work presented in this chapter focuses on investigating and quantifying the potential interaction effect between audio and video in low cost DVC, when the

transport mechanism carrying the two data is subjected to different levels of packet loss, separately.

Many studies have investigated the influence that video media 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) [67]. However, it has been suggested that the main use of the video link in DVC is psychological [100] 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 [65].

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? With this in mind, the research proceeded to investigate the end-to-end user's subjective opinion scores with respect to the different levels of audio and video impairments, for a given task performance.

The designated test experiments were carried out in order to evaluate;

- the effects of audio and video degradations – the objective is to investigate the combined effect of having different levels of audio and video degradations on

the perceived quality of IP media;

- the effects of video quality degradations – the study is aimed to investigate the effect gradually degrading the video quality while audio quality is unchanged, i.e. maintained at its best quality;
- the effects of audio quality degradations – the study is aimed to investigate the effect gradually degrading the audio quality while video quality is unchanged, i.e. maintained at its best quality;
- the effects of POOR video quality – the objective is to investigate the effect of having POOR video quality while the audio quality is being degraded from its ideal to worse quality;
- the effects of POOR audio quality – the objective is to investigate the effect of having POOR audio quality while the video quality is being degraded from its ideal to worse quality;
- the effects of having audio only as opposed to having both audio and video streams – to quantify the perceived multimedia quality under the influence of the presence of video channel;
- the effect of different talker languages – the aim is to investigate the impact of different talker language on the perceived quality of IP media.

In general, the main goal is to evaluate the interaction effect between the separate entity in the IP media, i.e. audio and video.

A more comprehensive experimental procedures, for each test scenario are elaborated in the following section.

6.2 The Experimental Approaches

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, whilst the quality of video is maintained at its best quality, 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 audio-visual of low bit rate videoconferencing. The method of assessment being used is the subjective test method, called Mean Opinion Score (MOS) which is the standard recommended by the ITU-T (1984), as previously explained.

As stated previously in Section 3.5.1, 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, position of the camera etc. Therefore, in the experiment, maintaining the above variables constant (for both end users), except packet loss, is vital to ensure the validity of the results. A study

by [101] stressed that the position of the camera can affect the perception of video quality. For example, viewing at a side angle can make one look 'shifty' and direct eye contact becomes difficult. Hence, within the context of the thesis, it is mandatory to ensure that the position of the camera are maintained across the studies.

Currently, transmission of audio and video over IP network uses the real time transport protocol (RTP) which runs on top of existing transport protocols, typically UDP (User Data Protocol), and provides real-time applications with end-to-end delivery services. In the experiments, a network emulation tool (NIST Net) was used to introduce different sets of impairments (packet loss) on each audio and video stream (for example, audio was degraded by 5% packet loss while video quality was unimpaired or vice versa).

The test bed configuration used was similar to that of Figure **4.1**. The two CPUs, 200MHz Pentium processors (64MB RAM) were placed in the separate rooms where the two subjects launched the Microsoft Netmeeting and evaluated the multimedia score throughout the series of different network congestions. The 200MHz CPU was being employed for the test since the 933MHz CPU (used in the previous experiments), inevitably could no longer be accessed as it belonged to the other party.

Like in the previous tests, the QCIF - 176x144 pixels frame size was used to provide better image quality and faster frame rates. Similarly, as explained in Subsection 4.2.1, the video setting was unchanged throughout the test, which was 'better quality' video. The setting is believed to be the best option to achieve a more com-

prehensive test result as the other option ('faster video') produces inadequate image quality.

For the audio CODEC, a G723.1, 6400bit/sec was employed. As previously explained in Subsection 4.2.1, the dual rate G723.1 CODEC was being used mainly for its popularity and it is considered as a newly and more advanced coding techniques, specially designed future DVC systems with improved QoS.

The test activities of the project were organised in a number of steps. First, tests were carried out under an error free network environment. Second, different sets of network impairments (packet loss), as in Table **6.1**, were introduced to the separate audio and video stream in order to assess their impact on the perceived multimedia quality. The intensive and comprehensive analysis of the collected data have led to the quantifying of the interaction effects between the perceived audio and video quality. The resulting observations are presented in Subsection 6.3.1.

Based on the results obtained from the various sets of audio loss, the potential impact of different spoken languages upon the perceived MOS was also investigated. Different nationalities, speak in different languages, producing different speech sounds in different frequency ranges. For example, most English speech sounds occupy the frequency limits of 2000Hz min and 8000Hz max, whereby most French speech sounds occur between 125–250Hz and 1000–2000Hz [102]. In a low cost videoconferencing environment, which is subjected to high levels of packet loss, the listener and speaker's language background can become a critical issue as it in-

troduces a novel type of degradation. Hence, the study also investigated the impact of different talker language on media quality.

Due to the limited number of varieties in the test subjects' spoken languages, English and Chinese (which constituted to the majority of test candidates) were selected for the test. There were 6 British and 8 Chinese participants took part in the interactive test as shown in Table **6.2**. To maximize the task performance, and ensure that the subjects were fully at ease during the conversation, they were encouraged to use their mother tongue while conducting the test. The results can be found in Subsection 6.3.2.

The conditions under considerations are presented in Table **6.1**.

Table 6.1: Packet Loss of Video (v) and Audio (a) Under Test (in Percentage)

video(v)/audio(a)	v/a	v/a	v/a	v/a	v/a	v/a	v/a
v deg/a no loss	0/0	a/0	1.5/0	2/0	2.5/0	3/0	4/0
v no loss/a loss	0/0	0/9	0/10	0/15	0/25	0/30	0/35
v loss/a loss	0/0	1/9	1/10	1.5/15	2/25	2.5/30	3/35
v poor/a deg	0/0	4/9	4/10	4/15	4/25	4/30	*
v deg / a poor	0/0	1.5/35	2/35	2.5/35	3/35	*	*

The test was conducted on two different task scenarios i.e.:

- **Interactive Test**: there were 20 adults involved in the test. They were allowed to select their own issue for discussion, with which they were comfortable, so as to maximise the interactions. It is stated that informal communication tends to be representative of individuals who are familiar with each other

[99]. Hence, to maximise task motivation and to ensure subjects are fully at ease with each other, individuals who were acquainted with one another were selected for the tests.

Table 6.2 shows the category of subjects involved in the interactive test.

Table 6.2: The Categories of Subjects

Number of Subjects	Nationality	Gender
8	Chinese	6 Males; 2 Females
6	English	Males
2	Greek	Males
1	Indonesian	Male
3	Malaysian	Males
Total = 20		Total = 18 Males 2 Females

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 to two minutes. For control purposes, tests were initially carried out under error-free condition, i.e. 0% packet loss.

- **Passive Test**: A number of 20 adults involved in 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.

Table **6.3** shows the category of subjects involved in the passive test.

Table 6.3: The Categories of Subjects

Number of Subjects	Nationality	Gender
1	African	Male
2	Arabian	Males
4	Chinese	4 Males
6	English	5 Males; 1 Female
2	Greek	1 Male; 1 Female
1	Indian	Male
2	Indonesian	1 Male; 1 Female
1	Malaysian	Male
1	Romanian	Male
Total = 20		Total = 16 Males 4 Females

6.3 Results and Discussions

6.3.1 The Interaction Effect between Audio and Video Media

All the Figures **6.1** to **6.10** below, show the results obtained from the tests and the observations made are described in this section.

Figures **6.1** and **6.2**, show the 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 conclusion derived from this observation is that the video quality was already so poor to make no significant differences between image impairments. The rating for video stays at approx. 2.5–2.6 (MOS) for audio loss ranging from 9%-35%. However, the MOS for the perceived quality of combined audiovideo for both test scenarios is approximately the same, i.e. approx. 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 the interactive test.

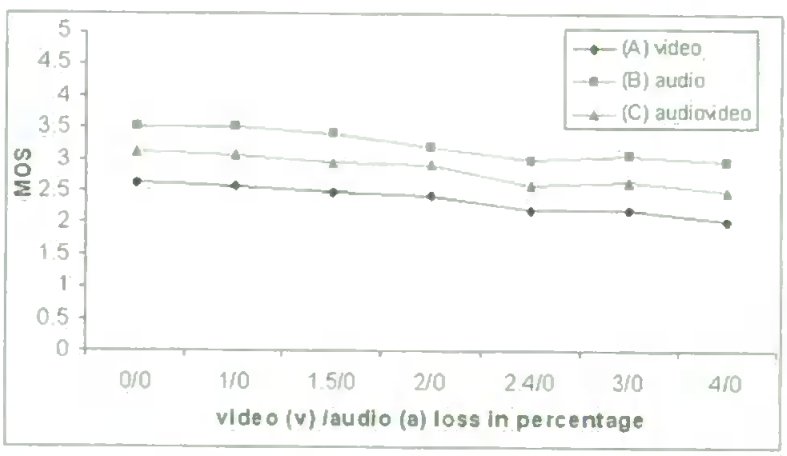


Figure 6.1: Interactive–Video-Degraded; Audio-Constant

Figures **6.3** and **6.4** show the impact of MOS packet loss on the perceived quality of (A) video, (B) audio and (C) combined audiovideo as obtained in the passive test. Unlike the interactive test, the MOS for audio is not affected by the degradation in

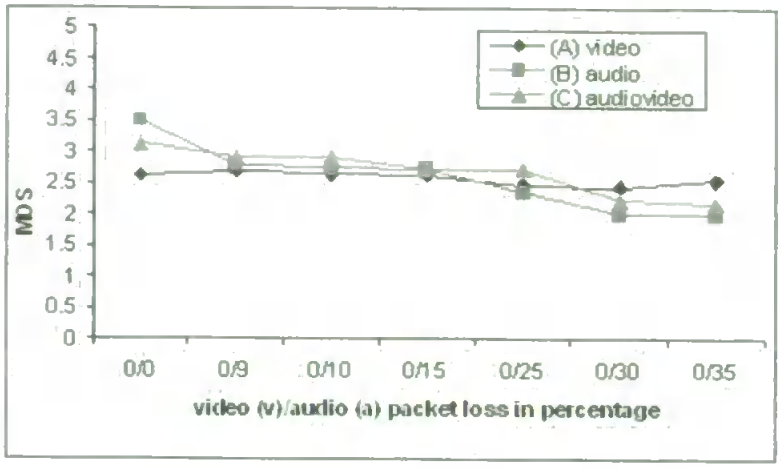


Figure 6.2: Interactive–Video-Constant; Audio-Degraded

video quality (see Figures **6.2** and **6.4**), in passive test. Also, by referring to Figure **6.4**, there is a slight drop in video score, i.e. 0.36 (MOS), when the audio loss is changed from 0%-35%. The MOS for combined audiovideo is affected severely by the change in video loss as compared to audio loss.

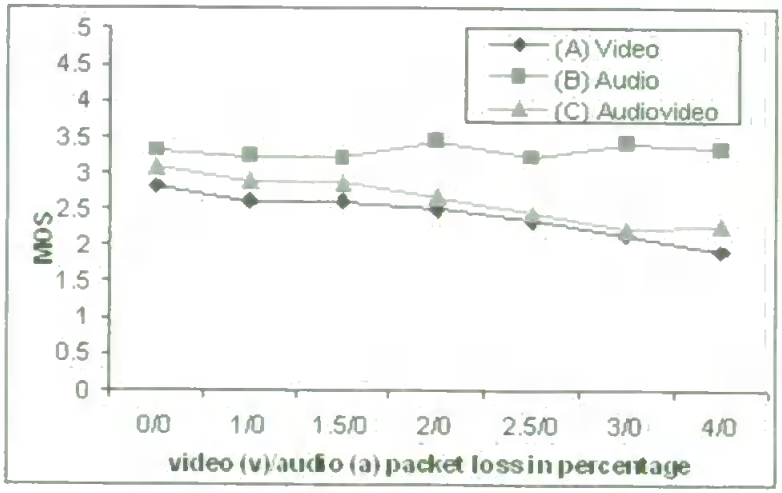


Figure 6.3: Passive–Video-Degraded; Audio-Constant

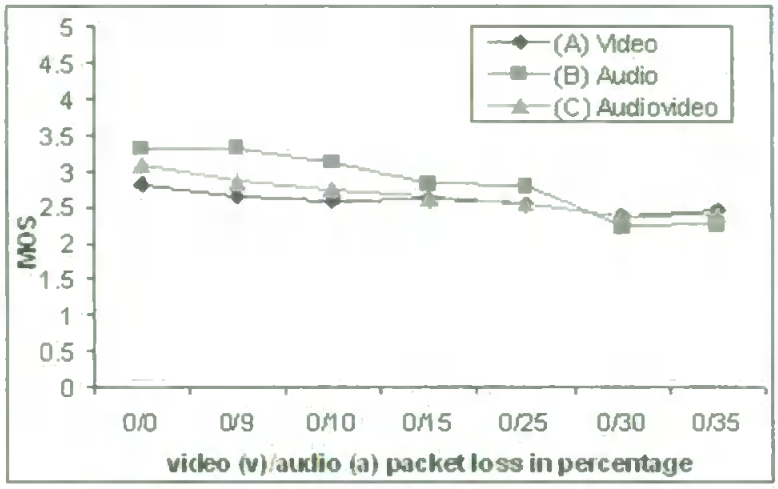


Figure 6.4: Passive–Video Constant; Audio Degraded

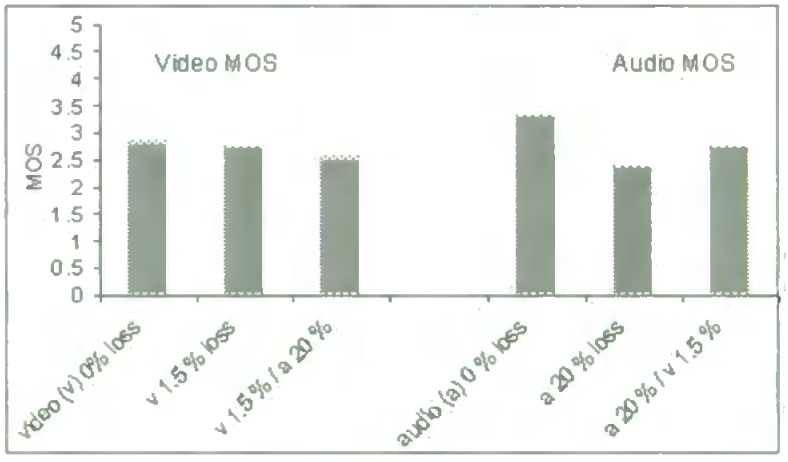


Figure 6.5: Passive–Packet Loss Impact on Audio and Video

Figure 6.5 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 with the audio and video when both are present during the test. The result indicates the strong interaction dependency between audio and video media. 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 6.5), the MOS is 2.3 without the presence of video. However, with the presence of video, the same audio sample gives

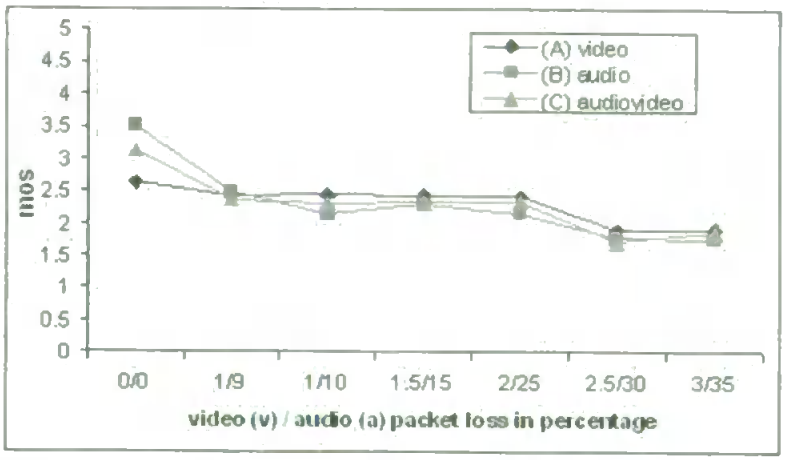


Figure 6.6: Interactive–Audio and Video Degraded

MOS rating of 2.7 (final column in Figure **6.5**). This indicates that video information enhances speech only communication. On the other hand, the perceived video quality degrades when poor quality of audio was present. Another example, Figure **6.1** shows how the 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 of multimedia quality.

Figure **6.6** 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 **6.2** (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 throughout the test.

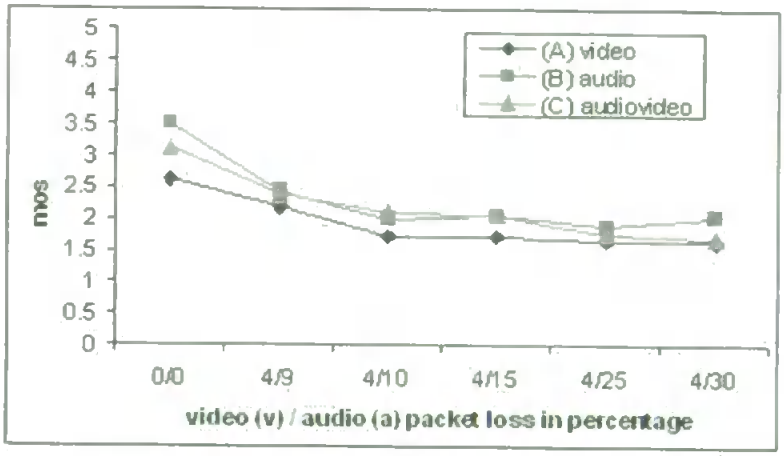


Figure 6.7: Interactive–Video Poor; Audio Degraded

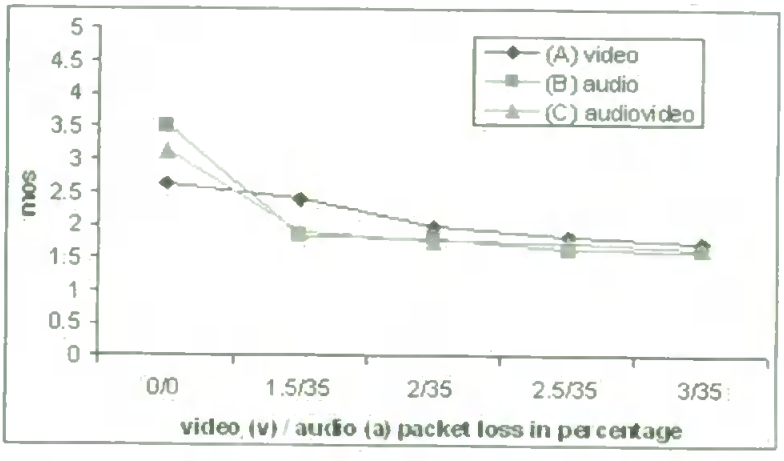


Figure 6.8: Interactive–Video Degraded; Audio Poor

Figures **6.7** to **6.10** 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%), as observed in the interactive test (Figure **6.7** and **6.8**) and passive test (Figure **6.9** and **6.10**). It can be seen that, when video is poor, i.e. 4% loss, the passive test (Figure **6.9**) gives a higher MOS rating for the perceived quality of video and audio only as compared to that in interactive test (Figure **6.7**). However, audiovideo rating for passive test shows lower scores as compared to that of the interactive test. Figure **6.8** shows that, when audio is very poor, the interactive test produces a very

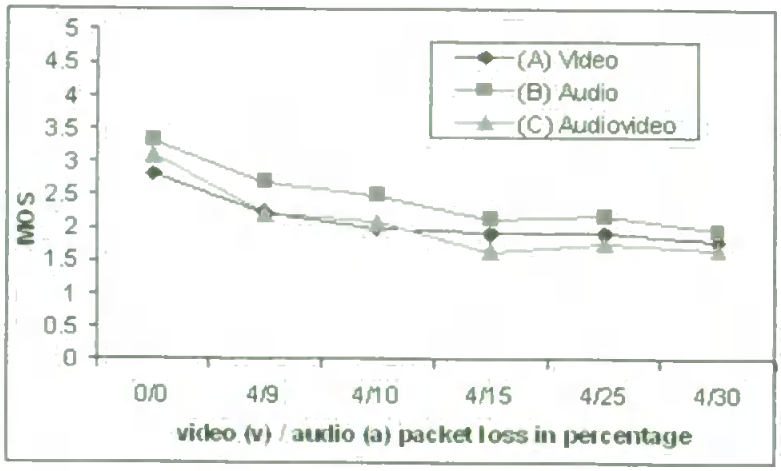


Figure 6.9: Passive–Video Poor; Audio Degraded

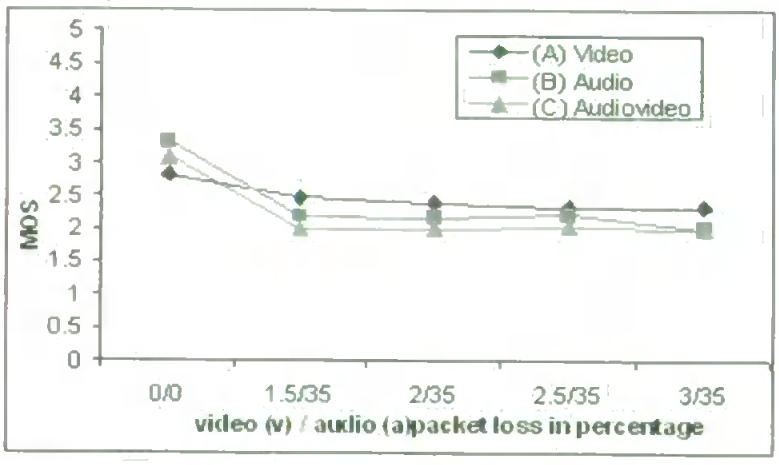


Figure 6.10: Passive–Video Degraded; Audio Poor

low MOS for the perceived multimedia quality. Hence, the interactive test severely depends on sufficient audio quality.

6.3.2 The Effects of the Different Talker Language on MOS

The MOS results for the different talker language are depicted in Figure 6.11 to 6.14 and the MOS data are available in Table 6.4 to 6.7.

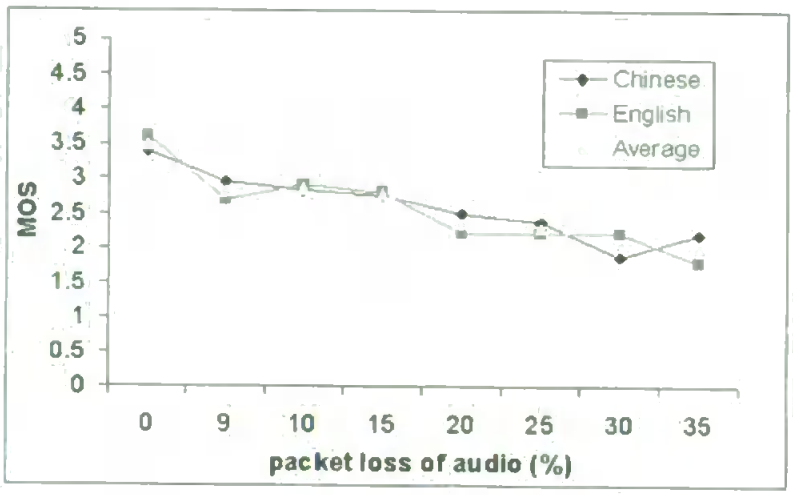


Figure 6.11: Diff. language; Audio Mos – Audio Degraded

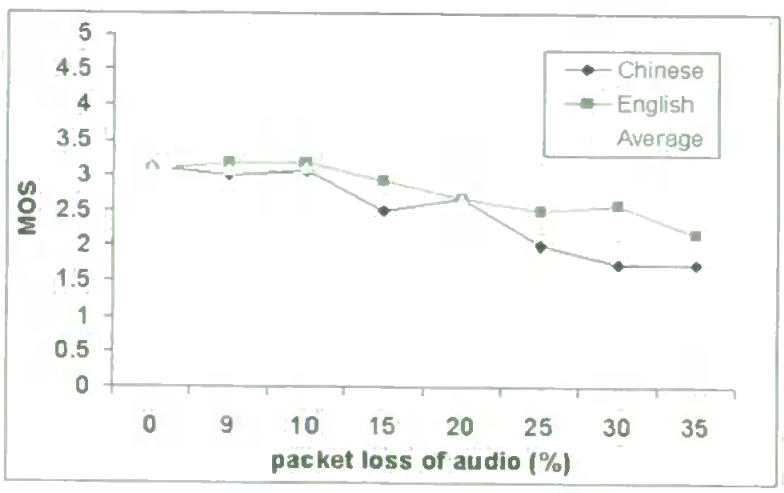


Figure 6.12: Diff. language; Audiovideo Mos – Audio Degraded

Table 6.4: Diff. Language; Audio Mos – Audio Degraded

Audio loss (%)	0	9	10	15	20	25	30	35	
Chinese	3.38	2.93	2.81	2.75	2.5	1.37	1.88	2.18	M
English	3.6	2.7	2.9	2.8	2.2	2.2	2.2	1.8	O
Average	3.49	2.81	2.86	2.78	2.35	2.28	2.04	1.99	S

Table 6.5: Diff. Language: Audiovideo Mos – Audio Degraded

Audio loss (%)	0	9	10	15	20	25	30	35	
Chinese	3.12	3	3.06	2.5	2.68	2	1.75	1.75	M
English	3.08	3.17	3.17	2.92	2.67	2.5	2.58	2.16	O
Average	3.1	3.08	3.12	2.71	2.68	2.25	2.16	1.96	S

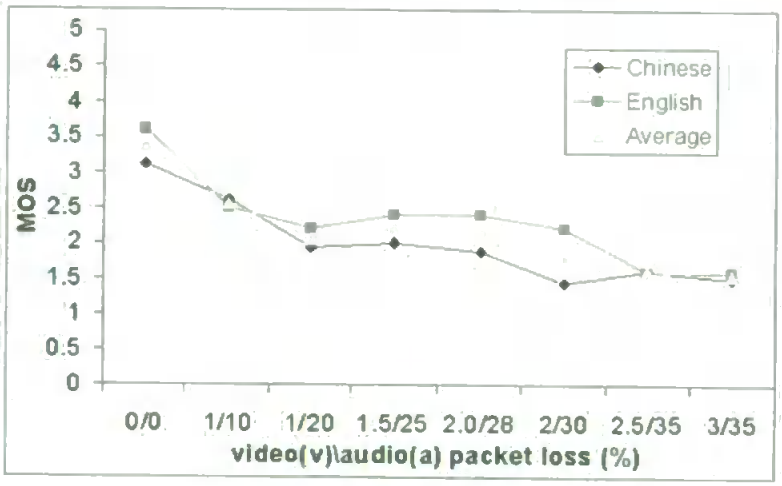


Figure 6.13: Diff. language; Audio Mos – Audio and Video Degraded

Table 6.6: Diff. Language; Audio Mos – Audio and Video Degraded

v/a Loss(%)	0/0	1/10	1/20	1.5/25	2.0/28	2/30	2.5/35	3/35	
Chinese	3.12	2.62	1.94	2	1.88	1.44	1.62	1.5	M
English	3.6	2.5	2.2	2.4	2.4	2.2	1.6	1.6	O
Average	3.36	2.56	2.07	2.2	2.14	1.82	1.61	1.55	S

In general, the results show the difference in the MOS for the given talker languages (i.e. Chinese and English), in that a lower range of scores were produced by the Chinese spoken language. The results are valid for both perceptual audio and au-

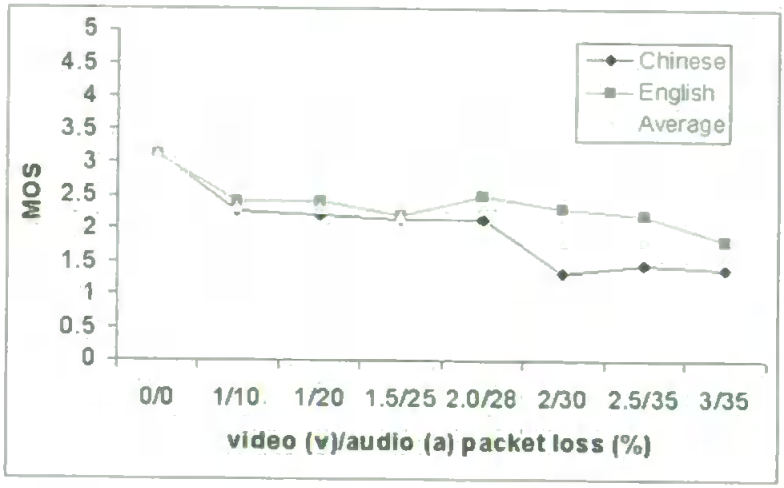


Figure 6.14: Diff. Language; Audiovideo Mos – Audio and Video Degraded

Table 6.7: Diff. Language; Audiovideo Mos – Audio and Video Degraded

v/a Loss(%)	0/0	1/10	1/20	1.5/25	2.0/28	2/30	2.5/35	3/35	
Chinese	3.12	2.25	2.19	2.12	2.12	1.31	1.44	1.38	M
English	3.1	2.4	2.4	2.2	2.5	2.3	2.2	1.8	O
Average	3.11	2.32	2.30	2.16	2.31	1.81	1.82	1.59	S

diovideo overall MOS, and for the tests where either audio (only) or both audio and video streams are degraded. As expected, in tests when both audio and video are degraded, as shown in Figure **6.13** (Table **6.6**) and Figure **6.14** (Table **6.7**), the MOS decreased, compared to that of when audio (only) is degraded.

6.4 Conclusion

The results concluded that there is strong interaction dependency between audio and video media. For example, it can be seen that the MOS for audio increases with the presence of video. It is also observed that, video adds subjective value to a

conference and enhances interactivity. Thus, it is evident that video is an important determinant in multimedia quality. As in the case of the interactive test, video scores are not affected by the audio quality. It can be concluded that the video quality was already too poor to be meaningful in that the subject did not notice any significant difference between each level of impairment. Due to the poor image quality the subject relied mainly on the audio channel and made some allowance for the low performance of the video channel, as found in DVC systems. It would then be important to state that in the context of low cost DVC applications, the video image is not the focus of attention in the same way that the picture is when we watch TV.

However, audio scores deteriorated as video is degraded. Therefore, it is justified to state 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 which is also reported in [65].

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 a passive viewer and listener (passive test), the person tends to view the image quality more closely; likewise, if a person is casually chatting (interactive communication), the quality of the video is of less important than that of the audio in that the viewer is more tolerant to the image degradations. Perhaps the candidates were deeply indulged in their issue of discussion that they did not notice the changed in the video quality. The finding also confirmed with the previous research results in Chapter

4, which indicated that the subjects are less susceptible to poor video quality while engaging in the interactive communications.

The perceived multimedia dependency on the different talker language has been reported by the test subjects, in Subsection 6.3.2. The average score reported by the native English spoken participants was higher than that of the Chinese. Therefore, it is concluded that the perceived quality differs with respect to the specific languages, under the same network conditions. This observation is in confirmation with the previous finding which also stated that the multimedia CODECS performance is clearly depended on the different talker language, as well as voice (i.e. male or female) [96]. Unfortunately, due to the small numbers of the female members of the test candidates, the task to clarify the effects of the different gender's voice on the MOS was dismissed. Research has shown that G723.1 CODECs (as used in NetMeeting) have the speech parameter filters designed for English speakers.

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

6.5 Summary

This chapter has discussed the issue of the correlation effect between the audio and video media quality on the subjective end-user's opinion of the low cost DVC applications based on the interactive and subjective test scenarios. The main conclusion derived from the study is that the perception of one media interacts and influences the perceived quality of the other. End user perception of a multimedia systems therefore cannot be accurately predicted by investigating the individual media in isolation. The correlation effect between these media must be taken into account, in the context of the task being performed by the participant. It is also important that the task scenario should be realistic and highly representative of the real world situation.

The next chapter investigates the issue of lip synchronisation and its effect on the perceived multimedia quality. Throughout the tests in this chapter, it was reported that lip synchronisation was difficult to obtain as the frame rates were generally very low, i.e. 2-5 frames per-sec.

Chapter 7

Investigating the Effects of Lip Synchronization

7.1 Introduction

The work presented in this chapter, outlines the test conducted to investigate the effects on lip sync of delay, jitter, and packet loss on the perceived quality of audio only, video only and audiovideo overall, using the subjective Mean Opinion Score method. The test experiments were again based upon passive and interactive tests.

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. As previously stated, 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 synchro-

nization (lip sync) error. In addition, due to the low frame rate in IP conferencing systems, i.e. 2-5 frame per-sec, [50], it is difficult to obtain lip sync. It is suggested that, the frame rate should exceed 8 frames per-sec to make lip sync a meaningful term.

Previous research has claimed that audio may be played up to 120ms ahead of video, whilst video can be played up to 240ms ahead of audio [2]. It is suggested that, as the nature of human beings are more used to perceiving an event before they hear it (i.e. light travel faster than sound), they are more tolerant to audio lagging video, rather than vice-versa. Also, it is claimed that the audio should be synchronized within +/- 90ms of the video (with a maximum range of +/- 160ms), before the lip sync error can be perceived [2]. In [3], it is stated that lip sync error becomes apparent as the delay variations between audio and video exceeds 80 to 100ms. Also, it is indicated that, audio delay above 400ms, would compromise the quality of two-way communication in IP conferencing, in that the nature of interaction is claimed to be significantly awkward and less than satisfactory [1], [2], and [3]. To date, considerable of work has been focused on implementing new techniques and approaches to minimize out of sync problems, such as, selective packet discard, buffering, and constant rate playback techniques at the receiving end [47], [103], [104] and [105].

However, those findings describe above are based on experiments involving higher frame rate video channel, such as, TV system and video on demand [2], [3], and [47]. To date, little research has been carried out on the assessment of lip synchronisation

effect over low cost DVC system, where the frame rates fall between 1–15fps. Therefore, the study proceeded to investigate the effect of lip sync error on the perceived multimedia quality in low cost videoconferencing systems.

The findings described in this chapter stated that, unlike the above results, the perceived multimedia quality (in the context of low cost DVC) is less affected by lip sync error. For example, the perceived multimedia quality started to deteriorate when the mismatch time between audio and video media reaches 600ms. The test results also 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. Hence, the level of the lip sync error and hence, the perceived multimedia quality, are determined by the different tasks performed by the end users.

The test has been design based upon four (4) different phases:

- investigating the effects of delay on lip sync;
- investigating the impacts of delay and packet loss on lip sync;
- investigating the effects of jitter on lip sync;
- investigating the effects of combined network constraints on lip sync.

The following section describes the experimental procedures, implemented for each phase. The outcomes obtained from the experiments are presented in the Result's section, followed by the discussion of findings, in the subsequent heading.

7.2 The Experimental Approaches

The subjects, participated in the tests, were mostly students (of multiple nationalities) of the University of Plymouth, aged between 20-45 years old.

As previously stated, the experiments were based upon investigating the effects of lip sync on the perceived quality of multimedia components, in two different task performances i.e., passive test and interactive test.

In the interactive test, the two communicative parties were already acquainted (and thus fully at ease with one another) to maximise the interaction. 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 conditions stated in International Telecommunications Union, ITU-R P500 [97]. For example, how to set the room conditions, equipments setting, and experimental procedures.

The testbed configuration is similar to that described in Section 4.2.1. A network emulation tool (NIST Net) is used to introduce the different sets of impairments i.e. packet loss, jitter and delay, as designed for each test, prior to transmission.

At the receiving end, the subjects were asked to evaluate the perceived quality of audio, video and the combined audiovisual components, in terms of MOS. The test candidates were also required to classify a perceived synchronization error based

upon 4 different categories, as the following;

- (i) audio is ahead of video;
- (ii) audio is behind video;
- (iii) not sure, whether audio is ahead or lagging video (random lead/lag - not specified);
- (iv) no synchronization error.

(Note: The questionnaires sheets are included in Appendix A)

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.

As explained in the previous test experiments, the 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 asked to maintain their movements throughout the test to minimise the dynamic variation in frame rates that could lead to the inconsistent in the image degradations.

7.3 The Experiments and Results

7.3.1 Investigating the Effects of Delay on Lip Sync

The study is divided into two experiments, namely, Experiment A and Experiment B, as described in the following paragraphs.

7.3.1.1 Experiment A

There were 38 subjects involved in the test. These subjects were aged between 20 to 40 years old and have normal sight and hearing. All subjects have the experiences with the multimedia applications over the Internet (such as movie clips and online-shopping) but have not directly involve in assessing multimedia quality before. Tests were divided into two sections, i.e. Section 1: Passive Test and Section 2: Interactive Test. There were 18 subjects involved in the passive test and 20 subjects participated in the interactive section. The categories of the subjects participated in these tests are shown in Table 7.1 and Table 7.2.

As explained previously, the NIST Net is used to introduce different amount of packets delay, on each audio and video stream (prior to transmission). Hence, the different levels of lip sync were produced. For each test, a delay within the range of 40-440ms was randomly introduced, to the audio and video streams, separately. These values were selected to give the significant effect, from the minimum to the

Table 7.1: The Category of Subjects (Passive Test)

Number of Subjects	Nationality	Gender
1	African	Male
1	Arabian	Male
6	Chinese	5 Males; 1 Female
3	English	2 Males; 1 Female
2	French	Males
1	Greek	Female
1	Indian	Male
1	Malaysian	Female
1	Mexican	Male
1	Spanish	Male
Total = 18		Total = 14 Males 4 Females

Table 7.2: The Category of Subjects (Interactive Test)

Number of Subjects	Nationality	Gender
1	Burmese	Male
3	Chinese	2 Males; 1 Female
6	English	3 Males; 3 Females
4	French	Males
3	Greek	Males
1	Indian	Male
1	Indonesian	Male
1	Malaysian	Male
Total = 20		Total = 16 Males 4 Females

maximum point where 440ms is believed to be the critical figure to ensure a smooth two-way interactions [1], [80], and [2]. A step of 40ms interval was selected due to the fact that the multimedia software and hardware are capable to refresh motion video data every 33/44ms. Each test lasted for approximately one minute and one test section would be completed in 30-40 minutes.

Table 7.3 describes the test scenarios implemented for Experiment A.

Table 7.3: Test Scenarios

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

7.3.1.2 Results: Experiment A – MOS

Figures 7.1 and Table 7.4 describe the results obtained from the experiment. Figure 7.1 shows the MOS of the perceived quality of audiovideo overall, obtained from the interactive test, when audio or video streams were delayed from 40ms up to 440ms. The MOS were in the range of 2.4 to 3.1, with video delay giving the higher scores than audio delay.

