Assessing the Quality of Audio and Video Components in Desktop Multimedia Conferencing

Anna Watson

A dissertation submitted in partial fulfilment
of the requirements for the degree of
Doctor of Philosophy
of the
University of London

Department of Computer Science
University College London

August 2001

Acknowledgements

I would like first to acknowledge the help and guidance of my supervisor Angela Sasse, who has inspired and encouraged me throughout the period of this research. Her enthusiasm, generosity and friendship are much appreciated. Thanks are also due to Philip Arden, my supervisor at BT Labs, Vicky Hardman, second supervisor at UCL, and Andrew Monk at the University of York for willing assistance and feedback.

I am indebted to many UCL colleagues for invaluable contributions to the studies documented in this thesis. Particular thanks go to Ismail Ismail for speaking proper, Orion Hodson for much and patient technical assistance, Colin Perkins for always being there to answer questions, and Yael Shmueli. Co-experimenters and friends Anne Adams, Anna Bouch, Sacha Brostoff, Julian Chesterfield, Jane Hughes, Isidor Kouvelas, Louise Sheeran and Gillian Wilson also deserve credit and thank yous.

I'd like to acknowledge the staff and students of the Language Centres at UCL and Exeter University, without whom there could have been no ReLaTe project; David Hearnshaw at Westminster University for allowing me to gather data from his willing and enthusiastic students; Daniel Pilon, formerly at CRC in Ottowa; and colleagues on the ETNA project at Glasgow University, in particular Rachel McEwan.

Thanks to family and friends, near and far, who have helped probably more than they realise, especially Judith Watson, Conor Watson, Nicola Mitchell, Laura Pettigrew and the 4-Pack.

Lastly, but definitely not leastly, thank you to my parents, who never really understood what I was doing, but were behind me all the way, and to Paul Conniff – thanks for being there, and now here.

Abstract

This thesis seeks to address the HCI (Human-Computer Interaction) research problem of how to establish the level of audio and video quality that end users require to successfully perform tasks via networked desktop videoconferencing.

There are currently no established HCI methods of assessing the perceived quality of audio and video delivered in desktop videoconferencing. The transport of real-time speech and video information across new digital networks causes novel and different degradations, problems and issues to those common in the traditional telecommunications areas (telephone and television). Traditional assessment methods involve the use of very short test samples, are traditionally conducted outside a task-based environment, and focus on whether a degradation is noticed or not. But these methods cannot help establish what audio-visual quality is *required* by users to perform tasks successfully with the minimum of user cost, in interactive conferencing environments.

This thesis addresses this research gap by investigating and developing a battery of assessment methods for networked videoconferencing, suitable for use in both field trials and laboratory-based studies. The development and use of these new methods helps identify the most critical variables (and levels of these variables) that affect perceived quality, and means by which network designers and HCI practitioners can address these problems are suggested. The output of the thesis therefore contributes both methodological (i.e. new rating scales and data-gathering methods) and substantive (i.e. explicit knowledge about quality requirements for certain tasks) knowledge to the HCI and networking research communities on the subjective quality requirements of real-time interaction in networked videoconferencing environments.

Exploratory research is carried out through an interleaved series of field trials and controlled studies, advancing substantive and methodological knowledge in an incremental fashion. Initial studies use the ITU-recommended assessment methods, but these are found to be unsuitable for assessing networked speech and video quality for a number of reasons. Therefore later studies investigate and establish a novel polar rating scale, which can be used both as a static rating scale and as a dynamic continuous slider. These and further developments of the methods in future lab-based and real conferencing environments will enable subjective quality requirements and guidelines for different videoconferencing tasks to be established.

Table of Contents

List of figures

List of tables

List of abbreviations and acronyms

ADPCM Adaptive Differential Pulse Code Modulation

CCIR Comite Consultatif International de Radiocommunication

CCITT Comite Consultatif International de Telephonie et Telegraphie

CCR Comparison Category Rating

CIF Common Intermediate Format

CSCW Computer-Supported Cooperative Work

CSN Circuit-Switched Network

DCR Degradation Category Rating

DSCQS Double Stimulus Continuous Quality Scale

DSIS Double Stimulus Impairment Scale

DVI Digital Video Interactive

EBU European Broadcasting Union

FTF Face to face

GSM Groupe Speciale Mobile

HCI Human-Computer Interaction

HDTV High Definition Television

ISDN Integrated Services Digital Network

IP Internet Protocol

ITU International Telecommunications Union

LPC Linear Predictive Coding

Mbone Multicast Backbone (of the Internet)

MMC Multicast Multimedia Conferencing

MOS Mean Opinion Score

MPEG Motion Picture Expert Group

ms millisecond

NTE Network Text Editor

PB Phonetically Balanced

PCM Pulse Code Modulation

PR Packet Repetition

PSN Packet-Switched Network

QCIF Quarter Common Intermediate Format

QoS Quality of Service

QUASS QUality ASsessment Slider

RAT Robust-Audio Tool

SCIF Super Common Intermediate Format

SDR Session Directory tool

SS Silence Substitution

SSCQE Single Stimulus Continuous Quality Evaluation

TCP Transport Control Protocol

UCL University College London

vat Video-Audio Tool

vic VIdeoConferencing tool

VMC Video Mediated Communication

wb WhiteBoard

Chapter 1: Introduction

1.1 Desktop conferencing over the Internet

Desktop multimedia conferencing entails the communication of two or more individuals across a network using a combination of audio, video and shared workspace. Users sit in front of their computers and communicate in real time via microphone, camera, and (usually) shared digital workspace. The communication can take place on a point-to-point basis, or involve many individuals and sites (via *multicasting* - Deering, 1988), and therefore offers many diverse communities the opportunity to collaboratively work across great distances. Uptake of the technology has been increasing in the higher education and research communities for applications such as distance education and remote project meetings, and the rapid growth of digital broadcasting and entertainment systems has ensured that the technology has spread throughout industrial and commercial communities in recent years. As network infrastructures improve, and equipment costs drop, it can be anticipated that multimedia conferencing will continue to grow in popularity.

However, as the popularity of the Internet is growing, and the uptake of multimedia conferencing technology is increasing, a cautionary note must be sounded. There is a danger that the growing Internet community could become a victim of its own success: as traffic increases, so does network congestion, which has particularly adverse effects on the successful and timely delivery of real-time data. This increase of network congestion has prompted calls for *Quality of Service* (QoS) guarantees, which could be achieved by mechanisms such as *bandwidth reservation* (see section 1.2). QoS means different things to different people. According to Fluckiger (1995), QoS in networking terms is a *"...concept by which applications may indicate their specific requirements to the network, before they actually start transmitting information data"*. This raises the question 'What requirements does the application have?' Ultimately, an application's requirements must be the application *user's* requirements, i.e. *"...a residential person, a professional individual, a subscriber organisation, a computer program, a computer system"*. In the context of this thesis, it is the requirements of the end-user of a system, the person who interacts with a network application to achieve a particular task, that determine whether the media quality is *good enough.*

The work reported in this thesis has been carried out in the context of Internet Protocol (IP)-based multicast multimedia conferencing (MMC), but many of the issues and findings should be relevant for other desktop conferencing systems and networks, and contribute to the design and development of multimedia conferencing applications in general.

1.2 Description of the research problem

The research focus of this thesis is how to establish, from an end-user perspective, what audio and video quality is required to make desktop videoconferencing usable.

There is an implicit assumption in the networking community that all audio and video quality problems in videoconferencing will be solved by increasing or reserving bandwidth (e.g. Jayant, 1990; Zhang et al., 1993). However, bandwidth is expensive, and to employ a scheme such as bandwidth reservation effectively, it would be sensible to establish what the different media quality requirements *actually* are, from an end-user's point of view, for different tasks. Designers of services and applications stand to benefit from knowing the *minimum* quality required (for a particular task to be achieved successfully, without causing user dissatisfaction and undue *user cost* - see section 2.1.2) and the *maximum* point beyond which increased quality offers no benefit to the user. Identifying these quality boundaries will be of importance to the networking community: even if available bandwidth can be maximised in most places, there will always be consumer demand for lower quality at lower cost (Podolsky et al., 1998).

Although bandwidth requirements undoubtedly need to be identified, many other factors can have an impact on user perception, especially the multimodal perception entailed in videoconferencing. For example, it is known that increasing audio quality can create an illusion of increased video quality (Negroponte, 1995), and that volume differences between speakers in videoconferences can cause negative quality perceptions (see section 8.2). Factors such as these need to be identified, and their impact on perceived quality should be explored systematically.

However, this raises a further problem - there is no established methodology in either the Human-Computer Interaction (HCI) or networking communities for assessing the subjective quality of audio and video delivered through real-time videoconferencing systems. This is partly due to the fact that until recently, much conferencing (over the Internet) was confined to the networking research community, but also because interactive, multi-person conferencing presents a non-trivial conglomeration of variables that can interact to affect perceived quality, rendering traditional speech and picture quality measurement methods insufficient. Increasing the bandwidth may reduce the effects of network congestion, but factors such as the hardware used, the method of encoding, the background noise, and the task being undertaken will all play a role in affecting the overall perceived quality. From Gestalt psychology we are told that "*the whole is greater than the sum of its parts*" (Kohler, 1930), but in evaluating perceived quality of multimedia conferencing, it is necessary to first identify and understand the role that each of those parts plays, and how they interact with one another.

Therefore, the main research questions addressed in this thesis are:

- How should the subjective quality of the component media (audio in particular) delivered in desktop videoconferencing be measured? (What existing measurement methods might be suitable, and where might these methods fail?)
- What are the fundamental factors that affect perceived media quality in desktop videoconferencing?
- What can and should be done to ensure sufficient subjective media quality for different videoconferencing tasks to be completed successfully?

1.3 The research approach

To date, surprisingly little empirical (either qualitative or quantitative) work has been carried out in the area of networked videoconferencing. Existing research into speech and video quality perception has tended to develop objective modelling techniques, or to investigate the subjective quality of the medium in isolated and brief segments. However, objective modelling techniques have not been perfected for the fluctuating conditions and degradations that can be common in networked communication, and it is therefore argued that they should not be relied upon to derive subjective quality requirements (see section 4.2). Although investigating the subjective quality of the medium in isolation permits a greater degree of control, with respect to MMC this approach can only be a part of the data-gathering, since there are so many other influencing factors in the real world. In addition, short segments listened to or viewed in isolation can tell us nothing about *task performance* (see section 2.1.2).

Central to this HCI thesis is the premise that effective evaluation of MMC perceived quality can only be achieved through grounding the research in a MMC *context,* i.e. *through conducting research with real users using MMC technology in real tasks.* It is argued that only through observation in the field can the issues most important to end users be identified. Experimental hypotheses about these issues can then be formulated and verified under laboratory simulation. This 'hand-in-hand' combined approach is believed to be particularly sensible in investigations of interactive networked communication, due to the multiplicity of factors that *could* be assumed to affect perceived quality.

In order to establish the feasibility and value of communicating via MMC, it is necessary to conduct field trials using the technology between different sites and across different disciplines. Both subjective and objective data must be collected, to increase understanding of the end user requirements and the optimal network configurations. Laboratory-based simulations alone would not be sufficient to achieve this aim within a reasonable timeframe.

The starting point of the research, therefore, is to gain an understanding of the audio and video quality issues involved in MMC. A review of the literature in the areas of spoken and visual communication, and how transmission of these media has traditionally been measured from a subjective point of view, is also required. Empirical research then needs to be undertaken to explore which of the existing methods can be successfully applied in the new research area of networked videoconferencing, and new methods need to be developed if necessary. On the basis of the research undertaken, conclusions and recommendations should be provided regarding the assessment of media quality, and the provision of sufficient levels of the same for different MMC tasks.

1.4 Scope of the thesis

The original aim of this thesis research was to establish, from an HCI perspective, what audio and video quality is required to make desktop videoconferencing systems usable. This means that the quality delivered should be sufficient to meet user requirements within the context of a particular task, such that the task goal can be achieved successfully (and be perceived as such), without provoking undue cost to the user (in terms of fatigue or stress). It was determined early on in the research, however, that there was no suitable existing method for determining perceived quality of audio and video experienced in interactive videoconferencing communication. As a result, the thesis research focus shifted to a more methodological one, exploring and developing techniques to help design an HCI method for assessing audio and video quality.

The research initially focused on identifying the issues that might have an impact on perceived quality of audio and video in networked videoconferencing environments. It rapidly became clear that the number of factors involved was far too great to investigate within the time period of the research (see section 2.8). As a result of the observations made in the ReLaTe field trial (see section 5.1), the research emphasis was placed on the audio channel rather than the video channel, since this was found to be the most critical factor to end users. The audio factors likely to play the largest roles in affecting perceived quality were focused upon, namely packet loss and repair schemes. However, the research approach meant that an awareness and consideration of all other factors was extant (see sections 5.1, 6.2 and 8.1).

In addition to limiting the scope of the variables to be investigated, the *means* by which to investigate their effect on perceived quality had to be constrained. Although one of the key goals of the research was to identify existing research methods that can be applied successfully in the evaluation of videoconference audio and video quality, it was clearly not possible to conduct an exhaustive investigation and comparison of all the available methods. Chapter 4 presents an overview of the methods identified and describes why some of them were not considered suitable for further research.

1.5 Contributions of the thesis

An understanding of *subjective* requirements is critical to the survival of real-time multimedia applications and services, since it is the end users that will determine whether an application is successful or not. For example, system developers may favour the implementation of one method of speech encoding over another if it reduces the amount of bandwidth required, but if users do not like the sound quality delivered, applications utilising the scheme are likely to be rejected. As mentioned in section 1.2, there are currently no suitable guidelines in the HCI literature for evaluating the conditions commonly experienced by users in videoconferencing applications. Traditional techniques for subjective assessment of audio and video quality have been developed from an engineering standpoint, with a view to determining whether degradations are detectable or not. From an HCI point of view, this level of investigation is almost irrelevant - what matters is whether the level of quality that is delivered is sufficient for a task to be carried out successfully, with acceptable cost to the user.

In addressing the need for an HCI method for assessing and establishing required audio and video quality, the research presented in this thesis makes a number of methodological and substantive contributions. Methodological contributions include:

1. An evaluation of the suitability of existing methods for subjective assessment of the audio and video quality delivered in networked videoconferencing environments (in theory in Chapter 4, and in practice in Chapter 5).
2. The development of new methods for the subjective assessment of this new technology, leading to a means of identifying different subjective quality issues and deriving subjective quality requirements for different videoconferencing applications (see Chapters 6, 7 and 8). For example, a new polar rating scale is explored and assessed in section 6.1.

Substantive contributions include:

3. A fuller understanding of the issues involved in subjective perception formation in MMC (in particular in Chapters 5 and 8). For instance, research in the field backed up by experimental data shows that volume differences can have a more profound effect on perceived quality than audio packet loss (see section 8.2).
4. Substantive knowledge about the audio and video quality requirements for some MMC tasks (in particular Chapters 7 and 8). For example, packet loss levels that will be tolerated in a conversational task are established in section 7.4.

The thesis contributes knowledge to the HCI community in terms of a methodology for assessing perceived quality of audio and video in MMC conferences, and offers benefits to QoS networking researchers and developers in terms of providing a means of establishing objective audio and video requirements for different tasks.

1.6 Overview of the thesis

In section 1.2 the research problems to be addressed were identified as:

1. What are the fundamental factors that affect perceived media (audio and video) quality in desktop videoconferencing?
2. How should the subjective quality of the component media delivered in desktop MMC be measured? (What existing measurement methods might be suitable, and where might these methods fail from an HCI viewpoint?)
3. Which methods aid in identifying critical quality factors, and upper and lower objective quality limits required for different videoconferencing tasks?

The fundamental factors that may affect perceived quality are described and explained in Chapters 2 and 3, and investigated practically in chapters 5, 6, 7 and 8. The existing methods for assessing subjective quality are investigated theoretically in Chapter 4, and practically in Chapter 5. New subjective assessment methods are explored in Chapters 5, 6, 7 and 8. The results of the studies undertaken in these chapters help answer the third research problem of which methods help establish quality requirements for different tasks.

The individual chapters are outlined below.

Chapter 2 presents a working definition of networked multimedia communication and provides descriptions of the multimedia tools used in the research presented in the main body of the thesis. The main issues involved in the subjective evaluation of the quality delivered through the constituent media of audio, video and shared workspace are discussed. Chapter 2 also presents an overview of the HCI discipline and outlines the academic approach undertaken in the thesis. This chapter therefore justifies the research strategy undertaken in the overall context of HCI research.