Table 7.4: Average MOS

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

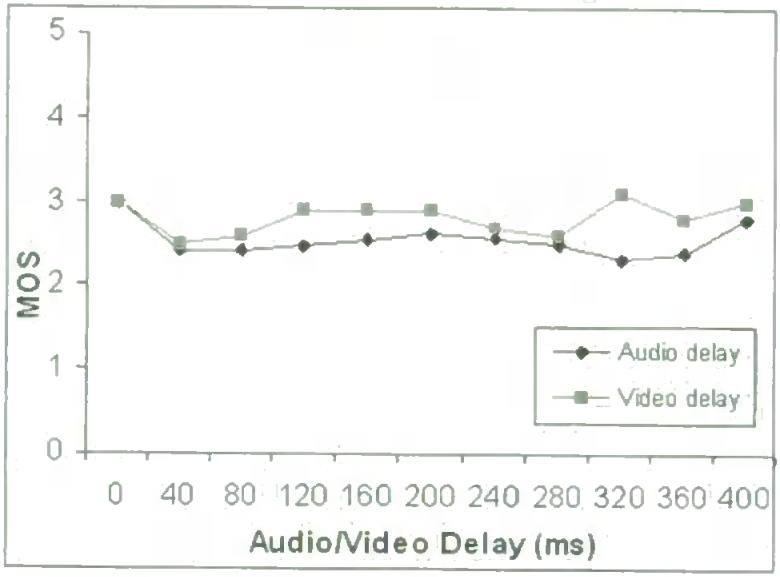


Figure 7.1: Interactive – Audiovideo Overall MOS

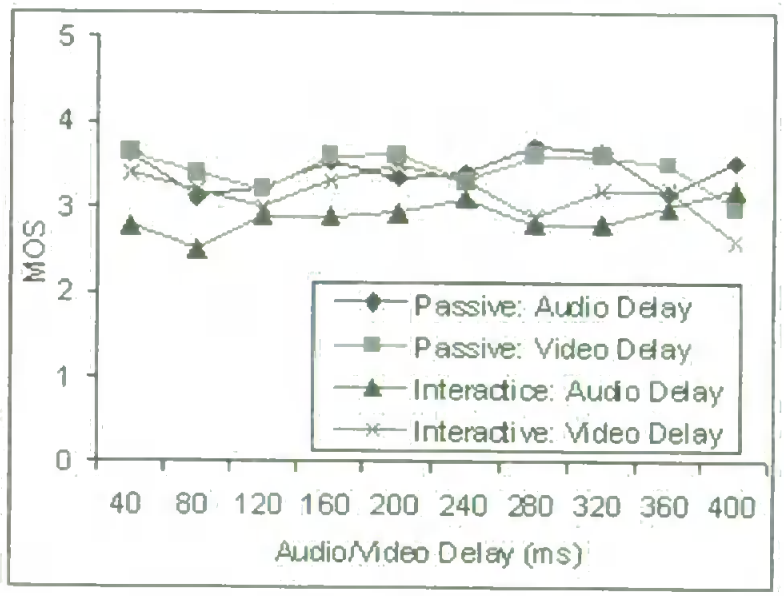


Figure 7.2: Audio MOS – Interactive Vs Passive

Figure 7.2 displays the MOS for audio, in both interactive and passive tests. The MOS of the perceived quality of Audio were generally higher, followed by audiovideo overall, while video scored the lowest MOS (see Table 7.4). The passive test gives higher MOS values than the interactive test, e.g. by referring to Figure 7.2

and Table **7.4**, 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 **7.4** and Figure **7.2**.

7.3.1.3 Results: Experiment A – 4-category Rating Result

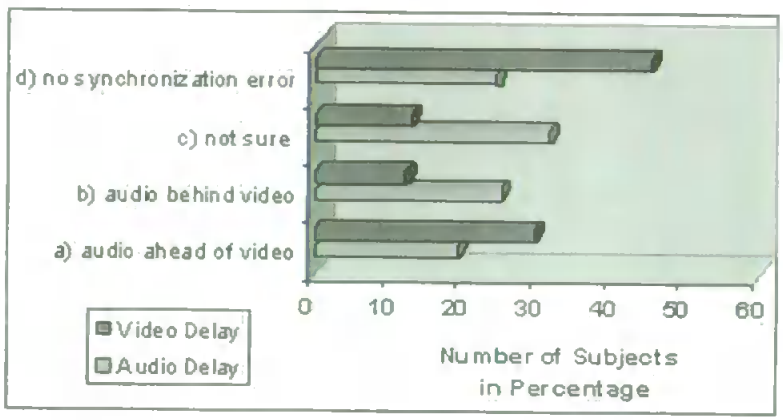


Figure 7.3: Passive – Audio Vs Video Delay

The results, based upon the percentage of students responding in each of the 4-category ratings, are shown in Figure **7.3** (passive test) and **7.4** (interactive test). As there is no distinctive variations in the effects of the audio and video delay within the range of 40-440ms, thus it was decided to take the average result.

The passive test (see Figure **7.3**), gives more accurate result, i.e. when audio was sent ahead of video, 29.7% of the subjects perceived that audio is ahead of video,

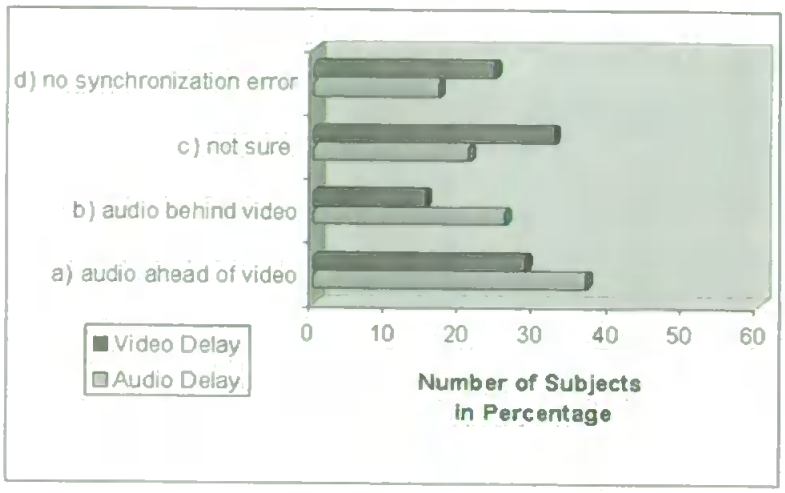


Figure 7.4: Interactive – Audio Vs Video Delay

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, a 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 **7.4**), a higher percentage of participants noticed the synchronisation error, i.e. 32.29% for video delay and 20.94% for audio delay. However, a majority of them were giving the wrong answer or not sure if audio was ahead of video 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. By

comparing these results with those when both audio and video were sent simultaneously using the same amount of delay, the latter score has shown a higher MOS, as depicted in Figure 7.5.

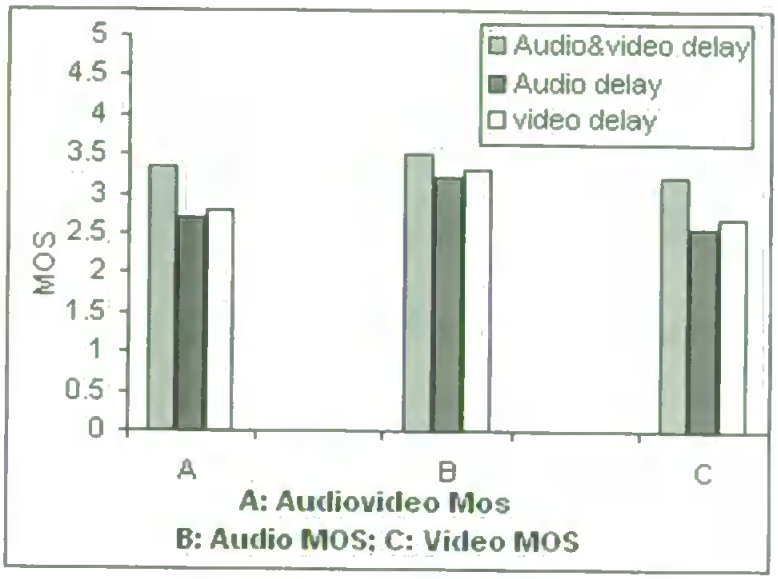


Figure 7.5: Combined Audio and Video Delay Vs Separated Audio and Video Delay

By referring to Figure 7.5, 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).

The results have shown that delaying both media concurrently, by the same amount has less impact on the perceptual media quality as compared to that when delaying either audio or video separately. Hence, as expected, it has been proven that lip sync

has more impact on the perceptual media quality. All the results in Figure 7.5 were deduced from the 400ms delay for both audio and video.

7.3.1.4 Experiment B

In Experiment B, the amount of delay between audio and video streams were increased further since the audio or video delay of 440ms did not make any significant difference (see the results for Experiment A). Therefore, the delay of 600ms, 700ms, and 800ms were selected for the test. There were 18 students and the academic staffs of the University of Plymouth participated in the test. The subjects were aged between 23–45 years old and have normal vision and hearing. None of the test candidates been directly involved in work connected with assessment of the performance of the multimedia systems before. However, they are widely exposed to the Internet and its applications.

Due to the lack of time and limited number of subjects, only the interactive test was considered for the test section. The interactive test was selected over passive test as it imitates the real world scenario more closely compared to the later. Moreover, based upon the results from the previous studies conducted in the thesis, the outcome of the passive test can be predicted in that it is slightly higher (i.e. by approx. 0.5 MOS) than the interactive test.

The category of the subjects participated in the test is shown in Table 7.5.

Table 7.5: The Category of Subjects

Number of Subjects	Nationality	Gender
1	Arabian	Male
1	Burmese	Male
11	English	9 Males; 2 Females
2	Greek	1 Male; 1 Female
1	Indonesian	Female
1	Romanian	Male
1	Spanish	Male
Total = 18		Total = 14 Males 4 Females

7.3.1.5 Results: Experiment B – MOS

Figure 7.6 illustrates the MOS of the perceived quality of audio, video and audiovideo overall, with respect to the separate audio and video delay, in the ranges of 600ms - 800ms.

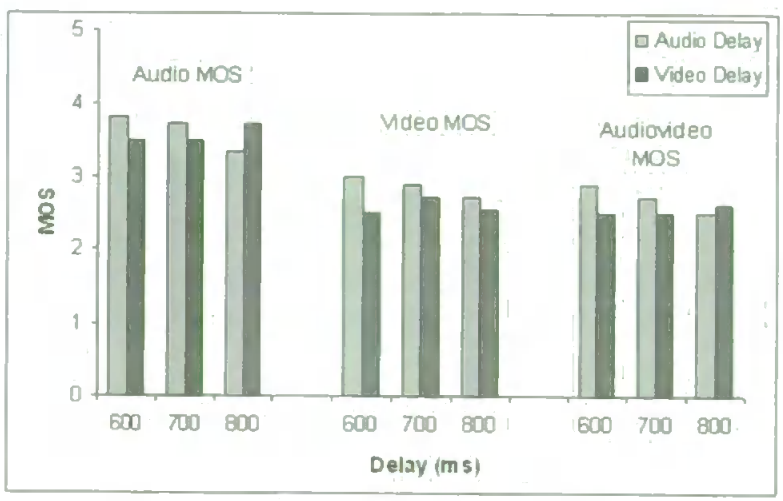


Figure 7.6: MOS Vs Audio and Video Delay

For audio delay of 600ms to 800ms, there is a slight degradation of the perceived audio quality. For example, at 600ms delay, the MOS for audio is 3.8, and 3.3 at 800ms delay. Thus, the MOS drops by 0.5. The same pattern is repeated for the MOS of audiovideo overall, where the MOS drops from 2.9, at 600ms delay to 2.5, at 800ms delay. The video score, however shows a much lower degradation.

When video was delayed, there was no effect reported on the perceived quality of audio, video and audiovideo overall. In fact, the subjective opinion scores improve as the video delay increases. For example, at 600ms delay, the MOS for video is 2.5 but at 700ms delay, the MOS is 2.7. The same pattern also occurred in the MOS of audio and audiovideo overall. Thus, delayed video media does not affect the perceived quality of audio and audiovideo overall.

7.3.1.6 Results: Experiment B – 4-category Rating Result

Table 7.6 shows the number of students (in percentage) responded to the 4-category rating test, as explained in the Experiment B.

Table 7.6: 4-category rating – Audio Vs Video Delay

	Audio delay			Video delay		
4-Catergory rating	600 ms	700 ms	800 ms	600 ms	700 ms	800 ms
Audio is ahead of video	50	61	28	78	72	56
Audio is behind video	17	28	22	0	11	5
Random lead/lag (not sure)	28	11	39	5	6	28
No synchronization error	5	0	11	17	11	11

It can be seen that, even when audio was delayed up to 700ms, the majority of the test subjects believed that audio is ahead of video (i.e. 50% for 600ms audio delay and 61% for 700ms delay)

When audio was delayed by 800ms, 38.9% of the subjects were unable to decide whether audio is lagging of video and vice-versa, while 27.7% of them still claimed that audio is ahead of video.

On the other hand, when video was delayed, the majority of the subjects perceived the correct settings. The result has shown that a number of 77.7%, 72.2% and 55.5% of the subject scored for audio ahead of video, when the the video stream was delayed by 600ms, 700ms, 800ms, respectively.

It is observed that the video frame rate in low cost videoconferencing systems is very low that delaying audio to as high as 700ms will not make any difference, in that video is always perceived as lagging that of video.

7.3.2 Investigating the Impact of Delay and Packet Loss on Lip Sync

In the study, there were only 10 subjects involved due to the limited number of participants and the lack of time. Unlike the previous study, apart from the delay (i.e. 320ms, 400ms, and 520ms), the packet loss was also introduced 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. The values of packet loss for audio and video media were carefully selected based on a number of exhaustive laboratory studies and previous experiments, such that they give significant effects to the perceived media quality, from the minimum to the maximum. The delays of 320ms, 400ms, and 520ms were employed in addition to the audio and video packet loss in order to investigate the combined effect of packet loss and delay. The category of the subjects participated in the test is shown in Table 7.7.

Table 7.7: The Category of Subjects

Number of Subjects	Nationality	Gender
3	Chinese	2 Males; 1 Female
1	French	Male
2	Indian	Males
2	Malaysian	1 Male; 1 Female
1	Mexican	Male
1	Spanish	Male
Total = 10		Total = 8 Males 2 Females

Since the test scenario was designed to last for approximately 30–40 minutes, hence, only audio delays of 320ms, 400ms, and 520ms were investigated. According to the previous findings, these figures are critical and would compromise the quality of two-way communication in IP conferencing [1] and [2]. The task performance was based on the interactive task where the two participants were casually interacting with each other and evaluating the quality, at the same time.

7.3.2.1 Results: MOS

Figures 7.7 to 7.12 describe the results obtained from the experiment. Figure 7.7 shows the perceived audio MOS for the Interactive Test, for audio packet loss of 5%, 10%, 15% and 20%. The MOS are ranging from 3.4 (GOOD/FAIR) and 3 (FAIR) when audio loss is between 5%-10% and the MOS drops by approx. 0.5, when audio loss reaches 15%. At 15% audio loss, however, for the 320ms audio delay, the MOS for audio is approx. 3 (FAIR) and drops to around 2.5 MOS, when the delays increase to 400ms and 520ms. At 20% audio loss, the scores are around 2.2, which are approaching the POOR threshold (i.e. 2 MOS). It is noticed that, the audio delay has no significant effect on the MOS as the audio loss reaches 20%. Therefore, it is concluded that, at 20% audio loss the audio quality was so poor that it was difficult to evaluate the precise quality. The MOS at this stage is claimed to be below 2.5 MOS.

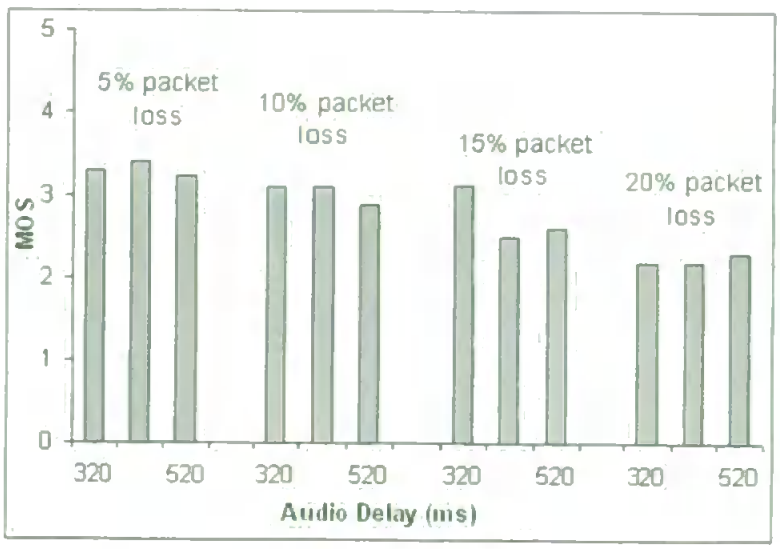


Figure 7.7: Audio MOS Vs Audio Delay and Loss

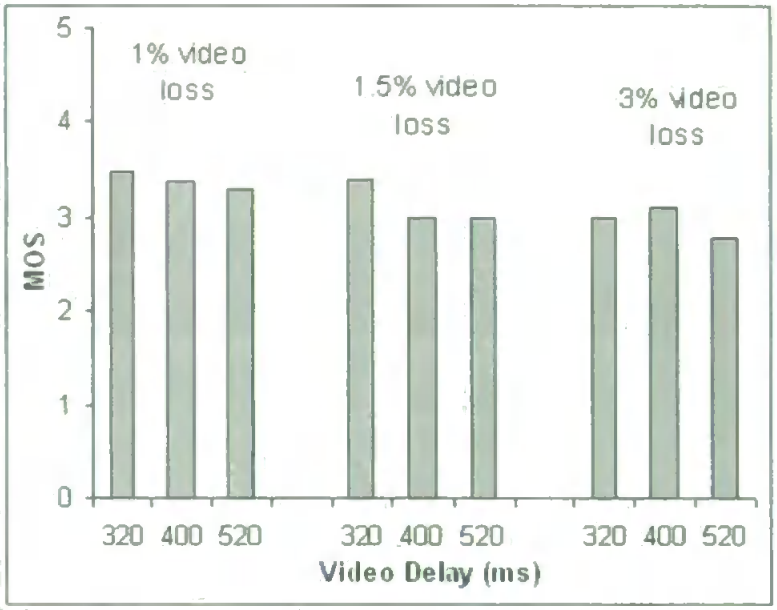


Figure 7.8: Audio MOS Vs Video Delay and Loss

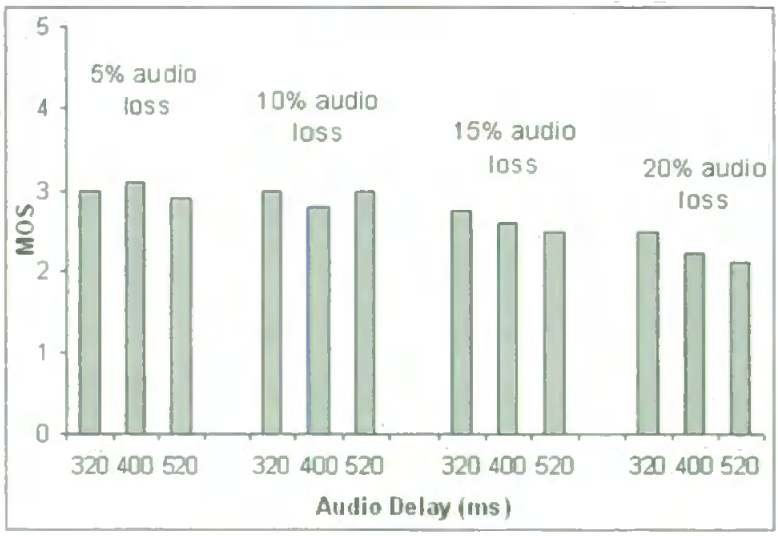


Figure 7.9: Video Mos Vs Audio Delay and Loss

Figure **7.8** 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% packet loss for video, 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.

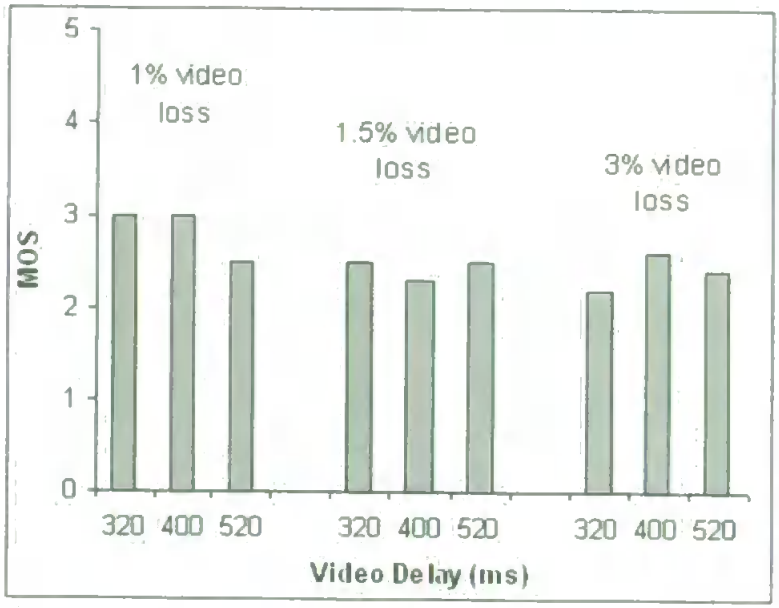


Figure 7.10: Video MOS Vs Video Delay and Loss

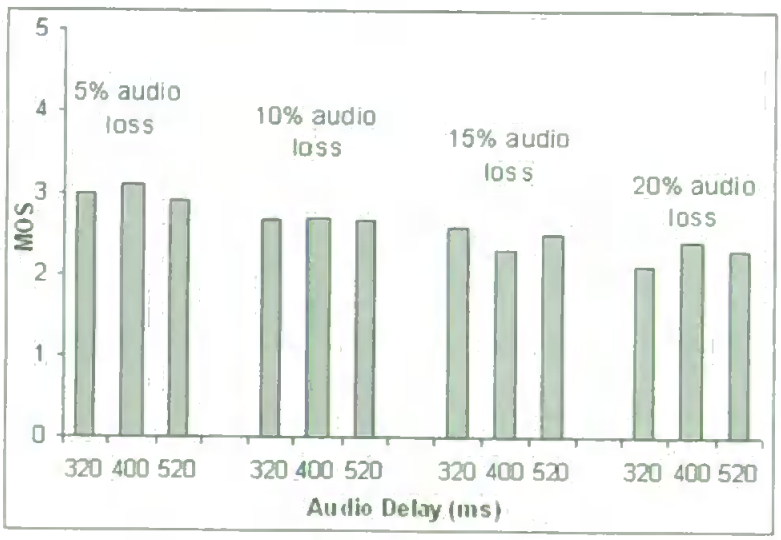


Figure 7.11: Audiovideo MOS Vs Audio Delay and Loss

The test candidates claimed that the 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 generally complicated since the degrees of deteriorations are constantly changing.

Figure **7.9** 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 decreases as the audio packet loss increases (i.e. from 15% to 20%), despite the fact that the quality setting of the video stream was unchanged for the test section. Hence, it is concluded that the MOS of video is affected by the quality of audio, i.e. the subjective opinion of the perceived quality of video is degraded in relation to the increased deterioration in the audio quality.

Figure **7.10** shows the MOS of the perceived quality of video, for video loss of 1%, 1.5% and 3%. It has been observed 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 also 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.

Figure **7.11** shows the MOS of the perceived quality of audiovideo overall, obtained from 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 2.7 (at 10% loss), 2.3-2.6 (at 15% loss) and 2.1-2.3 (at 20% loss).

Figure 7.12 shows the MOS of the perceived quality of audiovideo for the interactive test, for video loss of 1%, 1.5% and 3%. The scores 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 deteriorating gradually, with respect to video delay (i.e. from 320ms to 520ms). At 3% video loss, the average MOS is around 2.6, which is between FAIR and POOR quality threshold. However, the video delay (at this stage) does not shows any effect on the perceived audiovideo overall MOS. In general, the overall score is higher than that of the MOS of audiovideo where the audio losses and delays were introduced.

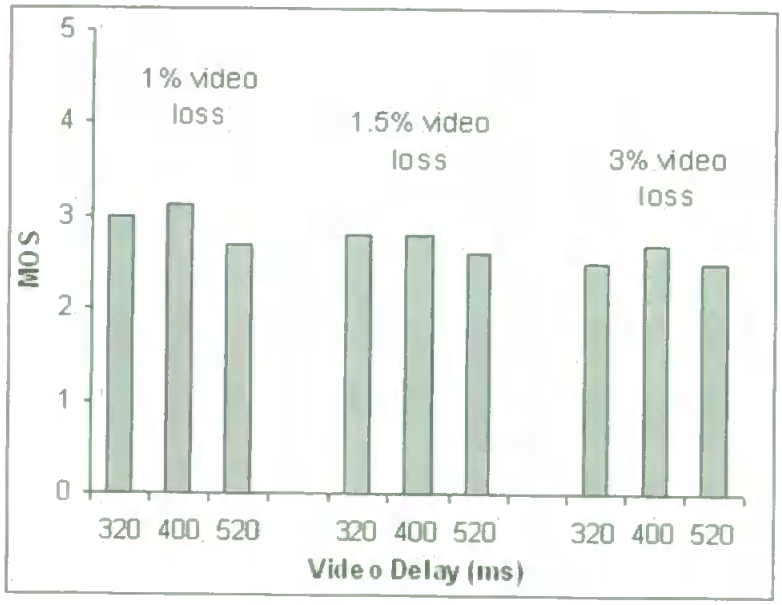


Figure 7.12: Audiovideo MOS Vs Video Delay and Loss

7.3.2.2 Result: 4-category Rating Results

The result, based upon the percentage of students responding in each of the 4-category, is shown in Figure 7.13.

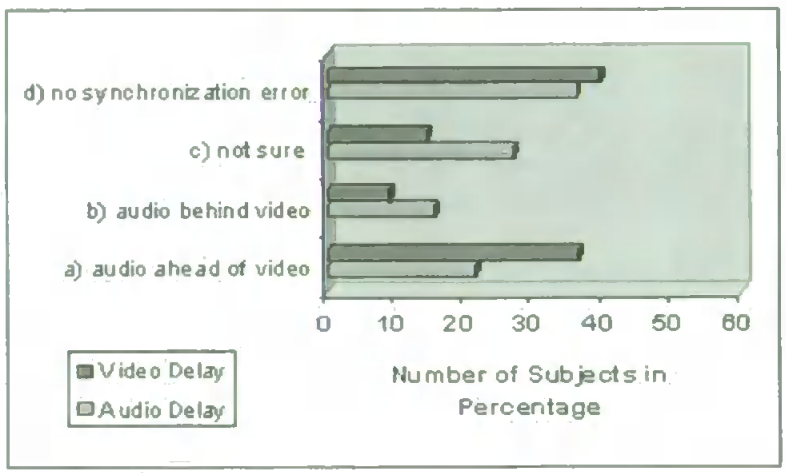


Figure 7.13: Lip Sync – Audio Delay Vs Video Delay

A high percentage of students stated that there is no lip sync error for both audio and video delay (i.e. 39.5% and 36.4%, respectively). In spite of the delayed audio(i.e. 440ms), as high as 36.8% of the students claimed that the audio is perceived to be ahead of video. Meanwhile, for the delayed video streams, 27.1% of the subjects were unsure which media is ahead of the other. However, only 21.5% of them stated that video is lagging of audio. It is observed that, throughout the interactive test section, the lip sync error is hard to define. Hence, it is then concluded that in the low cost desktop videoconferencing systems (e.g. Microsoft NetMeeting), the lip sync problem is a very complex issue and that it does not seems to affect the perceived multimedia quality, in the context of the informal interactive task performances.

7.3.3 Investigating the Effects of Jitter on Lip Sync

The work involved in the comprehensive evaluation of achievable audio and video quality, following the experimental design described in the previous test section, to investigate the effect of jitter on lip sync. Instead of delay and loss, a set of jitter values were introduced to the audio and video streams separately, in random order. For Audio, the jitter of 40ms, 80ms, 120ms, 160ms, and 200ms were selected to give the gradual impact to the perceived media quality, from minimal to worst. While for video, the values were 0.5ms, 1ms, 3ms, and 5ms. There were 18 subjects participated in the study. The same subjects as represented in Table 7.5 were involved in the test.

7.3.3.1 Results: MOS

This section describes the effects of jitter on the perceived multimedia MOS. Figure 7.14 shows the MOS results, with respect to 40ms - 200ms audio jitter. At 40ms audio jitter, the audio quality was reported to be considerately Good (i.e 3.22 MOS). However, on reaching 80ms to 200ms audio jitter, the perceived audio quality deteriorated significantly and the MOS drops to 2.11 - 1.7, i.e Poor quality.

Within these ranges, the video quality was not affected (i.e. the MOS fluctuated between 2.2 - 2.6). The MOS of audiovideo overall, however, has shown a great degradation which follows the same pattern as that in audio MOS.

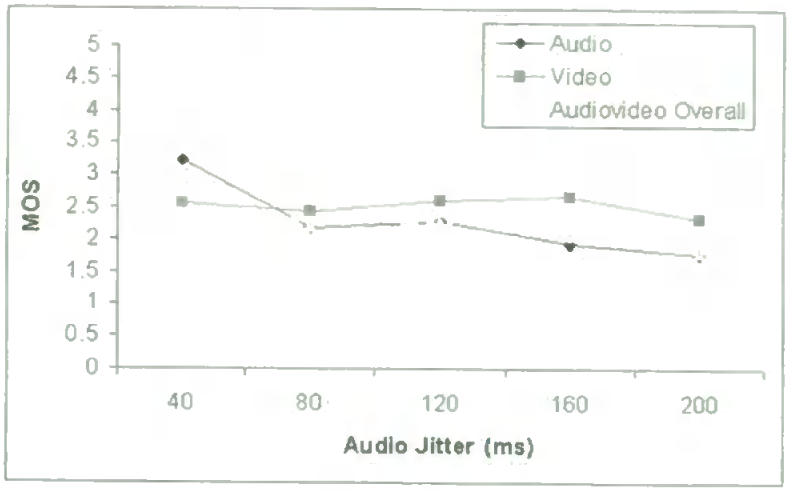


Figure 7.14: Audio Jitter Vs MOS

However, in the case of video jitter, for the range of 0.5ms to 5ms video jitter, the majority of the subjects claimed that there is no lip sync error on the perceived IP medias. The result is confirmed by the 4-category rating test, where a high percentage of the subjects, (i.e. 61%, 50%, 44%, and 67% for 0.5ms, 1ms, 3ms, and 5ms video jitter, respectively) did not notice any lip sync problem or strange effect. Consequently, the MOS of the perceived quality of audio, video, and audiovideo are maintained at around 3.3, 2.4, and 2.9, respectively (for the range of 0.5ms to 5ms video jitter). Thus, for the selected sets of video jitter, the perceived media quality is not affected.

7.3.3.2 Result: 4-category Rating

Table 7.8 shows the number of students (in percentage) responding to the 4-category rating, for the given jitter values.

Table 7.8: Audio Jitter - 4-Category Rating (Lip Sync Test)

Audio Jitter	40ms	80ms	120ms	160ms	200ms
Audio is ahead of video	22.2	16.7	16.7	11.1	11.1
Audio is behind video	5.5	27.8	16.7	38.9	33.3
Random lead/lag (not sure)	22.2	38.9	38.9	44.4	38.9
No synchronization error	50	16.7	27.8	5.5	16.7

At 40ms audio jitter, half of the number of the test subjects claimed that there was no strange effect or synchronization error on the perceived media quality. However, around 22% of them had chosen option C (Random lead/lag (not sure)).

As the jitter increased to 80ms, as high as 38.9% of the subjects noticed the lip sync error but could not decide which media is ahead of the other. As many as 27.8% of the subjects perceived that audio is played behind video.

At 160ms and 200ms jitter, 44.4% and 33.3% of the subjects, respectively, were unable to distinguish whether audio is ahead of video or otherwise. On the other hand, 38.9% (160ms) and 33.3% (200ms) of the subjects, did perceive that video is ahead of audio, which complies with the experimental settings.

7.3.4 Investigating the Effects of Combined Network Constraints on Lip Sync

The aim of the experiment is to investigate the combined network constraints Effect on Lip Sync. For the study, there are five (5) different sets of network constraints under investigation, which are:

- Network A: 0.5%–packet loss; 1ms–jitter; 100ms–delay
- Network B: 1.5%–packet loss; 5ms–jitter; 100ms–delay
- Network C: 3%–packet loss; 10ms–jitter; 200ms–delay
- Network D: 10%–packet loss; 15ms–jitter; 300ms–delay
- Network E: 20%–packet loss; 30ms–jitter; 500ms–delay

These figures were selected based on the laboratory results to represent network congestion impacts, from the minimum to the maximum (i.e. the point where the IP media becomes unusable). It is important to note that these network figures (traffics) are also realistic to real world.

These network conditions were assessed, in random order, using the MOS (in the context of the perceived multimedia quality), as well as the 4-category ratings – for the lip sync test. The task performance undertaken was based on the interactive test and there were 20 participants involved in the study. The same subject as shown in Table 7.5 were participated in the test.

7.3.4.1 Results: MOS

The MOS of the different network categories are illustrated in Table 7.9.

For Network A, B and C, the perceived quality of audio falls between Fair and Good MOS ratings which are 3.41, 3.5 and 3.28, respectively. On reaching network config-

uration D (i.e. 10%, 15ms jitter and 300ms delay) and network configuration E (i.e. 20%, 30ms jitter and 500ms delay), the MOS of the perceived audio quality drops to 2.67 and 2.33 respectively, which are categorized as between Fair and Poor quality.

The video quality is between Fair and Poor for Network A (2.76 MOS) and Network B (2.5 MOS). However, the quality drops to below Poor threshold for Network C, D and E. The same pattern is repeated in the audiovideo quality, but with slightly higher score.

For Network E, the overall quality is Poor, where the MOS are 2.33 (audio), 1.56 (video) and 1.94 (audiovideo). In Network D, the audio quality is between Fair and Poor rating scales (2.67 MOS), which is still acceptable. However, the MOS of the perceived quality of video and audiovideo overall is almost as bad as that in Network E.

The majority of the students claimed to have no problem of having the conversation over the Network A, B, C and D. However, as many as 55.6% of them experienced the difficulties in hearing and talking over the Network E connection (refer to the comm. diff. column in Table **7.9**). When asked, how would the lack of lip synchronization error affects the subjects (giving the scale between 1 to 5, where 1-very annoying and 5-less annoying), the majority of the subjects scored 3, 4, and 5, i.e. 26.5%, 41.2%, and 27.9% respectively over the Network A to D.

For the lip sync test in Network E, the subjects' opinions diverged (almost equally) from the rating scale of 1 to 5, i.e. 23.5% scored 1, 17.6% scored 2, 5.9% scored 3,

29.4% scored 4 and 17.6% scored 5.

Table 7.9: Different Network – MOS

Network	a (mos)	v (mos)	av (mos)	comm. diff. (%)
Network A	3.41	2.76	2.94	yes=35.3; no= 64.7
Network B	3.5	2.5	2.78	yes=16.7; no= 83.3
Network C	3.28	1.94	2.44	yes=27.8; no=72.2
Network D	2.67	1.61	2.22	yes=38.9; no= 61.1
Network E	2.33	1.56	1.94	yes=55.6; no= 44.4

note: a – audio; v – video; av – audiovideo

comm. diff. – communications difficulties

7.3.4.2 Results: 4-Category Rating

The overall results, as in Table **7.10** show that the majority of the subjects believed that there is lip sync error but they could not decide whether audio is lagging or ahead of video. For example, 44.4% for Network A, B and C, 38.9% for Network D and 66.7% for Network E.

However, for Network A, B, and C, a large number of students did not notice any lip sync error, i.e. 38.9% (Network A), 33.3% (Network B) and (Network C).

Table 7.10: Different Network – Lip Sync

Network	A	B	C	D	E
Audio is ahead of video	5.5	22.2	11.1	27.8	16.7
Audio is behind video	11.1	0	0	11.1	0
Random lead/lag (not sure)	44.4	44	44.4	38.9	66.7
No synchronization error	38.9	33.3	38.9	22.2	16.7

7.3.5 Discussion

It was observed that the video streams tend to arrive later then the played audio in the majority of the test sequences since it takes comparatively longer time to encode and decode video than it does for audio. The video data was also updated relatively infrequently due to the low bandwidth availability. As the results, the subjects perceived a low frame rate image that is not synchronised with the audio stream. However, as the remaining text will explains, the test candidates were surprisingly tolerant to asynchrony between the audio and video media.

In general, 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. The result suggested that video delay has less significant effect on the perceived quality of audio, video and audiovideo overall.

This could be due to the fact that, since the provision of video has often been viewed as being of secondary importance to audio (besides for its psychological benefits), little attention has been given to the change in video data. It is also apparent that, in addition to the low frame rates in low cost videoconferencing systems, the designated task performances may have low frame rates and hence, the subjects may not notice the delayed or missing frames.

A number of subjects claimed to notice the lip sync error but had the difficulties in

distinguishing 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. Perhaps, as they were so involved in the conversations, the mind no longer detected the lip sync error. In effect, some subjects did not even claim that every synchronization error is annoying and it is generally, ignored. Some observed that; audio, that is not synchronized with video, can be distracting or appeared strange due to the loss of lip sync. 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 (3) and GOOD (4) quality. In fact, the perceived multimedia quality was not affected even though the delay was as high as 600ms, when there is no packet loss. Hence, it is concluded that, in application scenarios where the subjects are having an informal conversation and that they are well acquainted with one another, lip sync error is not a critical issue. This finding is a contradiction with that of the ITU G.114 Recommendation [1], which stated that audio delays should be kept less than 200ms, for effective interaction.

On the other hand, a large number of students (approx. 40%) rated the same range of scores (FAIR–3 and GOOD–4), although they stressed that there is no lip sync error.

It was also observed that, more attention was given to the lip sync in passive test and hence, a larger number of subjects detected the correct sequence of the played media in passive test as compared to that in an interactive test. However, the passive test has shown higher MOS throughout the tests. It is then suggested that the subjective

evaluation of lip sync effect on the perceptual multimedia components is depended on the task performances.

Audio is perceived to be of good quality (3.33 MOS) at 40ms audio jitter. Beyond 40ms jitter (audio), the perceived quality of audio and audiovideo overall drop to the POOR quality threshold (2 MOS and below), gradually. However, the perceived quality of video media is not affected.