Although there has been little previous research on subjective quality assessment in networked videoconferencing, there has been much research on the process and product of video-mediated communication in general, and this research is surveyed in Chapter 3. The mechanisms and results of spoken and visual communication are considered separately, and then together in the context of the task being undertaken.

The final background chapter, Chapter 4, focuses on the methods that are traditionally used to measure the perceived quality of transmitted speech and video. A distinction is made between speech intelligibility and speech quality, since it is possible to gain high intelligibility without corresponding high quality. Subjective measurement methods recommended by the ITU are considered in relation to their suitability for assessing the novel conditions experienced in MMC. Although the ITU rating scales may in some cases be suitable, serious issues are brought to light: the labels on the category quality scales do not represent equal intervals (calling into doubt the legitimacy of results gained through them), and the scale does not allow for the multidimensionality of perceived quality to be measured. Chapter 4 also introduces the most recent ITU-recommended quality rating method, the Single Stimulus Continuous Quality Evaluation (SSCQE), which permits quality ratings to be recorded every second by means of a slider. The research presented in this thesis resulted in the development of a similar method for gathering subjective quality data, a tool named QUASS (QUality ASsessment Slider). However, as is argued in Chapters 7 and 9, QUASS is significantly different from the SSCQE method, and offers a more powerful functionality. Chapter 4 ends with a detailed discussion of the research agenda in the light of the literature review.

Chapters 5-8 present the empirical research undertaken. Field trials, exploratory studies, and larger experiments are described. The pros, cons and trade-offs involved in both experimental and real-world studies of MMC are elucidated and discussed. The motivations for, and methods used in, the studies are presented, and the results are considered in terms of methodological and substantive contributions to the thesis.

Chapter 5 describes a remote language teaching field trial, and presents the results of early research undertaken on the basis of observations made in the field trial. These early studies explored the use of traditional (ITU-recommended) rating methods in the context of MMC speech and video. Although the scales allowed substantive knowledge to be advanced, the vocabulary on the scales was found to be unsuitable for representing the range of MMC qualities, and the multidimensional aspect of quality perception. The need for a new rating method was identified, whereby the effects of different quality dimensions can be investigated.

Therefore, Chapter 6 presents studies, both in the lab and in the field, exploring the use of a novel, polar continuous rating scale. Although the novel rating scale proved useful and reliable, the drawbacks of one overall rating at the end of a session are highlighted, leading to the development of QUASS.

Studies using QUASS are discussed in Chapter 7. It was found that continuous real-time subjective quality rating can be performed well by subjects in an experimental setting, but not so well in a real-world conference, where the primary (conference) task demands all of the user's attention. This led to the development of a version of QUASS where the slider now *controlled* the objective quality that a user received i.e. the user could request the level of quality required in order to perform a task. In an interactive test environment the method was proven to work well, although individual differences appear profound. The results are analysed and discussed.

Chapter 8 presents the findings from a large-scale field trial where the impact of different types of quality degradations was clearly elucidated for the first time. The observations from the field trial were corroborated in an experiment, which also enabled data to be gathered on how users *describe* different types of degradation. Additional research into the vocabulary of MMC degradations, as used by different user groups, is summarised. The importance of establishing a common descriptive vocabulary between developers and users is discussed, and the application of different descriptive terms to different dimensions of quality, and to diagnostic help facilities, is discussed.

The final chapter, chapter 9, draws together all of the substantive results attained from the various methods employed throughout the research. The usefulness of the different approaches in terms of achieving the overall goals of developing an HCI evaluation methodology for videoconferencing applications, and of establishing quality guidelines for certain applications, are discussed. A summary of possible avenues for future research is presented.

Chapter 2: The research context

This chapter presents both the HCI context and research domain - videoconferencing across Internet Protocol networks - of the thesis. The first part of this chapter outlines the view of HCI within which the thesis research is conducted, and briefly introduces the methods used. The second part of the chapter introduces the research topic of networked multimedia communication in detail, and presents the issues involved in the evaluation of desktop videoconferencing. A definition of multimedia communication is first provided, and then a description of the characteristics of the networks over which multimedia communication occurs. The key components of audio, video and shared workspace are presented. Finally the key issues and difficulties involved in assessing the usability of, and evaluating the delivered quality in, MMC are discussed.

2.1 What is HCI?

Human-computer interaction can be defined as "*the discipline concerned with the design, evaluation, and implementation of interactive computing systems for human use and with the study of major phenomena surrounding them.*" (ACM SIGCHI, quoted in Dix et al., 1998). Since it is concerned with the technical design of the interacting machine in addition to the human-machine interface, HCI is vertically more extensive than human factors, but it is less extensive horizontally, since it is limited to a much narrower class of machines (Dix et al., 1998).

As a research discipline HCI is relatively young (dating from the early 1980s with the emergence of the personal computer), and has attracted researchers from a variety of disciplines (chiefly psychology, computer science and human factors, but also from areas as diverse as art and design, linguistics and social anthropology). This multidisciplinary make-up is a contributory factor to the discipline's continuing struggle to define its overall research strategy and methodology. Although it is agreed that the chief goal of HCI is to provide knowledge and methods for the design of usable computer technology, there is currently a lack of consensus over how the research should be conducted and its knowledge communicated to designers. Different approaches to the discipline aim to produce different levels of knowledge output e.g. heuristics, guidelines or engineering principles. The following section summarises the different positions with an aim to explaining the research framework of this thesis. For a more detailed discussion of these positions, see Sasse (1997).

2.1.1 Views of HCI

2.1.1.1 HCI as a science

In an attempt to prevent the 'soft' psychological input to HCI being pushed out by the 'hard' formalisms of the computer science and software engineering contributions, Newell and Card (1985) proposed a programme to turn psychological research relevant to HCI into a 'hard' science, by incorporating psychological knowledge into models and design tools to be used by designers. The framework for the construction of such models and tools would be based on:

1. *task analysis*: symbolic descriptions of tasks from which user behaviour can be predicted;
2. *calculation*: prediction of user behaviour through explicit operations on mathematical models;
3. *approximations*: acceptance that such calculations can only be approximate because of the complexity of human behaviour.

A number of criticisms can be levelled at this approach, however. It has been observed that restricting the psychological contribution to HCI knowledge to formal descriptions and approximate calculations would mean discarding a wealth of applicable information about users, and therefore actually *reducing* the overall psychological input to HCI. Carroll and Campbell (1986) suggest that, rather than reducing the psychological contribution to HCI to quantitative measurements and calculations, psychologists should ensure that the value of explanatory, conceptual HCI research is recognised and applied. Dix et al. (1998) agree with this, stating that "*the dearth of predictive psychological theory means that in order to test certain usability properties of their designs, designers must observe how actual users interact with the developed product and measure their performance.*" As Sasse (1997) argues, the "*application of quantitative methods to behavioural and cognitive phenomena which have not been sufficiently well described and understood can lead to the 'garbage in, garbage out' problem well known in computer science.*"

A further criticism of this approach is provided by Bellotti (1988), who observes that existing science-based modelling techniques such as Moran's Command Language Grammar (Moran, 1981) and Payne and Green's Task-Action Grammar (Payne and Green, 1986) have not been taken up by designers. It appears that these methods require too much effort and time to master, for too little gain. Mack and Nielsen (1994) also point out that in addition to being difficult to learn, formal methods "*are very difficult to apply and do not scale up well to complex, highly interactive user interfaces.*"

In response to these problems, an alternative to HCI as a science was formulated by Carroll and Campbell (1989), HCI as a design science.

2.1.1.2 HCI as a design science

Whilst Newell and Card (1985) argue that empirical HCI research should gather quantitative data from which to build models for designers to use, thereby following the traditional science approach, Carroll and Campbell (1986) stress the value and importance of conceptually deep, explanatory theories of human-computer interaction. They argue that it is necessary to collect "*all the information we can get*", since, like psychologists, HCI researchers are in no position to observe mental processes directly. In response to a criticism by Newell and Card that observation alone does not allow HCI research to advance beyond an informal, qualitative stage, Carroll and Campbell (1989) presented their *artifact theory*. The central premise of this theory is that it is only after a system or artifact has been developed that HCI research issues tend to be identified and addressed. Although Newell and Card would argue this as testament to HCI science lagging computing technology, Carroll and Campbell claim that the artifacts act in HCI in the same way as theories in more traditional scientific disciplines. By observing and evaluating artifacts in use in the real world, HCI research will eventually be advanced through the formulation of theories of user-system interaction, based on these observations and evaluations.

2.1.1.3 HCI as an engineering discipline

A third view of HCI is as an engineering discipline, formulated by Long and Dowell (1989). They define HCI as *"...the design of humans and computers interacting to perform work effectively."* The theory is that HCI can be decomposed into the domains of *software engineering* and *human factors*, forming two complementary problems to be solved:

- the design of computers interacting with humans (the domain of software engineering);
- the design of humans interacting with computers (the domain of human factors).

Both domains are responsible for constructing their own engineering principles. Software engineering is assumed to be close to establishing such principles, whilst human factors is not currently in a position to formulate them. Although the approach is an attractive one, with its long term goal of being able to apply engineering principles to HCI design problems (*'specify and implement'* the solution), this goal is a long way off: as previously stated, all of the requirements for an interactive system cannot be determined from the start because our models of the psychology and sociology of the human and human cognition are incomplete, and do not allow us to predict how to design for maximum usability. Sasse (1997) also cautions against the presumption that software engineering is close to establishing its engineering principles: "*The problem is the – often implicit – assumption that, because research on the "C" side of HCI is carried out in disciplines named* *computer science and software engineering, all research in these areas follows strict science and engineering procedures and yields "hard" results. This assumption is incorrect."*

2.1.1.4 Conclusions

As Sasse (1997) observes, although there are distinct philosophies of HCI, there are arguments why pursuing any of these approaches exclusively may not produce the integrated body of applicable knowledge which the discipline needs to establish. Subscribing to the 'HCI as a science' school of thought requires the formulation of models of the user that call on knowledge about human cognition that is far from complete. HCI as a design science, on the other hand, can be accused of producing craft knowledge which is only applicable to the system or artifact under investigation. The route to expressing concrete HCI knowledge from this approach is still not clearly mapped out. Engineering principles for HCI are an appealing concept, but as has been argued, the discipline is still some way off from establishing human factors principles. So how should an individual researcher plan his or her HCI research?

There are certain facts that are not disputed by any investigators, and it is these that provide a guide as to where to start in planning research. The HCI discipline seeks to construct a body of knowledge; therefore, any research undertaking should aim to contribute to this accumulation of HCI knowledge. It is both *substantive* and *methodological* support and knowledge that needs to be advanced and increased. Substantive (declarative) knowledge is advanced through research findings, with the aim of using these to prescribe guidelines or standards for design. Methodological support is provided through general system development methods and the refining and development of specific HCI techniques to suit the application under evaluation/development. Before undertaking research, the researcher should determine what related knowledge from other disciplines exists, whether it can be adopted in its existing form or whether an HCI-specific version needs to be constructed (Sasse, 1997).

HCI knowledge needs to be made accessible and applicable to designers in order to achieve its intended effect. Designers are users of HCI knowledge; therefore this knowledge has to be presented in a format that supports them in the task they are trying to perform. This means it has to be translated into accessible guidelines, efficient methods, or incorporated into tools. It can be argued that the best way to ensure the knowledge that is gained is rounded and comprehensive is for HCI to use both quantitative *and* qualitative methods, as advocated by Newell and Card (1986). As Sasse (1997) states, "*The semantic prejudice which associates experiments and quantitative data with good science and clear, valid and applicable results must be overcome. Experiments that are* *inappropriate, or badly planned or controlled, can set back research. Anecdotal science is avoided by studies that are well planned and executed, and meticulously documented for the inspection by other researchers."*

2.1.2 HCI evaluation

Evaluation can be defined as *"an assessment of the conformity between a system's actual performance and its desired performance"* (Whitefield et al., 1991). This performance can be measured from three HCI perspectives: *task performance*, *user satisfaction*, and *user cost*. Task performance is generally assessed in terms of *effectiveness* and *efficiency*. Effectiveness relates to whether users can complete the tasks given to them, and efficiency is concerned with aspects such as the time taken to complete the task, the frequency of errors, and the frequency of accessing help facilities. User satisfaction is subjectively measured, often by rating scales or interviews. User cost is determined mainly by objective measurements of factors such as degree of fatigue, stress, repetitive strain injury (RSI) potential, and mental effort. The perspectives that will be measured, and the methods used within these perspectives, will depend on the goals of the evaluation. The goals of the evaluation will in turn depend on the time available to carry out the evaluation, and the resources and skills of the investigators.

Two different types of evaluation can be undertaken: *summative* evaluation and *formative* evaluation. Summative evaluation tends to takes place after the implementation of an interface or system, and is designed to determine whether the design goals were achieved. The output of a summative evaluation is a report listing the issues/problems found. Formative evaluation, on the other hand, is an evaluation of an unfinished user interface, and aims to expose usability problems that exist in the current iteration. Formative evaluation is better positioned to help form the solution to a design problem.

Evaluation can be carried out in either the laboratory or the field. There are advantages and drawbacks to both types of evaluation. Laboratory studies offer a high level of control over the conditions in which the evaluation is carried out, but the lack of context (e.g. office environment and dĕcor) may mean that one is investigating a situation that never occurs in the real world. Dix et al. (1998) observe that it is *"especially difficult to observe several people cooperating on a task in a laboratory situation as interpersonal communication is so heavily dependent on context."* Field studies, on the other hand, allow the gathering of data on interactions between systems and between individuals which would be missed in a laboratory study, but are subject to greater interference from extraneous events such as phone calls and high background noise. Dix et al. conclude that, on balance, field study is to be preferred, since it permits the study of the interaction as it occurs in actual use, but they caution that even in this situation, the subjects are likely to be influenced by the presence of the analyst and/or recording equipment, so the situation will always be somewhat less than 'natural'.

As stated above, the evaluation methods that are used will depend on the goals and objectives of the evaluation. Their applicability/usefulness will also depend on where the data is being collected, and the type of system being assessed. Some common HCI evaluation methods include specialist reports, user reports and observational methods (Preece et al., 1994).

- **Specialist reports** such as an 'expert walkthrough' involve an HCI specialist operating a computer or system from the users' point of view and reporting the difficulties encountered/changes that should be made. These methods are also known as **usability inspection methods**, and are methods of evaluating software user interfaces using rules, guidelines or heuristics, rather than acquiring (and analysing) behavioural end-user feedback (Mack and Montaniz, 1994).
- **User reports** are based on users' memory of/inferences about a system. The techniques used to gather the data are questionnaires, interviews, surveys etc.
- **Observational methods** typically involve documenting user performance using a system, and can take place either in the real world or in controlled laboratory settings.

Mack & Nielsen (1994) include

- **Automatic assessment** (where usability measures are computed by running a system through evaluation software) and
- **Formal measurement** (using exact models and formulae to calculate usability measures)

but report that "*With the current state of the art, automatic methods do not work and formal methods are very difficult to apply and do not scale up well to complex, highly interactive user interfaces. Empirical methods are the main way of evaluating user interfaces, with user testing probably being the most commonly used method. Often, real users can be difficult or expensive to recruit in order to test all aspects of all the versions of an evolving design. In this respect, inspection methods are a way to 'save users'*".

2.1.2.1 Evaluation conducted in this thesis

The studies conducted in this thesis have considered task performance where possible, but the emphasis has been on subjective methods for user satisfaction. This was determined to be the most important aspect to investigate since establishing user satisfaction is critical for the uptake of any new technology. In the course of the thesis research, both summative (see sections 5.1 and 8.1) and formative (see, for example, sections 5.2, 5.3 and 5.7) studies are conducted. The research is conducted both in the field and in the laboratory: since there is very little HCI knowledge in the area of networked videoconferencing, observations from systems in use in the field allow the most critical HCI elements to be identified and explored further in a controlled setting.

2.1.3 The MMC research context

As discussed in section 2.1, there is currently no consensus in the HCI discipline as to what its overall research strategy and methods should be. Therefore a decision about these aspects needs to be made by the individual researcher for the specific topic under investigation.

Traditionally, HCI has been concerned with the user interface, the point at which the user and the computer meet. But with the advent of computer supported cooperative work (CSCW), this interface has become more complex - not only is it the interface between the computer and the human, but also between the human and the human (Gale, 1991). So whilst the human-computer interface must be evaluated with respect to task performance, user satisfaction and user cost, so too must the quality of the communication afforded between the two, or more, human communicants.

Within CSCW, networked videoconferencing is a new form of communication, and there are currently no HCI guidelines or recommended methods that explicitly address the subjective evaluation of media quality delivered through this technology. Guidelines that address videocommunication services, such as the RACE[1] ISSUE (IBC Systems and Services Usability Engineering) guidelines, tend to have been written with regard to high bandwidth, dedicated link services, where specialised room-based videoconferencing suites are common. Low bandwidth desktop conferencing systems over best-effort packet networks have not been addressed.