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. This 'blockiness' became more apparent as the subject movements increased. 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 video quality result has no meaningful term.

In the combined network constraints assessment test, the majority of the students reported to have the difficulties in the two-ways communications for network configuration E (see Section 7.3.4). As for Network C and D (see Section 7.3.4), although the quality of video and audiovideo overall are perceived to be around POOR quality threshold but with fairly good audio (refer to Table: 7.9), the students claimed to have no problem in having the conversation. Hence, it can be justified that good audio quality is essential to determine the IP multimedia QoS.

The overall MOS 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 quality of the other. The finding implies that, the QoS of one media can be predicted by the perception or the assessment of the other. Also, it be can suggested that the perceived quality of one media can be improved by upgrading the quality of the other and vice-versa.

7.4 Summary

The conclusion of findings from the evaluations described in this chapter can be summarised as follows.

- the mean opinion score of audio, video and audiovideo overall is unaffected by the loss of synchronization between audio and video media. It has been suggested that, lip synchronisation problem would compromise the quality of two-way communication in that the nature of interaction is claimed to be significantly awkward and less satisfactory when the mismatch between audio and video streams exceed 200ms [1], [2], and [3]. However, the extensive experimental results concerning these issues has concluded that the perceived multimedia quality in low cost DVC is unlikely to be affected by the lip sync error, in the context of passive and (informal) interactive communications. It is concluded that since the image frame rates are relatively slow (below 5 fps) to

obtain substantive lip sync, the subjects were making allowance for the asynchrony between the audio and video streams and concentrated on the audio to understand the sentences;

- media delay (audio and video), as high as 600ms (3.3 MOS) does not affected the two-way communications which is different from the previous finding, i.e. 500ms round trip delay is the upper limit that could be tolerated without major drops in conversational effectiveness [42]. However, beyond 600ms audio delay, the effect becomes discernable i.e. the MOS starts to decrease, in the context of interactive communication. However, the MOS for the perceived audio quality still reaches the FAIR rating score which is 3.0–3.3 MOS, while the perceived video quality is maintained at approximately 2.5–2.6 MOS (i.e. between FAIR and POOR rating scale). The explanation for this higher acceptable delay is that users also have the video image to look at and these far more tolerant than in a voice only test;

- the attention given to the lip sync assessment (and multimedia QoS, in general) is clearly depended on the task performance in that the subjects were observed to be more attentive while monitoring the quality and giving the score;

- the perception of one media is strongly influenced by the other, therefore: (i) the QoS of one media can be predicted by the perception of the other; (ii) the Qos of one media can be improved by upgrading the quality of the other;

- Audio jitter of 40ms is the minimum permissible jitter value to achieve satisfactory audio quality.

This chapter investigated the effects of lip synchronisation error, due to delay, jitter and packet loss, on the perceived quality of audio, video, and audiovideo overall in low cost videoconferencing systems, based on the subjective test method (MOS). The tests experiments conducted in Chapter 5, 6, and 7 are classroom-based experiments, with contrived tasks assigned to the subjects. The next chapter will presents the field study conducted between the University of Plymouth and Sarawak, Malaysia, which aims to make the comparisons between controlled classroom-based experiment and real-world environments, to provide actual real-world confirmation of the bench tests.

Chapter 8

Field Study

8.1 Introduction

The studies reported in this thesis so far have been conducted using NIST Net to provide an environment to model the performance of the Internet in a controlled manners. The next phase of the research actually employs the Internet as transmission medium for conducting subjective QoS measurements. Terminals were set up in UoP and Sarawak, Malaysia in identical manner to that described in the earlier tests.

Measurement of jitter, packet loss, and round trip time (rtt) were obtained for each connection using TCPdump [106]. QoS measurements obtained were compared with those obtained using NIST Net.

In general, the objectives of the field trial study are three-fold:

- to investigate the perceived multimedia quality in DVC and other related issues in real world scenario using the Internet as a transmission medium;
- to make the comparison of results between controlled classroom-based experiment and real world environment using the Internet as a transmission medium;
- to investigate the current IP network traffic behavior, i.e packet loss, round trip time, and jitter and how they affect the perceive media quality.

The studies describe in this chapter were divided into two main categories, namely, Study I and Study II. In study I, the communication links were conducted between the 1000MB backbone / 100 MB (University of Plymouth) and 56KB Voice MODEM (Sarawak). Whereas, in study II, the connections were made between the University of Plymouth and University of Malaysia Sarawak (UNIMAS) that has LAN's access speed of 100MB.

8.2 Study I: Experimental Procedure

8.2.1 Videoconference Configurations

Figure 8.1: Desktop Videoconference between Plymouth, UK (Site A) and Sarawak, Malaysia (Site B)

The systems specifications and configurations are outlined as follow:

- **Site A**: 1000MB backbone / 100MB access speed, PCI-based 10/100 Ethernet cards, Pentium III 700MHz (128.0 MB RAM) CPU, Microsoft Window 98 operating system. (Note: the data trace was captured at site A);

- **Site B**: 56KB External Data fax Voice Modem, PCI-based 10/100 Ethernet cards, Pentium IV 1.70GHz (128.0 MB RAM) CPU, Microsoft Window Me operating system. (Note: all the 12 participants are located at site B).

Both sites used the same videoconferencing tools, such as, Microsoft NetMeeting Version 3.01 and the USB video blaster Webcam Plus (capable of video capture up to 30 frames per second @ 352x288 pixels and 15 frame per second @ 640x480 pixels). To send and receive audio, two identical Platronic PC headsets (model LS1) were used. To make the systems configurations identical to the previous experiments in the thesis, the G723.1 (6400bit/s) audio CODEC and H.263 video CODEC, providing the Quarter Common Information Format (QCIF-176x144) frame size were used. As in the previous tests, the video setting was set to receive 'better quality'.

8.2.2 Test Subjects

There were 12 participants (6 males and 6 females), aged between 18 to 45 years involved in the test. These subjects have normal vision and hearing, and have never been involved in evaluating the multimedia quality before. The majority of them have very little or no knowledge about the Internet applications, and have never experienced videoconferencing over the IP network (except for one candidate), see Table **8.1**. Only a small number of them use the computer occasionally (approx. 48 hours a week). These subjects are all Malaysian, resided in a small town called Kuching, in Sarawak i.e. one of the provinces in Malaysia (on the island of Borneo).

The category of the subject is shown as follow:

Table 8.1: Subject Category

Subject	Gender	Professonal	Computer Literate	Ip Media Experience
1	Female	Student	Yes	Yes
2	Males	Student	yes	Very little
3	Male	Teacher	yes	No
4	Male	Police	Moderate	No
5	Female	Housewife	No	No
6	Female	Teacher	Yes	Very little
7	Male	Student	Yes	No
8	Female	Teacher	Moderate	No
9	Male	Teacher	Yes	No
10	Female	Housewife	No	No
11	Female	Housewife	No	No
12	Male	Teacher	Moderate	No

8.2.3 Task Performance

As previously mentioned, the videoconference was launched between the University of Plymouth and Sarawak (Malaysia), denoted as Site A and Site B respectively in Figure 8.1. The author from Site A, acting as the videoconferencing facilitator set up the desktop videoconference with each of the twelve participants located at Side B. The communications were based on one-to-one person group, as conducted in the previous studies. It is a requirement that the candidates are well acquainted with the facilitator to maximise the interaction and to ensure that they are fully at ease while communicating. Once the videoconference took place, the participant and the facilitator were involved in intensive informal communications for approximately 5-13 mins. At the end of the conversation, the participant was asked to mark the

scores on the answer sheets (see Appendix A).

8.2.4 Rating Method

Similar rating protocols as introduced in the previous chapter were repeated in this study in that the participants rated the perceived quality of audio, video and audio-visual overall components based on the ITU-T subjective 5-point rating, MOS.

The candidates were also required to classify the perceived synchronization error based upon 4 different categories, as the following;

- (i) audio is ahead of video;
- (ii) audio is behind video;
- (iii) not sure, whether audio is ahead or lagging video (random lead/lag - not specified);
- (iv) no synchronization error.

The level of annoyance caused by the lack of synchronisation between the audio and video media was also evaluated based on the 5-point rating scale, in that 5 being less annoying and 1 being very annoying.

8.2.5 TCP/UDP Data

The TCP/UDP data trace (containing TCP/UDP headers only) captured by tcpdump is shown in Figure 8.2. The beginning of each TCP/UDP segment is organised as follows:-

timestamp source ⟶ destination : flags

Tcpdump is a powerful network monitoring tool that allows all the packets going through the connection to be captured and statistically analyzed. The tcpdump output analysis was performed using the software developed by Bogdan Ghita [107] to obtain the network parameters, such as rtt, packet loss, and jitter as shown in Figure 8.3.

No.	Time	Source	Destination	Protocol	Info
1	0.000000	141.163.96.131	210.195.234.58	UDP	Source port: 49608 Destination port: 49608
2	0.063740	141.163.96.131	210.195.234.58	UDP	Source port: 49608 Destination port: 49608
3	0.092352	210.195.234.58	141.163.96.131	UDP	Source port: 49606 Destination port: 49606
4	0.135109	141.163.96.131	210.195.234.58	UDP	Source port: 49606 Destination port: 49606
5	0.135410	141.163.96.131	210.195.234.58	UDP	Source port: 49606 Destination port: 49606
6	0.189796	141.163.96.131	210.195.234.58	UDP	Source port: 49608 Destination port: 49608
7	0.252383	141.163.96.131	210.195.234.58	UDP	Source port: 49608 Destination port: 49608
8	0.287952	210.195.234.58	141.163.96.131	UDP	Source port: 49606 Destination port: 49606
9	0.361287	210.195.234.58	141.163.96.131	TCP	1116 > 1246 [PSH, ACK] Seq=0 Ack=0 Win=16316 Len=4
10	0.361670	210.195.234.58	141.163.96.131	UDP	Source port: 49606 Destination port: 49606
11	0.377836	141.163.96.131	210.195.234.58	UDP	Source port: 49608 Destination port: 49608
12	0.440836	141.163.96.131	210.195.234.58	UDP	Source port: 49608 Destination port: 49608
13	0.518027	141.163.96.131	210.195.234.58	TCP	1246 > 1116 [ACK] Seq=0 Ack=4 Win=8292 Len=0
14	0.553771	210.195.234.58	141.163.96.131	UDP	Source port: 49606 Destination port: 49606
15	0.566429	141.163.96.131	210.195.234.58	UDP	Source port: 49608 Destination port: 49608

Figure 8.2: Example of TCP/UDP Trace Data

```
RTCP SSRC 4138038665 210.195.234.58/141.163.96.131 SSRC 1317865026 RTT 1.10965 Lost 0 Jitter 0.11375
RTCP SSRC 1317865026 141.163.96.131/210.195.234.58 PktsSent 3499 RTT 0.000783062 Lost 27 Jitter 0.07725
RTCP SSRC 4138038665 210.195.234.58/141.163.96.131 PktsSent 1434 SSRC 1317865026 RTT 1.28729 Lost 0 Jitter 0.11375
RTCP SSRC 4138038665 210.195.234.58/141.163.96.131 SSRC 1317865026 RTT 0.465203 Lost 0 Jitter 0.11375
RTCP SSRC 1317865026 141.163.96.131/210.195.234.58 PktsSent 3555 SSRC 4138038665 RTT 0.00707225 Lost 27 Jitter 0.09175
RTCP SSRC 1317865026 141.163.96.131/210.195.234.58 PktsSent 3602 SSRC 4138038665 RTT 0.00633313 Lost 27 Jitter 0.09175
RTCP SSRC 4138038665 210.195.234.58/141.163.96.131 PktsSent 1446 SSRC 1317865026 RTT -4.42157 Lost 0 Jitter 0.11075
RTCP SSRC 4138038665 210.195.234.58/141.163.96.131 PktsSent 1461 SSRC 1317865026 RTT -4.51782 Lost 0 Jitter 0.11075
RTCP SSRC 1317865026 141.163.96.131/210.195.234.58 PktsSent 3652 SSRC 4138038665 RTT 0.00299525 Lost 29 Jitter 0.08675
RTCP SSRC 4138038665 210.195.234.58/141.163.96.131 PktsSent 1473 SSRC 1317865026 RTT 0.727069 Lost 0 Jitter 0.165625
RTCP SSRC 1317865026 141.163.96.131/210.195.234.58 PktsSent 3683 SSRC 4138038665 RTT 0.00247525 Lost 29 Jitter 0.069875
RTCP SSRC 4138038665 210.195.234.58/141.163.96.131 SSRC 1317865026 RTT 1.5773 Lost 0 Jitter 0.12875
RTCP SSRC 1317865026 141.163.96.131/210.195.234.58 PktsSent 3754 SSRC 4138038665 RTT 0.00273638 Lost 29 Jitter 0.097625
RTCP SSRC 4138038665 210.195.234.58/141.163.96.131 PktsSent 1482 SSRC 1317865026 RTT 0 Lost 0 Jitter 0.206125
RTCP SSRC 1317865026 141.163.96.131/210.195.234.58 PktsSent 3793 RTT 0.00273638 Lost 29 Jitter 0.091125
```

Figure 8.3: UDP Data – Showing Rtt, Lost, and Jitter

8.3 Results

This section is divided into three subsections:

- Traffic Monitoring Results – presents the traffic (i.e. rtt, packet loss, and jitter) analysis, collected between host Site A (Plymouth) and host Site B (Sarawak) for the individual 12 videoconferencing end users;
- MOS and Lip Synchronisation Results – presents the perceived quality of audio, video, and audiovideo overall for each of the participants, based on MOS rating method and the effects of lip synchronisation error;
- Traffics Vs MOS – marrying the traffics with the MOS to quantify their effects on the perceived quality of audio, video and audiovideo overall and to compare the outcomes with the findings obtained from the classroom-based studies.

8.3.1 Traffic Monitoring Results

Table 8.2 shows the date and time when the videoconference took place for each of the participants. It can be seen from the table that most of the desktop videoconference sessions were launched around 11:00 am to 13:00 pm, the UK local time which was seven hours lagging that of the Malaysian local time. The durations of the videoconferences varied from 5 to 13 mins, depending on the communicative party and the issues of the discussions. The videoconferences were conducted on three different days, as shown Table 8.2. The network traffic collected from the links were analysed and are presented in this section.

Table 8.2: Videoconference Time Allocation

Subject	Date Year 2003	Start Time (UK time)	Durations (Approx.)
1	30-Aug	12:23	8 mins
2	30-Aug	12:57	10 mins
3	30-Aug	12:31	13 mins
4	30-Aug	12:51	6 mins
5	13-Sep	13:00–14:00	5 mins
6	30-Aug	12:44	7 mins
7	13-Sep	13:00–14:00	5 mins
8	13-Sep	13:00–14:00	
9	15-Sep	11:48	6 mins
10	13-Sep	13:00–14:00	5 mins
11	13-Sep	13:00–14:00	5 mins
12	13-Sep	13:00–14:00	5 mins

8.3.1.1 Audio Jitter

Figure 8.4 shows the audio jitter versus packet sequence numbers experienced by Subject#_1, 2, 3, 4, 5, and 6, whereby Figure 8.5 represents the audio jitter obtained from Subject#_7, 8, 9, 10, 11, and 12. All the participants experienced fairly high audio jitter with the average of above 50 ms (*stdev* between 11%–20% of the average value), except for subject#_2 which had 49.89 ms jitter with the standard deviation *(stdev)* of 9.03. Subject#_7, 8, and 10 suffered from higher audio jitter compared to the rest of the group, i.e. 100.83 ms *(stdev 14.24)*, 99.37 ms *(stdev 12.85)*, and 100.18 ms *(stdev 11.14)* respectively, (see Figure 8.5).

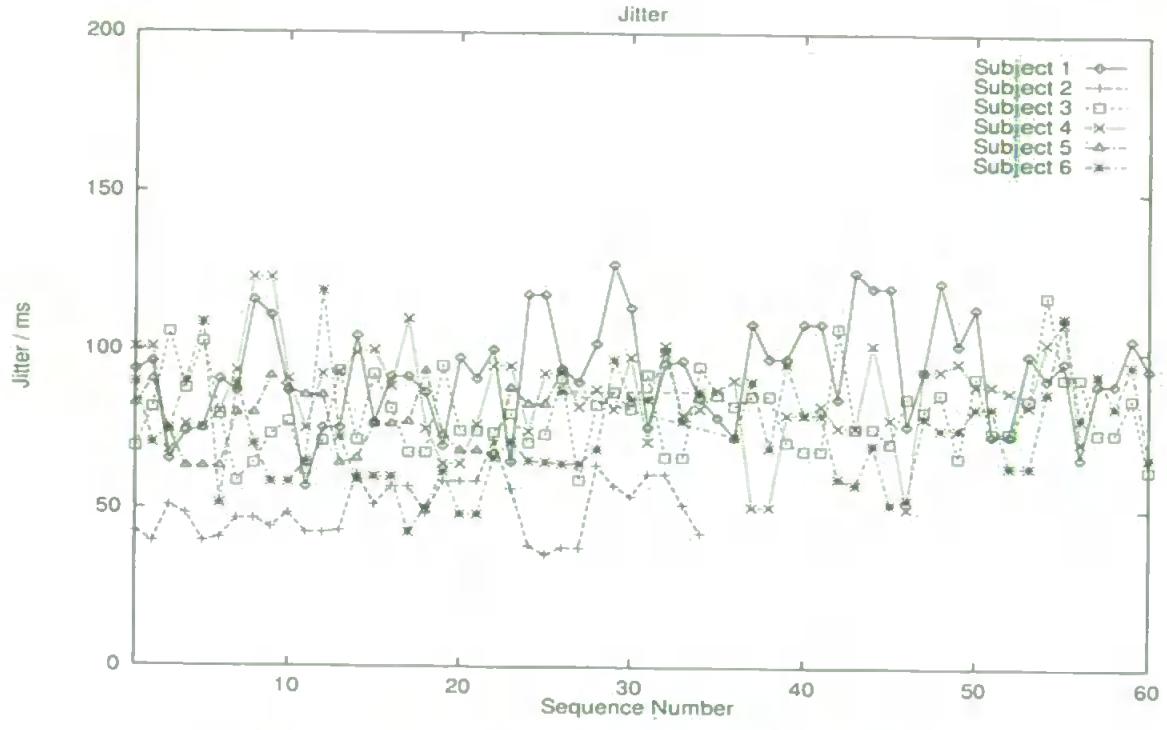


Figure 8.4: Audio Jitter (ms) – Subject#_1 to 6

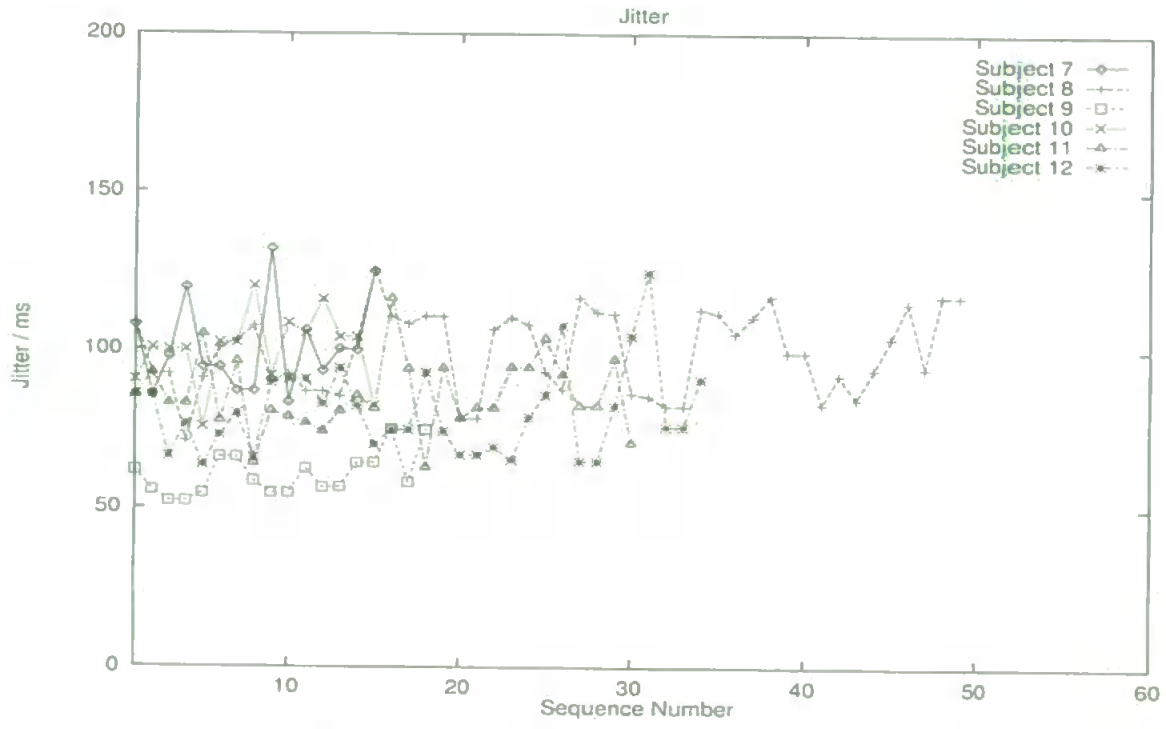


Figure 8.5: Audio Jitter (ms) – Subject#_7 to 12

8.3.1.2 Audio packet Loss

By referring to Figure 8.6, the audio packet loss experienced by the majority of the candidates was fairly small, i.e. around 3% and below. The audio packet loss captured on the 30th August is much lower than that obtained on the 13th of September. For example, on the 30th August 2003 Subject#_1, 2, 3, and 6 gained the audio packet loss of 0.89%, 1.97%, 1.32%, 1.83%, and 1.71% respectively. However, for Subject#_5, 7, 8, 10, 11, and 12, the audio packet loss (captured on the 13th September 2003) are 4.88%, 2.62%, 2.62%, 2.64%, 2.9, and 4.47% respectively. This implies that network traffic varies largely with respect to the time and date it is being captured. As shown in Figure 8.6 the standard deviation for all these lost packets are generally very small.

It is observed that the packet loss obtained from the links with Subject#_5 and Subject#_12 show higher packet loss, with larger standard deviation, i.e. in the average of 12.4% and 22.1% from the average values respectively. From the graph, it can be seen that the packet loss experienced by Subject#_12 decreases as the number of sequence (time) increases, i.e from approx. 7% to 3.5%. On the other hand, the packet loss in Subject#_5 increases after the 10th sequence number from approx. 4% to 5.8%. However, based on the previous results (from the laboratory based study), the perceived audio quality were not affected by these amount of packet loss, i.e. below 8%.

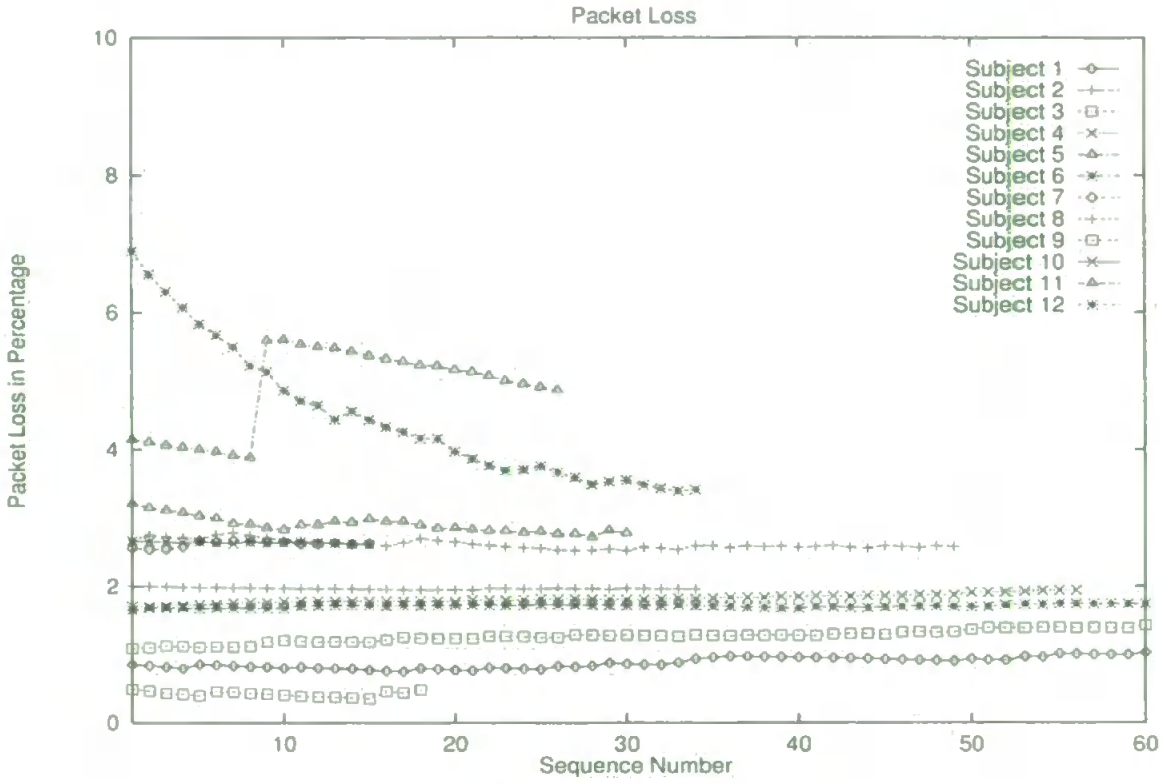


Figure 8.6: Audio Packet Loss (%)

8.3.1.3 Round Trip Time

The rtt values varied rapidly (as can be seen in Figure **8.7** and Figure **8.8**) which leads to very high rtt standard deviation. For example, the percentages of the standard deviations from the average rtt values were 72.57%, 75.96%, 87.48%, and 71%, taken from Subject#_1, 2, 3, and 7 respectively.

It can be seen from Figure **8.7** that Subject#_3 experienced a very high rtt with the average of 7.67 ms *(stdev 6.71)*. However, the amount of jitter and packet loss reported in Subject#_3 link were fairly small, i.e. 79.39 ms and 1.32% respectively. Hence, it is interesting to note that the figures for the audio jitter and packet loss were not affected by the rapid change in the number of rtt.

On the other hand, it can be concluded from the collected data that the amount of jitter and packet loss are inversely proportional to the rtt i.e. the traffic figures increased when the reported rtt values are fairly small.

Note that the rtt collected from Subject#_9 link were zero since the participant was able to receiving audio and video streams but only able to send the video due to an unknown reason. Since the participant had some other business to attend to, and could not wait for the problem to be solved, it was decided that the videoconference was to proceed although it was based on just the one-way audio channel, with video image at both ends. One of the disadvantages of field trial over a long distance, such as this one, is that the time allocation to conduct the communication link is restricted to the 7 or 8 hours time difference between the two countries.

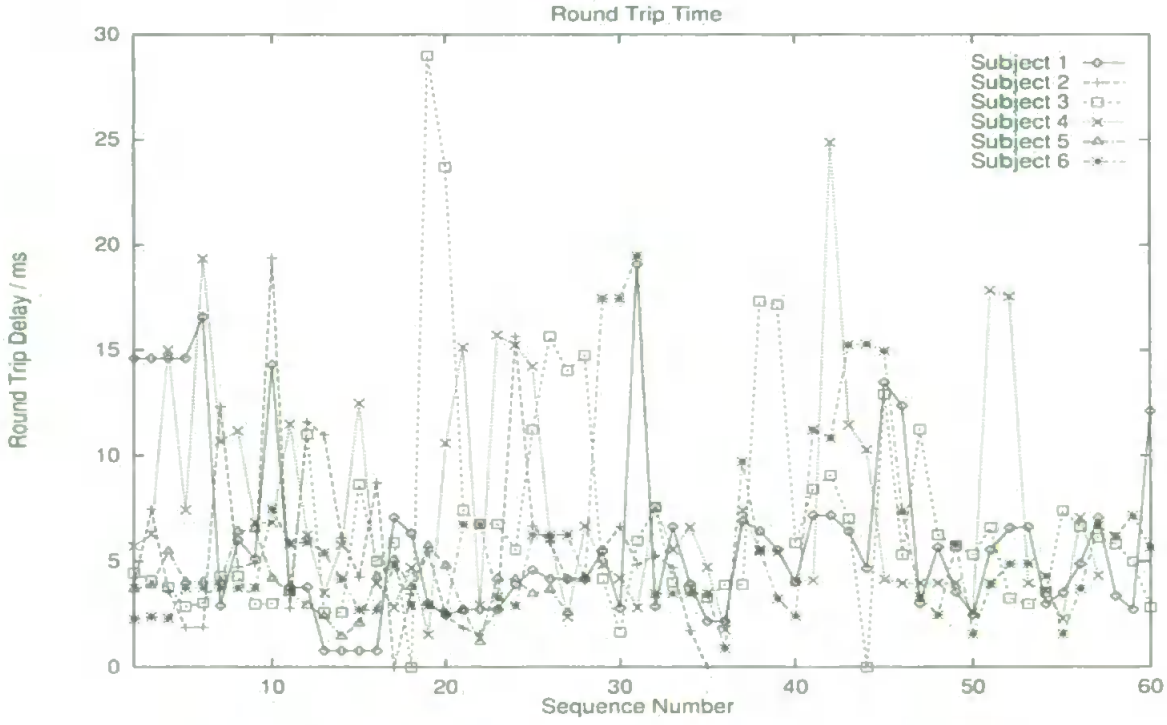


Figure 8.7: Audio Round Trip Time (ms) – Subject#_1 to 6

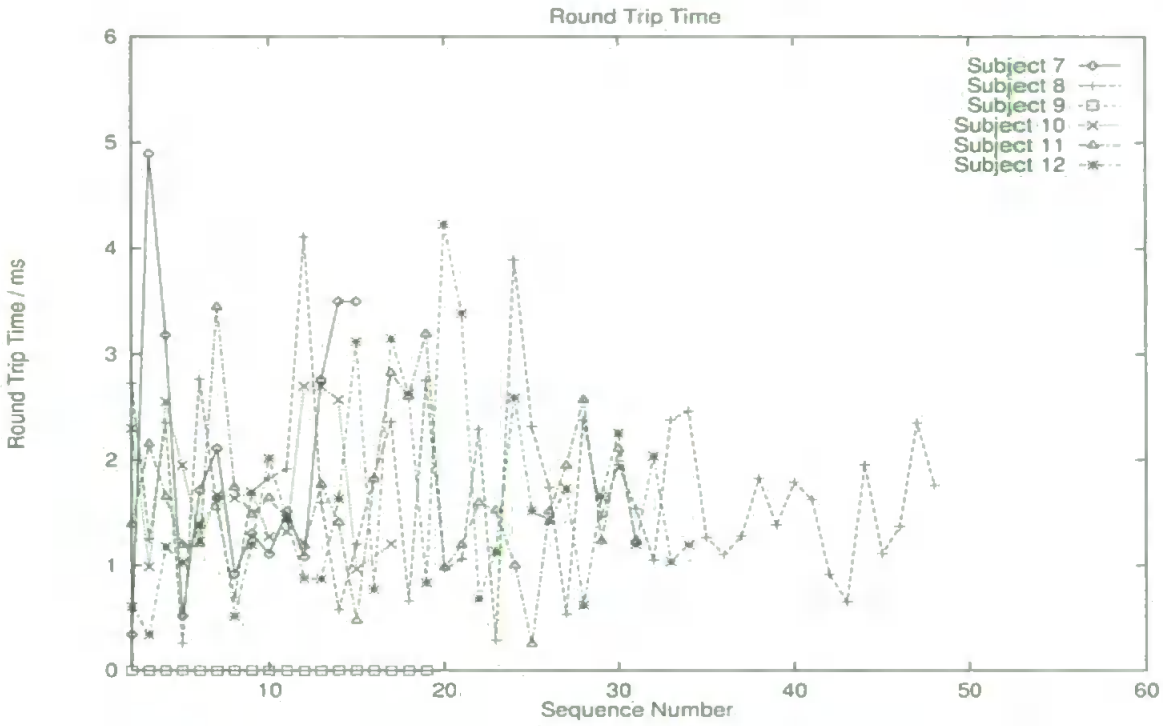


Figure 8.8: Audio Round Trip Time (ms) – Subject#_7 to 12

8.3.1.4 Video Jitter

Figure **8.9** shows the video jitter experienced by Subject#_1, 2, 3, 4, 5, and 6, whereby Figure **8.10** represents the video jitter obtained from Subject#_7, 8, 9, 10, 11, and 12. The jitter values varies constantly with time for all the communication links. This means that the standard deviation (stdev) of the video jitter are very high, i.e. between 15% to 30% of the average value. Also, the figures are much higher i.e generally in the range of 500 ms to 800 ms which are expected since the video data normally occupies a much bigger frame sizes. Therefore it introduces a longer delay and hence, a greater jitter can be produced. Among the group, Subject#_12 experienced the highest video jitter, i.e. 816.16 ms *(stdev 144.85)*, whereas Subject#_9, which used the one-way audio link, suffered the least, i.e 469.333 *(stdev 110.68)*.

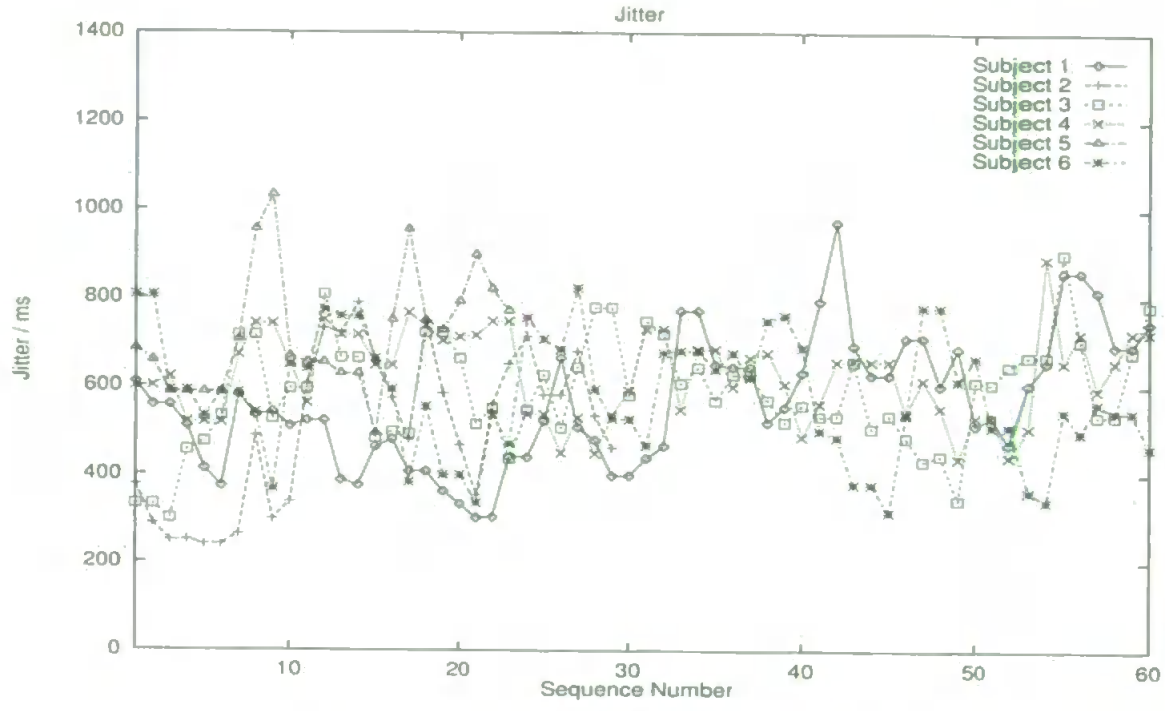


Figure 8.9: Video Jitter (ms) – Subject#_1 to 6

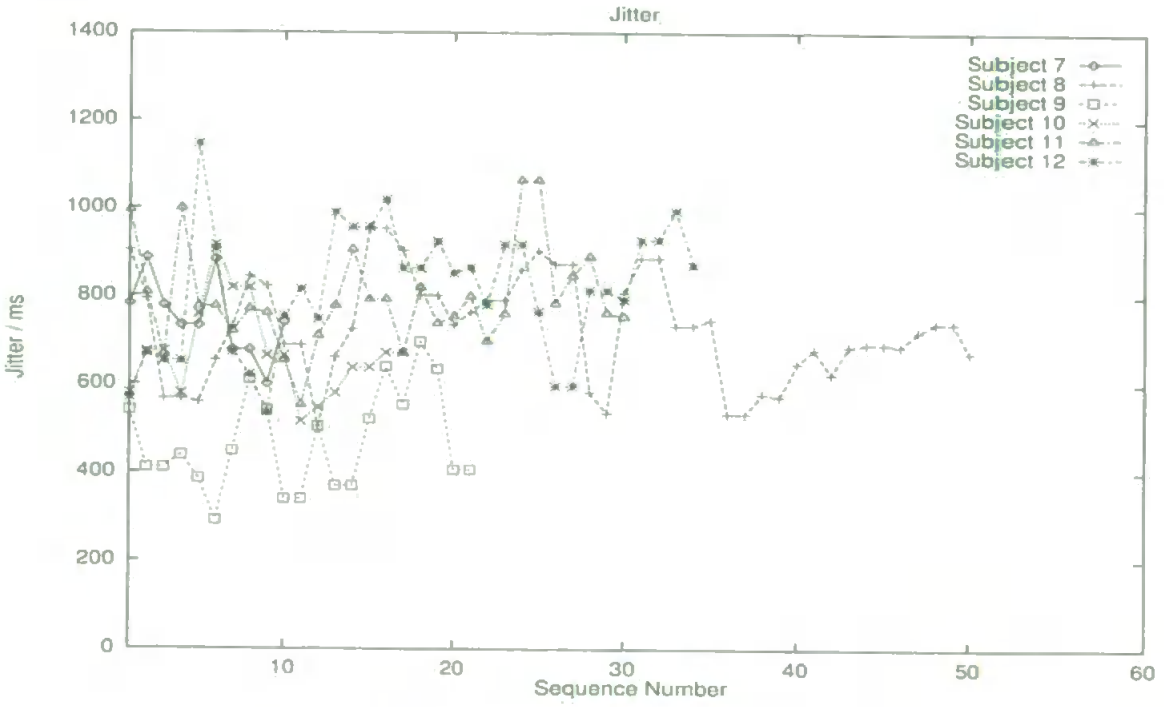


Figure 8.10: Video Jitter (ms) – Subject#_7 to 12

8.3.1.5 Video Packet Loss

Figure **8.11** shows the packet loss obtained from the video channels between the host in Site A and B, for all the twelve participants.