The approaches of HCI as a science (section 2.1.1.1) and as a design science (section 2.1.1.2) were discussed, where the contexts of enquiry are traditional scientific experiments and observation of real-world systems respectively. In this thesis a combination of these strategies is advocated and followed. It is argued that it is not possible to carry out networked videoconferencing research *purely* in either a controlled environment or in the field. There are a multitude of variables that may play a role in perceived videoconferencing quality. It is only by carrying out research in the field that the variables likely to have the biggest impact on perceived quality can be identified, and it is

[1] Research and development in Advanced Communications technologies in Europe (funded by the European Union).

only in the laboratory that the effect of these variables can be quantified and established. Using video telephones as an illustration, Nielsen (1994) comments:

"*As with many computer-supported cooperative work applications, video telephones require a critical mass of users for the test to be realistic: if most of the people you want to call do not have a video connection, you will not rely on the system. Thus, on the one hand field testing is necessary to learn about changes in the users' long-term behaviour, but on the other hand such studies will be very expensive. Therefore, one will want to supplement them with heuristic evaluation and laboratory-based user testing so that the larger field population does not have to suffer from glaring usability problems that could have been found out much more cheaply.*"

The research approach undertaken in this thesis is therefore similar to Carroll and Campbell's (1989) artifact theory, in that the technology must be observed in use in the real world first, but the theories that arise from this observation are verified in a fashion more akin to the traditional science approach advocated by Newell and Card (1986).

2.1.3.1 Evaluating MMC experiences

In the course of the research presented in this thesis, a number of different evaluation methods are employed, but their straightforward application in the context of MMC is shown to be problematic. In principle, many of the techniques mentioned in section 2.1.2 are suitable, but are pragmatically difficult to employ. The tasks that are commonly performed using MMC technology are necessarily complex, involving at least two, and often more, types of media. The task that is being performed is never performed *individually,* and frequently involves more than two people, rendering specialist reports/individual usability inspections inappropriate. Most importantly, the network over which the communication is taking place is in a state of constant flux, so quality does not remain constant, meaning that user reports are rendered, in field trials at least, insufficiently detailed. In addition, as will be illustrated in Chapter 8, care is required when asking users to describe the subjective quality they experience: different user groups use different descriptive vocabulary. These facts set the evaluation of MMC apart from most other systems that HCI practitioners are asked to evaluate.

As a result of these issues, practical steps towards developing a new subjective assessment methodology tailored to evaluating MMC communication are undertaken in this thesis. These steps involve adapting existing methods, developing new techniques, and simulating real-world conditions such that these new methods can be employed effectively. The research undertaken advances both substantive and methodological HCI knowledge.

2.2 Some technology definitions

As Fluckiger (1995) notes, the term 'multimedia' had become *"one of the most overused terms of the early 1990s"*, and its use certainly has not lessened towards the end of the decade. It now seems that any new software or hardware product requires the term 'multimedia' (prominently) displayed somewhere in its description if it is to stand any chance of being successful. It is therefore required to define more clearly what is meant by the term 'multimedia' when it is used in this thesis. Firstly, it is *digital* multimedia that is referred to:

"Digital multimedia is the field concerned with the computer-controlled integration of text, graphics, still and moving images, animation, sounds, and any other medium where every type of information can be represented, stored, transmitted, and processed digitally." (Fluckiger, 1995).

Secondly, and more particularly, the concern is with *networked* digital multimedia, whereby two or more multimedia end systems communicate with each other over a network. This entails the transmission of *independent* streams of data across a communications network. Principally, in multimedia conferencing, the data streams are audio, video and shared workspace, which will be considered in sections 2.5, 2.6 and 2.7 respectively.

What are the requirements to be met by networks in transmitting these data streams in order that *real-time* multimedia communication can occur?

2.3 Characteristics of digital networks supporting multimedia applications

There are two main types of networks: *circuit-switched* networks (CSNs) and *packet-switched* networks (PSNs). CSNs emulate the telephony model whereby a point-to-point connection is made across a dedicated link. Data is transmitted across this connection as a continuous stream of bits. ISDN (Integrated Services Digital Network) is an example of a CSN. PSNs, on the other hand, partition the data stream into small units (packets) and transmit these individually. The size of the packet will depend on the application tool in use, but often they will contain many hundreds of bytes. There is no set route for a packet to reach its destination. The underlying concept of PSNs is that more efficient use of the network can be made, since only the source that is transmitting will be using resources. The Internet is an example of a PSN.

Transmission of real-time information across digital networks is affected by three factors: the available *bit rate*, the *delay* incurred in sending and receiving across the network, and the *error rate*.

2.3.1 Bit rate

This is the rate at which binary information can be exchanged between two sites, per unit of time. (Rather misleadingly, it is also commonly referred to as the *bandwidth*, but technically bandwidth refers to the capacity of the underlying analog channel: depending on the encoding scheme, the channel can support a certain bit rate, but the bandwidth imposes a fundamental limit on this - Shannon, 1948.) Two types of bit rates can be supported: *constant* and *variable*, although variable bit rates are more common. The degree of variation depends on the application and network, and is known as *burstiness*. The bit rate/bandwidth between sites must be sufficiently large to cope with the requirements of the application.

2.3.2 Delay

End-to-end delay is defined as the capture time at the sender to the presentation time at the receiver. End-to-end delay is affected by the time it takes to encode and decode data, especially where compression is involved (see section 2.5.1). In PSNs, end-to-end delay is also affected by congestion (queue length) at the packet routers. Delay variation, or *jitter*, is the degree to which the time between sending and receiving varies, and depends on the technology used.

2.3.3 Error

There are three main types of error that can occur when transmitting data across networks: *loss*, *bit corruption*, and *out-of-order* delivery. Loss can be caused by errors in the transmission of a signal (e.g. 'crackles' on a phone line), or by congestion of the network or routers. In CSNs, corruption of the transmission signal results in *bit errors* in the received signal. Congestion does not occur in CSNs since the link between the end-systems is guaranteed. In PSNs without resource reservation, however, packet loss caused by congestion is the major type of error. The last type of error that can occur in PSNs is *out-of-order delivery*, which can happen when a stream of data is divided between two different routes. Since not all routes will have the same transit delay patterns, the data stream may not be received in the same order it was sent.

CSNs arose out of the digitisation of the original analog telephone transmission system, but PSNs were not originally designed to transmit real-time voice and video data. Transmission of real-time audio and video over PSNs will therefore result in very different *subjective quality* characteristics from those of CSNs. As will be argued in Chapter 4, traditional speech and video quality assessment methods, which were developed in the CSN world, are inappropriate for speech and video delivered across PSNs. (More importantly with respect to this thesis research, the methods are not HCI methods, meaning that the data gathered does not lend itself to interpretations about the *adequacy* of the quality for a particular task to be performed satisfactorily - see section 5.1.4).

The field studies (see sections 5.1, 6.2 and 8.1) presented in this thesis were carried out over PSNs using the Internet Protocol, IP and so the next section describes IP networks in greater detail.

2.4 Internet Protocol networks

Internet Protocol (Postel, 1981), IP, did not arise in the telecommunications domain, but rather the academic research community, and therefore it has different standards to those of a commercial telecommunications network provider. Principally, IP has not been designed with performance guarantees in mind: it is a flexible system, without 'hard-and-fast' connections. IP networks are constructed from end-systems and routers through which the data, known as *packets*, is sent. There is no set route between two end-systems.

In addition to not having set routes, there is also no *bandwidth allocation*: the bit rate of a certain communication stream is shared with other sources, and statistical multiplexing is relied upon to allocate bandwidth. In the case of a network overload, the network may discard packets: it does not guarantee a defined level of service. In this case, it is left up to the end-systems to cope with the loss of data.

There are three main causes of packet loss that affect real-time communication over the Internet:

- network congestion leading to dropping of packets at routers;
- network congestion leading to consecutive packets being sent by different routes, meaning that some arrive at the receiver too late to be played out, and are therefore discarded;
- overloading of the local machine, meaning that packets may not be decoded and played out in time.

2.4.1 The Multicast backbone (Mbone)

That IP-based networks function as 'best-effort' services rather than being able to offer a guaranteed level of service can be viewed as a disadvantage, but this is offset by their major advantage, the capacity to *multicast* information. Whereas a *unicast* connection can be compared to a telephone call where only two telephones are connected for the duration of the conversation, a multicast connection can in principle have an unlimited number of participants. This potential is the result of the ability of the network to replicate at routers the packets transmitted from a sender. The replicated packets can be sent on to as many end-systems that have requested the copies by virtue of being members of a multicast group with a certain address. Multicasting therefore differs from *broadcasting* in that broadcasting follows a *'one-to-all'* protocol, while multicasting is from *'one-to-however-many are in the group'*.

At present, some Internet routers are unable to deal with multicast addresses, and so special multicast routers are required, known as *mrouters.* Mrouters package multicast packets up as 'normal' IP packets and send them over the network to the next mrouter. This network of mrouters forms a physical multicast layer that overlays the Internet, and is known as the Multicast Backbone of the Internet, or Mbone for short (Deering, 1988). The Mbone grew out of a non-commercial (and therefore free) experiment in the research community, but multicasting is now present in commercial networking on a large scale. It is a scalable technology, meaning that it has great potential, especially in its key applications of transmitting real-time audio and video to and from multiple sources. There is currently a great deal of debate in the networking community over what the critical QoS delivered over the Mbone should be. To answer these questions, there is a need for a reliable method for assessing the usability/acceptability of the quality delivered (see sections 1.2 and 4.13.1).

2.4.2 Mbone packet loss

Some degree of packet loss over IP networks is inevitable. The Transmission Control Protocol (TCP) of the Internet determines the rate of data transmission by requesting acknowledgements from receivers that packets have been received (Postel, 1981b). When acknowledgements are not received, the transmission rate drops, since congestion has occurred. However, in order to ascertain that there is not enough bandwidth, the flow must be increased a little, which causes packet loss. In a multicast environment in which people are distributed at many various points on a network, packet loss can be more or less severe, depending on where on the network the participants are. Loss distributions are complex, and change and evolve as the network does, but it is possible to make some general observations.

Handley (1997) reported that most receivers in a multicast session tend to experience loss in the range of 2-5%, but others will experience a significantly higher range. Packet loss tends to be individual rather than in clumps, although loss of two or more packets does occur more frequently than chance would dictate. This finding is in broad agreement with that of Bolot et al. (1995). It has also been observed that in very large conferences each packet is likely to be lost by some participating receiver, which provides support for mechanisms such as redundant transmission (see section 2.5.2.5).

The scenario of real-time multimedia conferencing over the Mbone is now described. This is the research area under which the majority of the research reported in this thesis was carried out.

2.4.3 Multicast multimedia conferencing

There are many different uses of multicast technology, ranging from largescale broadcasts (e.g. NASA shuttle launches - Macedonia and Brutzman, 1994), to lectures and small seminars, in all of which the degree of interaction is limited. However, to cover the key issues of multicast multimedia conferencing (MMC), a scenario is presented of *interactive desktop conferencing*, in which the participant sits at a workstation and sends and receives three types of digital media: audio, video and shared digital workspace. Each workstation is equipped with a camera that sits either to the side or on top of the monitor, and speech communication takes place through a headset. A typical desktop conferencing environment is shown in Figure 1, illustrating on-screen video images of the conference participants, and shared electronic workspace.

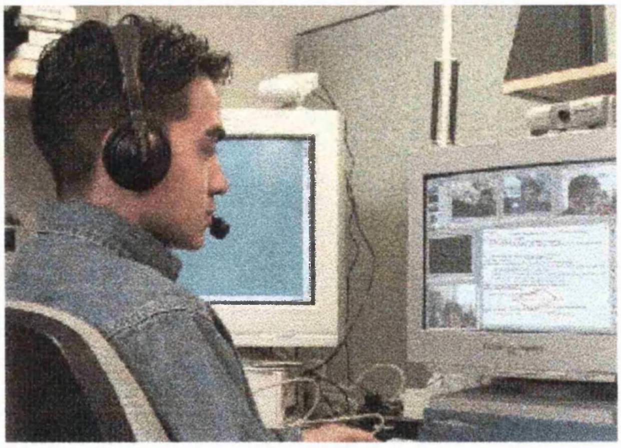


Figure 1: Typical desktop conferencing set-up

The characteristics of each of the three media streams are considered in turn, and representative software tools for each medium are described.

2.5 Audio

To send audio information across a network, the sound stream needs first to be digitised. There are a number of different digital speech coding algorithms. Speech coding algorithms can be divided into two types: *waveform* and *knowledge-based* coding techniques.

Waveform coding techniques attempt to preserve the speech signal. Speech characteristics are exploited relating to amplitude and time. The simplest of these techniques is *Pulse Code Modulation*, or PCM, defined by the ITU standard G.711. PCM is the simplest method by which an analog speech signal can be digitised. The signal is sampled 8000 times per second, allowing frequencies up to 4 kHz to be encoded. Although the frequency range that speech occupies is much larger than this (about 10 kHz), most speech sounds occupy the lower frequencies, and so this range is adequate for voice communication. PCM requires a bit rate of 64 kbit/s.

Although PCM is the simplest method, it is not the most efficient means of encoding speech, because it is not speech specific, and does not take advantage of the natural redundancy in speech and hearing. Advantage can be taken of the fact that the characteristics of the speech waveform do not change that frequently (Kent and Read, 1992), and thus speech can be compressed for digital transmission. ADPCM (Adaptive Differential Pulse Code Modulation) (ITU-T G.723) uses a predictor model that codes the difference between the signal at an instant and the prediction for that instant. This compression decreases the bit rate required from 64 to 32 kbit/s, with hardly any detectable drop in speech quality.

Knowledge-based coding techniques model the speech production process. Two examples are GSM and LPC. GSM is the codec (encoding/decoding device) used in mobile phones, and works by modelling the vocal tract (Mouly and Pautet, 1993). It is specially designed for voice signals, and occupies a bandwidth of 13.2 kbit/s. However, the processing power requirements are relatively high.

LPC (Linear Predictive Coding) is a synthetic speech encoder that fits speech to a simple model of the vocal tract and occupies a bandwidth of only 4.8 kbit/s. Although producing intelligible speech, it tends to sound robotic to the listener.

2.5.1 Bandwidth *vs.* compression, *vs.* processing power, *vs.* perception...

In transmitting speech across a network where available bandwidth is not guaranteed, it is obviously desirable to compress the speech signal as much as possible to minimise the bit rate required. It is possible to get good speech compression since, in essence, a large amount of the speech signal is unnecessary for comprehension. Speech occupies a frequency range of about 50 Hz to 10 kHz, yet it is possible to converse perfectly easily across the telephone network where the frequency range is 3.4 kHz. However, there are side effects to compression:

1. Compression requires processing power at both the sending and receiving end: generally the more compressed the speech signal, the more processing power is required.
2. In addition to processing power, increasing compression increases the end-to-end delay: it takes more time to both encode and decode compressed speech.
3. Depending on the compression method employed, even a small amount of packet loss can have a severe effect on the end user, since one packet can contain critical information for the reconstruction of speech information contained in both subsequent and preceding packets. The issue of user perception of different speech encoding methods and the effects of network loss will be discussed further in section 2.5.2.5.

The next section describes a software audio tool developed for use in multicast environments, **RAT** (Robust Audio Tool - Hardman et al., 1995; 1998), which offers different speech encoding methods and different means of dealing with the problematic network characteristics. The research reported in this thesis both contributed to the iterative development of **RAT** (through investigation of suitable packet loss repair mechanisms), and used the tool in MMC experiments and field trials.

2.5.2 RAT (Robust Audio Tool)

RAT was developed at UCL and is designed for use both in unicast and multicast audio conferencing environments[2]. The research reported in this thesis was carried out using version 3 of **RAT**, and therefore the settings for this version are described. The main **RAT** interface is shown in Figure 2.

RAT has been designed to cope with different conditions over the Internet, to interwork with other audio tools, and to be used across a variety of different platforms. It therefore has a number of different options that can be selected according to the needs of the user in a particular conference. The main features of **RAT** are covered below.

[2] More details, and the software, can be found at htpp://www-mice.cs.ucl.ac.uk/multimedia/projects/rat/

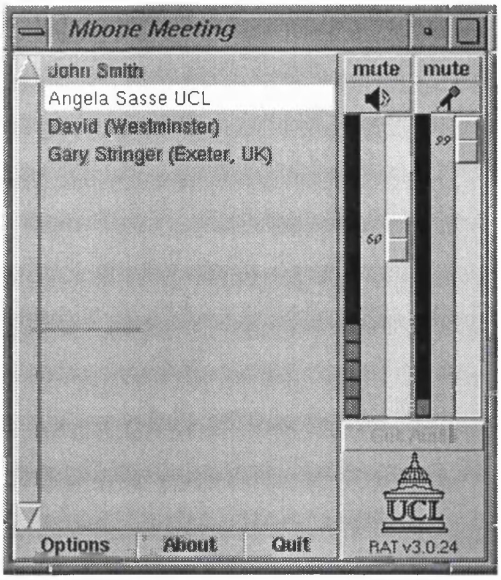


Figure 2: The RAT main interface

2.5.2.1 Encoding

There are five types of audio encoding currently available in **RAT**: 16-bit linear, PCM, DVI, GSM and LPC. All of these were described in section 2.5 apart from 16-bit linear and DVI. 16-bit linear is linear PCM at 16 bits per sample, and is therefore higher quality speech (regular PCM is 8 bits per sample). DVI is essentially the same as ADPCM, but without a linear predictor - it just uses the previous sample (IMA, 1992). A linear predictor uses much processing power, so DVI consumes less processing power than ADPCM. However, there is a trade-off between processing power and the quality of the speech signal: quality is reduced due to the absence of a predictor. Like ADPCM, DVI requires 32 kbit/s. DVI is the default setting in **RAT**, since it offers the compression benefits of ADPCM over PCM, but does not require as much processing power as ADPCM.

2.5.2.2 Packet sizes

RAT offers 4 different packet sizes: 20 milliseconds (ms), 40 ms, 80 ms and 160 ms. There is a trade-off associated with packet size as fewer, larger packets decrease the network load, but contain larger amounts of speech, which, should the packets be lost, is more detrimental to the communication. In addition to increasing the net number of packets transmitted, smaller packets increase the overall bandwidth required due to the fact that headers must be attached to each packet. These headers add 96 bits per packet. Therefore, transmitting 20 ms packets creates an overhead of 4.8 kbit/s, whereas 80 ms packets only incur an overhead of 1.2 kbit/s. By default **RAT** starts with 40 ms packets of DVI coded data.

2.5.2.3 Silence suppression

Silence suppression is the mechanism by which only audio above a certain volume is transmitted. The rationale for this is that users are able to leave their microphones unmuted whilst participating in a conference, without consuming bandwidth unnecessarily by transmitting background noise. It has been estimated that 50% of conversational time is filled with silence from a participant. Therefore only when users speak will their audio be transmitted. The default when using **RAT** is that silence suppression is turned on.

2.5.2.4 Full/half duplex

Full-duplex is the situation by which the user is able to talk and listen at the same time in a conference. In this situation it is necessary to converse via a headset or with an echo canceller to prevent feedback. If feedback occurs (through 'leaky' headsets or poor echo cancellation), it will be necessary to resort to muting the microphone when listening and unmuting it when talking, a function known as *push-to-talk*. Some audio cards are not enabled to cope with full-duplex, only half-duplex. In a half-duplex condition, there are two mode options that can be selected. *Net-mutes-mike* means that the speaker's microphone will be muted as soon as another participant speaks. *Mike-mutes-net* means that the speaker is unable to hear anyone else while talking. For audio cards that support full-duplex, this will be the default setting in **RAT**.

2.5.2.5 Repair methods

As discussed previously, some packet loss is inevitable in MMC. When audio packets are lost, the effect on speech perception can be quite severe, especially as the loss rate increases. In order to compensate for this loss of audio information, different repair methods are available. **RAT** offers two types of repair methods, those which are based at the receiver, and those which are sender-based.

Receiver-based repair schemes work by producing a replacement for a lost packet which is similar to the original, which is possible since speech waveforms exhibit large amounts of short-term self similarity. Two receiver-based techniques available in **RAT** are *silence substitution* and *waveform substitution* (commonly referred to as *packet repetition*).

In silence substitution, the gap left by a lost packet is filled with silence in order to maintain the playout order of the packets that do arrive. This technique is widely implemented in other audio tools (e.g **vat** - Jacobson, 1992) because it is cheap to implement in terms of processing power. However, the larger the packet, and the greater the loss rate, the more speech information is not heard by the user. As packet sizes increase, so does the probability that a phoneme, the smallest unit of speech intelligibility, is enclosed in that packet. When phonemes are lost, speech intelligibility suffers. It has been demonstrated that listening performance is worst when unexpected silence is encountered in speech (e.g. Miller and Licklider, 1950; Warren, 1982). Performance can be greatly improved by the insertion of noise of any type at the same frequency as the missing speech, since speech characteristics change relatively slowly. This is approximated by the technique of *waveform substitution* (or *packet repetition*), which fills in for the missing packet by repeating the last received packet. The default setting for receiver-based repair in **RAT** is packet repetition. This technique is more likely to work well where the packet sizes are small: when the packet is large, the speech signal is likely to have changed significantly within the missing packet, and repeating the previous packet may be more detrimental than helpful to perception.

In addition to receiver-based techniques, **RAT** implements two **sender-based repair schemes** to recover from the problem of packet loss: *interleaving* and *redundant transmission.* Interleaving disperses the effect of packet losses by placing originally adjacent units of packets into different packets for transmission, which are then returned to their original order at the receiver. This means that should a packet be lost, it contains small parts of many different packets rather than one large chunk of the speech stream (Perkins et al., 1998). Although interleaving does not increase the bandwidth requirements of a stream, it does increase the end-to-end delay. Therefore the technique is not suitable for interactive applications. Redundant transmission is, however, suitable for interactive applications. The method involves piggy-backing a (often more heavily compressed) copy of a packet onto the following packet. If the original packet is lost, the redundant copy can be used in its place. Because the redundant packet is very heavily compressed, sound quality suffers, but the method is an improvement over having no audio to play out in the place of the lost packet. There must be a trade-off between the amount of compression used for the redundant packet (and hence bandwidth overhead), and the quality of the resultant audio. The best compression can be achieved by using a synthetic version of the speech such as Linear Predictive Coding (LPC), which is a synthetic quality speech coding algorithm preserving about 60% of the information content of the speech signal (see section 2.5). Redundant transmission is illustrated in Figure 3.

Descriptions and auditory perception of the different repair schemes is shown in Table 1.

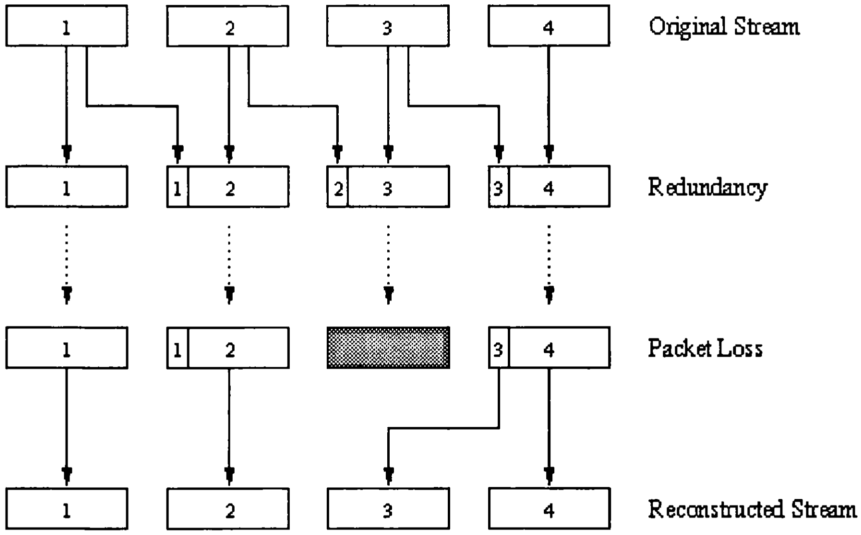


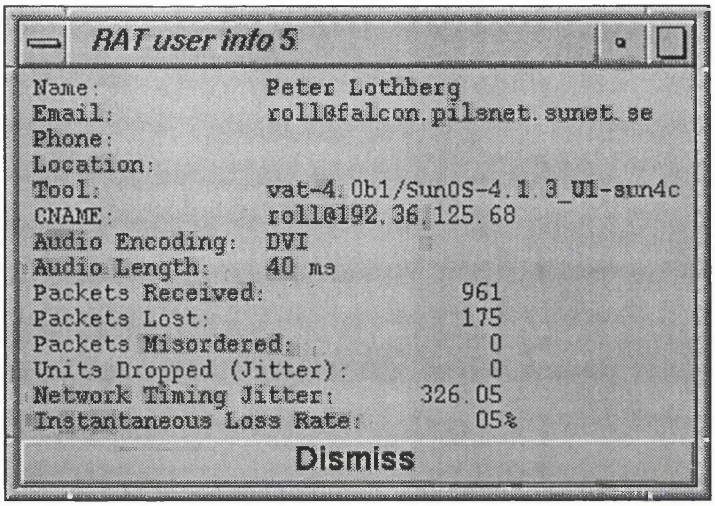
Figure 3: Redundant transmission

Speech reconstruction scheme	**Description of method**	**Auditory perception**
Silence substitution	*Replaces lost packets with packets containing zero amplitude (silence)*	*'Bubbly' at low loss rates, perceptible glitches/gaps with large packets*
Waveform substitution (packet repetition)	*Decodes previous packet again in place of the lost one*	*Clicks may be heard where frequency and amplitude change; 'stutter' effect with large packet sizes as phonemes get repeated*
Redundancy	*Compressed version of speech piggy-backed onto later packets*	*May sound unnatural (e.g. 'robotic') depending on compression algorithm used*

Table 1: Speech repair scheme characteristics: method and perceptual effect

2.5.2.6 Reception statistics

By clicking on the name of a participant in the main **RAT** window, a reception statistics box is brought up referring to the audio received from that person (see Figure 4). By this means it is possible to gain an idea of the amount of objective packet loss being received on that channel. However, it is not possible to pinpoint the cause of loss, which can be in the network or due to lack of processing power in either of the end-systems involved.



```
Name:                 Peter Lothberg
Email:                roll@falcon.pilsnet.sunet.se
Phone:
Location:
Tool:                 vat-4.0b1/SunOS-4.1.3_U1-sun4c
CNAME:                roll@192.36.125.68
Audio Encoding:       DVI
Audio Length:         40 ms
Packets Received:                    961
Packets Lost:                        175
Packets Misordered:                    0
Units Dropped (Jitter):                0
Network Timing Jitter:            326.05
Instantaneous Loss Rate:             05%
```

Figure 4: RAT reception statistics for a conference participant

2.6 Video

If it is desirable to compress speech for transmission over the Internet, it is an absolute necessity in the digital transmission of video. Historically, full-motion video is transmitted at 30 frames per second (fps) in the US and 25 fps in Europe. Given 24 bits per pixel in digital video, transmission of full size uncompressed digital video at 30 fps for only *one second* would generate 27 megabytes of data (Crowcroft et al., 1999).

Compressed digital video will always involve a trade-off between *frame rate* and *resolution*. With limited bit rate, increasing one implies reducing the other.

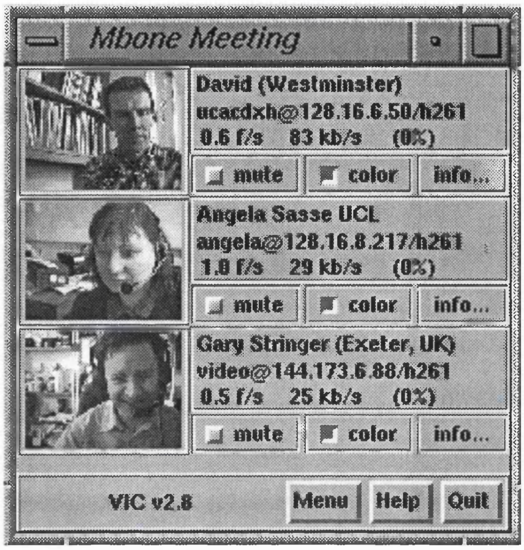
2.6.1 H.261 video coding

H.261 is one of the ITU H.320 suite of standards for videoconferencing and videotelephony, and is a standard for compressing colour motion video into a low-bit rate stream. The method was originally designed for ISDN (Integrated Services Digital Network), and is the most widely used standard for video compression.

H.261 supports three frame formats, QCIF, CIF and SCIF. QCIF (Quarter Common Intermediate Format) has dimensions of 176x144 pixels, CIF (Common Intermediate Format), fits an image into dimensions of 352x288 pixels, and SCIF (Super Common Intermediate Format) 704x576 pixels. The frames are in turn composed of blocks of pixels (there are 99 blocks in a QCIF image, and 396 in a CIF image). The whole frame is not updated in a single scan, but rather on a block-by-block basis depending on whether there is new information content in that block or not. This means that there can be partial updates of faces, especially at lower bit rates.

2.6.2 vic (VIdeoConferencing Tool)

vic (VideoConferencing Tool) was developed by Lawrence Berkeley Labs (McCanne and Jacobson, 1995). The main window of the interface is shown in Figure 5.



David (Westminster)
ucacdxh@128.16.6.50/h261
0.6 f/s 83 kb/s (0%)

Angela Sasse UCL
angela@128.16.8.217/h261
1.0 f/s 29 kb/s (0%)

Gary Stringer (Exeter, UK)
video@144.173.6.88/h261
0.5 f/s 25 kb/s (0%)

Figure 5: The main vic window, showing thumbnail images

Like **RAT**, **vic** is designed to interwork with a number of different conferencing environments and platforms, and therefore has a number of settings that users determine. It is designed to work with video encoding formats other than H.261, although H.261 is the one currently in most common use. Because of this, H.261 was the encoding format used in all of the trials and studies presented in this thesis.

2.6.2.1 Receiving video

It is possible to view four different image sizes in **vic**. Thumbnail images are displayed by default as soon as a video image is received. By clicking on this image, a larger image is displayed, the choice with H.261 encoding being QCIF, CIF or SCIF. The default setting is that the image enlarges to CIF. In addition to being able to change the size of a received image, it is possible to opt to receive it in greyscale, or to mute it, both of which save processing power on the local machine. The **vic** default is that the image will be displayed in colour.

2.6.2.2 Transmitting

The **vic** user is given a number of options when transmitting. Whilst the default setting is that video will be transmitted at 128 kbit/s and a frame rate of 8fps, both of these options can be changed on a slider bar. The bit rate available ranges from 10 kbit/s to the maximum for that session (determined by the distance the multicast packets have to travel in the conference, the time-to-live, ttl), while the frame rate slider ranges from 1 to 30. The rates selected are the upper rates possible, but as activity in the image increases, the frame rate is likely to drop (due to processing demands), and the bit rate increase. It is also possible to alter the quality (i.e. resolution) of the image that is transmitted, although as the quality increases, so will the bit rate.

2.6.2.3 Reception statistics

Beside the thumbnail images that are shown in the main **vic** window, information about the video stream received from that person is displayed. It notes how many frames and kilobits per second are being received from one individual, as well as the packet loss from that person (see Figure 5). As in **RAT**, these numbers may differ from the rate the person is actually sending, due to network loss and the processing power of both computers (sender and receiver) involved.

2.7 Shared workspace

In the scenario of desktop multimedia conferencing, it is likely that in addition to audio and video, the participants will also be using some form of shared workspace, such that they can view or work collaboratively on a document. In the multicast community there are a variety of shared workspace tools available, but two of the most reliable and common are **wb** and **NTE**.

wb was developed at Lawrence Berkeley Labs as a means for remote participants in a conference to write, draw and type on the same workspace at the same time (Jacobson, 1993)[3]. The contributions from participants are visible to everyone in the conference. The tool offers a variety of drawing/typing colours and the opportunity to import text or postscript documents. However, **wb** does not function as a word processor, and editing and saving facilities are limited. For example, it is not possible to edit a segment of text written by someone other than yourself. If a text editor is required, **NTE** is a more powerful tool (Handley and Crowcroft, 1997). Developed at UCL, **NTE** allows any contribution to be modified (unless locked) by any other member of the group. Text can be moved and pointed at by any member, and all actions (and actors) are visible to other participants, minimising confusion.

In general, shared workspace tools do not suffer too badly from the effects of packet loss, since they do not demand a large amount of bandwidth. When network congestion is very severe, it is sometimes the case that the letters arrive out of order on the workspace, but the text will eventually be completed.

The tools that have been described here have all been designed with the aim of making it easy for people to participate in multicast multimedia conferences, but as shall be seen, evaluating multimedia conferencing from a human factors perspective is by no means straightforward.