By referring to Figure **8.11**, the video packet loss experienced by the each candidate varies largely and can be divided into two group of range. The packet loss obtained on the 30th August 2003 were less than that collected on the 13th September 2003, as with audio. For example, on the 30th August 2003 Subject#_1, 2, 3, and 6 gained the packet loss of 2.7%, 3.78%, 2.76%, 3.45%, and 3.12% respectively. However, for Subject#_5, 7, 8, 10, 11, and 12, the packet loss (captured on the 13th September 2003) are 7.23%, 5.97%, 5.98%, 5.92%, 6.44, and 8.78% respectively.

By comparing Figure **8.6** and Figure **8.11**, it can be observed that the video packet loss of Subject#_12 follows the pattern of audio loss in that it decreases as the number of sequence increases, i.e from approx. 12.4% to 7.5%. Also, the video packet loss of Subject#_5 increases from approx. 6.5% to 8.2% on the 10th sequence number (approx.), where it then starts to decrease gradually with time.

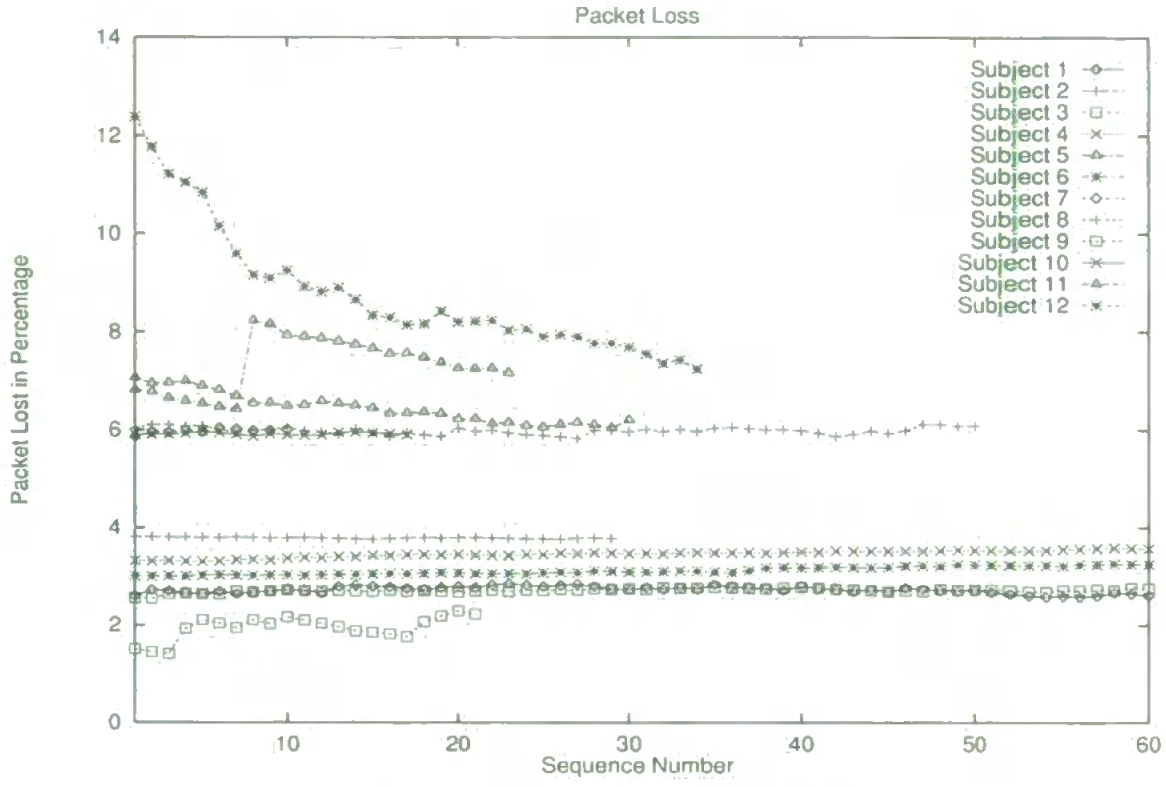


Figure 8.11: Video Packet Loss (%)

8.3.1.6 Video Round Trip Time

Figure **8.12** shows the video rtt experienced by Subject#_1, 2, 3, 4, 5, and 6, whereby Figure **8.13** represents the video rtt obtained from Subject#_7, 8, 9, 10, 11, and 12. Subject#_4 experienced a very high rtt standard deviation i.e. 120% of the average value, i.e. 9.32 ms. The number of jitter and packet loss reported by this link are 628.42 ms and 3.46% respectively. The rtt collected on the 30th August 2003, (i.e from the links of Subject#_1, 2, 3, 4, and 6) show the higher values that are within the range of 4.56 ms to 9.32 ms. The figures captured on the are much lower on the 13th September, i.e. 1.77 ms, 1.92 ms, 1.90 ms, 1.70 ms, and 1.89 ms, that can be seen in Subject#_7, 8, 10, 11, and 12.

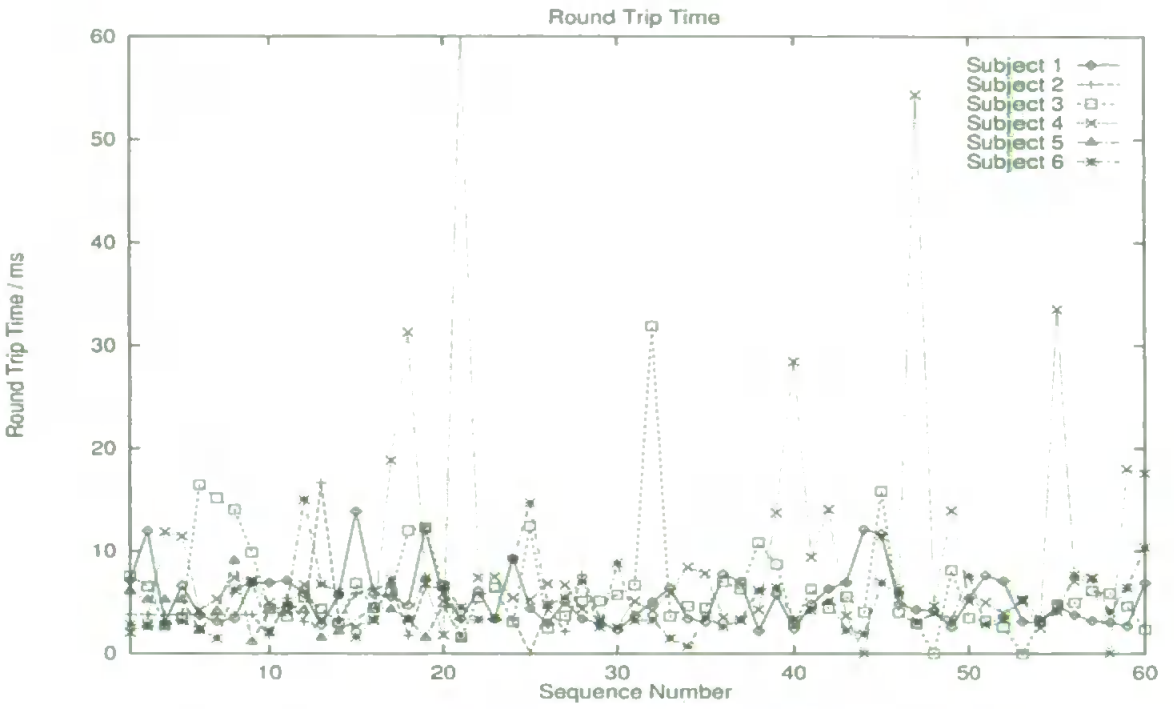


Figure 8.12: Video Round Trip Time (ms) – Subject#_1 to 6

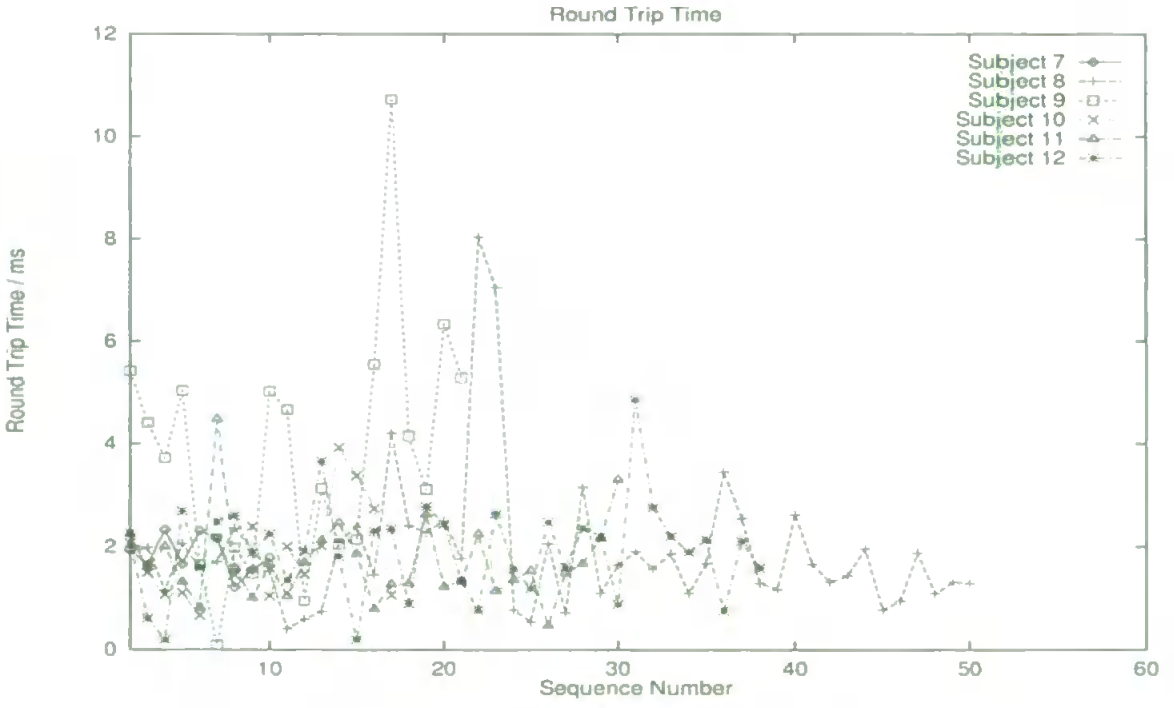


Figure 8.13: Video Round Trip Time (ms) – Subject#_7 to 12

8.3.2 MOS and Lip Synchronisation Results

Table 8.3 shows the subjective MOS results and lip synchronisation effects, reported by the individual participant involved in the study.

Table 8.3: Subject Scores

Subject	Audio MOS	Video MOS	Audiovideo MOS	4-Category Rating	Level of Annoyance	Conversation Difficulty
1	4	3	3.5	i. audio ahead	3	No
2	3	3	3	iii. not sure	3	No
3	3	3	3	iii. not sure	2	No
4	3	2	3	iii. not sure	3	No
5	4	3	4	iii. not sure	3	No
6	3.5	3	3	iii. not sure	3	No
7	2.5	2.5	2.5	iii. not sure	3	No
8	3	2.5	3	iii. not sure	2	No
*9	4	3.5	4	i. audio ahead	3	No
10	4	3	4	iii. not sure	3	No
11	4	3	3	iii. not sure	3	No
12	2.5	2.5	2.5	iii. not sure	3	Yes
Average	**3.32**	**2.77**	**3.14**		**2.83**	
Stdev	***(0.60)***	***(0.34)***	***(0.50)***		***(0.39)***	

By referring to the second column of Table 8.3, for the perceived audio quality, five out of twelve participants scored 4 MOS (GOOD) , four scored 3 MOS (FAIR), two scored 2.5 MOS (BAD/FAIR), and one scored 3.5 MOS (FAIR/GOOD), giving the average MOS of 3.32. For the perceived video quality, the majority of the participants scored 3 MOS (FAIR), i.e. seven out of twelve candidates, two scored 2.5 MOS (BAD/FAIR) and only one candidate scored 3.5 (FAIR/GOOD), giving the overall MOS of 2.77. For the perceived quality of audiovideo overall, the majority of them scored 3 MOS (FAIR), i.e. six candidates, three scored 4 MOS (GOOD), two scored 2.5 MOS (BAD/FAIR), and one scored 3.5 MOS (FAIR/GOOD), giving the average

of 3.14 MOS. In general, the candidates denoted as Subject#_1, 5, and 10 scored very high quality ratings that are above the average MOS result within the group (i.e. 3.5 – 3.7 MOS, i.e. between FAIR and GOOD quality scales) for all the multimedia components, whereas, Subjects#_7 and 12 scored the lowest (i.e. 2.5 MOS). It is interesting to note that the traffic analysis data, as previously explained were also in confirmation with these findings.

For the 4-category rating results, only two participants believed that audio is played ahead of video (i.e. Subject#_1 and 9), while the other ten subjects claimed to perceive the lip synchronisation error but were unable to distinguish between which media is ahead or lag of each other. When asked with respect to the level of annoyance, as many as ten participants stated that lips synchronisation error has moderate effect on them, although two participants (i.e. Subject#_3 and 12) claimed to be more affected, which gave 2-point score. With the exception of Subject#_12, all participants stated that they did not experience any conversation difficulty over the communication link.

8.3.2.1 Traffics Vs MOS

Table **8.4** and **8.5** show the audio traffic and video traffic, respectively, taken from Subject#_1 to 6 and Network (N/w) configuration A - from the study in Chapter 7 (see Section 7.3.4) and Figure **8.14** presents the relative MOS, given by the participants.

Table 8.4: Audio Traffic for Figure 8.14

Subject	Jitter (ms)	Network Traffic Packet Loss (%)	RTT (ms)
Classroom based N/w A	1	0.5	100
Field trial Subject#_1	92.48	0.89	5.98
Field trial Subject#_2	49.9	1.97	5.46
Field trial Subject#_3	79.4	1.97	5.46
Field trial Subject#_4	86.31	1.32	7.67
Field trial Subject#_5	78	4.88	3.82
Field trial Subject#_6	75.57	1.71	5.95

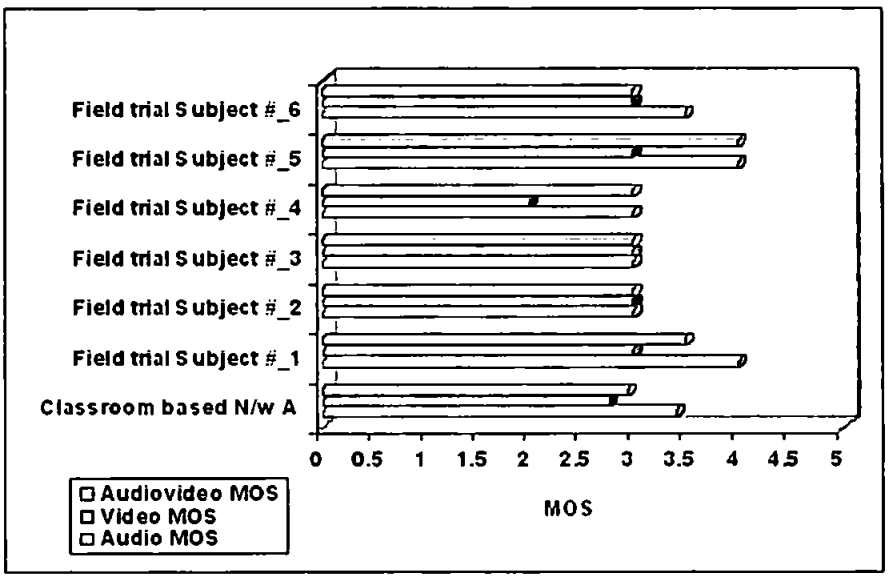


Figure 8.14: Subject#_1 to 6 vs network A

Table 8.5: Video Traffic for Figure 8.14

Subject	Jitter (ms)	Network Traffic Packet Loss (%)	rtt (ms)
Classroom based N/w A	1	0.5	100
Field trial Subject#_1	576.42	2.7	5.82
Field trial Subject#_2	500.18	3.78	4.56
Field trial Subject#_3	592	2.76	6.58
Field trial Subject#_4	628.42	3.46	9.32
Field trial Subject#_5	69.044	7.32	3.84
Field trial Subject#_6	590.38	3.12	4.64

Table 8.6: Audio Traffic for Figure 8.15

Subject	Jitter (ms)	Network Traffic Packet Loss (%)	rtt (ms)
Classroom based N/w B	5	1.5	100
Classroom based N/w C	10	3	200
Field trial Subject#_8	99.37	2.62	1.7
Field trial Subject#_10	100.18	2.64	1.72
Field trial Subject#_11	85.65	2.9	1.64
Field trial Subject#_12	80.47	4.47	1.57
Field trial Study I Average	84.3	2.54	3.89

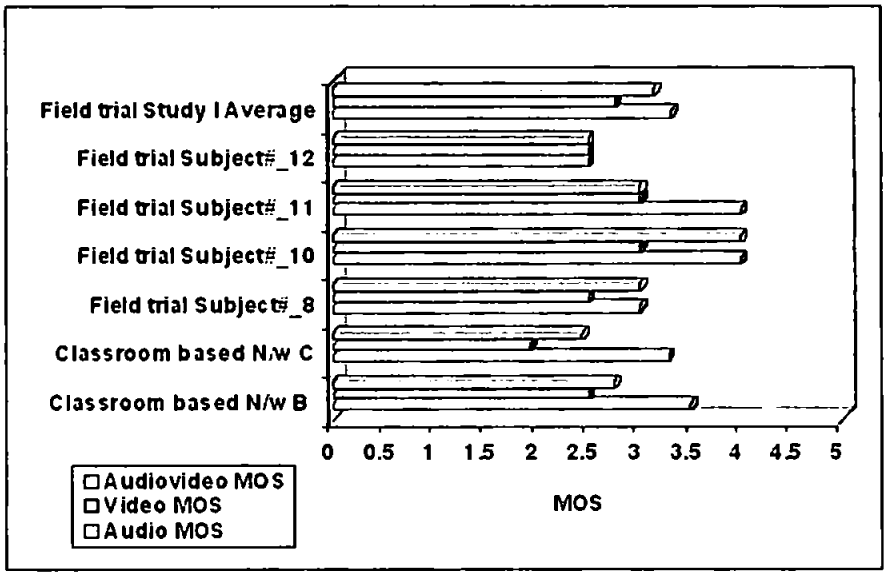


Figure 8.15: Subject#_8 to 12 vs network B and C

Table 8.7: Video Traffic for Figure 8.15

Subject/Study Type	Jitter (ms)	Network Traffic Packet Loss (%)	rtt (ms)
Classroom based N/w B	5	1.5	100
Classroom based N/w C	10	3	200
Field trial Subject#_8	728.31	5.98	1.92
Field trial Subject#_10	671.14	5.92	1.9
Field trial Subject#_11	795.15	6.44	1.7
Field trial Subject#_12	816.16	8.78	1.89
Field trial Study I Average	610.59	5.11	3.99

Table **8.6** and **8.7** show the audio traffic and video traffic, respectively, taken from Subject#_8, 10, 11 and 12 and Network (N/w) configuration B and C - from the previous study in Chapter 7 (see Section 7.3.4) and Figure **8.15** presents the relative MOS, given by the participants.

By comparing the MOS figures with respect to the traffic conditions, a number of observations have been made based on the end user's subjective opinion rating. Firstly, the MOS are generally higher with respect to packet loss (especially for video) and jitter as compared to that obtained from the laboratory based experiments.

In the previous studies, 80ms video jitter would severely affect the perceived multimedia quality in that the MOS was around 2 (POOR) for both audio and audiovideo overall (refer to Figure **7.14** in Section 7.3.3.1). However, the perceptual MOS of audio and audiovideo obtained from the field study are between 2.5 (POOR/FAIR) and 4 (GOOD) (Figure **8.14** and Figure **8.15**) with the average of 3.32 (for audio) and 3.14 (for audiovideo overall) even though the amount of the collected audio jitter are within the ranges of 50 ms to 100 ms, for audio (Table **8.4** and **8.6** and) and 470 ms to 816 ms, for video (Table **8.5** and **8.7**).

Similarly, the majority of the video scores captured from the field trial are generally higher, i.e MOS of 2.5% - 3%, as compared to the previous laboratory based studies presented in the earlier in the thesis, even though the packet loss for video channel obtained from the field trails are fairly high, i.e. mostly above 3% with the maximum number of 9% (see Figure **8.14** and Figure **8.15**).

8.4 Study II: Experimental Procedure

8.4.1 Videoconference Configurations

By referring to Figure 8.1 and Subsection 8.2.1, the systems specification and configurations and the videoconferencing tools used in Study II were the same as that in Study I, except for the systems specification and configuration of Site B that were given as follows:

- **Site B** – LAN 10/100Mbps, Realtek RTL8139/810x Family Fast Ethernet NIC, Pentium IV 2.4GHz (256MB RAM) CPU power, Windows 2000 operating system.

8.4.2 Test Subjects

The test candidates (and hence, the results) in this section can be categorized as Study IIa and Study IIb, depending on where they are located.

- **Study IIa**: all participants were located at Site A (UoP), ten PhD's researchers and lectures (all males), aged between 23–45 years old were involved, these participants are denoted as UoP_1 to 10;

- **Study IIb**: all participants were located at Site B (UNIMAS), twelve undergrade students (6 males and 6 females), aged between 20 and 25 years old were participated in the study. These participants are denoted as Unimas_1 to 12.

All these participants have normal hearing and vision. They had very little experience (if any) of using the software and have not been directly involved in the evaluation of other multimedia systems or related area. However, they are all familiar with the multimedia quality over the Internet.

8.4.3 Task Performance

To maximise task motivation and to ensure that the participant was fully at ease while communicating, a videoconferencing facilitator was introduced at the other end. It was anticipated that these participants should interact in pairs (instead of employing a facilitator), it is however, could not be done due to the language barrier. The communications were based on one-to-one person group, as conducted in the previous studies. Upon launching the videoconference across the two countries, the participant and the facilitator were then involved in intensive interactions for approximately five mins. At the end of the conversation, the participant was asked to mark the scores on the answer sheets (see appendix A).

8.4.4 Rating Method

Similar rating protocols as introduced in the previous chapter and Study 1 were repeated in this study. Once the test is completed, the results are statistically analysed, and graphically presented as in the following section.

8.5 Results

8.5.1 MOS and Lip Synchronisation Results

Table 8.8 shows the subjective MOS results and lip synchronisation effects, reported by the individual participant involved in Study IIa.

Table 8.8: Subject Scores – Study IIa

Subject	Audio MOS	Video MOS	Audiovideo MOS	4-Category Rating	Level of Annoyance	Conversation Difficulty
1	4	4	4	iv. No strange effect	4	No
2	4	4	4	ii. Audio behind video	3	No
3	4	4	4	iii. Not sure	4	No
4	5	3	3	i. Audio ahead	4	No
5	4	4	4	iii. Not sure	3	No
6	5	4	4	i. Audio ahead	3	No
7	4	3	2	iii. Not sure	3	No
8	4	4	4	iv. No strange effect	4	No
9	4	4	4	iii. Not sure	3	No
10	4	3	4	iii. Not sure	4	No
Average	**4.2**	**3.7**	**3.7**		**3.5**	
Stdev	*(0.42)*	*(0.48)*	*(0.71)*		*(0.53)*	

The MOS result for audio, video, and audiovideo components obtained from Study II are generally higher than that of Study I. For the perceived quality of audio, most participants scored GOOD (4) and two of them scored Excellent (5) point rating. For the perceived quality of video and audiovideo overall, the majority of the candidates rated FAIR (3) and GOOD (4) quality scores that give rise to the average of 3.7 MOS. It is interesting to note that these ratings are higher than that obtained from the classroom-based study (even in the ideal network environment). See Figure **5.1** and Figure **5.2**.

Out of the ten video conferencing participants, two believed that there is no strange effect (i.e. no lip synchronisation error), five were not sure which media is a head/lag of the other, two claimed that audio is played a head of video, and one claimed otherwise (audio behind video). From this observation, even at fairly good audiovisual quality (3.7 MOS), it is still hard to predict the lip synchronisation within the desktop videoconferencing environment.

Table **8.9** shows the subjective MOS results and lip synchronisation effects, reported by the individual participant involved in Study IIb. These results are taken on two different days, i.e. on the 29th January, 2003 – Subject Unimas_1 to 6, and 30th January, 2003. It can be seen that the MOS rating (in average) obtained between these two days are significantly different. For example, the perceived multimedia score taken on the 29th January are generally higher, i.e 4.17 (GOOD) - for Audio, 3 (FAIR) - for video, and 3.5 (FAIR/GOOD) - for audiovisual overall. For the lip sync error (level of annoyance) effect, the average of 4-rating scale was given (note: 5 – less

annoying). Two out of 6 participant claimed to have conversation difficulty over the link. Although, the perceived quality are considerably high, it is observed that the participants could not perceived the lip synchronisation between the audio and video media (refer to column 4-Category Rating in Table 8.9).

The perceived multimedia quality obtained from Unimas_7 to 12, i.e. taken on the 30th January 2003, are between POOR and FAIR quality. For example, the average MOS for audio, video, and audiovisual overall are 2.83, 2.5, and 2.83 respectively. For the level of annoyance due to lip sync error, the average of 3-rating score was given. Half of the total number of the participants claimed to experience the conversation difficulty over the communication link.

Table 8.9: Subject Scores – Study IIb

Subject Unimas	Audio MOS	Video MOS	Audiovideo MOS	4-Category Rating	Level of Annoyance	Conversation Difficulty
1	4	3	4	a	3	Yes
2	5	4	4	d	5	No
3	4	4	3	c	5	No
4	4	3	4	b	3	Yes
5	4	2	3	a	4	No
6	4	2	3	d	4	No
7	3	4	4	c	3	No
8	2	2	2	b	3	Yes
9	2	1	2	c	3	Yes
10	3	3	3	c	3	No
11	3	3	3	a	3	No
12	4	2	3	a	3	Yes
AVE Unimas_1-6	4.17	3	3.5		4	
Stdev	*(0.41)*	*(0.95)*	*(0.55)*		*(0.89)*	
AVE Unimas_7-12	2.83	2.5	2.83		3	
Stdev	*(0.75)*	*(1.05)*	*(0.75)*		*(0)*	
Total Average	**2.76**	**2.21**	**2.55**		**2.82**	
Total Stdev	***(1.26)***	***(0.94)***	***(1.04)***		***(1.21)***	

8.5.2 Traffics Vs MOS

Table **8.10** shows the average traffic parameters, i.e. jitter (jit), packet loss (pck), and round trip time (rtt) and the respective multimedia MOS obtained from Site A in Study IIa. It can be seen from the table that for audio streams, the jitter and packet loss values are fairly small, i.e. within the ranges of 4.84 - 6.5ms (jitter) and 0.3% - 0.46% (pck), that yielded to an average of 4.2 (GOOD) MOS. Whilst, for the video channel, the jitters values are within 52.75 - 71.44ms and the packet loss values are between 0.46% - 0.61%, and the MOS is 3.7 (FAIR/GOOD). The overall audiovisual MOS is 3.7 that is considered as a fairly good and acceptable quality.

Table 8.10: Study IIa – Traffic Vs MOS

Subject	audio			Audio	video			Video	Audio-visual
UoP	jit (ms)	pck (%)	rtt (ms)	MOS	jit (ms)	pck (%)	rtt (ms)	MOS	MOS
1	4.84 *(1.26)*	0.48 *(0.01)*	515.37 *(58.28)*	4	52.72 *(13.70)*	0.52 *(0.02)*	503.33 *(55.49)*	4	4
2	5.48 *(1.53)*	0.50 *(0.01)*	452.61 *(69.85)*	4	67.17 *(10.57)*	0.47 *(0.01)*	466 *(69.95)*	4	4
3	5.1 *(1.47)*	0.42 *(0.09)*	431.29 *(51.01)*	4	69.03 *(16.7927)*	0.65 *(0.04)*	435 *(68.51)*	4	4
4	5.92 *(3.22)*	0.46 *(0.01)*	503.80 *(64.50)*	5	56.7 *(10.19)*	0.46 *(0.01)*	525.38 *(142.315)*	3	3
5	6.5 *(2.38)*	0.39 *(0.08)*	519.52 *(50.72)*	4	71.44 *(12.10)*	0.58 *(0.10)*	514.99 *(64.68)*	4	4
6	5.29 *(1.67)*	0.44 *(0.03)*	491.80 *(67.59)*	5	69.32 *(15.19)*	0.56 *(0.07)*	483.49 *(69.43)*	4	4
7	5.69 *(2.00)*	0.54 *(0.01)*	543.10 *(66.67)*	4	65.92 *(13.71)*	0.61 *(0.04)*	535.13 *(58.26)*	3	2
8	5.45 *(1.06)*	0.53 *(0.01)*	552.40 *(38.79)*	4	63.25 *(31.66*	0.5 *(0.04)*	560.97 *(43.11)*	4	4
9	5.21 (1.42)	0.54 (0.01)	435.67 (61.72)	4	67.99 (11.47)	0.52 (0.02)	436.51 *(64.01)*	4	4
10	5.42 *(1.51)*	0.3 (0.03)	423.94 (42.81)	4	64.47 *(12.71)*	0.51 *(0.03)*	415.95 *(69.96)*	3	4
Average *Stdev*	**5.44** ***(0.50)***	**0.46** ***(0.08)***	**486.95** ***(57.19)***	**4.2** ***(0.422)***	**64.8** ***(5.90)***	**0.54** ***(0.06)***	**529** ***(70.57)***	**3.7** ***(0.48)***	**3.7** ***(0.67)***

Table **8.11** shows the average traffic parameters, i.e. jitter (jit), packet loss (pck), and round trip time (rtt) and the respective multimedia MOS obtained from Site B in Study IIb. It can be seen from the table that for audio streams, the packet loss experienced by the participants denoted as Unimas_1 - 6 that is taken on the 29th of January 2003 is very small (i.e. 0.29% in average), yielded to the average MOS of 4.17 i.e GOOD quality. Similarly, the video packet loss is around 2.85% (average) and the perceived quality is FAIR or 3 MOS. The audio jitter is 30.44ms in average, while, for video, it is 166.42ms in average.

By referring to the subjects denoted as Unimas_7 - 12, the figures for packet loss is higher i.e. 13.31% in average, that yielded to the average MOS of 2.83 (POOR/FAIR). The audio jitter is around 30.06ms in average. Whilst, for the video channel, the jitters values are within 208.59 - 233.08ms and the packet loss values are between 11.31% - 13.28%, yielding to the MOS of 2.5 i.e. between POOR and FAIR quality. The overall audiovisual MOS is 2.83 that is considered as POOR/FAIR quality. As previously mentioned, for Unimas_7 - 12, the videoconference is conducted on the 30th January 2003.

Table 8.11: Study IIb – Traffic Vs MOS

Subject	audio			Audio	video			Video	Audio-visual
Unimas	jit (ms)	pck (%)	rtt (ms)	MOS	jit (ms)	pck (%)	rtt (ms)	MOS	MOS
1	31.18 *(2.05)*	0.28 *(0.03)*	3.62 *(1.99)*	4	195.84 *(56.51)*	2.45 *(0.20)*	3.51 *(2.39)*	3	4
2	29.92 *(0.60)*	0.33 *(0.01)*	3.42 *(1.88)*	5	106.92 *(70.72)*	3.22 *(0.05)*	3.64 *(1.97)*	4	4
3	30.11 *(0.82)*	0.3 *(0.01)*	3.73 *(1.77)*	4	109.32 *(75.95)*	3.02 *(0.07)*	3.91 *(2.46)*	4	3
4	30.43 *(1.24)*	0.27 *(0.01)*	3.73 *(1.87)*	4	223.36 *(56.97)*	2.75 *(0.02)*	3.86 *(2.53)*	3	4
5	30.53 *(1.84)*	0.28 *(0.01)*	3.87 *(1.64)*	4	201.28 *(98.82)*	2.81 *(0.05)*	3.51 *(1.92)*	2	3
6	30.45 *(1.52)*	0.28 *(0.01)*	3.81 *(2.33)*	4	161.82 *(89.53)*	2.87 *(0.08)*	4.7 *(6.86)*	2	3
7	30.33 *(2.20)*	13.05 *(0.11)*	4.46 *(3.17)*	3	229.1 *(35.04)*	11.49 *(0.08)*	4.12 *(2.95)*	4	4
8	29.85 *(2.43)*	13.02 *(0.08)*	4.42 *(3.02)*	2	233.08 *(30.66)*	11.96 *(0.11)*	4.18 *(2.91)*	2	2
9	29.99 *(1.69)*	13.8 *(0.24)*	3.68 *(2.22)*	2	231.12 *(33.69)*	13.28 *(0.50)*	3.79 *(2.01)*	1	2
10	30.16 *(2.09)*	13.22 *(0.21)*	4.05 *(1.82)*	3	239.6 *(35.15)*	11.31 *(0.21)*	3.97 *(2.22)*	3	3
11	30.22 *(1.97)*	13.46 *(0.01)*	3.18 *(2.04)*	3	208.59 *(29.49)*	11.64 *(0.03)*	5.05 *(2.88)*	3	3
12	29.81 *(2.54)*	13.04 *(0.03)*	3.62 *(1.78)*	4	231.41 *(42.19)*	11.78 *(0.08)*	3.93 *(2.37)*	2	3
Average 29-Jan	**30.44 *(0.43)***	**0.29 *(0.02)***	**3.7 *(0.16)***	**4.17 *(0.41)***	**166.42 *(47.28)***	**2.85 *(0.26)***	**3.86 *(0.44)***	**3 *(0.89)***	**3.5 *(0.55)***
Average 30-Jan	**30.06 *(0.21)***	**13.31 *(0.32)***	**3.9 *(0.5)***	**2.83 *(0.753)***	**228.82 *(10.54)***	**11.91 *(0.71)***	**4.17 *(0.45)***	**2.5 *(1.05)***	**2.83 *(0.75)***

Table **8.12** and **8.13** show the audio traffic and video traffic, respectively, taken from Study I (Average) and Study II (Average) and Network (N/w) configuration A, B and C - from the previous study in Chapter 7 (see Section 7.3.4) and Figure **8.16** presents the relative MOS, given by the participants.

It can be seen from the graph that the perceived quality of audio, video, and audiovideo taken from the Study II (field trial) are generally higher compared to that taken from the classroom (lab)-based study. It is expected that the perceived quality of Study II is much higher than that of Study I (field trial) since the network configuration for Study II (100/1000MB backbone) are much greater than Study I (56KB modem).

Table 8.12: Audio Traffic for Figure 8.16

Study Type	Jitter (ms)	Network Traffic Packet Loss (%)	RTT (ms)
Classroom based N/w A	1	0.5	100
Classroom based N/w B	5	1.5	100
Classroom based N/w C	10	3	200
Field trial Study I Average	84.3	2.54	3.89
Field trial Study IIa Average	5.44	0.46	500

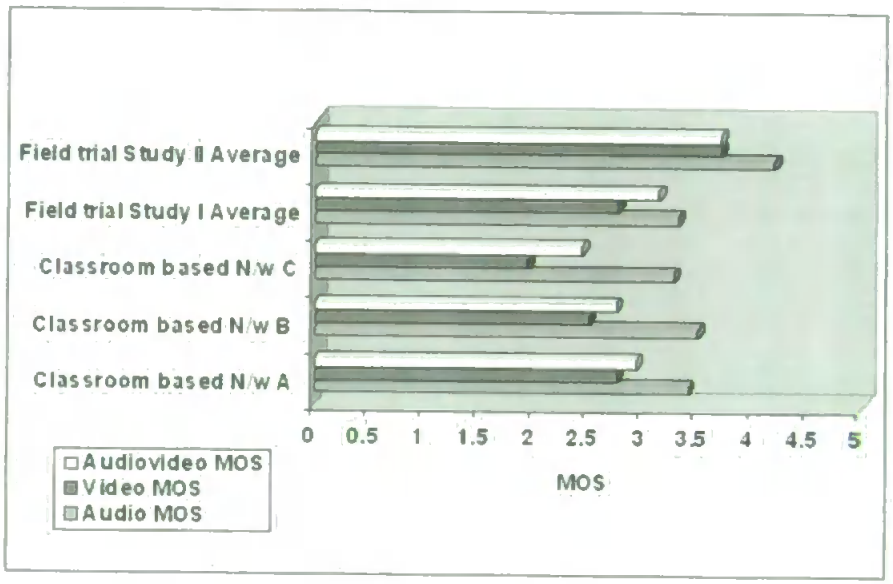


Figure 8.16: Average Study I and II Vs network A, B, and C

Table 8.13: Video Traffic for Figure 8.16

Study Type	Jitter (ms)	Network Traffic Packet Loss (%)	RTT (ms)
Classroom based N/w A	1	0.5	100
Classroom based N/w B	5	1.5	100
Classroom based N/w C	10	3	200
Field trial Study I Average	610.59	5.11	3.99
Field trial Study IIa Average	64.8	0.54	500

Table **8.14** presents the traffics across different studies, i.e. the lab-based and field trial, conducted in the thesis. The respective audio, video, and audiovideo subjective quality for each study is shown in Figure **8.17**, **8.18**, and **8.19**. From these figures, it can be seen that the perceived quality of IP media obtained from Study II is much higher than that taken from lab-based study, even under the ideal network condition, as shown in the Ideal Network - Int. (G723.1), Ideal Network - Pas. (G723.1), Ideal Network - Int. (PCM), and Ideal Network - Pas. (PCM) slots within the graphs.

By referring to Figures **8.18** and **8.19**, it can be seen that the subjective quality of video and audiovideo overall obtained from lab-based studies, i.e. the slots denoted by Lab-based - Int. (G723.1), Lab-based - Pas. (PCM), and Lab-based - Int. (PCM) are between FAIR/POOR rating (2.5 MOS) and POOR (2 MOS) that are lower than the scored taken from Study I, which had a greater network congestions (see Table **8.14**). Hence, again the rating score for the perceived multimedia quality in classroom-based study has proven to be lower than that in the field study.