2.8 Evaluating MMC communication

In a desktop MMC environment, participants are likely to be using audio, video, and shared workspace tools such as the ones described above. These tools have, in the main, been designed by, and for use in, the computer networking research community, and are therefore not necessarily particularly usable, at least by novice users. This has been demonstrated in a study carried out at UCL for the MERCI (Multimedia European Research Conferencing Integration) project, which assessed the usability of **vic**, and found that much of the **vic** options interface is confusing to novices, and even experienced users (Clark, 1997). Studies on other Mbone tools such as **SDR** (Session Directory[4]) and **RAT** have reported similar findings (Clark and Sasse, 1997; Bouch, 1997). In addition to the interfaces of the tools rarely being designed with novice users in mind, another usability problem is caused by the sheer number of windows that must be positioned on a screen in a conference in which there are a number of participants all sending video, conversing through the audio tool, and using shared workspace. Users often find that it is difficult to manage the screen 'real estate' effectively, especially when they are inexperienced in conferencing (Sasse et al., 1994b). To solve this problem, some researchers are beginning to develop integrated interfaces which organise the required windows in a sensible fashion, and hide the unwanted functionality. (However, this may cause a different set of problems, as will be discussed in section 5.1.)

However, usability of the tools is just one facet of evaluating MMC - more important than the tools themselves is the *quality of the communication* enabled by using the tools, and as Figure 6 illustrates, this overall quality can be affected by a myriad of factors.

Key factors affecting the overall quality of the communication can be usefully broken down into those of audio quality, video quality and the task being undertaken.

Perceived audio quality is affected primarily by the presence and degree of packet loss, and the method, if any, that is used to compensate for this loss. Perceived quality will also suffer from poor

[3] **wb** does not run on PCs. A similar tool, **WBD**, was developed for PC use at Loughborough University, and is now maintained at UCL.

[4] **SDR** is a session directory tool designed to allow the advertisement and joining of multicast conferences on the Mbone. Software available from http://www-mice.cs.ucl.ac.uk/multimedia/software/

quality headsets which may allow leakage from the headsets, resulting in echo and distracting background noise. In a conferencing environment subject to high levels of packet loss, the listener and speaker's language background will start to become more critical, since if people are trying to communicate in a non-native language, any degradation in the audio signal will make this process much harder. Female speech may not be transmitted as clearly as male speech, since female voices occupy a different frequency range from male voices, and it is the latter that have most often been modelled in synthetic speech codecs. Finally, if a large amount of compression is being used, delay will be increased, which can be detrimental to conversational behaviour.

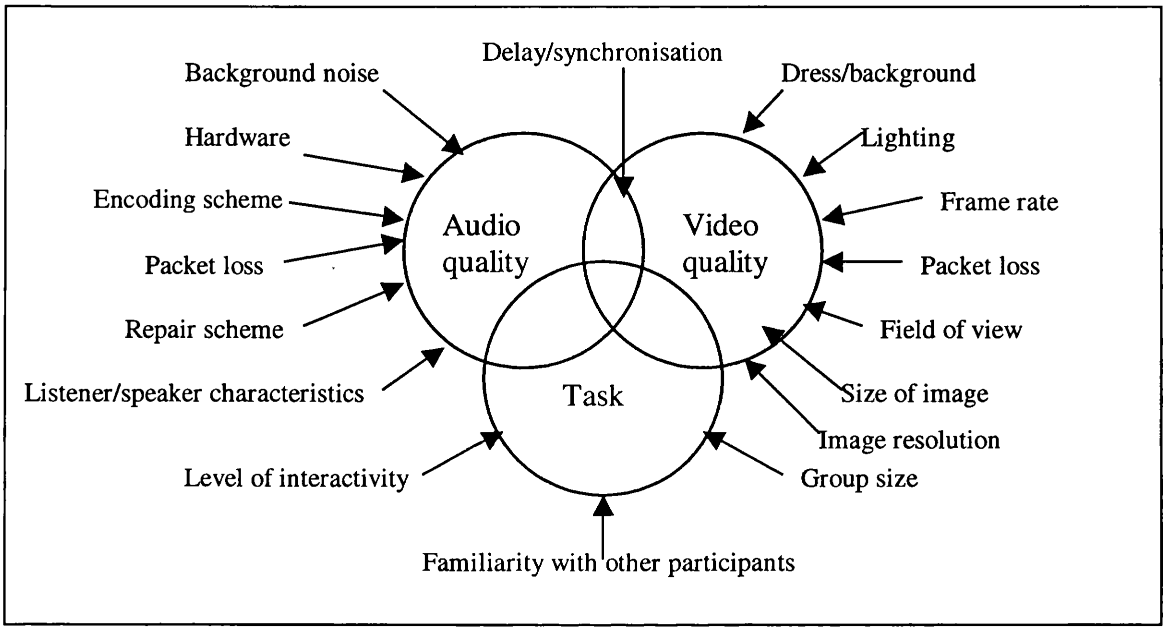


Figure 6: An overview of the main factors affecting perceived media quality in MMC

Delay is also a factor in perceived video quality. It takes comparatively longer to encode and decode video than it does for audio, and video tends to arrive at the receiver later than audio. The video stream is also updated relatively infrequently. These two facts combine to mean that a participant in a conference sees a low frame rate image that is not synchronised with the audio stream. Packet loss affects video transmission in terms of blocks not being updated at the same time as the rest of the image, so that the image can appear visibly 'blocky'. This 'blockiness' will become more apparent as the image size is increased, but in any case the image size is generally small, at least with H.261 encoding, and this is likely to affect perceived quality. Perception of the quality of the image will also be affected by the positioning of the remote camera - views from the side can make people look 'shifty' (Short et al., 1976) - and at no point will direct eye contact be possible. If the remote person is backlit then their features will not be visible at all, and with blocks of image being updated at different times, people are well advised to steer clear of checked or striped clothing.

In many desktop conferences, the focus of the conference participants will be the shared workspace, and often the audio or video quality will be commented on only if it becomes so poor as to be disruptive to adequate performance of the task. It is therefore important to ground the whole area of evaluating MMCs in the context of the task(s) being undertaken. It is easy to conceive of tasks that would require better quality than others. For example, a remote interview would demand higher audio and video quality than a routine project meeting between colleagues. In addition to the actual task or goal of the conference, there are certain other task factors that will impinge upon the perceived quality of the delivered data. The size of the group participating in the conference, for example, will determine how many video streams have to be decoded by the local workstation. If the participants are familiar to each other, then they are more likely to tolerate a lower quality than if they are unknown to each other, and of course if participants are not interacting to a great degree, the requirements will differ from a situation in which a high degree of interaction is engaged in.

These are some of the many variables that need to be considered when evaluating multimedia conferencing. Clearly, not all of these factors will carry the same weight in forming subjective opinion of the quality of a conference, but they are all likely to contribute, and so an awareness of them is necessary (especially when comparing experimentally controlled results with observations in the real world). A more in-depth consideration of some of the factors is presented in the next chapter.

This chapter has presented both the academic framework and the real-world setting within which the thesis research is undertaken. With the aim of assessing perceived quality of audio and video in MMC, it is necessary first to understand the nature of spoken and visual communication when it occurs through video-mediated systems. This topic is covered in Chapter 3.

Chapter 3: Speech and video communication

Although the research conducted in this thesis took place using multicast technology over IP networks, this chapter will discuss the issues involved in video-mediated communication (VMC) in general, since a thorough understanding of the current state of knowledge and issues entailed in mediated real-time communication is required. It is also contended that the observations and results arising from the thesis research will be applicable to video-mediated communication across all networks.

This chapter therefore refers to VMC mainly, and MMC when reference is being made specifically to multimedia conferencing in multicast environments.

3.1 Audio-visual communication

From Gestalt theory (Kohler, 1930), it is known that "*the whole is greater than the sum of its parts*", and this applies to the whole of multimedia conferencing and its parts, audio, video (and shared workspace). The interaction between audio and video is a complex one, and needs to be considered especially with regard to VMC. The effects on perception when both media are present can outweigh the effects of each medium considered in isolation. User perception of a multimedia system therefore cannot be accurately predicted by investigating the individual media in isolation. The effect of the interaction between the media must be taken into account, in the context of the task being undertaken by the participants.

That audio and video quality can influence the perception of the other media is a given. Ostberg et al. (1989) cite Risberg and Lubker (1978) as reporting that, when subjects were presented with video with no sound, they were able to correctly identify 1% of the test words. When they were presented with no video and a low-pass filtered version of the speech sounds, they correctly identified 6% of the material. When the video and degraded audio were presented together, the correctly identified material leapt to 45%. This clearly demonstrates the synergetic nature of presenting audio and video in combination. Earlier work by Sumby and Pollack (1954) showed that the percentage of words correctly identified in noisy environments was always greater when video accompanied the audio, than when audio alone was presented. Increasing or improving the quality of one medium can also lead to the perception of improvements in another, when there has been no actual increase or improvement. Neuman (described by Negroponte, 1995) showed that improving the quality of the sound accompanying high definition television (HDTV) video led to users reporting that the *video* had improved in quality even though there had in actual fact been no accompanying improvement.

Before continuing further it is important to gain an understanding of the individual facets of spoken and visual communication. Since VMC often entails degraded transmission quality of either one or both of the media, an understanding is required of what types of information are transmitted through the media, and how they suffer under degradation.

Audiovisual communication can be broken down into two sub-components: speech[5] and hearing, and visual communication (looking and seeing).

3.2 Speech and hearing

3.2.1 Speech

Speech is produced via the manipulation of structures of the vocal tract. The vocal tract is about 17.5 cm long in an adult male, and stretches from the larynx at the top of the trachea to the lips and nose.

Speech can be described by the discrete linguistic elements of which it is composed. The sounds of speech are represented linguistically by phonemes. The boundaries between phonemes in natural speech are often difficult to identify, but a useful definition of a phoneme is that it is the shortest segment of speech which, if substituted for another, would change the meaning of a word. /p/ and /k/ are two distinct phonemes, demonstrated by their presence in the words ‘pill’ and ‘kill’.

Phonemes are classified as consonants or vowels according to the articulatory gesture made in forming them. Vowels are produced when the vocal cords vibrate as air flows through the mouth in an open static configuration. The shape of the tongue and the lips is what distinguishes one vowel sound from another. When part of the mouth is constricted to a near or complete closure, a consonant is formed. This closure prevents the rush of expelled air through the mouth that would otherwise occur, causing a distinctive sound.

Consonants are the most important speech sounds (Voiers, 1977): they contain most of the important clues as to the identification of words. Miller and Licklider (1950) have reported that a monosyllabic word is likely to be perceived incorrectly if either its initial or final consonant is missing. Consonants differ from each other in 3 ways: place of articulation, manner of articulation and voicing.

[5] The sub-component of music is not addressed in this thesis.

- **Place of articulation** refers to the point where the oral cavity is constricted to reduce the flow of air in order to form the consonant.
- **Manner of articulation** refers to the articulatory gesture by which sound is produced.
- **Voicing** refers to whether the vocal cords are involved in the production of the consonant. When a consonant is produced via vibration of the vocal cords, it is called a voiced consonant. Otherwise it is known as an unvoiced consonant. This is often the only distinguishing feature between two consonants that share the same place and manner of articulation e.g.

  /t/ (unvoiced) - /d/ (voiced)
  /f/ (unvoiced) - /v/ (voiced)
  /p/ (unvoiced) - /b/ (voiced)

A summary description of these features is provided in Table 2.

Speech moves along through a stream of syllables, wherein the middle of each stands one or more vowels. As the tongue and mouth move from the production of one vowel to the next, any intervening consonants are pronounced. Therefore the vowel has great influence on how the consonant will be produced. Not only this, because the consonant is in such close proximity (both temporally and literally) to the vowel, what is produced is a consonant containing much of the information from the vowel. This phenomenon is known as *coarticulation.* Coarticulation is defined as occurring when the vocal tract can be determined to be adjusting to the production of two or more sounds at that one moment in time. The coarticulatory effect can be anticipatory or retentive, i.e. the vocal tract can show the characteristics of a future or a past phoneme at any one moment in time.

Manner **Place**	**Stops** *Closure of vocal tract at place of articulation, followed by release of air*		**Fricatives** *Incomplete closure at place of articulation*		**Nasals** *Passage of air through nasal cavity instead of mouth*		**Glides/semivowels** *Shaping of tongue in different ways; characteristics lie between consonants and vowels*	
	Unvoiced	**Voiced**	**Unvoiced**	**Voiced**	**Unvoiced**	**Voiced**	**Unvoiced**	**Voiced**
Labial *Two lips together*	/p/ ‘pin’	/b/ ‘bin’				/m/ ‘sum’	/hw/ ‘what’	/w/ ‘will’
Labiodental *Bottom lip against upper front teeth*			/f/ ‘fine’	/v/ ‘vine’				
Dental *Tongue against teeth*			/θ/ ‘thigh’	/δ/ ‘thy’				
Alveolar *Tongue against alveolar ridge*	/t/ ‘tin’	/d/ ‘din’	/s/ ‘sip’	/z/ ‘zip’		/n/ ‘sun’		/l/ ‘less’
Palatal *Tongue against hard palate*	/tʃ/ ‘char’	/dʒ/ ‘jar’	/ʃ/ ‘ship’	/ʒ/ ‘azure’				/j/, /r/ ‘yes’, ‘rim’
Velar *Tongue against soft palate*	/k/ ‘kilt’	/g/ ‘gilt’				/ŋ/ ‘sung’		
Glottal *Glottis in throat*			/h/ ‘hill’					

Table 2: Summary of consonant classification (adapted from Warren, 1982)

3.2.2 Hearing

The frequency range that people can hear is much larger than the range in which speech sounds are produced. A healthy young adult can hear sounds between 20 Hz and 20,000 Hz, while the general speech range is 50-10,000Hz. As people get older, they gradually lose sensitivity to high frequencies (a process known as *presbycusis*). By the age of 70, most people have trouble hearing frequencies above 6000Hz, impairing their ability to hear certain speech sounds[6]. In general, consonants consist of high frequency sounds, while vowels are primarily made up of low frequencies (Sekuler and Blake, 1994).

Just as different birds sing at different frequencies, when different nationalities speak, they occupy different frequency ranges. Most sounds of a language are located in one or several frequency zones. For example, most English speech sounds occur between 2000 and 8000 Hz, whereas most French speech sounds are found between 125-250 Hz and 1000-2000 Hz. The French ear is therefore used to perceiving different frequencies from an English speaker's ear (ARC report).

Hearing with two ears (binaural hearing) allows sounds to be located in space (*sound localisation*). The brain compares the sounds received by both ears in two ways: *interaural time difference* (difference in time of arrival of sounds at the two ears) and *interaural intensity difference* (sound energy will be more intense at the ear located nearest the source). Sound localisation depends on interaural time differences at low frequencies and interaural intensity differences at high frequencies. These effects are not found in multimedia conferences where mono, rather than stereo, transmission is normal, and sound is delivered through headphones. Therefore confusion may arise in conferences involving many participants, since there is no means of determining from which (virtual) direction a speaker's voice is coming.

3.2.3 The impact of degraded speech on understanding

As was discussed in section 2.4, one of the main problems with communication over packet networks is the possibility of packet loss. In the case of audio transmission, packet loss can mean the loss of phonemes or syllables[7], implying a disruption of speech intelligibility. As will be seen, however, the ear and brain are surprisingly resilient to different types of speech degradation: there is

[6] This loss of hearing is not critical, as demonstrated by the fact that we able to carry out conversations with ease on the telephone, the upper limit of which is 3.4 kHz (which explains why it is harder to recognise familiar voices over a phone line).

[7] Although there remains some debate over whether it is the phoneme or the syllable that is the critical unit of speech intelligibility (Segui, 1984), it is certain that the critical elements are very short in duration, and as such, packet loss will affect the perception of these.

no direct relationship between speech intelligibility and degradation. This is mainly due to the effects of the surrounding speech context.

Miller and Licklider (1950) performed an early investigation into the intelligibility of interrupted speech. Employing a speech-silence ratio of 50%, they found that when speech is interrupted less than 10 times per second, intelligibility is impaired, but when the interruptions occur *more* than 10 times per second, intelligibility does not suffer. This means that when the bursts of silence were greater than 50 ms in duration, the intelligibility of the speech was adversely affected, but when the bursts of silence were shorter than this, speech intelligibility was not affected. The explanation is that critical features of speech e.g. phonemes, are being blocked out by bursts of silence larger than 50 ms, but when the bursts are smaller, critical speech constituents are not being eliminated.

There is a considerable body of literature that shows that intelligibility can be boosted by the effects of context. Warren's (1970) *phonemic restoration effect* is a case in point. Warren extracted the phoneme /s/ from the word 'legislature' and played subjects the sentence *"The state governors met with their respective legi*latures convening in the capital city*", replacing the missing phoneme with a cough. It was found that subjects perceived both the /s/ and the cough. There was no possibility of coarticulation accounting for this finding because the relevant portions of the adjacent phonemes were also removed. None of the subjects were able to identify where in the speech stream the cough had occurred, illustrating that the perception of speech from its constituent elements is an unconscious process. The effect of context is also illustrated by Warren et al. (1969), who found that subjects were unable to determine the order of a hiss, buzz, tone and vowel when these were played to them at a rate faster than 1.5 sounds per second. Speech sounds, on the other hand, are' produced at a rate of 10 per second and fully comprehensible, so the importance of a syntactic/semantic context is clear.

It is not only the smaller elements of speech such as phonemes that are subject to context effects. Words and sentences are processed more efficiently if there is some kind of surrounding context. Miller et al. (1951) presented subjects with words isolated in noise or within sentences in noise. Words within the sentence context were identified better. Miller and Isard (1963) gave subjects a recognition task and found that both syntax and semantics aided recognition.

Pollack and Pickett (1963) demonstrated another aspect of the importance of context. They extracted individual words from everyday conversational speech and played them to subjects. They found that subjects were able to correctly identify single words less than 50% of the time, despite the fact that the speech was considered perfectly intelligible when played in full. It appears that listeners compensate for 'sloppy' speech by unconsciously picking up cues from the phonetic, syntactic and semantic information contained in the surrounding speech.

Although people are relatively poor at identifying isolated words, it has been demonstrated that this performance drops further when the words are produced synthetically. Replicating findings by Luce et al. (1983), Waterworth and Thomas (1985) showed that synthetic words are harder to recall than naturally produced words in a serial recall test. They then went on to demonstrate that this is due to the fact that people are poorer at identifying and encoding synthetic words than natural words. It is the encoding difficulty that leads to the poor memory performance: synthetic words require more processing power by the listener than natural words. This has implications for the method chosen to repair packet loss in MMC - the more synthetic sounding the speech, the more likely that it will cause memory and other difficulties, as a person expends more effort on merely discerning the words.

3.2.4 Speech and hearing in MMC

In MMC, participants either wear headsets or converse via a microphone and speakers. With either option, there are currently no available cues for sound localisation, and the speech bandwidth is in any case narrow (although not quite as narrow as the telephone). In addition, echo and feedback can be caused through 'leaky' headsets or faulty echo cancellation using speakers and microphone. Speaking into a microphone can also lead to the feeling that the channel is 'dead' since there may be no *sidetone*[8] functionality. This impression can be particularly acute when using a headset, since with over-the-ear sets it can be hard to hear one's own voice. This can also lead to either speaking too loudly or too quietly.

In addition to these hardware issues, speech and hearing in MMC is affected by network conditions, primarily packet loss and delay. Packets tend to contain 40 or 80 ms worth of speech information, which is approximately the size of a phoneme, so the effects of packet loss will obviously have an impact on speech intelligibility. As discussed in section 2.5.2.5, there are different means of compensating for packet loss, but the perceptual consequences arising from these different methods can be very varied. Silence substitution, for example, is the easiest method to employ, but gaps in the speech stream usually signal that a speaker has come to the end of a talkburst, and that it is time for another person to take a turn. The use of silence substitution, therefore, may increase the number of interruptions between participants in a dialogue, and impair the smooth flow of the conversation. A better means of repairing packet loss can be gained through exploiting the phenomenon of *phonemic restoration* (as afforded by waveform substitution, for example). By inserting *noise* in place of the missing packet, the brain will hopefully hear the missing speech. However, this method will only suffice when the degree of packet loss is low: once the speech characteristics begin changing, the missing information will become more apparent. A third method, synthetic redundancy, offers a means of providing a synthetic copy of the *actual speech* that has been lost. However, since the literature suggests that people find it harder to process synthetic speech, it may be that this method could be more harmful than helpful once a certain level of packet loss has been reached.

The impact that end-to-end delay can have on speech and listening is striking. In particular, turn-taking becomes problematic, and the arrival of inappropriately timed *backchannel* responses (e.g. "*mmm*", "*uhuh*") can disrupt the smooth flow of a conversation. In an investigation into the effects of delay on different conversational tasks, Kitawaki and Itoh (1991) concluded that the impact of delay depended on the task, and the degree of understanding of the subject as to why the delay was occurring. They concluded overall that a round-trip delay of 500 ms (end-to-end 250 ms) was the upper limit that could be tolerated without serious drops in conversational effectiveness. The impact of delay will be discussed further in section 3.3.7.

The main factors involved in speech and hearing, and how they might be affected by transmission in MMC environments, have been presented. However, one major contributor to the perception of speech is the information that can be gleaned from the *visual* channel, and a discussion of visual communication is now provided.

3.3 Visual communication

When people communicate using the visual channel as well as speech, a wealth of additional information becomes available to the listener, including vocal tract and lip movements, facial expressions, gaze, gestures and body language. The following sections describe these sources of information in more detail and discuss their support in VMC.

3.3.1 'Reading' speech from the face

The importance of lip movements in helping to discern speech in noisy environments has long been known (e.g. Sumby and Pollack, 1954). For example, consonantal phonemes which may be difficult to distinguish aurally, such as /f/ from /s/, and /n/ from /m/, are often readily distinguished

[8] Sidetone refers to the practice of diverting sound from the microphone to the earpiece of a telephone, such

visually from movements of the lips. However, 'speech-reading' information is not only provided by the lips. Bruce (1996) reports that schematic faces (where teeth and lips are represented) produce recognition performances of 50-57% accuracy for vowels, compared to 78% accuracy with a real face, providing evidence that 'lip'-reading information relies on more than just the lips. Summerfield (1992) suggests that this additional information may include perception of the tongue or wrinkling and protrusion of the lips, but it is likely that movement of many other parts of the face, such as the chin, cheeks and eyebrows also give clues to what people are saying. Gross movements of the head such as nodding or shaking also provide more obvious visual communicative cues (see section 3.3.8).

The impact that vision has on speech perception is perhaps best illustrated by the McGurk effect. McGurk and MacDonald (1976) dubbed the soundtrack for one syllable onto the visual information for another syllable. When the audio and visual information was played at the same time, they found that subjects sometimes heard what the lips were saying, but more often found that a third unrelated syllable was heard. For example, when the syllable /ba/ was dubbed onto the lip movements for the syllable /ga/, the subjects nearly all reported hearing /da/. Otherwise, subjects reported the visual 'sound' rather than what was heard. Dodd (1977) found that the effect was not restricted to syllables - subjects sometimes heard the word 'towel' when the sound of the word 'tough' was dubbed over the lip movements for 'hole'.

The McGurk effect is generally taken as evidence of the multimodal nature of speech perception (and the dominance of the visual channel on perception), and therefore its production can be taken as evidence that the visual and auditory channels in a communication system are working 'as normal'. However, in VMC any mismatch between the timing of the audio and video streams would increase the chance of the effect occurring unintentionally, and perhaps affecting the speech communication adversely, and therefore it is an effect one should seek to avoid.

3.3.2 Speech-reading in VMC

As discussed in section 2.2, in networked communication audio and video channels are transmitted as separate digital streams, and the video frames are often slowly updated, meaning that facial expressions may not be synchronised with accompanying speech. However, people are surprisingly tolerant to asynchrony between the audio and video streams, especially when the audio lags behind the video. This can be explained by the fact that sound travels slower than light in the everyday world (Munhall et al., 1996; Bruce 1996) and also perhaps by the fact that there is visible motion in

that the speaker hears his own voice at the same level as the listener.

the place of articulation before the onset of the sound (Munhall et al., 1996). Nevertheless, there is a limit to the asynchrony that people will tolerate. Jardetzky et al. (1995) reported that subjective studies have indicated that the mismatch in audio and video streams can be in the region of 80-100ms before a lack of synchronisation is perceived. Roy (1994) states video-to-voice lags are tolerated in the range between -90 and +120 ms. As a rough guide, then, plus or minus 100 ms seems to be the limit of perception (Fluckiger, 1995).[9]

Despite the tolerance of viewers to asynchrony, it must be assumed that any occurrence of audio/video asynchrony increases the likelihood of a perceptual effect such as the McGurk Effect. Further investigations of this effect have revealed aspects that are especially pertinent to VMC, since temporal coincidence of information from the auditory and visual channels is not necessary for the McGurk effect to occur: the effect can be produced even when the audio information lags the visual information by as much as 180 ms (Munhall et al., 1996). Although the effect is weaker when video lags audio (the more likely playout order in VMC), Massaro and Cohen (1993) showed that syllabic misperception can be generated when auditory stimuli precede visual stimuli by as much as 200 ms.

Recent findings, however, suggest that the likelihood of the McGurk effect happening in VMC is not high. Jordan and Sergeant (1998) found that people are more *resistant* to the McGurk effect (i.e. they correctly identify more auditory stimuli) when images are very small, as they tend to be in desktop VMC. In addition, Bruce (1996) reports that schematic faces combined with auditory speech sounds also leads to a reduced incidence of the McGurk effect. Since many VMC systems have low quality video, this finding suggests that a low quality picture might actually be 'safer' than a good quality picture when the audio and video streams are not synchronised.

Overall then, with respect to the low frame rates, small images and low quality currently common in most MMC systems, and the fact that video tends to lag audio, the McGurk effect may not seem very threatening. However, considering the Munhall et al. (1996) finding that temporal coincidence is not necessary for the effect to occur, it is a consideration to be aware of. The literature suggests that 15-17 frames per second are sufficient to perceive actual speech/facial movements (see next

[9] In the psychological literature the perceptual limit identified has tended to be longer than that identified in the networking literature. For example, Munhall et al. (1996) cite the work of Dixon and Spitz (1980) who showed that audio lag had to be greater than 250 ms before subjects noticed an audiovisual discrepancy for the test syllables. This is likely to be the key to the discrepancy between psychological and networking findings: Munhall et al. (1996) reports that the psychological research has tended to focus on syllabic perception rather than continuous speech. The figures arrived at by the networking community investigating the asynchrony tolerated in real speech should therefore be taken as a more realistic guide.

section), and frame rates this high are already commonplace in some applications, meaning that the effects of a mismatch between audio and video could become critical. One important proviso though is the fact that the vast majority of research into perceptual effects such as the McGurk effect has been passive, i.e. subjects are asked to identify what was said without having to interact with the speaker, allowing attention to be fully focused on the speaker (Anderson et al., 1997a). What happens in an interactive setting is as yet unclear.

Regardless of the McGurk effect, there is much evidence to suggest that visual information plays a more significant role in speech perception at low frame rates than might initially seem likely. This issue is considered in the following section.

3.3.3 The impact of frame rate on speech reading

In the natural world, the human face is in almost continuous motion - facial expressions are dynamic, and can be as brief as 50ms in duration (Bruce, 1996). A reduced frame rate may therefore be misleading for the correct perception of emotions, especially if the audio in use is compressed or synthetic sounding such that prosodic cues are missing or altered. A person's momentary expression of faint surprise may represent a point of increasing or decreasing amazement, so its relationship to both prior and subsequent expressive movements, and to concurrent events in the world, needs to be known in order to interpret it properly.

Full motion video requires 25/30 frames per second, but perceptual benefits to communication can be gained from frame rates far lower than this. Vitkovitch and Barber (1994) report that a rate of 16.7 frames per second may be sufficient for the transmission of speech/facial movements, whilst Frowein et al. (1991), in studying videotelephony for the hard of hearing, found there was no difference in speech perception when 15 or 30 frames per second was used. This may be because vocal tract movements are relatively slow - less than 20 Hz (Munhall et al., 1996). These results seem to indicate that a minimum of 15 frames per second is required for adequate speech reception. However, Frowein et al. (1991) found that the addition of a video signal even at very low frame rates (5 fps) led to a significant improvement in speech reception[10]. Barber and Laws (1994) report that, while lip-reading performance at 25 fps is 70% accurate, at 8.3 fps the performance is 55%, which is still a surprisingly high figure. Ostberg et al. (1989) found that *any* increase in visual representation of the speaker increases the listener's tolerance of noise. Nevertheless, it should be stressed that the frame rate required is likely to be highly dependent on the task being undertaken.

The RACE TELEMED project (1992) investigated the effect of frame rate on non-native lip-reading performance, since anecdotal evidence suggested that non-native speakers of English found it easier to comprehend English when they could see the lips of the speakers. Interestingly, the researchers found that there was a difference in performance between French and German speakers (French speakers performed better), and that with French and English lipreading studies, there seemed to be a slight advantage for a frame rate of 16.6 over 25 fps. This latter finding might be due to the fact that subjects perform better when they have to concentrate harder, a claim Reeves and Nass (1996) make with respect to impaired audio.

3.3.4 The impact of frame rate on interactive task performance

It has been suggested by some researchers that video at very low frame rates is not useful for enhancing task performance, and might actually be detrimental to communication (see Whittaker, 1995). However, the effect of low frame rate video has tended to be measured in terms of the *process* of the interaction, i.e. number of turns taken by participants and the degree of simultaneous speech. Studies that have investigated the performance of a task, on the other hand, have found little evidence that a low frame rate *impairs* performance. In studies investigating remote education using desktop VMC, Kies et al. (1996) and Hearnshaw (1999) observed no impact of low frame rate on learning. The RACE TELEMED (1992) project found no significant difference in comprehension performance between a frame rate of 25 or 8.3 fps. A possible interpretation of these low frame rate findings is that people find low frame rates distracting, and therefore rely on the audio channel. This is supported by a TELEMED finding that there was an advantage for audio-only over audio plus video. However, in this study the audio link was of high quality: when the investigators repeated the experiment with background office noise, they found that the advantage for audio-only disappeared. The result was interpreted to suggest that as auditory interference increases, people will rely more on visual cues and information, a conclusion consistent with the work of others (e.g. Frowein et al., 1991; Sumby and Pollack, 1954). In the Kies et al. and Hearnshaw studies, however, the audio was *not* of high quality (i.e. bandwidth), and still no impact of frame rate was observed.

Other findings (Anderson et al., 1997b) suggest that a *high* quality of video can actually impair conversational behaviour (in terms of turn-taking and amount of speech generated). They investigated the task performance and communicative data for two different tasks, the Map Task

[10] Interestingly, this frame rate also seems to be the level at which the reading of finger-spelling and signing is just receivable.

and the Travel Game[11]. They found that with high quality video and the Map Task, turn-taking was increased by 11% and 10% more words were used than in face-to-face (FTF) or audio-only. However, when the Travel Game was played with only 4-5 frames per second, there was no difference in behaviour between VMC, FTF and audio-only. Like so much of the literature on VMC, it is likely that the impact of frame rate on task performance depends on the task that is engaged in.

Leaving aside the issue of frame rate, the presence of a visual channel affects communication in several other ways.

3.3.5 Speech production

Anderson et al. (1997a) report that face-to-face words are enunciated less clearly than words where there is no visual contact. This effect occurs regardless of whether or not participants are gazing at each other, and thereby utilising visual cues to help decipher the speech. The interpretation is that, if speakers know that their listeners can see them, they are sloppier in their delivery of speech. If a more careless delivery of speech were to be found in situations where the quality of video is not good enough for visual aids to speech perception, such as lip movements, to be transmitted, the implications for successful speech communication would be dire. Interestingly, however, Blokland and Anderson (1998) report that when frame rates are low (5 fps), subjects enunciate *more* clearly than in face-to-face settings. Although the authors interpret this as a negative finding (since it does not emulate FTF behaviour), it should be interpreted as a positive finding in terms of MMC communication, especially when there is a degraded audio link.

3.3.6 Gaze and taking turns

As Bruce (1996) states, *"The prolonged stare means something different from the brief glance."* The issue of eye contact and gaze is an important one for VMC, and a more detailed discussion of gaze is now turned to, and how its presence or absence affects communicative behaviour.

There are two types of gaze: *regular gaze* (speaker looking at listener or listener looking at speaker, usually focusing on the mouth) and gaze in which eye contact is made, *mutual gaze* (speaker and

[11] The Map Task (Brown et al., 1984) is a collaborative problem solving task where participants must agree on a route from two slightly different maps. The Map Task was developed in order to compare FTF with audio-only communication. The Travel Game was developed to capture more of the social interaction that can occur in VMC, and involves having to plan holiday routes (see Anderson et al. 1997b).

listener looking at each other). Research has shown that neither type is used as often as might be supposed during conversations. In face-to-face situations without visual distractions, Argyle (1990) noted that gaze occurs only 50% of the time at the most, and the occurrence of mutual gaze is much lower. Anderson et al. (1997a) suggest three reasons for this surprising lack of gaze in conversations:

- eye contact can be taken as an unwanted indication of intimacy;
- looking at an interlocutor's face can distract the speaker from thinking and planning speech;
- there may be a more pressing need for other visual information.