Table 8.14: Traffics Across Different Studies - for Figure 8.17, 8.18, and 8.19

Study Type	Traffic Analysis
Ideal Network - Pas. (PCM)	No packet loss (loss), jitter, delay
Ideal Network - Int. (PCM)	No packet loss (loss), jitter, delay
Ideal Network - Pas. (G723.1)	No packet loss (loss), jitter, delay
Ideal Network - Int. (G723.1)	No packet loss (loss), jitter, delay
Lab-based - Int. (G723.1)	3% loss, 10ms jitter, 200ms delay
Lab-based - Pas. (PCM)	5% loss, 20ms jitter, 400ms delay
Lab-based - Int. (PCM)	5% loss, 20ms jitter, 400ms delay
Study I - Int. (G723.1)	audio: 2.54% loss, 84.3ms jitter, 3.89ms RTT video: 5.11% loss, 590.38ms jitter, 4.64ms RTT
Study IIa - Int. (G723.1)	audio: 0.46% loss, 5.44ms jitter, 500ms (RTT) video: 0.54% loss, 64.8ms jitter, 500ms (RTT)

Note:
Pas. = Passive
Int. = Interactive
delay = end-to-end delay

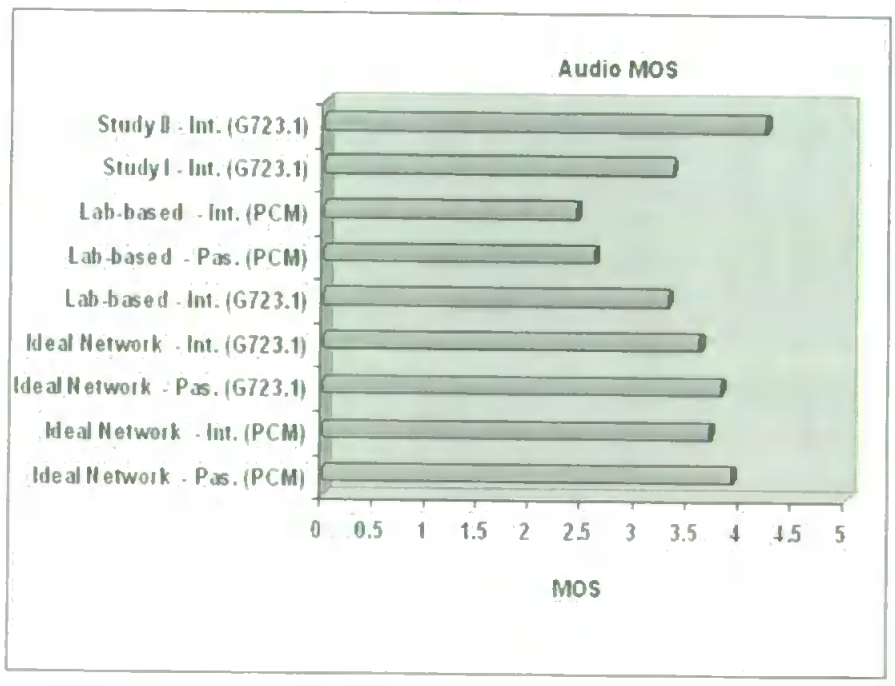


Figure 8.17: MOS of Audio Across Different Traffics

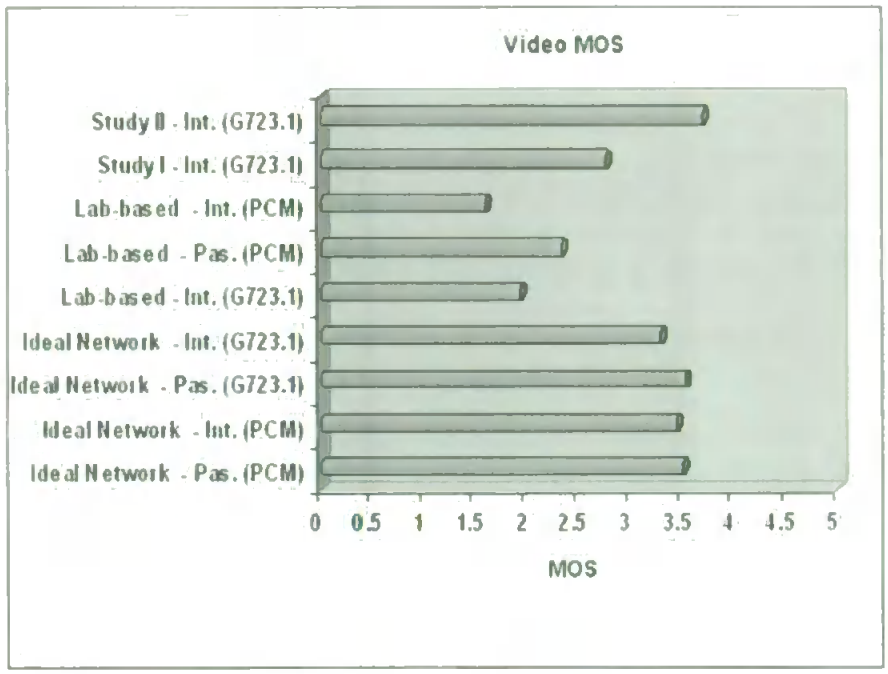


Figure 8.18: MOS of Video Across Different Traffics

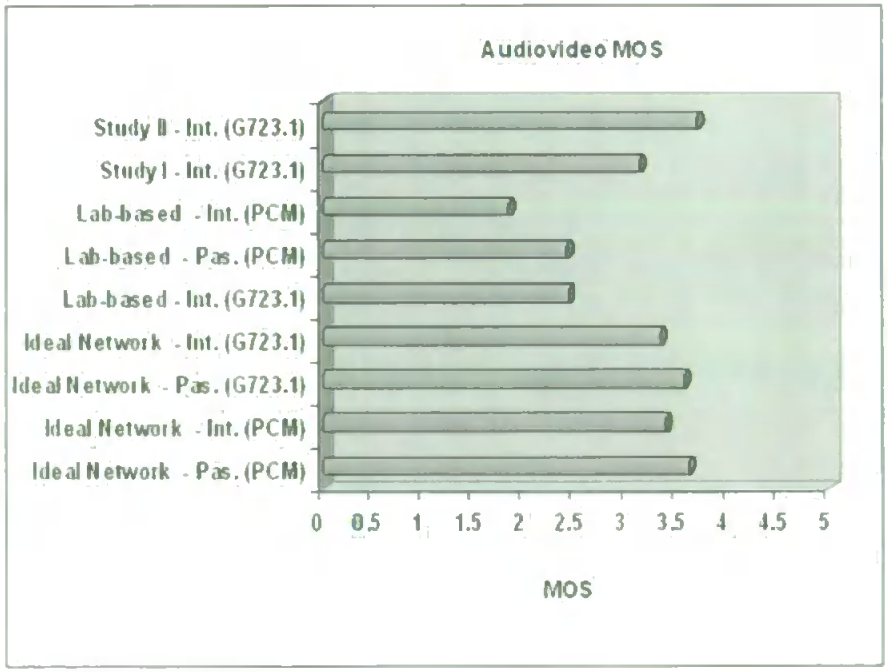


Figure 8.19: MOS of Audiovideo Across Different Traffics

Figure **8.20** show various examples of image quality of a participant (in Site B) captured at Site A using Microsoft NetMeeting. The top-middle image is suffered from video artefacts (color splotches and distortion) and the bottom-left corner image is appeared to be jagged or fuzzy. It can be deduced from the images that the video quality was generally good and acceptable.

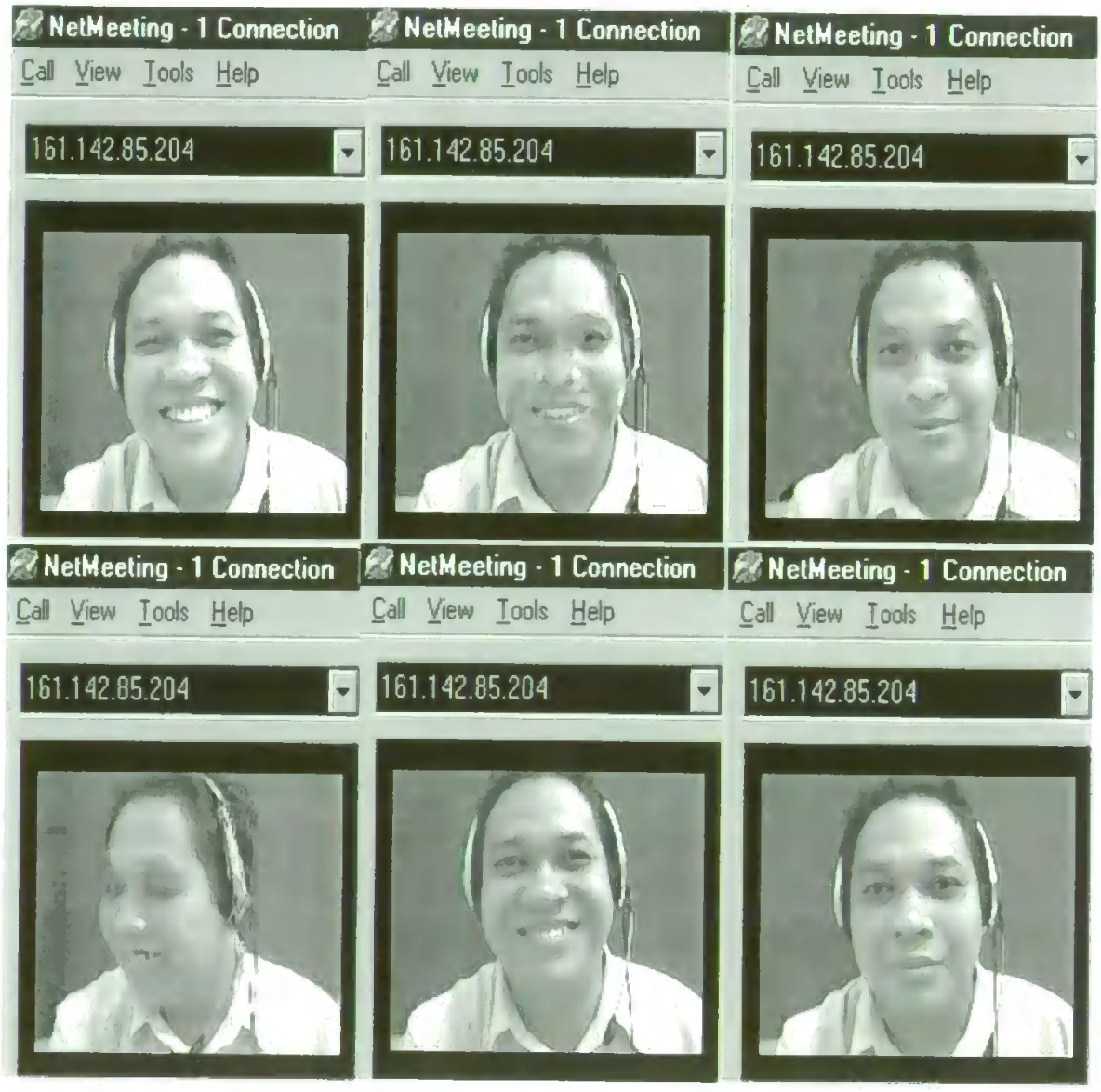


Figure 8.20: Image Quality of a Participant (from UNIMAS) Captured at UoP using Microsoft NetMeeting

8.6 Discussion

The traffic collected on the two different day varies immensely. Therefore it is suggested that for future reference, the work in the similar area should be conducted in a specific time and date to minimise the high variation in the traffic across the links.

It is important to note that the the network over which the communication is taking place is in the state of constant flux as shown in Figure **8.4** to Figure **8.13** in this chapter. This means that the quality varies rapidly with time. Therefore, ideally, the perceived multimedia quality should be rated several times (or continuously) during the period of launching the desktop videoconferencing systems.

Initially, it was anticipated that the participants were required to rate the score several times (i.e. every 1 min) during the conversation, so that the average of the results can be obtained. However, it was observed that from several attempts in the earlier trials that the suggested rating method could interfere with the task being performed. This inarguably would affect the end user's subjective opinion on the perceived media quality. Therefore, the participant was asked to rate the overall score of the perceived multimedia quality over the duration of the conversation, at the end of the task.

The number of packet loss captured from the field study was generally small, i.e. in the range of 0.8% to 4.88%. From the classroom based study in the previous chapters, below 8% packet loss, the audio quality was unaffected and considered as

good quality (3.5 – 4 MOS). It can be concluded that the main deteriorating factor of the audio quality in the field study was due to the high jitter values (i.e. 60 – 100 ms) between the links which produced 'glitches' sounds. However, based on the end user's subjective quality scores, it can be clarified that they are more tolerant to audio jitter compared to that from the previous laboratory-based studies.

For the perceived video quality, like audio, the majority of the candidates in the field trial scored 3 MOS (FAIR) even though the data analysis showed high video packet loss, i.e. between 2.7% to 8.77%. From the classroom-based results, at 4% packet loss, the viewer's perceptual rating is between 2 to 2.5 MOS, i.e between POOR and FAIR scales (see Figure **6.3**, **6.1**, and **5.4**). Therefore, it can be concluded that the end user's perception of IP media QoS in the field trial are generally higher than that of the classroom-based studies. This may be due to several reasons:

- good audio quality – the majority of the field trial participants believed that the perceived audio quality is good or fair. The traffic analysis also confirms the statement in that the packet loss of audio is generally small (below 5%). From the studies in the previous chapters it has been justified that the perception of one media can interacts and influences the perception of the other. Therefore, in effect, the multimedia components subjective rating scores as perceived in the field trial fall between FAIR (3 MOS) and GOOD (4) quality, in general;
- subject background – as suggested in Table **8.1**, the majority of the participants has very little knowledge about Internet and its application. Most

of them have never heard of desktop videoconferencing technology and appeared amazed and very satisfied with the technology (as suggested by their remarks in Appendix B: Study I), since they have never experienced any other form of multimedia quality apart from the TV broadcasting systems. As the result, this may have influenced their perceptual IP media quality in that it is generally higher than that obtained from the classroom-based studies. As mentioned before, most candidates participated in the classroom-based studies were the from the undergraduate background and were already familiar with the Internet multimedia quality (eg. steamed audio and/or video);

- geographical distance – the fact that the communications links were performed between the two countries that are geographically separated over a great distance could have had some psychological influences on the end users perception of the media quality in that they were more tolerant to the channels degradations.

It is observed from the 4-category rating results that even at fairly good audiovisual quality (3.7 MOS) as in Study II, it is still hard to predict the lip synchronisation within the desktop videoconferencing environment.

One of the shortcomings of the field study is that the number of the participants were not as high as in the previous studies in the previous chapters, due to the time constraints. The seven hour time difference between the two countries had restricted the time allocated to set up the videoconference, and hence, minimised the number

of potential attendants. The previous chapters indicated that the number of subjects was generally around 20 while in the filed study there were only 12 participants. This unfortunately prevents a firm conclusion being drawn from the results due to the imbalance number of subjects across the studies. As such, it was decided the result should be analysed individually rather than considered the average result among the subjects.

The fact that any plan, problem solving (technical), user guidance, or any form of verbal instruction can only be done 'virtually' during the study, these in effect had prevented the task to be done in a more contrived way. These are considered as another major problems encountered while conducting the field study, besides the seven hour time gap between the two countries.

8.7 Summary

The findings obtained from the field trial presented in this chapter were of primary importance to the lab-based research since they enabled the identification of the involving factors and problem aspects in the real environment which could then be explored and verified in laboratory settings. For example, the subjective rating results obtained from the field trial are likely to be higher than that of the classroom based and task managing factors were also considered as one of the major problems. From the traffic analysis it can be concluded that the network congestion varied immensely depending on the time and day, and therefore, it is suggested that the

work in the similar area should be conducted in the specific time and date of the year to minimise the variation of the network factors.

This chapter has thoroughly described the tasks, observations, and conclusion of findings derived from the field trials. Thus far, the data-gathering approach that is advocated in this thesis involves both the field and laboratory study.

The next chapter presents a new adaptive architecture method that is arising from all the findings described in this thesis. The new method is intended to enable the performance of IP based DVC of a particular session to be predicted for a given network condition and task, and hence, can be used as a means of delivering the best possible audio and video quality over IP-based videoconferencing.

Chapter 9

An Adaptive Architecture for IP Videoconferencing Systems

9.1 Introduction

From the earlier chapters within the thesis, it has been learnt that when real-time audio and video streams are sent through a network, they experience different delays, jitter, and packet loss due to the unpredictable nature of IP network. Real-time multimedia flows (such as videoconference) are inevitably highly sensitive to traffic fluctuation and therefore, a means of transmission control mechanism would be advantageous to achieve the best possible delivery of audio and video quality in accordance to certain network situation and task performance.

Furthermore, the increasing use of multimedia videoconferencing technology has led to the difficulties of employing a best-effort services for communication, and this has necessitated higher bandwidth demands or some kind of bandwidth reservation

to be employed, so that the quality can be guaranteed for a particular conference. Since bandwidth is a valuable resource, it is extremely important to make the media transmission properties or settings adaptable to the available bandwidth to maintain or upgrade the perceived quality.

This chapter presents a new adaptive architecture method that is intended to enable the performance of IP based DVC of a particular session to be predicted for a given network condition. By inferring the multimedia quality scores (MOS) and the congestion control information from a network monitoring system, the proposed technique (that can be implemented within the videoconferencing architecture) can automatically adapt to the network change by negotiating for the best configuration within the adaptive architecture tools to give the best quality and improved bandwidth utilization. For example, changing the codec, packetisation interval, bit rate, frame rate, error concealment technique, play-out buffer size, etc. are possible adaptations that could be implemented within such an adaptive approach. The schematic diagram of the model of the proposed adaptive architecture is illustrated in Figure 9.1.

The proposed technique can be used to minimize the effect of delay variation and packet loss when highly congested traffic is reported from the monitoring systems. The adaptive system can also be used to estimate in realtime the received audio quality, at both ends so that each user can know how the other user perceives what he/she sends and receives.

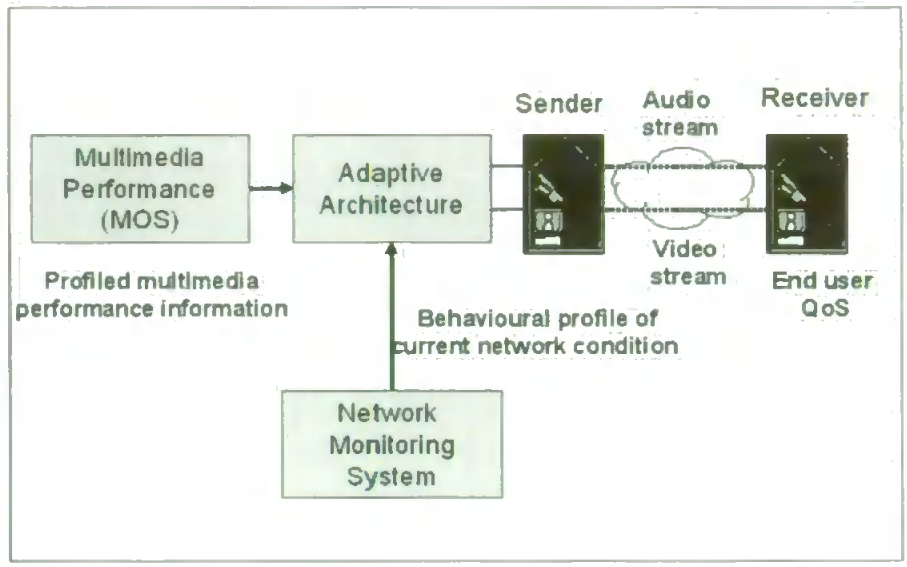


Figure 9.1: Adaptive Architecture Model

As shown in Figure **9.1**, the model of the adaptive architecture is based up on three main subsystems, i.e. Profiled Multimedia Performances unit, Network Monitoring System, and Adaptive Architecture tool. The role and function of each of these subsystems is now outlined in the sections that follow.

9.2 Profiled Multimedia Performances

The studies conducted in the previous chapters successfully investigated the quality threshold for audio and video required for specific task performance. The findings obtained from the studies can be used as a basis of a control mechanism to predict the transmission quality of audio and video in IP multimedia applications.

On the basis of MOS results and observations from a number of field trials, experiments and other reported literatures, a set of criteria is established that can be used

to determine user's perception of multimedia quality for different application scenarios (i.e. passive and interactive communications) and network conditions. The collection of data has enabled a taxonomy of quality boundaries for audio and video for a specific task to be defined.

Figure **9.2** shows a taxonomy of multimedia performance analysis employed in the thesis. The analysis yielded into two main categories, i.e. multimedia MOS results and multimedia behavioral profile, taken from a number observations and comments reported by the subjects.

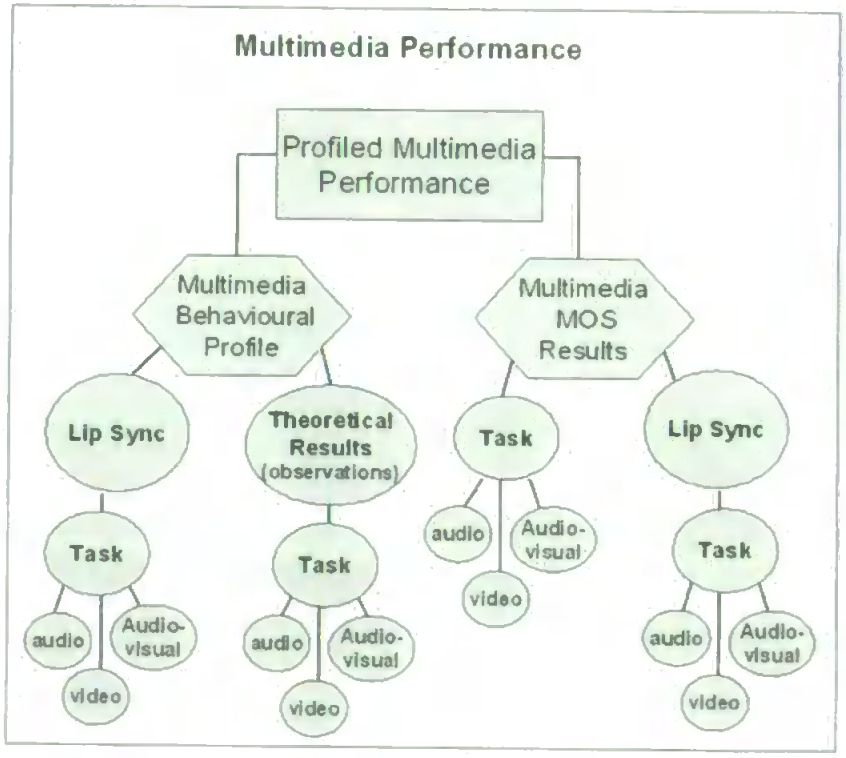


Figure 9.2: Profiled Multimedia Performance

9.2.1 Multimedia MOS Results

Table 9.1 shows the summary of the quality boundaries derived from the experiments in Chapter 5, where both audio and video media were subjected to the same amount of packet loss and jitter. The Pentium III 933MHz (128MB CDRAM) systems, PCM (μ-law) and G723.1, and QCIF-176x144 picture format were used for the passive and interactive tests.

From a series of results, observations and experiences while conducting the tests, it is suggested that the MOS of 2.5 point-rating is the minimum acceptable quality without any serious deterioration. The MOS of 2 to 2.4 represent that the degraded quality is still usable, but it required some effort to understand the perceived media. Below 2 MOS rating, the perceived quality is considered as bad or unacceptable.

By referring to Table 9.1, it can be seen that the interactive test gives a lower quality threshold than that of the passive test. For example, a packet loss of 10% (G723.1) would achieve an acceptable audio quality in the passive test, whereas, in the interactive test, the 8% (G723.1) packet loss would give the same level of quality. Similarly, in the passive test, the perceivable audio and audiovisual quality are acceptable at 40ms audio and video jitter but in the interactive test, to achieve the same quality level, the jitter value of 30ms is reached. The video scores are generally low, with a network threshold of 4% loss and 20ms jitter to achieve an acceptable quality (2.5 MOS and above). It is reported that the assessment video quality is hard to predict since the perceived quality is constantly fluctuated with time. Also, it is also

hard to draw the line as to where the quality is not usable anymore.

Throughout the experiments, the PCM CODEC produced better audio (and even video) quality, which is expected since it has a higher bit rate (64Kbit/s). By considering these results in the adaptive architecture implementation, the best possible real-time multimedia quality can be guaranteed in that the multimedia CODEC could be selected or re-engineered (scalable CODEC) to adapt to the bandwidth change, user's expectation, and task scenario.

From the study, it was observed that the delay of 300-600ms (without the presence of packet loss and jitter) have no significant effect on the perceived multimedia quality across the studies, in that the quality scores were maintained at around FAIR level (3-3.5 MOS). However, at 800ms delay, the perceived audio MOS in the Interactive test drops by approx. 0.3 MOS than that of 600ms delay score. Hence, it is suggested that delay parameter is not a critical issue in designing the proposed adaptive architecture.

Network traffic type	IP Media	Acceptable Quality Usable MOS	Acceptable Quality Usable MOS with effort	Unacceptable Quality Unusable MOS	Maximum threshold			
					Passive		Interactive	
					PCM	G723.1	PCM	G723.1
packet loss	video	2.5	undefined		4%	4%	4%	4%
	audio	2.5		< 2	10%	10%	<10%	<8%
			2 to 2.5		<20%	15%	<20%	15%
	Audiovideo	2.5		<2	10%	<8%	<8%	<8%
			2 to 2.5		10%	10%	10%	10%
Jitter	video	2.5	undefined		10ms	10ms	<10ms	<10ms
	Audio				40ms	40ms	40ms	40ms
		2.5		<2	30ms	20ms	<40ms	20ms
			2 to 2.5		40ms	40ms	40ms	30ms
	Audiovideo	2.5		<2	<20ms	<20ms	<20ms	<20ms
			2 to 2.5		40ms	40ms	40ms	30ms

Table 9.1: Summary of Results of Study:- Investigating the Effects of Network Constraints

Figures **9.3**, **9.4**, and **9.5** show the summary of the multimedia performance across the studies within the thesis, obtained from the classroom-based experiments, where each of the audio and video media is subjected to a specific impairments. The perceived multimedia quality taken from the field trial is generally higher than that of the classroom-based study, as explained in Chapter 8. Therefore, only the summary of results obtained from the classroom-based experiment is presented here, as a baseline comparison. Furthermore, the field trial results are derived from a comparatively small number of subjects (compared to that of the classroom-based experiments) to yield a more concrete conclusion.

These figures (Figure **9.3**, **9.4**, and **9.5**), show the results obtained from the interactive test. From a number of results and observation, the perceived quality obtained from the passive test has proven to be greater (i.e. by 0.3-0.5 MOS, in average) than that of the interactive test.

In general, the audio component has shown a greater quality threshold, in that a 10% audio packet loss and 40ms audio jitter would score a FAIR quality rating. While for video, 3% packet loss is the permissable values to achieve fairly good quality. At 4% packet loss, the perceived quality is still acceptable although the image degradation is discernable. The effects of jitter on the perceived video quality is difficult to predict. This is because the video quality is constantly changing with time. The work to access the video quality needs more effort and time, since the perceived quality is very sensitive to a number of variables, such as, the subject's movements, room background, and the type of session.

For audiovideo overall, to produce FAIR quality level, the audio packet loss and audio jitter values should be around 5% and 40ms, respectively. Whilst, at 20% audio packet loss and 120ms audio jitter, the perceived audiovisual overall is acceptable although some effort is required to understand the spoken words.

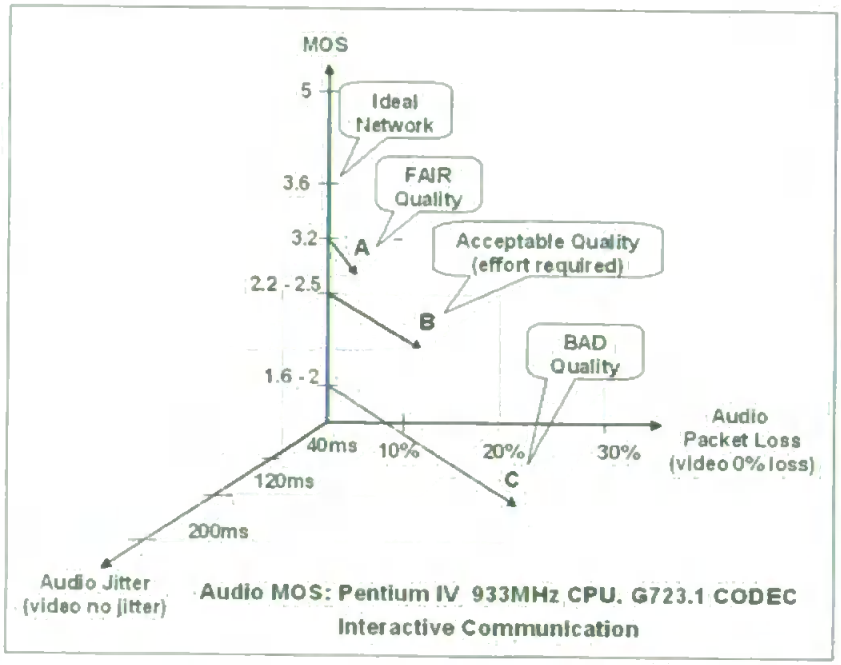


Figure 9.3: Audio MOS vs Audio Jitter vs Audio Packet Loss

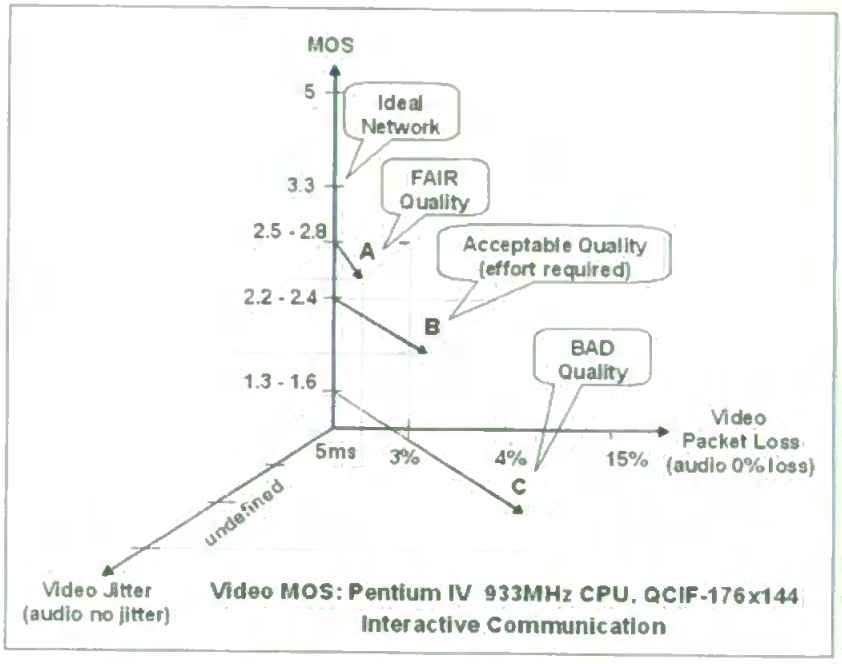


Figure 9.4: Video MOS vs Video Jitter vs Video Packet Loss

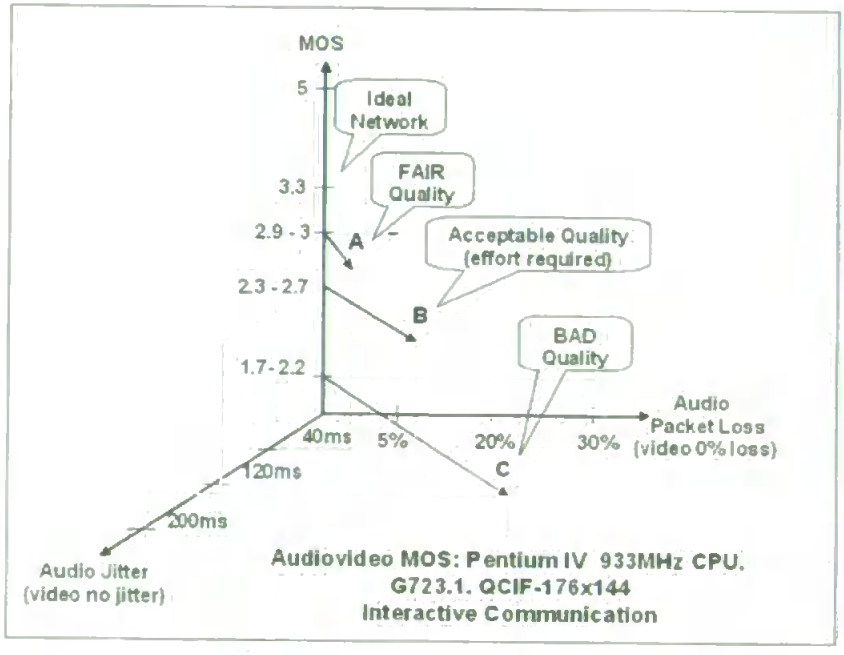


Figure 9.5: Audiovideo MOS vs Audio Jitter vs Audio Packet Loss

In summary, the MOS results yielded an appropriate audio/video quality ratio based on Quality of Service (QoS) requirements, and can be used by the system developer as a guideline to achieve a better bandwidth utilisation while giving the best possible delivery of audio and video quality in accordance to certain network situation and task performance. Hence, the findings have yielded an adaptive architecture which has been proposed to enable the MOS of a particular section to be predicted for a given network conditions.

9.2.2 Multimedia Behavioral Profile

The following summarize the series of observations made from the entire studies in the previous chapters:

From Chapter 5:

- the perceived multimedia are susceptible to packet loss and jitter, but are less susceptible to delay;
- in general, the multimedia rating is between FAIR (3) and GOOD (4), although GOOD is seldom given. None score EXCELLENT (5);
- the multimedia performance is affected by CODEC used and task performance.

From Chapter 6:

- there is a strong interaction dependency between audio and video media. Thus, it can be concluded that: (i) the QoS of one media can be predicted by the perception of the other; (ii) the Qos of one media can be improved by upgrading the quality of the other;
- the attention given to a particular aspect of performance is clearly content-dependent;
- the perceived multimedia dependency on the different talker language has been reported;
- due to the poor image quality, the subject relied mainly on the audio channel and made some allowance for the low performance of the video channel;
- despite the poor image quality, the subjects prefer to have both media as the video channel enhances interactivity;
- the task performance should be realistic and and highly representative of the real world situation.

From Chapter 7:

- the multimedia performance is unaffected by the loss of synchronization between audio and video media;

- the attention given to the lip sync assessment is clearly depended on the task performance;
- audio jitter of 40ms is the minimum permissible jitter value to achieve GOOD audio quality. Beyond 40ms audio jitter, a 'glitches' sound is perceived.

From Chapter 8:

- the multimedia performance rating obtained from the field trial are likely to be higher than that of the classroom based studies;
- from the traffic analysis it can be concluded that the network congestion varied immensely depending on the time and day.

These observations help in providing added information to successfully design a more effective adaptive architecture.

9.3 Network Monitoring System

Figure **9.6** shows a schematic diagram of the network monitoring system proposed for the predictive modelling architecture. The network monitoring system is based up on the traffic monitoring and traffic prediction that would provide the user with a better understanding of the current status of the network condition and TCP (Transmission Control Protocol) performance prediction.

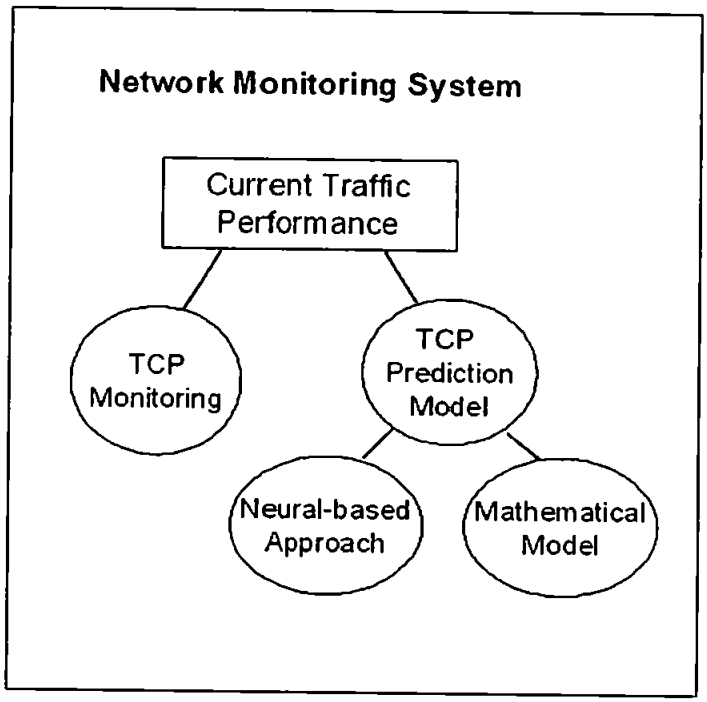


Figure 9.6: Network Monitoring System

The current traffic prediction techniques that employ either the traditional mathematical models or neural network-based approach can be found in [108], [109], [110], and [111].

The existing mathematical model, although theoretical is very successful in examining and describing network characteristic, it may perform poorly when dealing with real traffic. Hence, the current traffic prediction techniques are more inclined toward adopting the neural-based method.

The neural-based approaches are generally based upon two aspects: a classification one, i.e. the mapping between the parameters involving the TCP transfer and the quality; and a function approximation, i.e. the evaluation of the quality as a function of the quality-affecting parameters in an operational environment [107], [96], [112].

By inferring a behavioral profile of current network condition and the multimedia quality scores (MOS), a control algorithms (both transmitter-based and receiver-based) that could adapt to the network changes by negotiating for the best configuration within its architecture, could be applied to yield the best QoS requirement with respect to the multimedia session in progress.

9.4 Adaptive Architecture

Figure 9.7 shows the schematic diagram of an adaptive architecture that comprises a Transmission Control Unit and Receiver Adaptation Unit, the operation of which will be explained in this section.