(Indeed with respect to this latter point, Anderson et al. (1997a) showed that, when visual distraction is provided (in the form of the Map Task), only 30.2% of the speech time was concurrent with gaze, and mutual gaze was found only 2.7% of the time. This finding is important since it can be safely assumed that many VMC tasks will involve visual attention being directed at the shared workspace.)

In VMC, however, gaze - especially mutual gaze - cannot easily be afforded. The key problem is that of *visual parallax* (Short et al., 1976). Ostberg et al. (1989) expressed this as follows: "*... a camera positioned at the periphery of the display screen introduces a significant parallax - seeing into the (virtual) eyes of your interlocutor does not mean that he/she will be seeing into your (virtual) eyes. Direct eye contact is lost, and the resulting eye-gaze may actually be a source of misinformation*".

However, even when eye contact is enabled in VMC (using an arrangement of mirrors to produce a 'videotunnel'), different conversational behaviour is found in VMC compared with FTF communication. O'Malley et al. (1996) compared frequency of gaze in high-quality VMC with that of FTF communication and found that subjects using video gazed around 56% *more often* than subjects in FTF communication. They postulate that this finding might be due to the novelty value of VMC, or that the lack of co-presence leads to more gaze in an effort to gain more visual cues. Whatever the reasons, it is clear that enabling eye contact in VMC is not sufficient to replicate FTF communicative patterns of behaviour.

In most desktop VMC systems, though, eye contact is not possible, and the assumption is that the communicative benefits that arise from gaze will be lost. Anderson et al. (1997a) summarise three main functions for gaze in FTF communication:

- indicating that an interlocutor is attentive to a message;

- emphasising a particular word through brief eye contact;
- facilitating turn-taking in a dialogue.

With respect to VMC, it could be argued that attentiveness and emphasis might be indicated by facial expression other than eye contact, for example by nodding and raising eyebrows. However, irregular frame update rates may be misleading: what moment in time is being shown by the video image? The facilitation of turn-taking without recourse to eye contact poses a greater problem though.[12] Sellen (1992) reports that speakers tend to terminate their turn with a sustained gaze. In VMC this behaviour can be problematic: when frame rates are low it could appear that a speaker is maintaining a long gaze, when in fact the frame has not been updated recently enough. Additionally, since every viewer receives the same view of, and visual cues from, the speaker, how can one tell whom the speaker is gazing at or handing over to (Gale, 1991)?

This might suggest that what would arise is a chaotic series of interruptions, but in fact the consequences of lack of eye contact in VMC are quite the opposite: turn-taking tends to become more structured (Sellen, 1992). Overall, participants in VMC tend to take more turns and use more words, but interruptions are fewer (see Whittaker, 1995; Anderson et al., 1997b). Blokland and Anderson (1998) attribute this to *"the general fact that communication becomes more formal when it is mediated"*. There seems to be something intrinsic to VMC that makes it more formal: with eye contact enabled through a videotunnel, O'Malley et al. (1996) found that even more turns were taken and more words used than in a condition *without* eye contact or audio-only. This is an interesting result, since Boyle et al. (1994) reported that in FTF communication *fewer* words and turns were required compared to audio-only communication.

Although turn-taking in VMC becomes more formal, task outcome is not usually affected, with one exception: when the link between participants is *delayed.*

3.3.7 Delay

In section 3.2.4 the impact of delay on audio-only conversations was discussed. Although the effects of delay are most critical with respect to audio, how does the presence of a delayed video channel affect communication?

Anderson et al. (1997b; also reported in O'Malley et al. 1996) hypothesised that a delayed video channel in addition to a delayed audio channel might lessen the detrimental impact seen in audio-

[12] In MMC the issue of turn-taking is especially important since the technology is scalable and there is, in theory, no limit to the number of participants in some conferences (Handley et al., 1993).

only delay studies. However, even though the delayed channels were synchronised (so that the overall delay was 500 ms), Map Task performance was significantly impaired, and the number of interruptions even *greater* than in an audio-only delay condition. The finding suggests that a visual channel does little to ease the negative effect of delay. However, the study involved the use of a videophone for the delay conditions, which had slow frame rates, so even though the delayed streams were synchronised, conflicting communicative signals may have been received from the video image. This makes teasing out the effects of delay from those of frame rate difficult, and the validity of the comparison should perhaps be questioned.

There can be little doubt that delay does impair the smooth flowing of conversations, however. Surveying the somewhat meagre research into the perceptual effects of audiovisual delay in the networking literature, Roy (1994) concludes that ideally end-to-end delay should be less than 300 ms in duration.

3.3.8 Image size and gesture

In videotunnel experiments, the image size is generally large (life-size). In desktop conferencing, on the other hand, there is usually not enough space for large images, and the bandwidth and processing requirements in any case tend to restrict the size of the image. Investigating the effect of absolute image size on gaze behaviour, Monk and Watts (1995) reported that a small video image (40 x 65 mm) results in a less fluent verbal interaction (in terms of turn-taking), but that gaze behaviour is unaltered compared to a larger video image (103 x 140 mm). There has been relatively little other research into the effects of image size outside the field of remote education. These results have shown that although people prefer larger image sizes (i.e. CIF-sized) on the whole, their objective learning has not been affected by the smaller image sizes (i.e. QCIF) (e.g. Kies et al. 1996; Hearnshaw, 1999). Reeves and Nass (1996) found that people recall more from larger images. However, this recall is likely to be linked to the larger image itself, and not necessarily to do with the purpose of a conference.

It is, of course, not just the size of the video window that will have a bearing on perception, but the *content* of the window. Gestures and posture can be valuable sources of communicative information, and in lieu of direct eye contact they may become more important in VMC. However, the camera position and angle common in desktop VMC means that full body views are uncommon. Heath and Luff (1991) observe that in FTF communication gestures catch attention because they are generally in the periphery of the visual field. Gestures in VMC, on the other hand, are not effectively supported since they either are not visible (due to the field of view or camera angle, or because frame updates are slow enough to miss them), or they constitute a very small area of the total visual field (i.e. the computer screen).

In order to maximise the sense of presence of the remote participant(s), Reeves and Nass (1996) recommend that in desktop situations the camera should zoom in as close as possible to the person. Other researchers have found, however, that head and shoulders views are preferred over head only (e.g. Frowein et al., 1991). When listeners look at a speaker, they take in additional information to that available from the face. Support for this comes from O'Malley et al. (1996), who found that more words were produced in a head-only condition, suggesting that the additional visual cues available in a head and shoulders image condition might enable more effective communication.

3.4 Video - to use, or not to use?

Much research into VMC appears to be based on the premise that the use of a video channel can only be justified if FTF communication is emulated. Surveying the findings from studies comparing users working on a FTF group task with groups using audio and video channels and solely audio channels, Whittaker (1995) concludes that video adds little or no value to task performance, and that audio is the main medium of communication. He argues the finding that even high-quality videoconferencing fails to replicate face-to-face communication processes is a result of the fact that "... *most videoconferencing systems do not support directional sound or visual cues. They tend to present sound and picture from a single monitor and speaker which may compromise sound direction, head turning and gaze cues in group interactions.*" Others (e.g. O'Malley et al., 1996; Anderson et al., 1997b) postulate that the video channel causes a cognitive load that impairs conversational processes. Whatever the underlying reasons, there is ample evidence that FTF behaviour cannot be adequately emulated by VMC. But how much does this matter?

It could be argued that the main utility of video in VMC lies not in the determination of speech sounds or the maintenance of eye contact, the speed of task outcome or the amount of turn-taking, but rather the *sense of presence* that it allows. People simply prefer the video link to be there, even when the frame rate is very low (5fps), as reported by Tang and Isaacs (1993). Since it is known that the presence of video rarely impairs task outcome (Whittaker, 1995; Anderson et al., 1997b), it is not clear that its use should be avoided simply because it does not allow FTF communication to be emulated. Studies have shown that users make more use of the video channel than they think they do (suggesting that it may be "... *intuitive and transparent to the communication process*"

Gale, 1991)[13]. Indeed, Tang and Isaacs (1993) report that users are disinclined to use desktop conferencing when there is no video element. They concluded that the users' desire for video arises from the impact video has on the *process* of their interpersonal interaction, rather than from any perceived effect on the *product* of their interaction on a task.

Many studies assert that users perceive the video channel to add value by offering a sense of presence to the remote participants. Although people can translate cognitive and turn-taking cues across to the audio domain, video remains much better at communicating affective cues: images on a screen reassures that other participants are present (Whittaker, 1995). Gale (1991) found that perceived *social presence* (Short et al., 1976) increases as a function of the number of communication channels, i.e. shared whiteboard had less than whiteboard and audio, which in turn had less than the full complement of whiteboard, audio and video. In general, people tend to believe that the addition of video to an audio channel renders conversation more natural, and increases the ease with which interruptions can be made and attention tracked, regardless of objective findings that may counter this (Whittaker and O'Conaill, 1997). This distinction between subjective and actual behaviour is an important one. Subjective opinion will make the difference between a system being used and not being used, so affective and emotional elements should not be marginalised or ignored.

Although little evidence has been found to support that the addition of a visual channel helps in many collaborative tasks, it is accepted that it *does* help in tasks that involve conflict resolution and negotiation (Short et al., 1976). Veinott et al. (1997) showed that the definition of 'conflict resolution' could perhaps be broadened too, by showing that video improved performance when non-native English speakers performed the Map Task together in English. No such advantage was found for native speakers. It can therefore be safely said that researchers are still working to determine which situations benefit from a video channel, and in lieu of evidence that it harms task outcome, it would seem safer to leave it in real-time mediated communication, as long as it is not at the cost of audio quality (see Whittaker and O'Conaill, 1997)[14].

[13] This seems to contradict the suggestion of O'Malley et al. that people use video more because they are *not* used to it. There may, however, be no contradiction here: it is quite possible that people adapt to the use of video very quickly (until they are unaware that they are using it). The O'Malley study involved subjects using the equipment for one hour only, whereas the Gale study involved was a large repeated measures study over a period of 3 weeks, meaning that Gale's subjects would have had more chance to habituate.

[14] Whittaker and O'Conaill comment that some VMC systems may have concentrated on the video channel at the expense of the audio channel. In MMC, audio packets get prioritised.

3.4.1 VMC and tasks

One of the key reasons that researchers are as yet unsure of where the benefits (or otherwise) of a video channel lie arises from the wide variety of environments in which the data has been gathered. Ethnographic and experimental studies have necessarily involved different criteria for measuring the effects of a video channel. In laboratory-based studies, many different types of task have been used, ranging from collaborative tasks such as decision making and problem solving tasks (e.g. Anderson et al., 1997b) to tasks with a more social content such as negotiating and bargaining (e.g. Short et al., 1976). This range and variety of tasks makes it difficult to draw comparisons and conclusions from the findings in general, and as Anderson et al. (1994) comment, "... *not only have different tasks been used in evaluation studies, but different methods of analysis have been adopted*". As has been shown, these methods have included number of interruptions, efficacy of turn-taking, and length of dialogue. There is an element of subjectivity as to how these are coded and measured across studies, which has led to a difficulty in making a coherent whole out of the result.

Another reason so many different methods of data gathering and analysis have been used arises from the fact that there is no accepted methodology in the HCI community for the evaluation of VMC. Commenting on the issues involved in evaluating new CSCW systems, Gale (1991) states that: "*Such evaluation marks a change of emphasis, away from human-computer interaction and towards an approach which analyses human-human interaction and considers the technology merely as a mediator. This causes problems for traditional HCI methodologies, such as video taping users as they interact with their computers, or performing keystroke analysis. These are not effective ways of studying how groups work together.*"

The situation is particularly acute in the MMC environment, where the technology is already being widely used in the academic and research communities, and there is still no agreed evaluation methodology (see section 4.13).

3.5 How MMC communication differs

The vast majority of the VMC studies that have been discussed have been carried out across experimental links where it has been possible to guarantee quality. An IP network, on the other hand, offers no guarantees of quality, and indeed it is likely that quality will fluctuate throughout a conference. Many of the studies reported have sought to replicate FTF communication through the implementation of videotunnels in order to enable eye contact. Not only is eye contact not enabled in typical MMC communication, but the image size will generally be far smaller (i.e. QCIF or CIF) than that enabled by videotunnels. Finally, the studies reported here have, in the main, involved high quality audio and no headsets. MMC audio is narrowband, subject to packet loss, and typically heard through headsets.

In MMC frame rates tend to be low, but some research suggests that this may not be a negative factor: although it has been demonstrated many times with high quality video that the number of words and turns required is greater in VMC than FTF or audio-only (e.g. Anderson et al. 1997b; O'Malley et al., 1996), Anderson et al. (1997b) found that with a rate of 5 fps, there was no difference in these measures compared to FTF and audio-only. The implication of this result is that lower frames rates may actually be less 'harmful' (if a measure of successful task performance is turn-taking; but see Monk and Watts, 1995) than higher frame rates in VMC. Taken in conjunction with the Blokland and Anderson (1998) finding that speech is enunciated more clearly when video frame rates are low, the prognosis for successful communication using video in MMC settings is not bad. However, it does raise the issue of at what point, if any, increasing the video frame rate might actually become detrimental to the performance of a task.

Most of the findings discussed in this chapter have looked at the impact on conversational behaviour and task performance of VMC *between two end-users only*. One key advantage to MMC is that more than two remote people (or groups) can work on a task together. In this type of environment, some of the issues discussed may become more critical. For example, the lack of eye contact and directional audio may be of more importance in larger groups.

3.6 Summary

This chapter has highlighted insights and findings from the VMC literature, and discussed how the conditions of MMC communication differ from other videoconferencing environments and usage. Bearing these facts and differences in mind, a clear high level research objective can be established: the quality of the audio and video that is delivered in MMC communication must be assessed from a subjective point of view, in the context of the MMC task that is being undertaken. But how can the perceived quality be assessed in meaningful terms? The following chapter presents a review of existing methods of assessing speech and video quality, with a view to establishing their suitability or otherwise with respect to MMC conditions.

Chapter 4: Measuring perception of speech and video

This chapter focuses on the existing methods available to assess speech and video perception. Assessment of speech perception has traditionally been divided into *speech intelligibility* and *speech quality*. First, the methods used for the measurement of speech intelligibility are discussed, followed by a discussion of objective and subjective speech and video quality measurement methods. The conclusion of the investigation is that the most widely used methods for subjective quality assessment of speech and video transmission, recommended by the International Telecommunications Union (ITU), are not suitable for assessing the quality delivered in MMC systems. However, a new continuous scaling method has just recently been recommended by the ITU, and this is discussed in greater detail. The suitability of quality rating scales in general is considered. The chapter ends with a discussion of what precisely the concept of 'quality' means in the context of MMC.

A summary of many of the issues covered in this chapter was published in Watson and Sasse (1998).

4.1 Measuring speech intelligibility

Intelligibility tests have generally been devised for the assessment of speech that is likely to be less than *toll* (i.e. telephone) quality. The tests are suitable for use in assessing systems that are synthetic or do not cater for the entire speech range (Kryter, 1972). According to the requirements of the test, intelligibility test material can be syllables, words, sentences or passages of speech. The task of the listener is to record or answer questions about what was heard. This description might appear to suggest that these sorts of tests are suitable for assessing degraded IP speech intelligibility. However, there are a number of problems that need to be considered, the major one being that IP audio packet loss is unlike any other degradation previously encountered in the telecommunications world. The four main types of intelligibility tests, and their immediate suitability for assessing the intelligibility of speech affected by packet loss, are considered in the following sections. A summary of this information is presented in Table 3.

4.1.1 Syllable tests

Syllable tests involve the training of listeners. Two main types of syllable tests are found, CV (Consonant-Vowel) syllables, and nonsense syllables (constructed in the manner consonant-vowel-consonant to form a nonsense word such as 'geed' - Kryter, 1972). Both tests require the training of listeners in phonetic spelling. Syllable tests are generally used to highlight small differences between systems - they are sensitive to small degradations. Packet loss over an IP network such as the Mbone cannot be classed as a small degradation: when packet sizes are large, there is a danger that most of a syllable will be lost entirely.

4.1.2 Word tests

Word tests can be divided into those which offer the listener a choice of responses, such as the rhyme tests (House et al., 1965; Voiers, 1977), and those which are free response (Egan, 1948). Word tests use monosyllabic words in order to minimise the redundancy found in longer words. The rhyme tests offer a good means of assessing which spoken features are transmitted successfully through a system (e.g. voiced sounds, nasals etc.), but the tests rely on critical consonants, and as such are not useful for studying the effects of random packet loss. Egan's (1948) monosyllabic word lists provide a more useful test battery. Comprising lists of 50 phonetically balanced (PB) words (i.e. the frequency of speech sounds such as plosives and fricatives are proportional to those found in everyday speech), the listener must record which word was heard. There are no choices of answer, and the test does not rely on the perception of a critical initial or final consonant.