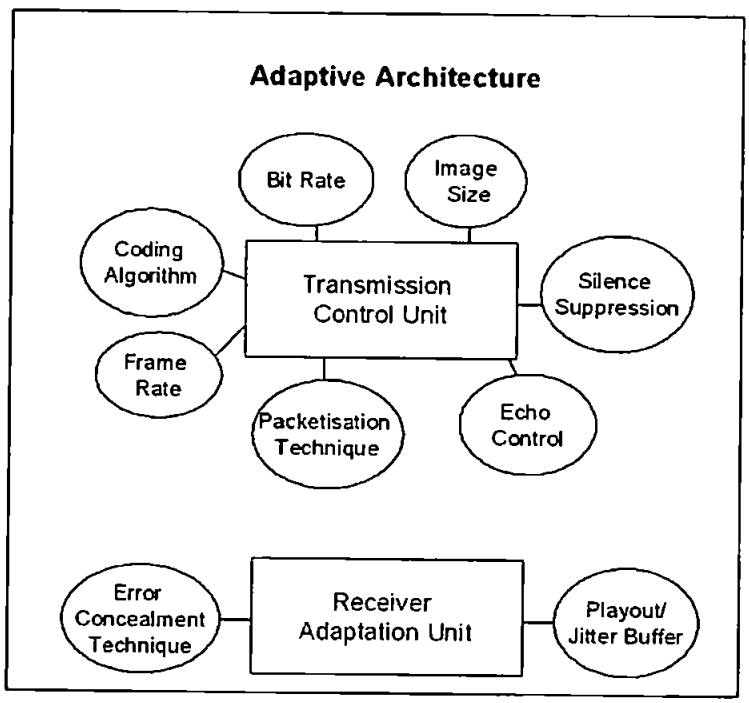


Figure 9.7: Adaptable Factors at Transmitter and Receiver Ends

9.4.1 Transmission Control

9.4.1.1 Coding Algorithms

In the current videoconferencing tools and settings, the encoder format can be changed at any time either before transmission or during a videoconferencing session to get the best possible perceived quality in a given network state. This is usually done manually by an expert user when he/she starts to suffer from a loss of quality due to heavy congestion in the network. The multimedia CODECs used in NetMeeting were presented in Section 4.4. Indeed, different types of encoding format will require different amounts of bandwidth. Similarly, a coding format that employs a higher sampling rate, more bits per sample, and greater coding complexity improves the multimedia quality but it also requires more network resources and increase processing delay.

By adapting state of the art coding algorithms to the available bandwidth and profiled multimedia quality assessment the optimum delivery of media quality could be achieved for a specific task performance.

9.4.1.2 Packetisation Interval

The packetisation interval is defined as the duration of the sampled speech of the access channel that has been collected, coded, and packetised [113]. The Timestamp field in the RTP packet header is used to determine when to play-out received

data. At the receiving end, frames are decoded in the order specified by the RTP timestamp.

To improve the perceived multimedia quality, the packetisation technique can be changed at any time, without requiring additional signalling. For example, to minimise the effect of variation in timestamp, and hence jitter, it is suggested that the stream should be packetised differently prior to transmission [114].

The packetisation technique is commonly used to maximize bandwidth utilization in the network. By packetising a single frame into one packet, a large overhead is introduced by the UDP header [50] and the achievable quality is heavily bandwidth dependent [70]. A drawback of the packetisation technique is that the loss of a single packet affects multiples frames that can be minimized by the interleaving process (a process which changes the order of a sequence of packet prior to transmission).

Chapter 5 concentrated on investigating the interaction effect of the two media of audio and video. The findings suggested that the quality varies according to the task undertaken and user expectation also varies accordingly [115]. This outcome is also in agreement with the findings in [59].

Also, it is observed that increases in task difficulties have the effect of decreasing the subjective video and audio quality. For example, in the passive test, where users are required to understand the read material, the overall scores for the combined audio and video quality were much lower than that in the interactive test. Therefore, it may be necessary to prioritise video over audio, or vice versa, depending on the

type of session and user expectation. This can be done during the packetisation.

9.4.1.3 Bit Rate and Video Image

In most videoconferencing environments, the size of image that is being transmitted can be change at any time during the interaction. The actual size of the image depends on the encoding format and signal type (sampling rate) used. For example, H261, as used in Microsoft NetMeeting, provides two frame sizes, i.e. CIF (352x288) and QCIF (172x144). Obviously, the larger the frame size the more bandwidth is required to transmit it. In a limited bandwidth allocation, a larger frame size will results in a very low frame rate (sometimes frozen) compared to that of smaller image size. The frame rates as perceived in Microsoft NetMeeting are in the range of 1 to 5 frames per second, depending on the traffics.

In low bandwidth applications, the frame rates are affected by the degree of movement there is in the image that is being transmitted. It is observed that increased subject's movements would result in the frame rate going down and the bit rate going up.

In the ReLaTe (Remote Language Teaching Over SuperJANET) [116] application, the user can adjust the maximum amount of bandwidth (kbit/s) that he/she is using and the maximum number of frame per-second (fps) that he/she is transmitting.

From the studies in Chapter 5, it is observed that, 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 than that of the audio [115]. This finding is also confirmed with the previous research result which stated that 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 [117]. From this aspect, the proposed adaptive system can adjust the frame rate accordingly, depending on the application and profiled QoS obtained from the studies within the thesis to achieve the best quality.

9.4.1.4 Echo Control Technique and Silence Suppression

In the echo suppression technique, the audio transmission of one end is suppressed whilst the other communicative party is active, in that an echo suppressor disables the channel in one direction or the other, depending on who is talking. This prevents problems with feedback or echo when using headphones.

Whereas, Echo Cancellation is a technique used with voice circuits to isolate and filter unwanted signal energy which accompanies analog transmissions. An echo canceller, however, does not turn off the voice channel (as echo suppressor), but electronically removes unwanted echo, while maintaining a full-duplex channel.

Silence suppression technique is used to ensure that periods of silence within a conversation (including background noise) will not be transmitted, hence reducing network traffic.

These audio enhancement method can be applied in applications where good audio quality is highly demanded and where the network is considered to be under heavy load.

9.4.2 Receiver Adaptation

9.4.2.1 Playout or Jitter Buffer Adaptation

Most existing real-time multimedia applications are designed with the objective of minimizing end-to-end delay, i.e. to keep the adaptive play-out buffer size as small as possible, which in effect, maximises packet drop due to buffer overflow or long delay arrival. Many methods are being employed to reduce delay and delay jitter [118], [119], [46], [47]. Resizing of the jitter buffer under varying network conditions is one of the techniques to eliminate variation in arrival delay.

From the observations made in this research, the thesis proposes a new methodological technique to minimise the issue of jitter. In applications where play-out delay is not critical as opposed to the QoS issue, such as in passive communication, the buffer size can be made larger to obtain better quality. Likewise, in applications where the effect of variation of delay (jitter) becomes critical, and is likely to occur

such as in intensive two-way communications, an adaptation of jitter buffer size can be used to minimize the deterioration effect.

9.4.2.2 Error Concealment Technique

The error concealment technique is widely used in real-time videoconferencing systems to minimise the effect of error at the receiving end. The traditional encoding techniques, such as Forward Error Correction (FEC) and Automatic Repeat Request (ARQ) that employ added redundancy codes and negative acknowledgement (NACK) mechanisms, respectively, to overcome packet loss problem are not practical when applied in the real-time interactive services such as videoconferencing applications due to the unacceptable long delay they imposed.

The existing receiver-based repair schemes have been investigated in [29], [30], [27], and [31] and have been proven to be beneficial to help minimise the packet loss effect. Error concealment techniques for speech or audio are silence substitution, noise substitution, packet repetition, and waveform repetition (see Subsection 2.5.1. Error concealment techniques for video are block replacement, linear interpolation, motion vector, and hybrid technique [120] and [121].

These error concealment methods can be implemented in the adaptive architecture, within the videoconferencing environment that can be selected, depending on best suited bandwidth and predicted QoS requirement.

9.5 Adaptive Architecture Versus Objective Methods

With respect to how the objective testing method is implemented, as described in Subsubsection 3.5.2.2 (Chapter 3), it can be concluded that the proposed adaptive architecture would works in the similar way.

Given that, all the criterions for IP multimedia performances have been defined (based on the subjective ratings) and acts as a lookup table, the adaptive architecture would just be looking up the best-fit match, based upon the information currently being received from the network monitor (see Figure **9.1**).

9.6 Summary

To summarize, the ultimate goal of this chapter is to utilise the model produced, as shown in Figure **9.1** as a means of delivering the best possible audio and video quality for a given task and network conditions.

A subsequent application developer could apply the defined taxonomy of audio and video quality to design a new multimedia over IP system with improved bandwidth usage and quality.

In addition, the taxonomy could also be used by service providers to infer objective or subjective QoS requirements for particular services, and hence, to charge accordingly.

As such, the proposed work is considered likely to provide a new and valuable reference for the realisation of future multimedia systems over IP networks.

The next chapter concludes all the findings obtained throughout the research programme, and outlines limitations, and makes recommendations for further investigations and developments in the similar area.

Chapter 10

Conclusion and Recommendations

This chapter concludes the thesis, by highlighting the achievements and contributions, while reflecting upon a number of the limitations and obstacles experienced during the research programme. Finally, the chapter concludes by discussing potential further developments, proposed for future research areas and directions.

10.1 Research Achievements and Contributions

10.1.1 Research Achievements

The study has achieved all the objectives, previously described in Section 1.1, through comprehensive studies undertaken during the entire research programme.

To summarise, the thesis has proposed a new approach to assessing the perceived quality of audio and video, delivered in networked videoconferencing. A number of achievements derived from the study can be summarised as follows:

- the fundamental factors that influence user's perception of multimedia quality (see Chapter 3), have been investigated and acknowledged. For example, it is observed that, the attention given to the media quality is severely affected by task performance and other surrounding factors, and hence, the same evaluation prototype was carried out throughout the study;

- the task to benchmark the current state of the art of the existing DVC systems was undertaken, both in controlled-based research and field-work study and the comparisons of the outcomes were made. Benefits and disadvantages of both approaches have been identified. The informal interactive communications between two friends has been considered the most appropriate task for the test as it is crucial that the task scenario models the type of interactions for which the application is being designed;

- in depth knowledge of the existing methodologies and techniques for assessing the end-user QoS have been established, and the subjective test has been confirmed as being the most reliable method for assessing multimedia quality, both in the passive and interactive tasks;

- the study observes a strong interaction effects between the perceived IP media (especially audio and video), and hence, it can be concluded that the quality of one media can be predicted by the perceived quality of the other. Subsequently, by upgrading the quality of one media could benefit the quality of the other and vice-versa;

- despite being considered as having a secondary role to audio in assessing IP media quality issues, perceived video quality was investigated and analysed comprehensively in this thesis. It is observed that video adds subjective value to a conference in that it ameliorates psychological effects and enhances interactivity. It is reported that the candidates greatly appreciate being able to see their conversational partner, even at a very poor image. The phenomenon indicates that there are clear advantages to including video streams in real-time videoconferencing applications, even at low bandwidth. Thus, it is evident that video is an important element in multimedia quality. This finding negates the statement which claimed that there is a little benefit of adding video, and audio quality is considered to be of high priority [68]. This thesis concludes that, the importance of video at the expense of audio cannot be underestimated, as it often is;

- thorough investigation has proven that the impact of lip sync error on the perceived multimedia quality is less than what it is claimed to be in low budget DVCs. The research concludes that, in applications involving low bandwidth systems, such as Microsoft NetMeeting which provides relatively poor video, the lip synchronisation error does not seem to matter. Thus, the mismatch time between audio and video streams can be neglected in this context. It is also observed that, the attention given to the lip sync assessment is clearly content dependent. However, the MOS of audio, video and audiovideo overall show that the perceived quality is unaffected by the loss of synchronization between

audio and video media. The reason why the lip sync is not important in low cost DVC systems is that the picture quality is so poor that the users are not using lip movement to enhance audio understanding - as would be in the case in high quality videoconferencing. As a result the quality of the audio is very important in low cost DVC. On the other hand, the findings obtained from the field study in Chapter 8, concluded that although the rating score of the perceived multimedia quality is generally GOOD (4), the subjects claimed that it is very difficult to predict the mismatch time between the audio and video streams, i.e. to clarify which media is ahead of or lagging the other;

- the overall result has led to the establishment of the quality boundary for audio and video, required for specific task performance, based up on the end-user's QoS requirements.

The results of this study have been disseminated via a series of five peer-reviewed publications in international journals and conferences. Copies of these papers are provided in Appendix D.

10.1.2 Research Contributions

The benchmarking of the current desktop videoconferencing QoS that is advocated in this thesis involves both the field and laboratory study. By addressing the major drawbacks in realizing the real-time multimedia applications and services, as explained in Chapter 1, the research has contributed to new knowledge and improved

understanding of the various factors that could affect the perceived quality of IP media, and hence, propose techniques as to how it can be improved. These contributions can be summarized as follows:

- the research conducted in this thesis illustrates a novel approach of assessing IP media, in that it exhaustively investigates the impacts of a wide range of different network conditions, interpolated into audio and video streams, both simultaneously and separately (prior to transmission), on the perceived quality of audio, video and audiovideo overall. By undertaking multidimensional research of the individual and combined impacts of the various type of network conditions (advocated in this thesis), on videoconferencing quality, and providing explicit information of its effect and nature, a valuable taxonomy for audio and video quality requirements across different network congestions and tasks has been derived and defined;

- the results obtained, in terms of the mean opinion scores have enabled the minimum and maximum quality thresholds, as well as an appropriate audio and video quality ratio (based on end-users' QoS requirements) for a given network condition to be defined. The findings would assist the system and network developers to enhance bandwidth utilization, while delivering the best possible media quality in accordance with task performance, network condition and available bandwidth;

- the network service providers could stand to benefit from knowing the minimum and maximum quality requirements for each specific task (based on the MOS results) and to charge accordingly, without causing user dissatisfaction and undue cost;

- the subjective assessment technique and methodology approaches, implemented in this thesis provide a guideline towards developing the standard consensus to determine the QoS of multimedia over IP (and mobile) networks, which is one of the major limiting factors of the services today.

The fact that the subjective tests were carried out, and that the same benchmark prototypes were maintained throughout the entire programme, the work provides new knowledge from results that were derived, compared and generalized across the studies. It is relevant to the growing popularity of videoconferencing over packet network, and there is an urgent need to establish a reliable technique to define the the level of standardized quality required to successfully complete a specific task.

10.2 Research Limitations

Despite achieving all the original objectives of the research program, the work faced several challenges, which are described below:

- assessing video quality in low cost videoconferencing environment is a very complicated issue since the frame rates are constantly changing due to the nature of IP networks and subject's movement. As such, some subjects felt dissatisfied in giving only one score, as their opinion fluctuated between two or more scores from one moment to another during the test. Refer to Section 5.3.3 for further explanations;

- although subjective methods are widely used to evaluate multimedia quality, it is the end-user who will determine whether a service meets their satisfaction, and the technique inevitably contains a number of limitations. There is a rising concern regarding the validity of MOS test results, which mainly stems from a number of inadequacies and problems, as thoroughly explained in Section 3.5.3. However, it has been proved to be more reliable than the objective test method, as explained in Subsection 3.5.4;

- the difficulty of getting test subjects, time consuming, and high cost are some of the reasons why subjective test has become unpopular in spite of its advantages. The number of attendances for some of the tasks in the thesis are not as high as one would have liked, and thus, it is hard to draw firm conclusions;

- the difficulty in maintaining certain conditions in a prolonged study is one of the shortcomings of the research programme. The degree of the inter-observer interaction is a very critical issue, due to numerous variables that can affect the end users' perception of multimedia quality, as perviously described in Section 3.2.6. For example, at the start of the fourth year of the study, the research group's office moved to a new university building. A few changes were inevitable – for example, the test room environment and system's network cards, in order to adapt to the new conditions. These changes may have affected some of the results (mostly in Chapter 7). For example, under the same network conditions, the MOS of the perceived audio quality in the most recent test has increased by approximately 0.5;

- familiarization of subjects to the test procedure has proved to be another critical issue in realising the project, as the subject could misunderstand their task. Since, it is essential that the designed task should be sustained, the subjects must be fully informed so that the task will be conducted coherently in order to obtain a more reliable result. The implication of this mater is described further in Section 5.3.3.

In spite of these limitations, the achievements and contributions of the research are still considered to be valid.

10.3 Research Recommendations

In further developing the research objectives, there are a number potential recommendations to ameliorate future investigations in the similar area, that are illustrated in the following.

10.3.1 Guidelines for Future Researchers

Some guidelines for future researchers in the similar area are presented as follows:

- subjective test method – subjective testing, based on MOS, is best suited for assessing the multimedia quality in videoconferencing system and fellow researches in the similar area should implement the same technique as it is the end users' opinion that will determine the success of an application and for other reasons, described in Section 3.5.4.

  The continuous rating scale methods, such as the Quality Assessment slider (QUASS), Simultaneous Double Stimulus Evaluation (SDSCE), and Single Stimulus Continuous Quality Environment (SSCQE), are inadequate for assessing the perceived multimedia quality in the real-time interactive communication, since they may interrupt the nature of the conversation. On the other hand, these methods are effectively used to assess the impact of sparse degradations, e.g. transmission errors, on the perceived quality of video [77];

- use of interactive test scenarios – the specific task performance must be carefully designed, and the interactive task performance is considered to be more practical and highly recommended when evaluated the end-user QoS of multimedia systems. This is mainly because it imitates real world scenario more closely than the passive or the intelligibly test, since informal conversational scenario between two (or more) users is the norm in videoconferencing applications;

  Also, one of the biggest problems in assessing IP multimedia quality is due to the unstable nature of packet networks in that the quality can fluctuate rapidly and to a certain degree from one minute to another. This means that assessing the IP media quality based on short sentences or short segments of media streams may not be meaningful since it does not represents the real world occurrence.

- familiarization of test procedure – the subjects should be fully informed of their tasks, to minimise discontinuity of task performance, since it is assumed that different application scenarios will have different subjective quality requirement;

- test environments – it is mandatory to maintain the test environments, such as room condition (background noise), systems specification, systems hardware, and etc. throughout the entire study to ensure the validity of the results;

10.3.2 Suggestion for Additional Research

The following outlines some of the possible extensions of the work presented in this thesis.

- one-to-many – all of the findings discussed in the thesis are based on the impact on conversational behavior between one-to-one person group only. It is suggested that in the future, the research should considerer the impact of having more than one-person group, for example, many-to-many, one-to-many, and etc. Some issues like the lack of eye contact, turn taking, and directional audio may become more critical in this type of conversational nature;

- use larger number subject groups and categories – it is essential to note that one of the main obstacles of the research programme is to find the test candidates. Due to time constraints and a limited number of subjects, the research was unable to conduct an in-depth assessment of the impact of different levels of society, in term of age, gender, language (nationality), professional, experience and etc., on the perceived media quality. There is unavoidably, great advantages of investigating the subjective opinion of IP-based multimedia quality from a multiple backgrounds of test candidates. Therefore, it is suggested that future work should strive more towards the issue;

- evaluation of Third Generation (3G) mobile videoconferencing – since, the popularity of multimedia over 3G mobile applications is advancing very fast, it is highly important to determine the multimedia quality over mobile net-

works and to investigate the problems inherent in clarifying the end user's QoS. Although the research mainly focuses upon benchmarking the low cost desktop videoconferencing system, the test methodology illustrated in this thesis could be employed in assessing multimedia quality over 3G mobile, since they are a number of similarities between these systems, such as, the range of CODECs used and picture size. The difference with mobile systems would arise in the issue of terminal heterogeneity problem, which is one of the major challenges facing the mobile services today. Due to the physical characteristics of cellular mobile networks, the quality the data rate of an ongoing connection varies, contributing to the heterogeneity problem.

- implementation and evaluation of the proposed adaptive architecture – the continuation of the research work should focus upon the implementation and evaluation of the proposed adaptive architecture, and validation of its applicability in the practical communication scenarios.

10.4 The Future of Multimedia Over IP

The usage of real-time multimedia information over the IP network has increased rapidly over the past few years. The application of DVC systems has become increasingly popular both professionally and personally due to its many advantages, namely, improved services and reduced cost. Furthermore, the emergence of 3G mobile has made videoconference-over-mobile become a reality. 3G provides suf-

ficient high bandwidth to transport the very large quantities of video data needed for real time videoconferencing. The technology and standards for multimedia services on mobile networks have continued to develop in recent years and many new services have increasingly became available.

However, until very recently, there have been no explicit study focuses upon the assessment of the perceived multimedia quality over low cost DVC or 3G mobile, but rather more inclined towards the audio and video quality over more sophisticated and higher bandwidth communications and entertainment (broadcasting) systems. To date, a standard consensus to define the IP media QoS, has yet to be implemented. Therefore, the research has addressed this problem by investigating a new approach to assess the perceived quality of the IP media components, i.e. audio, video, and audiovisual overall. Based on the exhaustive study, a novel method has been proposed to optimize the perceived multimedia quality.

Although addressing the technological impact of sending high quality real-time data with minimal bandwidth over a conference is undoubtedly of the primary importance, the respective parties (i.e. network provider and system developer) should have the equally important task of understanding and measuring the subjective quality score of the end product (and other determining factors e.g. hardware factor), since it is the end user's opinion that will determine the success of an application. Only then will the problem inherent in determining the end's QoS can be addressed.

The work presented in this thesis has been carried out in the context of low cost desktop videoconferencing, but many of the issues and findings should be relevant for the benchmarking of wireless conferencing, as in the 3G Mobile applications and other desktop conferencing systems and networks, in general. It provides guidelines to the service providers to design their own tests to evaluate the performance of multimedia services, as perceived by the users and it can be applied at either laboratory experimental prototypes or in-service applications. The designer of services and applications stand to benefit from knowing the minimum and maximum quality requirement for each specific task (based on the MOS results) and to charge accordingly, without causing user dissatisfaction and undue cost. It is anticipated that, the future trend in 3G mobile would charge for its services in terms of the bandwidth availability, and hence, QoS.

There are numerous similarities between the low cost desktop conferencing (e.g. Microsoft NetMeeting) and 3G mobile communication systems, such as, the range of multimedia quality obtained in both networks are highly time varying (due to low bit rate and the type of error profiles), picture display size and low bandwidth CODECs used. ITU-T H.263 video CODEC is being used in both networks. However, 3G mobile employed a better performance audio CODEC i.e. AMR (12.2Kb/s) as opposed to G.723.1 (6.3 Kb/s), for voice implemented in the thesis. The fact that the work had been carried out with lower bit rates for audio has an advantage, in that the subjective MOS obtained would provide a baseline comparison to determine audio quality, as perceived in mobile videoconferencing systems. In addition,

there is a need to identify the minimum end-user's QoS requirements in multimedia services, as there will always be consumer demand for lower quality at lower cost [122].

In summary, the study has been designed and focused upon the subjective evaluations of low bandwidth multimedia services that can be applied for all types of IP-based networks. The test methodologies and findings of the research described in this thesis could contribute to the design and development of future multimedia videoconferencing applications in general.

References

[1] International Telecommunications Union-Telecommunication (ITU-T). One way transmission line. *ITU-T Recommendation G.114*, Study Group 12, February 2001.

[2] Steinmetz R. Human perception of jitter and media synchronization. *IEEE Journal on Selected Areas in Communication*, 14(1), 1996.

[3] Jardetzky P.W, Sreenan C.J., and Needham R.M. Storage and synchronisation for distributed continuous media. *Multimedia Systems*, 3:151–161, 1995.

[4] et. al. Hoffman. RFC 2250: RTP Format for MPEG1/MPEG2 Video. Internet Draft, 1998.

[5] Schulzrinne H., Casner S., Frederick R., and Jacobson V. RTP: A transport protocol for real-time applications. Technical report, RFC 1889, Internet Engineering Task Force (IETF) [on line] ftp://ftp.ietf.org/ rfc/ rfc1889.txt, January 1996.

[6] Kouvelas I., Hardman V., and Watson A. Lip synchronisation for use over the

Internet: Analysis and implementation. In *Proceedings of IEEE Globecom '96,* London UK, November 1996.

[7] Tanenbaum A. *Computer Networks.* Prentice Hall, fourth edition, 1996.

[8] Picone J. Signal modelling technigues in speech recognition. In *Proceedings of the IEEE,* volume 81, pages 1215–1247, 1993.

[9] Huerta J.M. and Stern R.M. Distortion-class weighted acoustic modeling for robust speech recognition under GSM RPE-LTP coding. In *Proceedings of the Workshop on Robust Methods for Speech Recognition in Adverse Conditions,* Tampere, 1999.

[10] European Telecommunications Standards Institute (ETSI). Digital Cellular Telecommunications System (Phase 2+); Adaptive Multi-Rate (AMR) speech transcoding (GSM 06.90) version 7.2.1 release 1998. *ETSI EN 301 704,* 2000–04.

[11] ITU-T. Dual rate speech coder for multimedia communication transmitting at 5.3/6.3 kbps. *ITU-T Recommendation G.723.1,* March 1996.

[12] ITU-T Recommendation G.729. Coding of speech at 8 kbit/s using Conjugate-Structure Algebraic-Code-Excited Linear-Prediction (CS-ACELP). Technical report, March 1996.

[13] Zatsman A. G.728 compression. *[on-line] ftp://dspsun.eas.asu.edu/ pub/ speech/ ldcelp.tgz.*

[14] Borko Furht. *Handbook of Internet Computing*. CRC Press, Boca Raton, Florida, 2000.

[15] Watson A. and Sasse M.A. Multimedia conferencing via multicast: Determining the quality of service required by the end user. *Proceedings of AVSPN '97 - International Workshop on Audio-Visual Services over Packet Networks*, pages 189–194, 15–16 September 1997.

[16] Brady N. MPEG-4 standard methods for the compression of arbitrarily shaped video objects. *IEEE Transactions on Circuits and Systems for Video Technology*, 9(8):1170–1192, December 1999.

[17] MPEG Video CD Editoral Committee. MPEG Video Committee Draft, ISO-IEC/JTC1/SC2/WG11/MPEG90/176. Technical report, December 1990.

[18] Ghanbari M. An adapted H261 two-layer video codec for ATM neworks. *IEEE Transactions on Communications*, 40(9):1481–1490, September 1992.

[19] M Walker and M Nilsson. A study of the efficiency of layered video coding using H.263. *Proceeding od Packet Video 99, the 9th International Packet Video Workshop, New York*, April 1999.

[20] Wu S. and Gersho A. Rate-constrained picture adaptive quantization for JPEG baseline coders. In *International Conference of Acoustics, Speech and Signal Precessing*, volume 5, pages 389–392, April 1993.

[21] Arya S. and Mount D.M. Algorithms for fast vector quantization. In J. A.

Storer and M. Cohn, editors, *Proceedings DCC'93 (IEEE Data Compression Conference)*, pages 381–390, Snowbird, UT, USA, 1993.

[22] Aldred B. *Desktop Conferencing*. McGraw-Hill Book Company, London, 1996.

[23] Okubo S., Dunstan S., and Morrison G. ITU-T standardization of audiovisual communication systems in ATM and LAN environments. *IEEE Journal on Selected Areas in Communication*, 15(6):965–981, August 1997.

[24] Kostas T. and Borella M. Real-time voice over packet switched networks. *IEEE Network*, 12(1):18–27, Jan/Feb 1998.

[25] Lindberg D. The H.324 multimedia communication standard. *IEEE Communications Magazine*, 34(12):46–51, December 1996.

[26] Hassan M., Nayandoro A., and Atiquzzaman M. Internet telephony: Services, technical challenges, and products. *IEEE Communications Magazine*, 38(4):96–103, April 2000.

[27] Perkins C., Hodson O., and Hardman V. A survey of packet-loss recovery techniques for streaming audio. *IEEE Network Magazine*, 12(5):40–48, Sept/Oct 1998.

[28] Bolot J., Fosse-Parisis S., and Towsley D.S. Adaptive (FEC)-based error control for internet telephony. In *INFOCOM (3)*, pages 1453–1460, 1999.

[29] Anand M. An empirical study of VoIP. Master's thesis, Department of Electrical Engineering, Indian Institute of Technology, Mumbai, India, July 2002.

[30] Perkins C. and Hodson O. Options for repair of streaming media. *RFC 2354, Internet Engineering Task Force (IETF)*, June 1999.

[31] Wah B., Su X., and Lin D. A survey of error-concealment schemes for real-time audio and video transmissions over the internet. In *In Proc. IEEE International Symposium on Multimedia Software Engineering*, pages 17–24, Taipei, Taiwan, December 2000.

[32] Babich F. and Virez M. A novel wide-bande audio transmission scheme over the Internet with a smooth quality degradation. *ACM SIGCOMM Computer Comunication Review*, 30(1), Jan 2000.

[33] Philip A. RFC 1349 - type of service in the Internet protocol suite. Technical report, Network Working Group, [on-line] http://www.faqs.org/ rfcs/ rfc1349.html, July 1992.

[34] Tang J.C. and Isaacs E.A. Why do users like video?: Study of multimedia supported collaboration. *Computer Supported Cooperative Work 1*, pages 163–196, 1993.

[35] Schmidt D.A. Telemedicine and telecommunications: How carriers can play a more prominent role in the telemedicine market. *http://www.teleconferencemag.com/html/issues/issues1998/july 1998/798telemed.html*, July 1998.

[36] Skiar D.L. And telejustice for all. *http://www.telconferencing.com/html/issues/issues2000/j* July/August 2000.

[37] D.M. Stanek. Modeling perceptions and preference of home-based and center-based telecommuting. *Master's Thesis Department of Civil and Environmental Engineering, Institute of Transportation Studies, University of California, Davis,* December 1995. Research Report No. UCD-ITS-RR-95-12.

[38] Claypool M. and Tanner J. The effects of jitter on the perceptual quality of video. *ACM Multimedia Conference,* 2, 30 October–5 November 1999.

[39] Claypool M. and Riedl J. End-to-end quality in multimedia application. *Chapter 40 in Handbook on Multimedia Computing,* 1999.

[40] Sasse M.A., Biltung U., Schulz C.D., and Turletti T. Remote seminars through multimedia conferencing: Experiences from the mice project. *INET'94/JENC5,* 1994.

[41] Myers M. Predicting and measuring quality of service for mobile multimedia. *The Human Perspective,* August 2000.

[42] Kitawaki N. and Itoh K. Pure delay effects on speech quality in telecommunications. *IEEE Journal on Selected Areas in Communication,* 9(4):586–593, 1991.

[43] Bruce V. The role of the face in communication: Implications for videophone design. *Interacting with Computers,* 8(2):166–176, 1996.

[44] Munhall K.G, Gribble P., Sacco L., and Ward M. Temporal constraints on the McGurk effect. *Perception and Psychophysics,* 58(3):351–362, 1996.

[45] Zhang H. and Keshav S. Comparison of rate-based service disciplines. In *ACM SIGCOMM'91*, Zurich, Switzerland, September 1991.

[46] Bashandy A. A protocol architecture for guaranteed quality of service in collaborated multimedia applications. In *In the Proceedings of the IEEE Symposium on Application -Specific Systems and Software Engineering and Technology (ASSET'99), IEEE Computer Society*, pages 120–127, Los Alamitos, CA,, 1999.

[47] Joe I. Packet loss and jitter control for real-time MPEQ video communications. *Computer Communications*, 19:901–914, 1996.

[48] Internet Engineering Task Force (IETF). IETF website. *[on-line] http:// www.ietf.org*.

[49] Jardetzky P.W, Sreenan C.J, and Needham R.M. Storage and synchronization for distributed continuos media. *Multimedia Systems*, 3:151–161, 1995.

[50] Rudkin S., John A., and Whybray M. Real-time application on the Internet. *BT Technology Journal*, 15(2), April 1997.

[51] Marc H. Willebeek-LeMair and Zon-Yin Shae. Videoconferencing over packet-based networks. *IEEE Journal on Selected Areas in Communication*, 15(6):101–114, August 1997.

[52] Vitkovitch M. and Barber P. Effect of video frame rate of subjects' ability to shadow one of two competing messages. *Journal of Speech and Hearing Research*, 37:1204–1210, 1994.

[53] Frowien H.W., Smoorenburg G.F., Pyters L., and Schinkel D. Improved speech reconignition through videotelephony: Experiments with the hard of hearing. *IEEE Journal on Selected Areas in Communication*, 9:611–616, 1991.

[54] Cote G., Erol B., Gallant M., and Kossentini F. H.263+: Video coding at low bit rates. In *IEEE Transactions on Circuits and Systems for Video Technology*, volume Vol.8.No.7, November 1998.

[55] ITU-T. Objective quality measurement of telephone-band (300-3400 hz) speech codecs. *ITU-T Recommendation P.861*, February 1998.

[56] Wang Shufeng and Zhang Junyou. A coding algorithm of image compression – Recommendation H.263+. In *Proceedings of 99 International Conference on Agricultural Engineering*, Beijing, China, December 1999.

[57] Anderson A.H., Sotillo C. Bard E.G., Newlands A., and Doherty-Sneddon G. Limited visual control of the intelligibility of the speech in face-to-face dialogue. In *Perception and Psychophysics*, volume 59 of 4, pages 580–592, 1997.

[58] Hollier M.P and Voelcker R.M. Towards a multimodal perceptual model. *BT Technological Journal*, 15(4):162–171, 1997.

[59] Bouch A. and Sasse M.A. It ain't what you charge it's the way that you do it: A perspective of network qos and pricing. In *IFIP/IEEE International Symposium on Integrated Network Management*, pages 24–28, Boston, USA, May 1999.

[60] Wilson F. and Descamps P.T. Should we accept anything less than tv quality: Visual communication. International Broadcasting Convention, Amsterdam.

[61] International Telecommunications Union-Telecommunication (ITU-T). Control of talker echo. *ITU-T Recommendation G131*, August 1996.

[62] International Telecommunications Union-Telecommunication (ITU-T). Echo cancellers. *ITU-T Recommendation G131*, March 1993.

[63] Finger R. Measuring video quality in videoconferencing systems. *Techonline Review, [on-line] URL: http:// www.techoline.com*, 2(5), October 1998.

[64] Kawalek J. A user perspective for qos management. In IS & N 1995, editor, *Proceeding of 3rd International Conference on Intelligence in Broadband Services and Network Proceeding of 3rd International Conference on Intelligence in Broadband Services and Network*, Crete, Greece, 1995.

[65] Sellen A.J. Speech patterns in video-mediated conversations. *Proceedings of CHI '92, ACM*, pages 49–59, 1992.

[66] Finholt T., Sproull L., and Kiesler S. Communication and performance in ad hoc task groups. *Galegher J., Kraut R.E., and Egido C. (editors), Intellectual Teamwork*, pages 291–325, 1990.

[67] Wilson G. and Sasse M. A. Do users always know what's good for them? utilising physiological responses to access media quality. In *In S. McDonald, Y. Waern and G. Cockton (Eds.): People and Computers XIV - Usability or Else! Proceedings of HCI 2000*, pages 327–339, Sunderland, UK, September 5th–8th 2000(2).

[68] Whittaker S., Frohlich D., and Daly-Jones O. Informal communication: What is it like and how might we support it. *Proceedings of CHI'94 Conference on Computer Human Interaction*, pages 130–137, 1994.

[69] Isaacs E. and Tang J. What video can and cannot do for collaboration: A case study. *Multimedia Systems*, 2:63–73, 1994.

[70] Watson A. and Sasse M.A. Evaluating audio and video quality in low-cost multimedia conferencing system. *Interacting with Computers*, 8(3):255–275, 1996.

[71] Ostberg O., Lindstrom B., and Renhall P-O. Contribution of display size to speech intelligibility in video-phone systems. In *International Journal of Human-Computer Interaction 1*, volume 1, pages 149–159, 1989.

[72] Negroponte N. Being digital. Hodder and Stoughton (eds.), 1995.

[73] Mullin J., Jackson M., Anderson A., and Smallwood L. Assessment methods for assessing audio and video quality in real-time interactive communications. *[on-line] http://www-mice.cs.ucl.ac.uk/ multimedia/ projects/ etna/ assessment-methods.pdf.*

[74] Beerends J., Hekstra A.P, Rix A., and Hollier M. Perception evaluation of speech quality (PESQ), the new ITU standard for end-to-end speech quality assessment, Part II–Psychoacoustic model. *Journal of Audio Engineering Society*, 50(10):765–778, October 2002.

[75] Bouch A., Watson A., and Sasse M.A. QUASS - a tool for measuring the subjective quality of real-time multimedia audio and video. In *May J., Siddigi J. and Wilkinson J. eds., HCI'98 Conference Companion*, pages 94–95, Sheffield, UK, 1st–4th September 1998.

[76] Wilson G. and Sasse M.A. Investigating the impact of audio degradations on users: Subjective vs. objective assessment methods. In *Paris C., Ozkan N., Howard S. and Lu S. (eds.) Proceedings of OZCHI 2000: Interfacing Reality in the New Millennium*, pages 135–142, Sydney, Australia, December 4th–8th 2000(1). ISBN 0-643-06633-0.

[77] EURESCOM Project P905-PF. AQUAVIT - assessment of quality for audio-visual signals over Internet and UMTS: Methodology for subjective audio-visual quality evaluation in mobile and ip networks. *Deliverable 2*, August 2000.

[78] International Telecommunications Union-Telecommunication (ITU-T). Perceptual evaluation of speech quality (PESQ), and objective method for end-to-end quality assessment of narrowband telephone networks and speech codecs. *ITU-T Recommendation P.862*, February 2001.

[79] Watson A. and Sasse M.A. Measuring perceived quality of speech and video in multimedia conferencing. In *Proceedings of ACM Multimedia'98*, pages 55–60, Bristol, 1998.

[80] ITU. Methods for subjective determination of transmission quality. *ITU-T Recommendation P.800*, August 1996.

[81] International Telecommunications Union-Telecommunication (ITU-T). Subjective video quality assessment methods for multimedia applications. *ITU-T Recommendation P.910*, 1999.

[82] Voran S. and Sholl C. Perceptual-based objective estimators of speech quality. *IEEE Workshop on Speech Coding*, 20–22 September 1995.

[83] Rix A., Reynolds R., and Hollier M. Perceptual measurement of end-to-end speech quality over audio and packet-based networks. In *AES 106th Convention*, pages 8–11, Munich, May 1999.

[84] Psythecnics Limited. PESQ: An intorduction white paper. *[on-line] http://www.agilent.com/ comms/ onenetworks*, 2001.

[85] Whittaker S. Rethinking video as a technology for interpersonal communication: Theory and design implications. *International Journal of Human-Computer studies*, 42(5):501–529, 1995.

[86] William Y.Z. and Philip J.C. Methods for evaluation of digital television picture quality. *Doc. G-2.1.6/28, [on-line] grouper.ieee.org/ groups/ videocomp/ 1997g216/ zou970501.pdf*, 29 April 1997.

[87] Wolf S. Features for automatic quality assessment of digitally transmitted video. Technical report, National Telecommunications and Information Administration (NTIA), Report 90–264, US Department of Commerce, June 1990.

[88] Webster A. An objective video quality assessment system based on human perception. *SPIE Human Vision, Visual Processing, and Digital Display IV*, 19(13), February 1993.

[89] Hollier M.P., Rimell A.N., Hands D.S., and Voelcker R.M. Multi-modal perception. *BT Technology Journal*, 17(1):35–46, Jan 1999.

[90] El-Zarki M. Video over IP. *IEEE INFOCOM 2001*, April 2001.

[91] Voran S. Estimation of perceived speech quality using measuring normalizing blocks. *IEEE Workshop on Speech Coding for Telecommunications Proceeding*, pages 83–84, 1997.

[92] Rix A. and Hollier M. The perceptual analysis measurement system for robust end-to-end speech assessment. *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing – ICASSP*, pages 1515–1518, 2000.

[93] Flanagan J.L. Speech analysis, synthesis and perception. *Springer-Verlag*, 1965.

[94] Carson M. Nist net home page. *(URL: http://www.antd.nist.gov/ nistnet)*, 2000.

[95] Microsoft NetMeeting. Microsoft netmeeting home page (url: http://www.microsoft.com/windows/netmeeting).

[96] Samir M., Francisco C., and Hossam A. Intergrating networks measurement and speech quality subjective scores for control purposes. In *IEEE Infocom 2001*, volume 2, pages 641–649, Anchorage, Alaska, 22–26 April 2001. USA IEEE. ISBN 0-7803-7016-3.

[97] ITU-R. Recommendation BT. 500 - 7, method for the subjective assessment of the quality of television pictures. *RBT*, 1997.

[98] ITU-T Recommendation P.920. Interactive test methods for audiovisual communications. *[on line] http://www.itu.int/ publications/ itu-t/ itutrec.htm*, 1996.

[99] Isaacs E.A. and Tang J.C. What video can and cannot do for collaboration: A case study. *Multimedia Systems*, 2:63–73, 1994.

[100] Hardman V., Sasse M.A., and Kouvelas I. Successful multiparty audio communication over the internet. *Communications of the ACM*, 41(5):74–80, 1998.

[101] Short J., William E., and Christie B. *The Social Psychology of Telecommunications*. Wiley, 1976.

[102] Association de Recherche en Communication (ARC). On retient mieux ce que l on entend bien. *Association de Recherche en Communication et Formation Langues, Paris*, (date – not available).

[103] Ravindran K. Real-time synchronization of multimedia data stream in high speed packet switch networks. *Workshop on Multimedia Information Systems (MMIS 92), IEEE Communications Society*, pages 164–188, 1992.

[104] Ohmori T., Maeno K., Sakata S., and Watabe K. Cooperative control for sharing application based on distributed multiparty desktop conferencing system: MERMAID2, SuperCOMM/ICC'92. *Discovering a New World of Communications*, 2:1069–1075, June 1992.

[105] Ravindran K. and Bansal V. Delay compensation protocols for synchronisation of multimedia data streams. *IEEE Trans. On Knowledge and Data Engineering*, 5(4):574–589, August 1993.

[106] Jacobson V. Tcpdump source code. ftp://ftp.ee.lbl.gov/tcpdump.tar.Z, 2003.

[107] Ghita B., Lines B, Furnell S, and Ifeachor E. Non intrusive ip network performance monitoring for tcp flows. In *Proceedings of IEEE ICT 2001*, 2001.

[108] Bakircioglu H. and Kocak T. A survey of random neural network applications. European Journal of Operational Research 126, 2000.

[109] Chen B.S., Peng S.C., and Wang K.C. Traffic modelling, prediction, and congestion control for high-speed networks: A fuzzy AR approach. *IEEE Transactions on Fuzzy Systems*, 8(5):491–507, 2000.

[110] Youssef S.A., Habib I.W., and Saadawi T.N. A neurocomputing controller for bandwidth for bandwith allocation in ATM networks. *IEEE Journal on Selected Areas in Communication*, 15(2):191–199, February 1997.

[111] Yu E.S. and Chen C.Y. Traffic prediction using neural network. *IEEE GLOBECOM*, 991–995, 1993.

[112] Rumelhart D.E., Wildrow B., and Lehr M.A. The basic ideas in neural networks. *Communications of the ACM*, 37(3):87–92, March 1994.

[113] T1 Committee ANSI [T1.312-1991]. Voice packetization-packetized voice pro-

tocol. Technical report, American National Standard Dictionary of Information Technology (ANSI), 1991.

[114] Perkins C. S. and Crowcroft J. Effects of interleaving on rtp header compression. *IEEE INFOCOM*, March 2000.

[115] Mued L., Lines B., Furnell S., and Reynolds P. Investigating the interaction effect of audio and video as perceived in low cost videoconferencing. In *Third International Network Conference (INC 2002)*, Plymouth, UK, 16–18 July 2002.

[116] Hughes J. and Sasse M. A. Internet multimedia conferencing - results from the ReLaTe project. *ICDE World Conference*, June 1998.

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

[118] Ramjee R., Kurose J., Towsley D., and Schulzrinne H. Adaptive playout mechanisms for packetized audio applications in widearea networks. *IEEE INFOCOM*, June 1994.

[119] Kouvelas I. and Hardman V. Overcoming workstation scheduling problems in a realtime audio tool. *USENIX Annual Technical Conference*, January 1997.

[120] Zhang J., Arnold J.F., and Frater M.R. A cell-loss concealment technique for MPEG-2 coded video. *IEEE Transactions On Circuits and Systems for Video Technology*, 10(4):659–665, June 2000.

[121] Cuenca P., Orozco-Barbosa L., Garrido A., and Quiles F.J. Loss-ressilient ATM protocol architecture for MPEG-2 video communications. *IEEE Journal on Selected Areas in Communication*, 18(6):1075–1086, June 2000.

[122] Podolsky M. G. A study of speech or audio coding on packet switched networks. UCB–ERL Technical Memorandum (Number M96), Dec 1996.

APPENDIX A

Instruction and Questionnaire Sheets for Subjective Test

The Instruction for the Mean Opinion Score (MOS) Test

The goal of this test is to evaluate the quality of video, audio and combined audiovideo that will be displayed on the screen that is in front of you.

You and your partner will have an informal discussion using desktop videoconferencing i.e. Microsoft NetMeeting. You are asked to discuss with your friend about any topic of your own interest.

The discussion should last approximately 5-10 minutes. You are requested to watch with attention the perceived quality of video, audio and combined audiovideo during the whole duration and, at the same time, to interact with your partner. After the section, you are asked to rate the perceived quality of:

(a) Video ONLY;
(b) Audio ONLY;
(c) Combined Audiovideo, on the answer sheets in front of you.

To express your score, you will use the MOS rating scales that is given in front of you. If anything is unclear, please do not hesitate to ask for clarifications.

Thank You!

Answer Sheets

Name: .....
Date: .....
Section: .....

Questionnaires: (please tick the appropriate box)

Section 1

1. Giving a scale from 1-5, how would you assess this quality?

AUDIO	VIDEO	AUDIOVIDEO OVERALL
(a) Scale 5, i.e. **Excellent** ☐	(a) Scale 5, i.e. **Excellent** ☐	(a) Scale 5, i.e. **Excellent** ☐
(b) Scale 4, i.e. **Good** ☐	(b) Scale 4, i.e. **Good** ☐	(b) Scale 4, i.e. **Good** ☐
(c) Scale 3, i.e. **Fair** ☐	(c) Scale 3, i.e. **Fair** ☐	(c) Scale 3, i.e. **Fair** ☐
(d) Scale 2, i.e. **Poor** ☐	(d) Scale 2, i.e. **Poor** ☐	(d) Scale 2, i.e. **Poor** ☐
(e) Scale 1, i.e. **Bad** ☐	(e) Scale 1, i.e. **Bad** ☐	(e) Scale 1, i.e. **Bad** ☐

Section 2

1. Do you notice any strange effect or lip synchronization error?

(a) Yes, I notice that audio is ahead of video → ☐

(b) Yes, I notice that audio is behind video → ☐

(c) Yes, but I am not sure if audio is ahead of video or behind video → ☐

(d) No, I don't notice any strange effect or synchronization error → ☐

2. Giving the scale from 1-5, how would the lack of lip synchronisation error affect you?

Given that: **1 – very annoying**
5 – less annoying

Your score: 1→☐ 2→☐ 3→☐ 4→☐ 5→☐
(very annoying) ----------> (less annoying)

3. Did you or your partner have any difficulty in talking or hearing over the connection?

Yes → ☐
No → ☐

Section 3

1. Do you prefer to have audio and video or just audio only?

(a) I prefer to have **audio and video** → ☐
(b) I prefer to have **audio ONLY** → ☐

WHY? (Please give reason(s) to your answer, above)

2. Please write down your general observation, comment, suggestion or any input regarding the test.

THANK YOU!!!

	Name:						
	Date/time:						
No	Assessment of:	MOS Score					Comments (if any)
1	NetMeeting in general	1	2	3	4	5	
2	audio quality	1	2	3	4	5	
	video quality	1	2	3	4	5	
	combined audio and video quality	1	2	3	4	5	
3	audio quality	1	2	3	4	5	
	video quality	1	2	3	4	5	
	combined audio and video quality	1	2	3	4	5	
4	audio quality	1	2	3	4	5	
	video quality	1	2	3	4	5	
	combined audio and video quality	1	2	3	4	5	
5	audio quality	1	2	3	4	5	
	video quality	1	2	3	4	5	
	combined audio and video quality	1	2	3	4	5	
6	audio quality	1	2	3	4	5	
	video quality	1	2	3	4	5	
	combined audio and video quality	1	2	3	4	5	
7	audio quality	1	2	3	4	5	
	video quality	1	2	3	4	5	
	combined audio and video quality	1	2	3	4	5	
8	audio quality	1	2	3	4	5	
	video quality	1	2	3	4	5	
	combined audio and video quality	1	2	3	4	5	
9	audio quality	1	2	3	4	5	
	video quality	1	2	3	4	5	
	combined audio and video quality	1	2	3	4	5	
10	audio quality	1	2	3	4	5	
	video quality	1	2	3	4	5	
	combined audio and video quality	1	2	3	4	5	
11	audio quality	1	2	3	4	5	
	video quality	1	2	3	4	5	
	combined audio and video quality	1	2	3	4	5	
12	audio quality	1	2	3	4	5	
	video quality	1	2	3	4	5	
	combined audio and video quality	1	2	3	4	5	
13	audio quality	1	2	3	4	5	
	video quality	1	2	3	4	5	
	combined audio and video quality	1	2	3	4	5	

APPENDIX B

Observations, Comments, and Suggestions

Subject's Observations, Comments, and Suggestions (refer to Section 3 on the answer sheets)

Classroom-based experiments

Subject 1 Prefer to have audio and video; *"It is useful to be able to gauge their mood and response to what you are saying. However, it can be distracting at times. I would opt for audio and video but would not object unduly if only audio were available."*

General Observation: *"I found it very difficult to tell if audio and video were in synchronised, mainly because of my partner beard. As a result, it was hard to comment on whether I found this annoying. The size of the video was really to small to be able to see many of the effects clearly. A video size perhaps double this size would be better."*

Subject 2 Prefer to have audio and video; *" .. but I could live with audio only. We are visual creatures and enjoy looking at human features particularly - it is part of the feedback of mood and expression which helps a conversation and the engagement of attention."*

General Observation: *"I would be concerned about the subjective nature of my responses especially after a number of tests where I found it difficult to to distinguish any difference - I think I tended to start marking to a 'low average'. To me, the quality of video was consistently lower than the quality of audio and this might be a distraction."*

Subject 3 Prefer to have audio and video; *"Gives me something to look at in terms of observing the other person's reaction to things."*

General Observation: *"Lip sync was difficult to assess from the size of the webcam window, but my opinion was that it did not really matter in the context of this form of one-to-one communication."*

Subject 4 Prefer to have audio and video; *" Under the test circumstance full audio and video was better as I used a number of visual cues to communicates. When audio went bad, I switched to video for information and vice-versa."*

General Observation: *"Some of the initial tests were very close and difficult to weight accurately. I had to concentrate to give a correct assessment or to really distinguish the difference between the test sequence."*

Subject 5 Prefer to have audio and video; *"Assuming same quality would like video for non verbal communication, but if audio quality would improve, audio only"*

General Observation: *"Background noise disturbing, headset quite/loud. Didn't know if I could adjust the sound or if it was part of the test. Very difficult to determine quality difference. Electronic test on screen might have been useful. a script or material to read may have helped or useful to know how easy/effective to convey instruction via this medium"*

Subject 6 Prefer to have audio only; *"Actually it would be preferable to have both if they work well, but the video is so bad that sometimes it is better not to have the video because in real life there could be misinterpretation from the other side. Having said that, in some of the test, it was very nice to have both, but only when the distortion is not to much and you can still see the other side."*

General Observation: *"Without knowing what parameters you had done during each test, it is difficult to have any comment. The only thing noticeable was there was sometimes echoes, changing video/audio quality from very good to very bad just in one second (during one test). Sometimes audio is going very high/low quickly, but there's no place to give such comments on each test."*

Subject 7 Prefer to have audio only; *"When audio and video are not synchronised, it produce an annoying effect (if the lag is large enough), which distracts me. As the image is normally static, good quality audio could be enough."*

General Observation: *"Sometimes the video quality is so bad that it is impossible to appreciate the synchronisation, and then the answer is 'I could not say'. I would give a better assessment if first I would be given different image and sound references, from bad quality to the best"*

Subject 8 Prefer to have audio and video; *"... because it is good to see expressions accompanying the voice (even if there is a delay between the two)"*

General Observation: *"It is very difficult to find variation in the audio and video. The findings are similar in every test. My picture clearer than my partner's but there seems to be more of a time lag with the voice before the video."*

Subject 9 Prefer to have audio and video; *"I prefer to see my (communicative) partner"*

General Observation: *"Video quality was not good. There is some audio delay slightly and some test was not clear."*

Subject 10 Prefer to have audio only; *"Poor video quality can be very annoying and it's better to have good audio"*

General Observation: *"Good audio quality, with good image but low refresh rate should be alright".*

Subject 11 Prefer to have audio and video; *"Prefer audio and video but only when both are working correctly. If the bandwidth is low I would prefer audio only rather than have poor audio and poor video."*

General Observation: *"a little worried about the highly subjective nature of the test. Sometimes difficult in obtaining a fair comparison of results for the different compression algorithms."*

Subject 12 Prefer to have audio and video; *"It's more realistic, enjoyable simply because you got to see the person you're talking to."*

Subject 13 Prefer to have audio and video; *"Even if video isn't always good (fast) it's better to interact with the other by seeing their reactions."*

Subject 14 Prefer to have audio only; *"In this context the video does not add useful information."*

General Observation: *"At the start of test - it's difficult to 'calibrate'."*

Subject 15 Prefer to have audio and video; *"I like to see the person I am talking to. Seeing helps my hearing."*

General Observation: *"The impairments are very much CONTENT DEPENDANT. With desktop videoconferencing the audio is the most important component."*

Subject 16 Prefer to have audio and video; *"Better interaction"*

General Observation: *"Too many variables: text to read, image movement, etc. There was some delay between audio and video even at the original image."*

Subject 17 Prefer to have audio only; *"Video did not add anything to communication, just an annoying distraction as the quality was not good enough."*

Field Trials: Study I

Subject 1 Prefer to have audio and video; *"more effective (conversation)"*

General Observation: *"A lot of synchronisation error when connected. Anyway, its a good test - gain some experience and knowledge through this test."*

Subject 2 Prefer to have audio and video; *"we can see and make the real 'connection' with our partner."*

General Observation: *"simple, easy to use - (videoconferencing) tools but very effective (useful)..."*

Subject 3 Prefer to have audio and video; *"it is more interesting to see the person you're talking to."*

General Observation: *"It was pretty good considering the other person was in another continent."*

Subject 4 Prefer to have audio and video; *"the conversation - more real and effective."*

General Observation: *"Simple test but difficult to arrange or set the connection."*

Subject 5 Prefer to have audio and video; *"more effective (conversation)"*

General Observation: *"Good idea to have a connection (communication link) like this. Problem when dialling."*

Subject 6 Prefer to have audio and video; *"makes a real conversation."*

General Observation: *"take time to get the connection right."*

Subject 7 Prefer to have audio and video; *"As long as I can see the (communicative) partner, it's a very good thing for me - more effective (conversation) and almost feel real*

General Observation: *Regarding the test, I do't have any comments. Anyway, it's good to make some improvement regarding the trouble when connected, but good attempt."*

Subject 8 Prefer to have audio and video; *"easier to understand"*

General Observation: *"Got a lot of interrupt and problem when dialing"*

Subject 9 Prefer to have audio and video; *" Adds to the richness of the communications."*

General Observation: *"The delay made the communication difficult."*

Subject 10 Prefer to have audio and video; *" - can communicate more effectively - real conversation."*

General Observation: *"Feel bad when suddenly (the line) - disconnected"*

Subject 11 Prefer to have audio and video; *"because visual (information) is important."*

General Observation: *"Good experience, needs more time to make a good judgement."*

Subject 12 Prefer to have audio only; *"without video, the flow of conversation is smoother."*

General Observation: *"This kind of communication link is excellent."*

Field Trials: Study II

Subject 1 Prefer to have audio and video; *"occasionally, problems with audio were easy to ignore as the video helped to compensate."*

General Observation: *"more structured conversation - good attempt."*

Subject 2 Prefer to have audio and video; *"It was interesting that this time I did prefer to have video and I think this was because I did not already known the other person."*

Subject 3 Prefer to have audio and video; *"because easy to communicate."*

General Observation: *"This is the first time (using desktop videoconferencing) and maybe the quality of video can be improved for the time being"*

Subject 4 Prefer to have audio and video; *"Depending on current situation and environment, for better view or explanation video is a compliment because the body language is best under certain circumstances..."*

General Observation: *"The quality of audio and video is of good but video frame rate is an acceptable state - only with certain distortion other qualities are satisfactory."*

Subject 5 Prefer to have audio and video; *"At least I can save ISDN video conferencing (if I have one considering the bill and the equipment is very very expensive) cost."*

General Observation: *"The quality is good except for the lagging of the audio."*

Subject 6 Prefer to have audio and video; *"because whenever you talking to somebody, if using video conference, you need to see the face of the person you're talking to."*

General Observation: *"Sometimes, the video image is distorted."*

Subject 7 Prefer to have audio and video; *"because I like talking with someone when I can see their faces. I think this is more effective."*

General Observation: *"From my observation, the quality of audio and video was poor..."*

Subject 8 Prefer to have audio and video; *"because I feel much closer to person whom I talking to."*

General Observation: *"In my observation, the overall audio/video test performance was bad. Maybe it need a little bit more improvement in audio and video quality."*

Subject 9 Prefer to have audio and video; *"because I prefer talking or communicate with visual right in front of me. The distance between us was much closer if I can see the person..."*

General Observation: *"From my observation the quality of both audio and video was poor... Maybe it need a little bit more improvement in audio and video quality."*

Subject 10 Prefer to have audio and video; *"Because I like to talk with someone and see his face too. It is easier to understand the information from the sender."*

General Observation: *"From the test, I notice that there is error/noise in video and audio output."*

Subject 11 Prefer to have audio and video; *"Make better communication and easy to understand."*

General Observation: *"Audio and video must be set properly to get better audio and video. Suppose students are allowed to use video conferencing at least two times in semester to learn more knowledge from the other side (UK) - to exchange data or knowledge."*

Subject 12 Prefer to have audio and video; *" More comfortable, and also I can know the situation there."*

General Observation: *"Good video can help enhances the conversation."*

APPENDIX C

Further information on NIST Net
(obtained from http://snad.ncsl.nist.gov/itg/nistnet/)

National Institute of Standards and Technology NIST

... working with industry to develop and apply technology, measurements, and standards

How to use the NIST Net emulation package

Before running NIST Net, the kernel emulator module must be installed through **Load.Nistnet** or **insmod nistnet**. You can add this to */etc/rc.d/rc.modules* if you wish.

To control the emulator, several tools are included in the package.

Cnistnet

Cnistnet is the new command-line interface. It takes the following arguments:

-u up (on)
-d down (off)
-a src[:port[.protocol]] dest[:port[.prot]] [cos]
add new
[--delay delay [delsigma[/delcorr]]]
[--drop drop_percentage[/drop_correlation]]
[--dup dup_percentage[/dup_correlation]]
[--bandwidth bandwidth]
[--drd drdmin drdmax [drdcongest]]
-r src[:port[.prot]] dest[:port[.prot]] [cos]
remove
-s src[:port[.prot]] dest[:port[.prot]] [cos]
see stats
-S src[:port[.prot]] dest[:port[.prot]] [cos]

see stats continuously
[-n] -R
read current settings (-n numerical format)
-D value
debug on (value=0 none, 1 minimal,... 9 maximal)
-U debug off
-G global stats
-K kickstart the clock
-F flush the queues
-h this help message

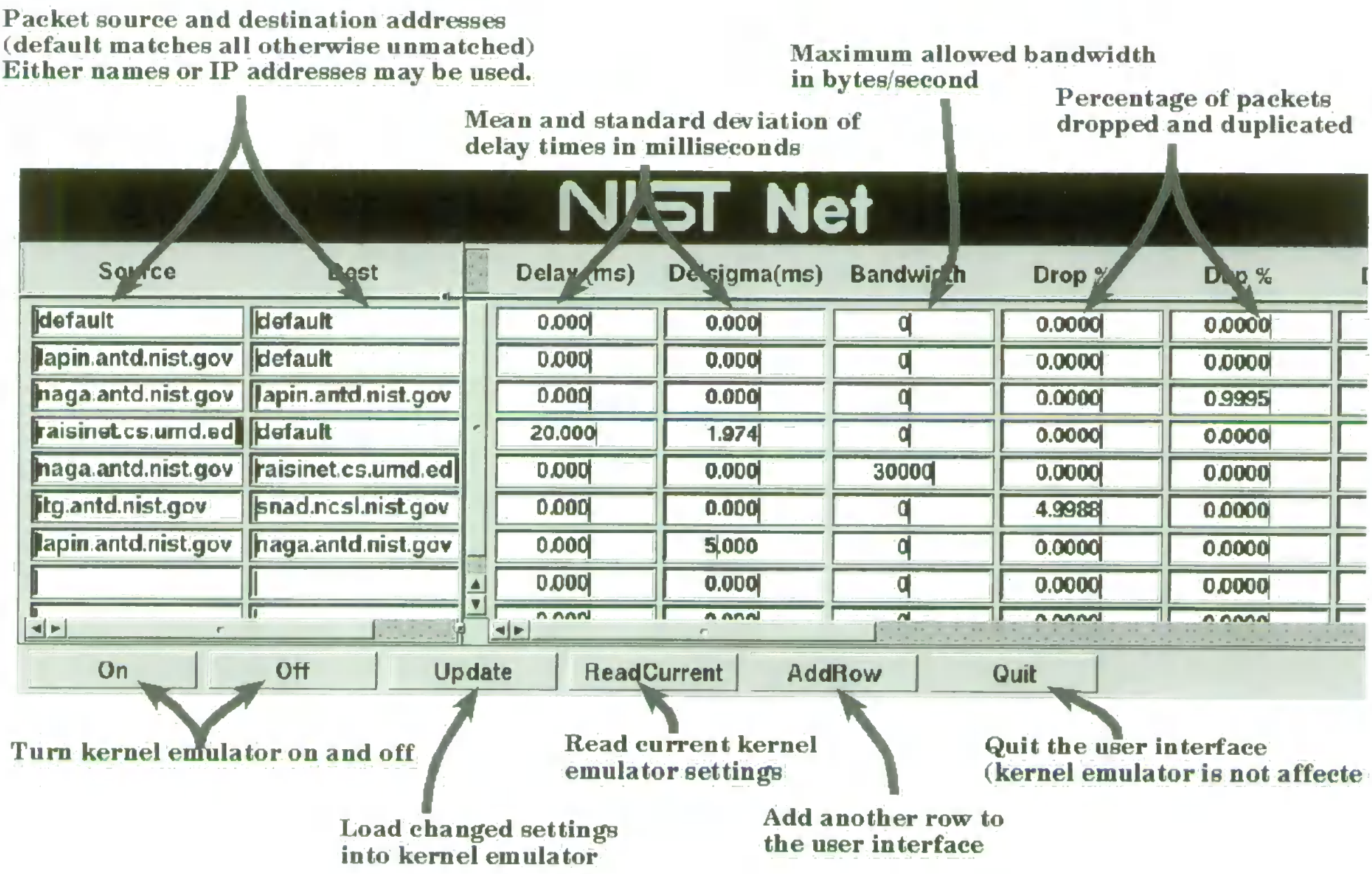
Hitbox

Hitbox is the old command-line interface. It takes the following arguments:

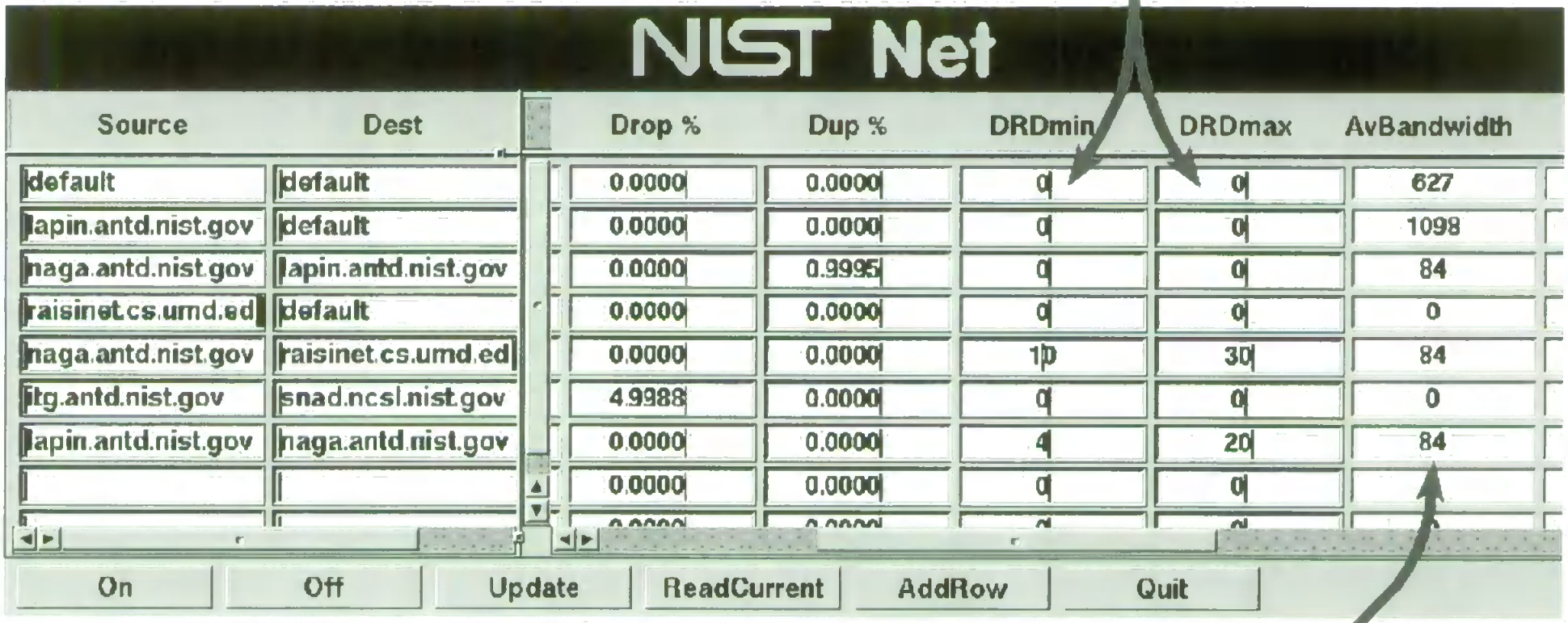
-u up (turn emulator on)
-d down (turn emulator off - entries are retained)
-a src dest delay delsigma bandwidth drop dup drdmin drdmax
add new entry
-r src dest
remove entry
-s src dest
see stats for entry
-S src dest
see stats for entry continuously
-R read current list of entries in kernel
-D debug on
-U debug off
-G see global stats

Xnistnet

Nistnet is the GUI version of the user interface. It provides for control and monitoring of multiple entries. These diagrams show the controls **xnistnet** offers. *Note: these need to be updated!!*


Packet source and destination addresses
(default matches all otherwise unmatched)
Either names or IP addresses may be used.
Mean and standard deviation of
delay times in milliseconds
Maximum allowed bandwidth
in bytes/second
Percentage of packets
dropped and duplicated
NIST Net
Source
Dest
Delay (ms)
Delsigma(ms)
Bandwidth
Drop %
Dup %
default default 0.000 0.000 0 0.0000 0.0000
lapin.antd.nist.gov default 0.000 0.000 0 0.0000 0.0000
naga.antd.nist.gov lapin.antd.nist.gov 0.000 0.000 0 0.0000 0.9995
raisinet.cs.umd.ed default 20.000 1.974 0 0.0000 0.0000
naga.antd.nist.gov raisinet.cs.umd.ed 0.000 0.000 30000 0.0000 0.0000
itg.antd.nist.gov snad.ncsl.nist.gov 0.000 0.000 0 4.9988 0.0000
lapin.antd.nist.gov naga.antd.nist.gov 0.000 5.000 0 0.0000 0.0000
0.000 0.000 0 0.0000 0.0000
On
Off
Update
ReadCurrent
AddRow
Quit
Turn kernel emulator on and off
Load changed settings
into kernel emulator
Read current kernel
emulator settings
Add another row to
the user interface
Quit the user interface
(kernel emulator is not affecte

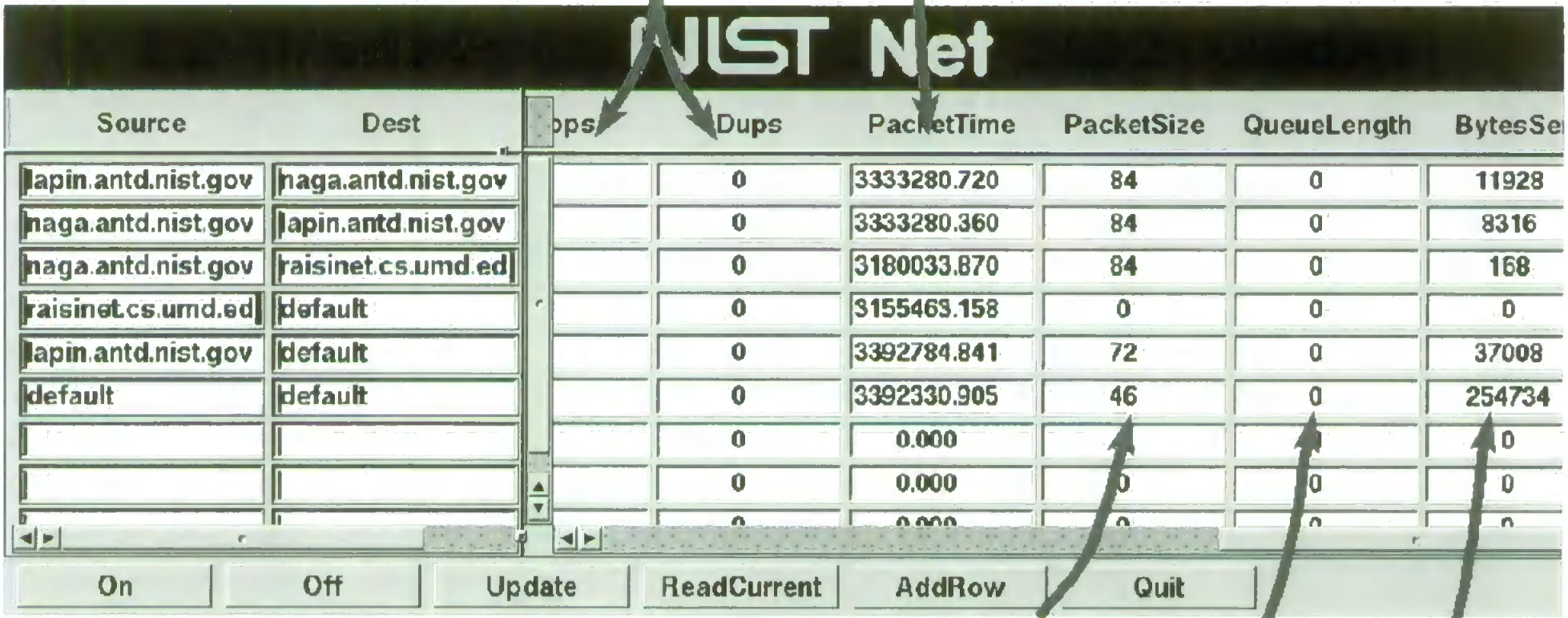
Derivative random drop (DRD)-style parameters:
No packets dropped if queue length under DRDmin
95% dropped if queue length greater than DRDmax
Drop percentage ramps up between two values



Running 10-second average of observed active bandwidth utilization

Kernel time last packet was received in milliseconds
(Displayed as 32-bit scaled microsecond value, hence wraps around every 72 minutes. Actually stored as a 64-bit value.)

Number of packets dropped and duplicated



Size of last packet received in bytes (including all headers)

Number of packets in wait queue

Total number of bytes handled for t source/destination pair

APPENDIX D

List of Publications

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.

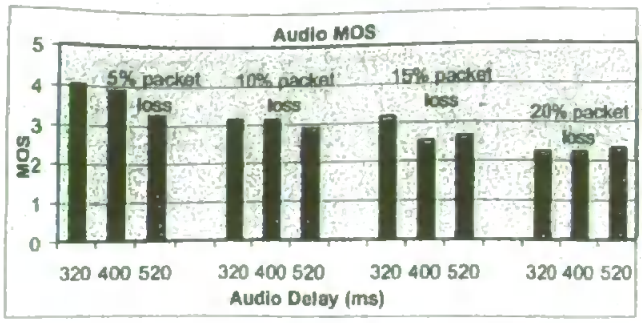
RESULT

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.

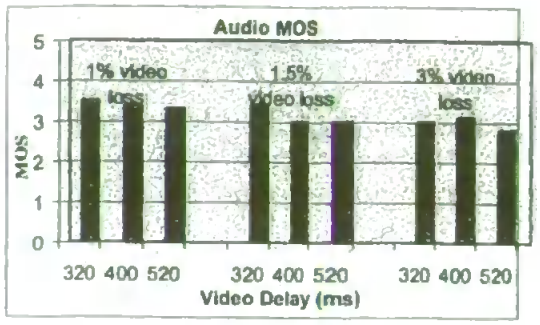
Graph 2 shows the MOS of the perceived audio for the Interactive Test, for video packet loss of 1%, 1.5% and 3%. It can be seen that the degradation of video quality, due packet loss and delay, has a significant impact on the perceived Audio quality. At 1% video loss, the MOS drop from 3.5 to 3.3, i.e.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

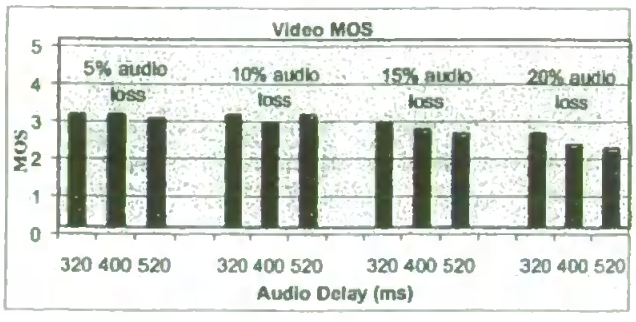
Graph 3 shows the perceived Video MOS for the interactive test for audio packet loss of 5%, 10%, 15% and 20%. The MOS are around 3 (FAIR) for audio loss of 5% and 10%. It was observed that the video MOS decrease as the audio packet loss increases, i.e. from 15% to 20%, while the quality settings of the 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.



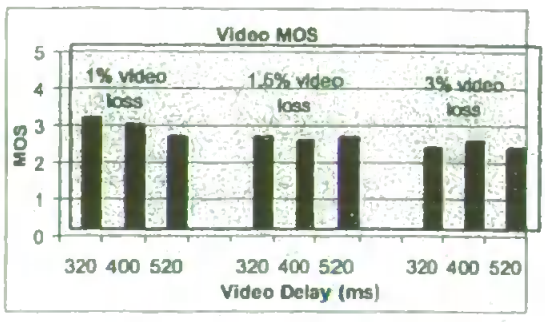
Graph 1: Audio MOS Vs Audio Delay and Loss



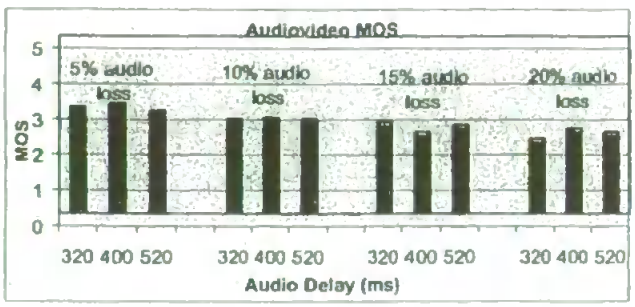
Graph 2: Audio MOS Vs Video Delay and Loss



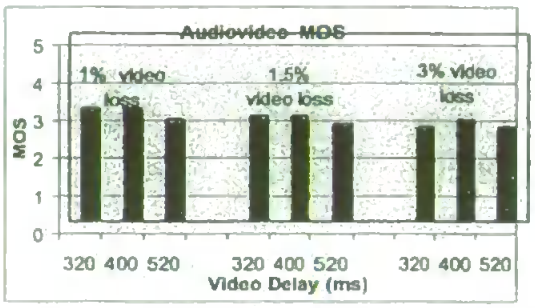
Graph 3: Video MOS Vs Audio Delay and Loss



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

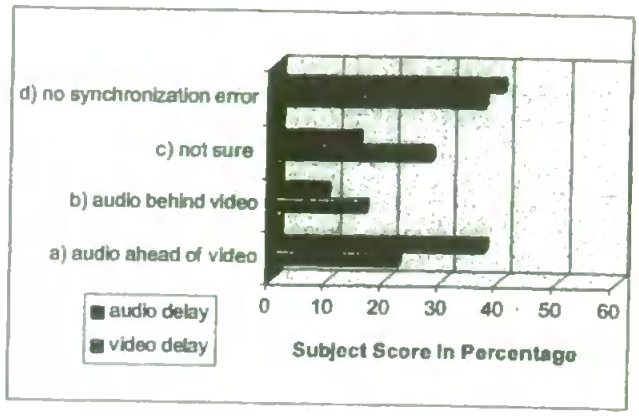
Graph 4 shows the MOS of the perceived quality of video for video loss of 1%, 1.5% and 3%. It has been noticed that, for 1% video loss there is a gradual degradation of the perceived video score as the video delay increases from 320ms to 520ms. It is al so suggested that the result becomes less meaningful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

THE EFFECTS OF LIP SYNCHRONIZATION IN IP CONFERENCING

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

University of Plymouth, United Kingdom

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

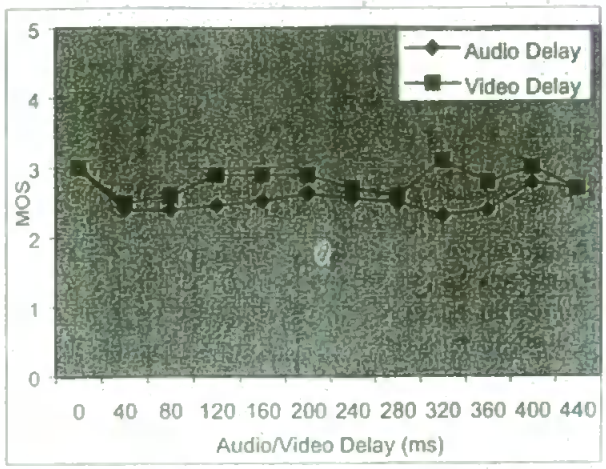


Figure 2: Interactive Test-Audiovideo Overall MOS

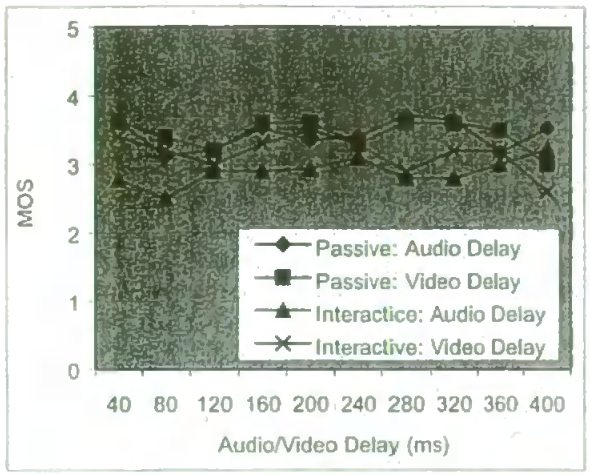


Figure 3: Audio MOS-Interactive Vs Passive Test

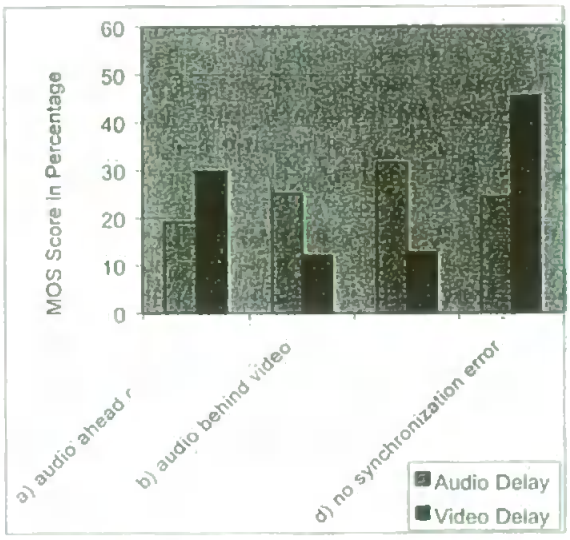


Figure 4: Passive Test – Audio Vs Video Delay

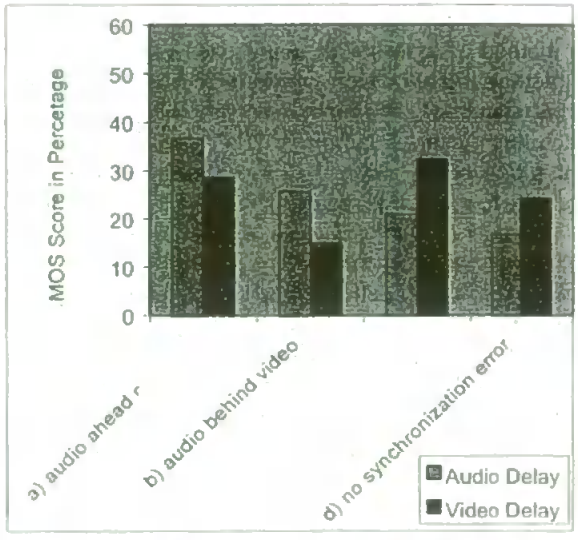


Figure 5: Interactive Test – Audio Vs Video Delay

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.

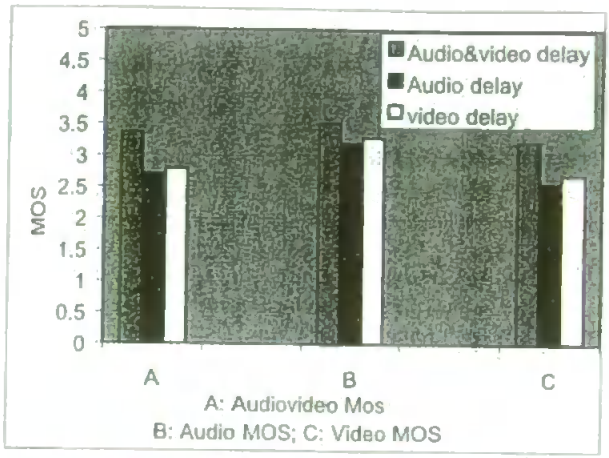


Figure 6: Combined Audio and Video Delay Vs Separate Audio and Video Delay

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.

The effects of audio and video correlation and lip synchronization

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

The authors

L. Mued, B. Lines, S. Furnell and **P. Reynolds** are all in the Network Research Group, Department of Communication and Electronic Engineering, University of Plymouth, Plymouth, UK.

Keywords

Multimedia, Task analysis, Correlation analysis, Audiovisual media, Sensory perception

Abstract

This paper investigates the interaction effect of audio and video, and studies 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.

Electronic access

The Emerald Research Register for this journal is available at
http://www.emeraldinsight.com/researchregister

The current issue and full text archive of this journal is available at
http://www.emeraldinsight.com/1065-0741.htm

Campus-Wide Information Systems
Volume 20 · Number 4 · 2003 · pp. 159-166

· ISSN 1065-0741
DOI 10.1108/10650740310491333

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 the 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 synchronization.

Current desktop videoconferencing systems transmit between two and eight frames of video per second (quarter common interchange format, QCIF 176 × 144 pixels/common interchange format, CIF 352 × 288 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 interaction (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 classes 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

Campus-Wide Information Systems
Volume 20 · Number 4 · 2003 · 159-166

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 the 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 prioritize 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, first, Experiment A: investigate interaction effects between perceived quality of audio and video and second, Experiment B: study the effects of lip sync on multimedia quality.

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

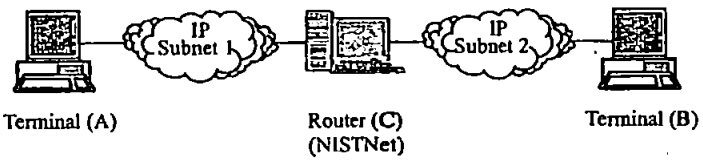
As previously stated, the experiments were based upon investigating a potential interaction effect between audio and video media in DVC systems in the presence of packet loss. The approach is to send the audio and video component with respect to the assigned quality for each media, in two different task 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 depicts the VoIP (voice over IP) test bed configuration used for the experiments, and the various elements illustrated are described below.

Terminal A and 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: 200MHz Pentium processors, 64MB RAM, were used. The QCIF – 176 × 144 pixels

Figure 1 Test bed configuration



frame size is used. The video setting was unchanged throughout the test, which was "better quality" video. For the audio CODEC, a G723.1, 6,400bit/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 and 2: IP networks
The test activity of the project is organised in a number of steps. First, tests are carried out under an error free network environment. Second, different sets 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.

The test was conducted on two different task scenarios, i.e. interactive test and passive test.

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 maximize 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 audio, video and combined audiovideo. The discussions were limited to two minutes. For control purposes, initially, tests were carried out under error-free conditions, i.e. 0 per cent packet loss.

Passive test
A total 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 audio, video and 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 configurations 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. audio is ahead of video; audio is behind video; cannot tell if audio is ahead or lagging; and, 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.

Table I Packet loss of video (v) and audio (a) under test, in percentage

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		

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 five-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 at 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 per cent 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. Figures 2-9 show the results obtained from Experiment A, while Figure 10 is deduced from Experiment B.

Figures 2 and 3 show the MOS of packet loss impact on the perceived quality of video, audio

Figure 2 Interactive test – video degraded, audio constant

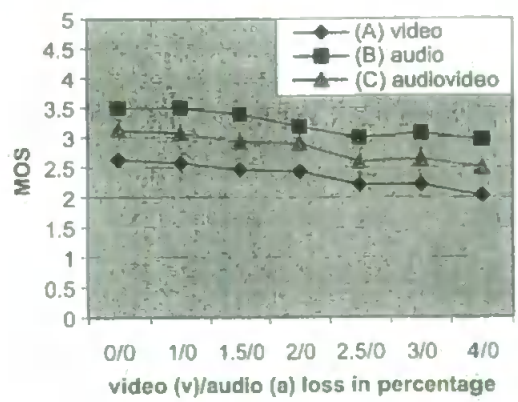


Figure 3 Interactive test – video constant, audio degraded

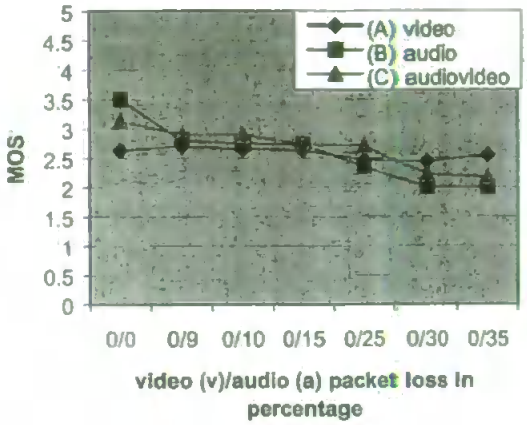


Figure 4 Passive test – video degraded, audio constant

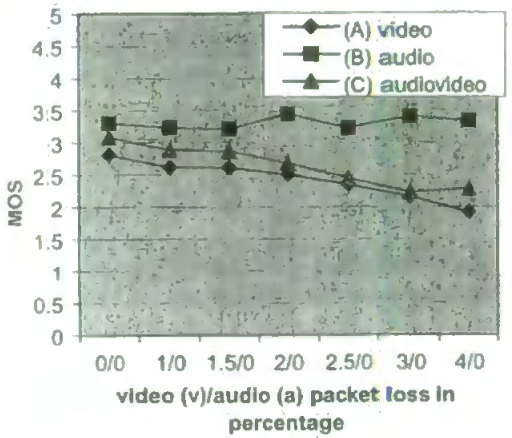


Figure 5 Passive test - video constant, audio degraded

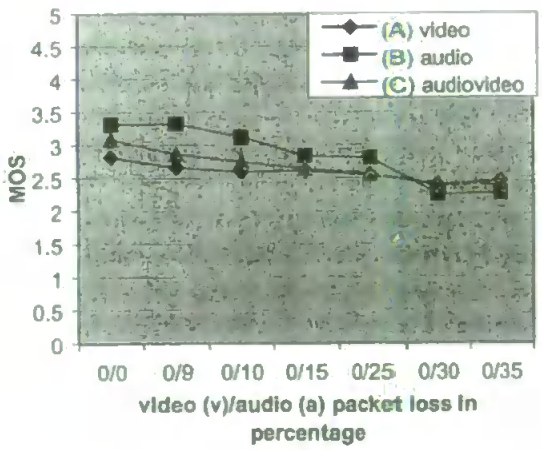


Figure 6 Passive test – packet loss impact on audio and video

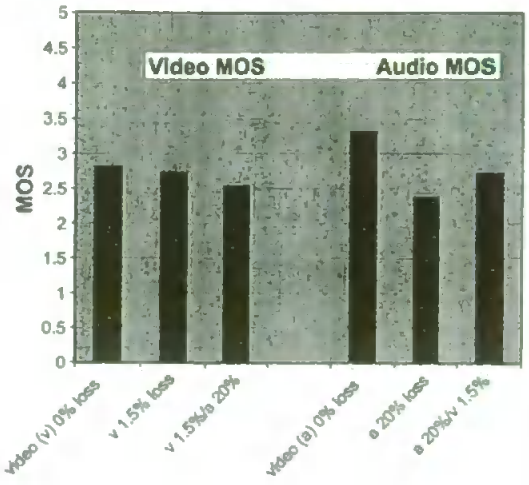


Figure 7 Interactive test – audio and video degraded

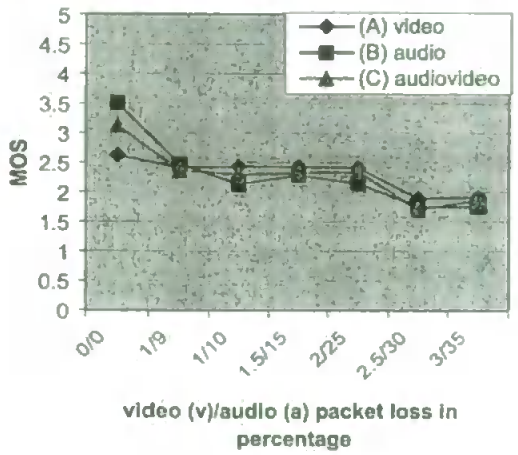


Figure 8 Interactive test – video poor (4% loss), audio degraded

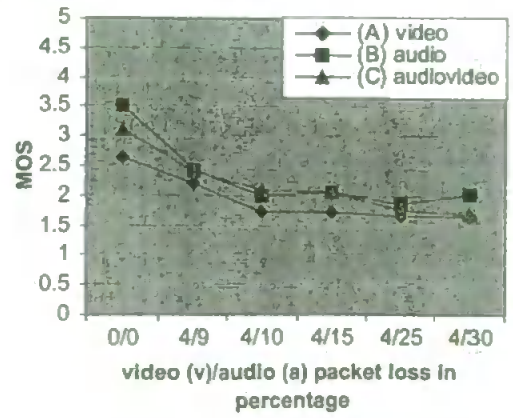
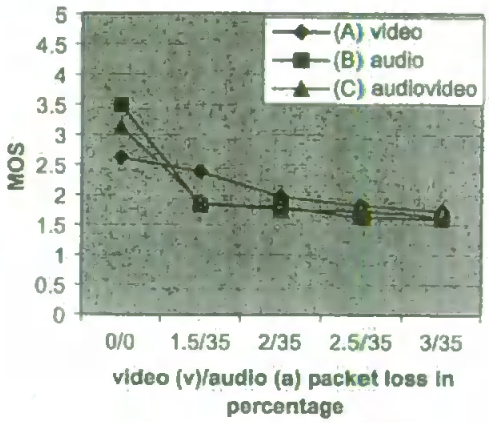


Figure 9 Interactive test – video degraded, audio poor (35% loss)

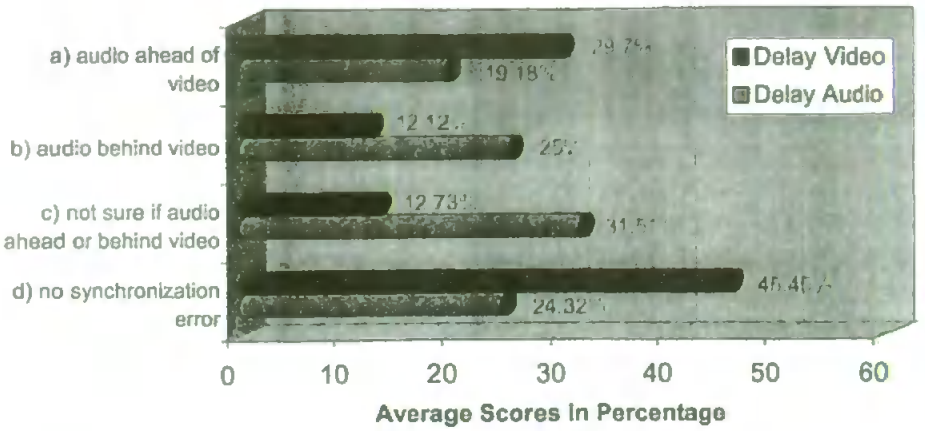


and combined audiovideo as obtained in the interactive test. It can be seen that when video is degraded, audio scores also decreased by 0.5 (MOS), for a video packet loss in the range of 1-4 per cent, even though the audio quality was kept constant. However, the MOS for video, while its quality being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

Figures 4 and 5 show the 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 ranges from 0-35 per cent. The MOS for combined audiovideo is affected severely by the change in video loss as compared to audio loss.

Figure 6 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 when audio and video are both present during the test. The

Figure 10 Effects of delay video or audio on lip sync



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 per cent audio loss (the fifth 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 an 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 per cent 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 a higher rating with good video (i.e. 0 per cent loss), even though the audio was degraded by the same amount of loss through out the test.

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 per cent) and high audio loss (35 per cent). Figure 9 shows that, when audio is very poor, interactive test scores very low MOS for the perceived multimedia quality. Hence, the interactive test heavily depends on sufficient audio quality.

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

For the test where audio was delayed up to 440msec, 25 per cent of the subjects noticed that audio was played behind video and only 19.18 per cent of them claimed that audio was ahead of video. Whereby, a number of 31.51 per cent of the subjects were not sure if audio was ahead of video or behind video and 24.31 per cent stated that there was no synchronization 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 the 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 importance than that of the audio. This finding also confirmed the previous research result which states that 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, the increase in task difficulties have the effect of decreasing the subjective video and audio quality. For example, in the passive test, where users are required to understand the read material, the overall scores for the combined audio and video quality in the passive test are much lower than those in the interactive test.

A number of problems were encountered while conducting the tests. For example, task performance was dynamically varying. This could lead to varying frame rates that could result in inconsistencies 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 differing sensory and perceptual abilities 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 of 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 four 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 vice 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*, Vol. 52 No. 1, pp. 165-87.

Carson, M. (2000), *NIST Net Home Page*, available at: http://snad.ncsl.nist.gov/itg/nistnet/

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. (Eds), *Intellectual Teamwork*, Lawrence Erlbaum, Hillsdale, NJ, pp. 291-325.

Hardman, V., Sasse M.A. and Kouvelas, I. (1998), "Successful multiparty audio communication over the Internet", *Communications of the ACM*, Vol. 41 No. 5, pp. 74-80.

International Telecommunications Union (ITU) (1996), "Methods for subjective determination of transmission quality", *Recommendation P.800*, ITU-T.

Isaacs, E. and Tang, J. (1994), "What video can and cannot do for collaboration: a case study", *Multimedia Systems*, Vol. 2, pp. 6-73.

Kawalek, J. (1995), "A user perspective for QoS management", *Proceeding of the 3rd International Conference on Intelligence in Broadband Services and Network*, IS & N, Crete.

Kitawaki, N. and Nagabuchi, H. (1998), "Quality assessment of speech coding and speech synthesis systems", *IEEE Communications Magazine*, October, 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.

PERFORMANCE EVALUATION OF DESKTOP VIDEOCONFERENCING

Licha Mued, Benn Lines, Steven Furnell and Paul Reynolds

Network Research Group, Department of Communication and
Electronic Engineering, University of Plymouth, United Kingdom
Email: lmued@jack.see.plym.ac.uk, nrg@jack.see.plymouth.ac.uk

Abstract

This paper discusses the evaluation of multimedia conferencing systems with respect to audio and video quality. We investigated various aspects that affect the perceived audio and video quality in desktop videoconferencing (DVC), including network constraints (packet loss, delay jitter and delay), CODEC, and task performance. The paper outlines the benchmarking of Microsoft NetMeeting, using the subjective assessment method of Mean Opinion Score (MOS) and concludes that the various factors mentioned above have different effects and that video is the main determinant of the overall perceived audiovisual quality.

1. Introduction

The state of the art in audio and video assessment has generally focused on finding the point at which degradation is not discernable. Audio and video quality can be measured either subjectively or objectively. It is generally agreed that subjective methods are more reliable [1], but recent research findings suggest that the subjective method alone is inadequate to determine the audio and video quality in videoconferencing [2,3,4].

At present, despite the increased popularity of low cost DVC, it is often questioned whether the quality of the audio and video provided is adequate to the tasks that users wish to perform performance. Some findings suggest that the perceived quality of audio and video varies according to the task undertaken and user expectation also varies accordingly. In the meantime, the issue of determining multimedia conferencing quality has certain difficulties, as there is no recognized industry standard of what really determines audio and video quality. In addition, assessing the quality of multimedia over the Internet is further complicated due to its constantly changing and unpredictable nature [2]. Many research efforts are now being directed toward developing new approaches in assessing audio and video quality in IP multimedia [4,5,6].

The authors are conducting research into quality of service in IP videoconferencing scenarios. The research to date has investigated the current state of the art in desktop videoconferencing. This paper focuses upon benchmarking the performance of the popular Microsoft NetMeeting with respect to the related issues that affect the perceived audio and video quality, such as network congestions, computing resources, tasks performance, CODEC, and conferencing hardware. NetMeeting was selected over other existing IP telephony tools due to its readily available software and its highly demand in the current market. The associated study had three main aims:

(1) To investigate the performance of NetMeeting. This involved assessment of audio quality, video quality and combined audio and video quality under a real network environment, and also under assigned network constraints.

(2) To investigate the task performance effects on audio quality, video quality and the combined quality of audio and video under a real network environment, and also under assigned network constraints.

The purpose of the test in (1) and (2) was to quantify traffic related phenomena, such as packet-loss (i.e. the number of lost packets, reported at the total traffic), delay (i.e. the time passed between the sending of a packet and its arrival at the destination), and delay jitter (i.e. the variance of the delay value) and to present the data in the form, which could be related to perceived quality of service experienced using NetMeeting software.

(3) To investigate the impact of the two speech CODECs (PCM and G723.1) on perceived audio and video quality.

The method of assessment being used is an objective test method, called Mean Opinion Score (MOS) and is the standard recommended by the ITU-T (CCITT, 1984). The MOS is typically a 5-point rating scale, covering the options Excellent (5), Good (4), Fair (3), Poor (2) and Bad (1).

2. The experiment

2.1 The NetMeeting System

NetMeeting is a Windows application that allows real-time communications, offering audio, video, and data conferencing functionality [7]. In NetMeeting, like most videoconferencing systems, the audio and video signals are encoded and transmitted using two separate TCP/IP sessions, and then reassembled at the receiving end. It is designed for use on the Internet and other IP networks. Video transmission requires camera and video capture card. NetMeeting uses the H.263 standard for video compression and, for our experiment, we used the Quarter Common Interchange Format (QCIF-176x144) frame size, as Common Interchange Format (CIF-352x288) provided an almost static picture.

2.2 Method

Ten pairs of volunteers from the university were involved in the test. The subjects had very little experience (if any) of using the software. Previous research implies that informal communication tends to be representative of individuals who are familiar with each other [8]. Hence, to maximise task motivation and to ensure subjects were fully at ease with each other, subjects who were acquainted with one another were selected and they were allowed to select their own issues for discussion.

The tests can be classified as below:

(i) Stage 1; Passive Communication (i.e. viewing and listening to a 'talking head', without interaction), was to benchmark the quality of: (a) Audio only, (b) Video only and (c) Audio and video overall.

(ii) Stage 2; Interactive Communication (where a variety of contents were introduced for example, intensive informal communications, moving object and using the shared workspaces, for example; sending and receiving large text files), was to benchmark the effects of content on the quality of: (a) Audio only, (b) Video only and (c) Audio and video overall.

In this stage, Stages (1) and (2) above were repeated, but tested under various network constraints. Different aspects of network congestion under investigation were packet lost, delay and delay jitter. Stages (1) and (2) were also conducted using two different audio CODECs (μ-Law, PCM and G723.1 6400bit/s). The video quality was unchanged throughout the test.

(iii) Stage 3 was to evaluate the performance of NetMeeting in a real network environment. The performance of two configurations was compared, operating firstly with no network constraints (Config. 1 - no packet loss, delay or delay jitter) and then subsequently with impediments introduced (Config. 2 - 5% packet loss, 400msec delay and 20msec delay jitter).

2.3 Apparatus and procedure

The subjects were seated in the two separate rooms provided with 15 inch monitors. For the experiments, Pentium III 933 MHz systems with128MB (CDRAM) were used and a USB video blaster Webcam Plus (capable of video capture up to 30 frames per second @ 352x288 pixels and 15 frame per second @ 640x480 pixels), was mounted on each monitor. To send and receive audio, two identical Platronic PC headsets were used. The audio channel was a 64kb/s full duplex.

The audio and video were transmitted from one end to another using a testbed containing the NIST Network Emulation Tool (NISTNet) - a general-purpose tool for emulating performance dynamics in IP network [9]. Each subject was provided with a self-view window, a remote view window and a talk window to send text to the remote partner.

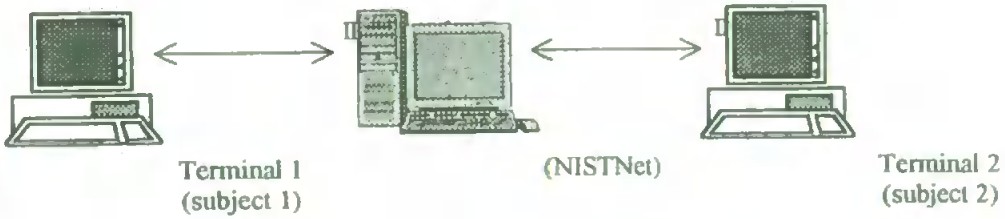


Figure 1: Testbed configuration

3. Result

Figure 3.1 shows loss effects on video. Video was found to be error free for packet loss below 3%, (MOS is between 2.9 and 3.5). At 4%, MOS drops to 2.6 or 2.7, and video degradation can be perceived. On approaching 8% to 30% packet loss, error in the video then became apparent and became unusable. Overall MOS for video, obtained by system using G723.1, shows lower results although video settings are not changed. This raises suspicions as to whether the change in audio quality would cause the change in perceived video quality and vice-versa.

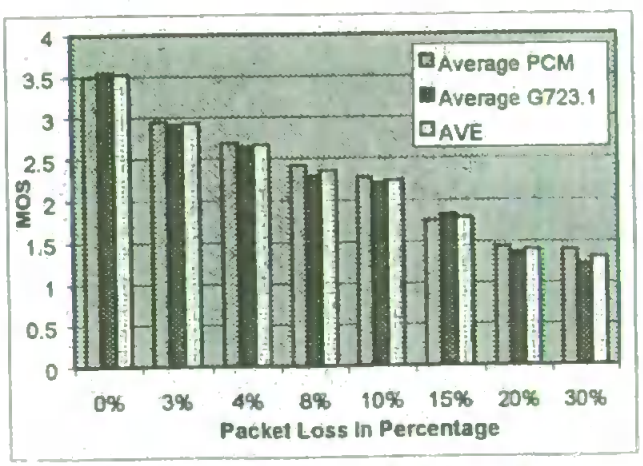


Figure 3.1: Packet Loss Effects on Video

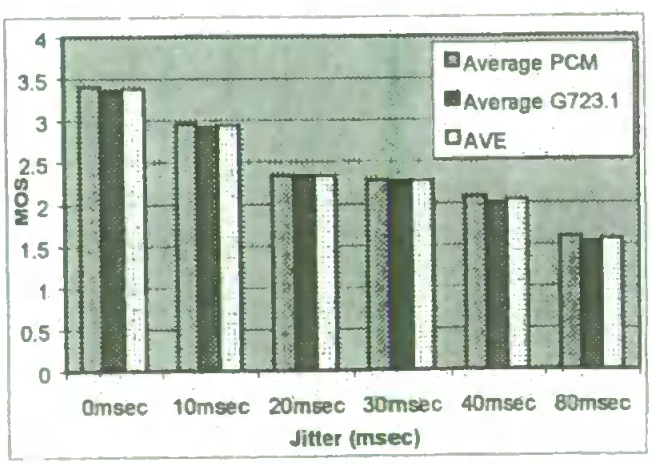


Figure 3.2: Jitter Effects on Audiovisual Overall

The MOS rating for loss effects on audio, for system that using PCM (μ-Law), is Good (MOS 3-4) at 0% to 8%. Whereby, the system employing G723.1 produced lower MOS (2.5) at 8% loss. This indicates that systems employing PCM (μ-Law) are more tolerant to packet loss. Packet loss on audiovisual overall is generally the same as that on video.

Figure 3.2 shows jitter effects on perceived audiovisual quality overall. Between 0msec-10msec, audiovisual quality is Good, i.e. MOS 3.4-2.9 for PCM and 3.3-2.7 for G723.1. At jitter between 20msec – 30msec, both systems score barely above Poor quality (i.e. 2.2-2.3 MOS). At 40msec jitter and above, the quality is Poor. For jitter effects on video, at 10msec, the scores are just below the Good (3) threshold, i.e. 2.81 (PCM) and 2.79 (G723.1), although the quality is still acceptable. At 20msec, MOS are 2.5 (PCM) and 2.4 (G723.1). At 30msec jitter, both systems produce Poor video. As expected, PCM gives better results, for jitter effects on audio (i.e. at 0msec–30msec, MOS are 2.7-3.9). At 40msec, audio started to degrade and became poor at 80msec.

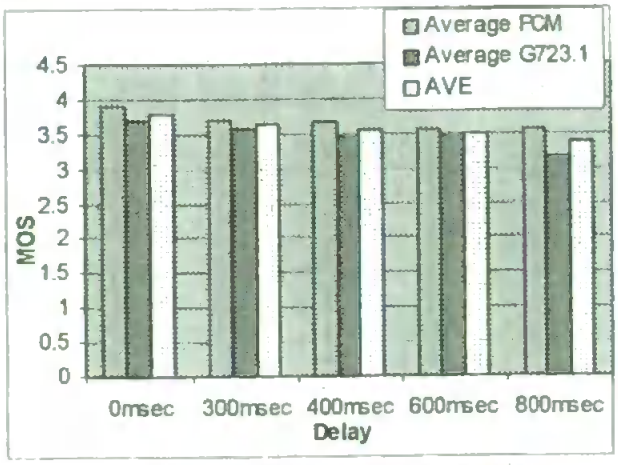


Figure 3.3: Delay Effects on Audio

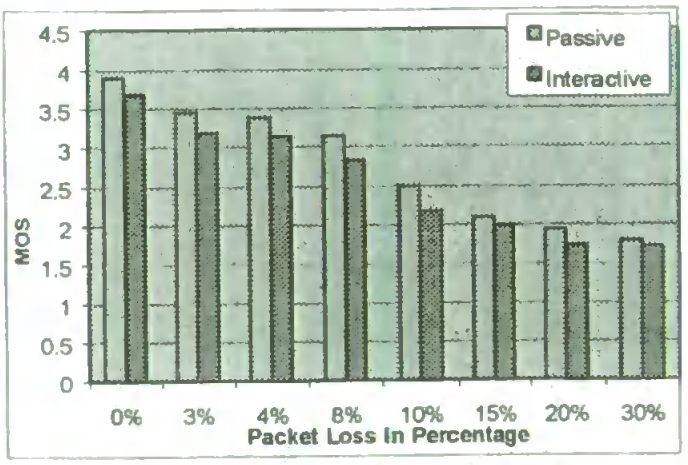


Figure 3.4: Packet Loss Effects on Audio Passive Vs Interactive Communication

Figure 3.3 shows delay effect on audio. As can be seen from the result, audio is less susceptible to delay as compared to packet loss and jitter. The MOS are around 3.2 to 3.7 for the range of 0msec to 800msec delays. Delay also has little impact on video. For example, for the range of 0msec to 800msec delay, MOS are around 2.9 and 3.6. The same pattern of results is repeated for audiovisual overall.

Figure 3.4 shows loss effects on audio, comparing the results obtained from passive communication and interactive communication, as suggested in Stage 2. It is evident that passive communication produces higher MOS than interactive communication. Results given by video quality and audiovisual quality overall follow the same pattern of results given by loss effects on audio, but with lower MOS.

Figure 3.5 shows the result of Stage 3. As expected, system Config. 1, as in Stage 3 produces much higher MOS (3.3-3.8). However, the overall score for System Config. 2, under congested network, is Poor (MOS 1.8-2.4). The overall rating for general performance of NetMeeting under ideal network is either Good (4) or Fair (3).

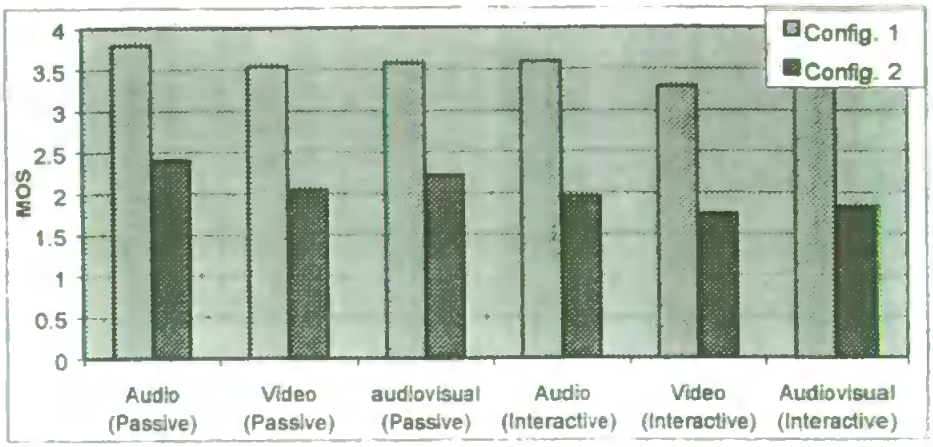


Figure 3.5: System Config. 1 vs System Config. 2

4. Discussion

Our observation indicates that, to assess video quality in videoconferencing is a very complicated issue since the frame rates are constantly changing. Subjects found it hard to give the score for video in a limited period of time. Some subjects felt dissatisfied in giving only one score, as their mind fluctuated between two or more scores from one moment to another during the test. This is due to the fact that the frame rate is high if the subject is relatively still, but as the movement of the subject increases, then either frame rates will vary inconsistently or the amount of video artifacts are likely to increase. It was observed that frame rate varies with the motion content, the level of detail, and the percentage of image that changes from one frame to another. However, assessing audio quality is less complex since the audio degradations are relatively significant as the network constraint increases.

When evaluating audio quality only, the MOS rating is high but when evaluating the combined quality of audio and video, the rating drops to almost similar to video rating. This implies that video quality contributes an important element in benchmarking the overall performance of desktop conferencing system.

By considering the result obtained in Stage 2 (refer to Figure 3.4), task performance has small effects on both audio and video, with the difference is only below 0.5 MOS. Our conclusion here, the task performance designed to carry out the assigned task was insufficient to obtain a more outstanding result. Another possibility is that the subjects were not fully trained and were therefore unable to perform the task exactly as required. In the future, each specific task performance must be carefully designed and the subject must be well trained so that the task will be conducted coherently in order to obtain a more reliable result. Generally, the result shows that the task performance effects on both audio and video became apparent for packet loss around 8% - 10%, for audio and around 4% - 10%, for video. Whereby, for jitter it is around 20msec and 40msec.

5. Conclusions and Future Work

Observations so far indicate that audio quality, video quality, and overall audiovisual quality are susceptible to packet loss and jitter, but are less susceptible to delay. Throughout the test, audio quality is higher when compared to video quality and overall audiovisual quality. Assessment of audio is very straightforward. Assessment of video, however, is very complex as its quality varied during the study, from very acceptable to almost useless.

Throughout the test, the best MOS rating for both audio and video is between Fair (3) and Good (4), although Good (4) MOS is seldom given. None of the subjects gave an Excellent (5) MOS rating. Thus, it is evident that NetMeeting or IP videoconferencing in general, is in its infancy with substantial improvements needed to achieve higher performance. We also learnt that traffic related network factors (such as packet loss, jitter, and delay), CPU power, CODEC, and task performance are all vital in maintaining the quality of video and audio service in NetMeeting.

The work presented in this paper will eventually lead to the characterisation of the factors that are necessary for audio and video optimisation in multimedia conferencing system. Future work will include, investigate more on issues such as lip synchronization. Currently, it is understood that lip synchronization is difficult to achieve in most low cost videoconferencing system due to the fact that the audio and video signals are transmitted via separate channels, and then reassembled at the receiving end. With this in mind, we will study a different approach in sending audio and video streams i.e. to combine them in the same packets. The work will then proceed to investigate the problems inherent in multi-modal transmission. A prototype system will be implemented in order to enable a baseline comparison against existing conferencing systems. The associated results will be the focus of future publications.

References

[1] Flanagan, J.L. (1965), Speech Analysis, Synthesis and Perception. Springer-Verlag, Berlin

[2] Gillian M. Wilson and M. Angela Sasse (2000), Do Users Always Know What's Good For Them? Utilising Physiological Responses to Assess Media Quality. In S. McDonald, Y. Waern & G. Cockton (eds.) Proceedings of HCI200: People and Computers XIV - Usability or Else! Springer, pp. 327-339, September 5th - 8th, Sunderland, UK. ISBN 1-85233-318-9.

[3] A. Watson & M. A. Sasse (1996), Evaluating Audio and Video Quality in Low-Cost Multimedia Conferencing Systems. Interacting with Computers, Vol. 8 (3), pp. 255-275.

[4] Wilson, G. & Sasse, M.A. (2000), Investigating the Impact of Audio Degradations on User, 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. ISBN 0-643-06633-0.

[5] Hollier, M.P., Rimell, A.N., Hands, D.S., Voelcker, R.M. (1999), Multi-modal perception, British Telecom Res. Labs., Ipswich, UK BT Technology Journal, vol.17, no.1, Jan. 1999

[6] Bouch, A., Watson, A. and Sasse, M.A. (1999) QUASS – A Tool for Measuring the Subjective Quality of Real-Time Multimedia Audio and Video. Proceeding of HCI '98, 1-4 September 1998, Sheffield, UK

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

[8] Isaacs, E. and Tang, J. (1994), 'What Video Can and Cannot Do for Collaboration: A Case Study', Multimedia Systems 2, 6~73

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

Author's Declaration

At no time during the registration for the degree of Doctor of Philosophy has the author been registered for any other University award. This study was financed by the University Malaysia Sarawak (UNIMAS) with support from the Network Research Group at the University of Plymouth. Relevant scientific seminars and conferences were regularly attended at which work was often presented. Several papers were published in the course of this research project. The work presented in this thesis is solely that of the author.

Signed .....

Date 5/05/04