4.1.3 Sentence tests

Sentence tests also fall into two categories. Harvard (phonetically balanced) sentences, for example, are those which are constructed with 5 key words arranged in a variety of syntactic structures such that the sentence is semantically correct. Listeners must perceive the key words correctly (IEEE, 1969). A second type of sentence test has been devised so that the contextual information of the sentence is minimised. The Haskins Laboratories SNST (Syntactically Normal Sentence Test) has been constructed so that syntax is present but the sentence has no semantic context (Nye and Gaitenby, 1974). Again, there are key words which must be recorded correctly. Sentence tests make it easier to make speech quality judgements (since there is a longer stretch of speech from which to form an impression), but the contextual information present in them mean that intelligibility judgements tend to be less sensitive.

4.1.4 Passage tests

Prose passages are usually used as a test for the intelligibility of synthetic speech (Ralston et al., 1991). Listeners hear a passage, and then answer questions about the material that was heard. It is hard to avoid placing some emphasis on memory in tests of this sort, and as such prose passages may not be the best measure for intelligibility. In addition, the means by which the passage was recorded is important: normal everyday speech is spoken at a slower 'conversational' rate than the 'clear' speech with which people read aloud. Traditionally, passages used in prose tests have been read from newspapers or journals, meaning that intelligibility results gained from the studies may not translate well to the real world.

4.1.5 Assessing packet speech intelligibility

A search for the experimental methods that have been used to investigate the perception of packet-switched network speech was carried out. The main finding is that the vast majority of studies have asked listeners to rate the *quality* of the speech rather than the intelligibility (e.g. Gruber and Strawczynski, 1985; Goodman et al., 1976; Suzuki and Taka, 1989; Wasem et al., 1988), because the studies have generally addressed networks where degradation of the size and frequency experienced on the Mbone is not common. Many studies have been carried out over ATM networks (e.g. Nagabuchi and Kitawaki, 1992; Roy, 1994), where the packet sizes and losses are so much smaller (ATM audio packets contain only 6 ms of speech information; the limit of human perception is 4 ms) that when intelligibility *is* explicitly investigated, fine-grained tests such as the CV (Consonant + Vowel) syllable test can be used (Nagabuchi and Kitawaki, 1992). As already mentioned, however, syllable tests are not suitable for the assessment of Mbone packet speech intelligibility.

When the perceived quality has been investigated, the rating method that has been used has typically been that recommended by the ITU, the 5-point Listening Quality Scale, giving rise to the Mean Opinion Score (MOS). This scale and the recommendations from that organisation are discussed in section 4.4, but first a brief discussion of the difference between objective and subjective speech quality measurement methods is provided.

Type of test material	Examples	Test method	Suitable for IP packet speech?
Syllables	CVC (Consonant-vowel-consonant) nonsense syllables e.g. thawf, zayk, geed	Trained listeners must record what they hear.	No, since the duration of a packet can match or exceed the length of a syllable.
Words	Rhyme tests e.g. MRT (Modified Rhyme Test). Selection choices might be fang, rang, gang, bang, hang or sang.	Listeners must select which word was heard.	No, since rhyme tests rely on the perception of a critical consonant. In packet speech the consonant is likely to either be heard or lost completely.
	PB (phonetically balanced) monosyllabic words e.g. jell, rope, chew, queen.	Listeners must record the word they hear.	Possibly, since training of subjects is not required and there is little contextual redundancy, but large packets could exceed the duration of a word.
Sentences	Harvard phonetically balanced sentences e.g. Feed the white mouse some flower seeds.	Listeners record what they hear. Scoring relies on the correct perception of key words.	Yes, but the contextual redundancy present in the sentences may mean that the sensitivity of this test is reduced.
	Haskins syntactically normal, semantically abnormal sentences e.g. The sick seat grew the chain.	Listeners record what they hear. Scoring relies on the correct perception of key words.	Yes, although the contextual redundancy is reduced, the sentences are composed of monosyllables only, and always follow the same structure: 'The (adjective) (noun) (verb, past tense) the (noun).'
Passages	Extracts from magazines/ newspapers, usually about 30 seconds in duration.	Listeners answer questions on the material after it has been heard.	Possibly: reservations include the fact that assessing perception of passages places emphasis on memory and interpretation as well as hearing.

Table 3: Speech intelligibility test methods

4.2 Objective versus subjective speech quality assessment methods

There are both objective and subjective methods of obtaining perceived speech quality. Objective methods allow subjective quality to be predicted on the basis of psychoacoustic modelling (e.g. Beerends and Stemerdink, 1994; Hollier and Cosier, 1996), whereas subjective methods involve playing sample stimuli to listeners and gathering opinion data. This thesis will consider only *subjective* methods, for the following reasons:

- Psychoacoustic models can only be validated through correlation with subjective results. Since multi-way communication over IP networks is a relatively new research area, there is a dearth of subjective results against which to measure any new model. The degree of degradation that can be experienced over IP packet networks is larger than that commonly experienced in other networks, so there is no justification for using current perceptual modelling techniques that have been validated against subjective results for other 'less lossy' networks.
- As discussed in section 2.8, the number of factors that can affect perceived quality in multicast multimedia conferencing is very large. The different perceptual weights of these factors need to be ascertained. This is particularly crucial in terms of the *task* that people are undertaking in the conference. Although the importance of task and other factors such as *expectation* has been recognised (e.g. Hollier and Cosier, 1996), as yet their impact has not been added into the models. This is because investigating the impact of task and expectation on perceived quality is non-trivial.
- That objective modelling techniques have been validated against subjective opinion ratings is of course proper and laudable. However, as will be made clear in the latter part of this chapter, there is evidence that the subjective opinion scales themselves may not be the most sensible or valid method of gathering data.

For these reasons this thesis considers only subjective methods in detail.

Commonly used subjective measurement methods for speech and video are now presented and discussed.

4.3 Subjective measurement methods

Measuring subjective quality of speech and video images has been a field of enquiry ever since military and commercial bodies first began to develop the technologies to transmit these media. Over the years, the need to standardise different testing methods and conditions emerged, in order that different laboratories in various parts of the world could compare results and attain the same standards. This movement culminated in the establishment of bodies such as the ITU (formerly the CCITT and CCIR) and the European Broadcasting Union (EBU). Today, it is the ITU recommendations that are in most widespread use, and as such it is to the ITU that most researchers turn when seeking quality assessment methods.

Until very recently, there have been no explicit ITU recommendations for the subjective assessment of *multimedia* applications, but rather the ITU-T and ITU-R recommendations which address speech transmission over telephone networks, and audio and video quality over entertainment and broadcasting systems, respectively. The recommendations are designed for application in controlled testing environments with defined viewing and listening conditions. It is normal practice for the test stimuli to be *anchored* by playing pre-test examples of the ranges that the subjects can expect to hear or see.

In the following sections the most widely used subjective measurement methods for speech and image quality are presented and discussed with respect to applying them to MMC speech and video.

4.4 Speech quality scales

There are a number of ITU-T recommendations pertaining to the assessment of speech transmission, gathered together in the P series. These are contributed to and updated every few years by international study groups.

Of the various recommended methods for the subjective determination of transmission quality covered in Recommendation P.800 (ITU-T P.800), the two key types are *conversation opinion* tests and *listening opinion* tests. The recommended rating scale for both is a 5-point category scale commonly known as the quality scale. In conversation tests, a binary difficulty scale follows the (connection) quality scale. Listening-only tests can also be assessed via the listening effort scale. These scales are shown in Figure 7, (a)-(c).

The recommended material for listening tests is groups of short sentences such that the duration of each individual test stimulus is somewhere between 5 and 15 seconds (ITU-T P.800). In conversation tests, it is recommended that each conversation should have a natural beginning and ending and be of sufficient duration that a full impression of the quality of the connection can be obtained. In many cases this requirement necessitates some kind of conversational task to be provided, since people are notoriously poor at talking to order.

Results from category scales such as the quality and listening effort scales are averaged across subjects in order to provide a Mean Opinion Score (MOS).

4.4.1 Other speech quality tests

There are a number of other ITU-recommended listening test scales, which are listed briefly below and illustrated in Figure 7, (d)-(g).

- The loudness preference scale is a 5-point category scale used to investigate volume issues, and is shown in (d). With respect to MMC speech, volume is a factor most often under the control of the users, not the system itself, which renders the scale not particularly useful. (However, the issue of volume does become problematic in MMC when there is more than one sound source - see section 8.2.)
- The quantal-response detectability method is used to determine whether a particular attribute of a sound, for example echo, is detectable or not. Different degrees of detectability can be assessed by adding in more points to this type of scale. The most basic scale is illustrated in (e).
- The degradation category rating (DCR) method compares the system under test with a high quality fixed reference, and the degradation is rated on a 5-point scale (f).
- The comparison category rating (CCR) method involves comparing an unimpaired speech sample with an impaired one and evaluating the difference between them on a comparison scale shown in (g).
- The threshold method entails direct comparison of a transmission system again with a reference system. A pair of stimuli are presented to the listeners, consisting of the test and the reference condition, and the listeners are asked to indicate which of the stimuli had the highest quality - a preference rating. The listener *must* choose whether stimulus A or B was better. This method is therefore a forced choice comparison.

The issues that are of import in the application of these methods to subjective measurement of MMC speech are now examined.

4.4.2 Applying the scales to the assessment of MMC speech

As discussed in section 3.2.4, the main characteristics of MMC speech are that it is (in the main) narrowband, subject to network loss and different repair methods, and therefore generally lower than telephone quality. Its quality is also *unstable*, since at present there is no guaranteed level of service. Consideration of these facts in relation to the ITU testing methodologies described above can be broken down into the key areas of:

- the vocabulary of the scale labels;
- the length of the recommended test material;
- the conversation difficulty scale.

Quality of the speech/connection	*Score*
Excellent	5
Good	4
Fair	3
Poor	2
Bad	1

(a) Listening quality scale

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

Yes	1
No	0

(b) Conversation difficulty scale

Effort required to understand the meaning of the 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

(c) Listening effort scale

Loudness preference	*Score*
Much louder than preferred	5
Louder than preferred	4
Preferred	3
Quieter than preferred	2
Much quieter than preferred	1

(d) Loudness-preference scale

Degradation is inaudible	5
Degradation is audible but not annoying	4
Degradation is slightly annoying	3
Degradation is annoying	2
Degradation is very annoying	1

(f) Degradation opinion scale

A	Objectionable
B	Detectable
C	Not detectable

(e) Detectability opinion scale

The quality of the second compared to the first is:

3	Much better
2	Better
1	Slightly better
0	About the same
-1	Slightly worse
-2	Worse
-3	Much worse

(g) Comparison category rating scale

Figure 7: ITU-T recommended speech quality measurement scales

MMC speech is subject to a range of network and environmental degradations. Given these facts, the labels on the listening quality scale (i.e. Excellent, Good, Fair, Poor and Bad) are inappropriate. It is unlikely that listeners would rate the quality of the speech as Excellent, even after training. It is far more likely that responses would be concentrated at the lower end of the scale, and this likelihood has been borne out in both experimental and field studies carried out as part of this thesis (see, for example, sections 5.1 and 5.2). The listening effort scale is a better measure of conversational quality since it encompasses a measure of both intelligibility and quality. However, with respect to the category labels on this scale, it is even easier to see how a bias towards the lower end of the scale might occur, since ‘complete relaxation’ when listening to MMC speech is rarely possible. The labels on the DCR actually appear to be the most suitable in terms of assessing the effect of packet loss and repaired speech. However, the DCR has been found to be most effective when the impairments to be measured are *small*, which is certainly not the case in best-effort packet network speech. The issue of scale vocabulary will be discussed further in section 4.10.

The variable network conditions that affect some real-time services mean that speech quality can, and often does, change rapidly and unpredictably[15]. In listening-quality tests, the recommended test material is short in duration (5-15 seconds). This length of time does not afford the opportunity to experience the unpredictability of some networks, or - if loss rates are low - the full potential of the resulting impairment. Obviously, if the duration of the material needs to be longer than 15 seconds, then this makes the use of double-stimulus methodologies such as the CCR, forced choice and DCR problematic, since the subject will begin to make decisions based more on cognitive than perceptual effects, i.e. memory effects will start to play a role.

Finally, it is recognised that the best measure of subjective quality will be gained from people engaging in a conversation over a connection. In laboratory settings these conversations can be quite artificial, but there are obvious benefits over passive listening-only tests. However, the connection quality is unlikely to remain constant throughout the conversation in MMC communication, which raises the question of what *part* of a conversation people are rating. In addition, the binary difficulty scale is patently unsuited for the assessment of MMC conversations, since even a small amount of packet loss is likely to cause difficulty in hearing or talking, even if short-lived.

As can be seen, then, there are important issues to be aware of that render the straightforward application of ITU-recommended methods to the measurement of MMC speech problematic.

The next sections present the ITU-R recommended methods of assessing image quality, and discuss which methods might be appropriate for the assessment of MMC video quality.

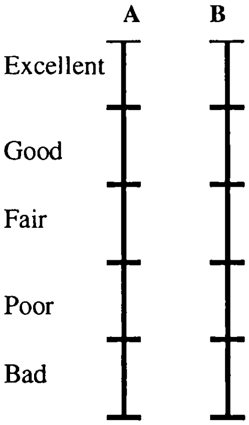
[15] Variation of quality has its own effect, independent of the quality level – it has been demonstrated that if users have a low expectancy of the quality delivered, they will rate the same level of objective quality higher than those with a high expectancy. However, this remains true only when the quality level is *predictable* to users (Bouch and Sasse, 2000).

Image quality	*Score*
Excellent	5
Good	4
Fair	3
Poor	2
Bad	1

(a) Image quality scale

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

(b) Image impairment scale



(c) Double stimulus continuous quality scale

-3	Much worse
-2	Worse
-1	Slightly worse
0	The same
1	Slightly better
2	Better
3	Much better

(d) Stimulus comparison scale

Figure 8: ITU-R recommended image quality measurement scales

4.5 Image quality scales

The subjective assessment of image quality falls under the brief of the ITU-R, and in particular the recommendation ITU-R 500. In the assessment of image quality, single stimuli are rated using the quality scale or impairment scale, and comparisons to reference conditions are made using the double stimulus impairment scale (DSIS), the double-stimulus continuous quality scale (DSCQS) or the stimulus comparison scale (ITU-R BT.500-8). These scales are shown in Figure 8, (a)-(d).

The DSIS (also known as the EBU) method presents the viewer with two images per test. One is the test image, and the other is a reference image against which to rate the test image. The presentation of the stimuli is cyclic: first the unimpaired reference picture, and then the same picture, this time impaired. Viewers are asked to judge the *impairment* of the second image compared to the first (Figure 8, b).

In the DSCQS method, the viewer is presented with a pair of pictures, one of which is the test picture, and the other a reference. The viewer is not told which is which. The viewer is allowed to switch back and forth between these pictures before making a quality judgement for both pictures on the scale (Figure 8, c). The viewers insert a mark anywhere on the scale, and in scoring the scale is translated into a 0-100 range.

The recommended duration of video test material is 10 seconds.

4.5.1 Additional image quality methods

In addition to the adjectival categorical methods (the quality and impairment scales) described above, numerical categorical and non-categorical methods can also be used. Numerical categorical judgement methods involve the user assigning a number as a rating e.g. a number between 1 and 11 on a scale. An 11-grade numerical scale such as this has proved to be more reliable and sensitive than adjectival methods when there is no reference available. Non-categorical judgement methods require the viewer to assign a rating value either on a point on a line drawn between semantic labels (continuous scaling), or to assign a number that reflects a value on a specified dimension. Both forms result in a distribution of numbers for each condition.[16]

The utility of these scales with respect to MMC video is now considered.

4.5.2 Fitting the scales to MMC video

As discussed in section 2.8, one of the prime characteristics of MMC video is that it tends to be of low frame rate with irregular updates. The image tends to be blocky and relatively small. These traits render the use of the ITU-R recommended assessment methods inappropriate on the basis of the following reasons:

- the magnitude of the impairment;
- the vocabulary of the scale labels;
- the duration of the test material;
- the artificiality of assessing video without audio.

The ITU-R recommendations are primarily concerned with establishing the subjective performance of *television pictures.* This means that in terms of colour, brightness, contrast, frame rate etc., the quality component under investigation is already assumed to be of a high standard, which is simply not the case for MMC video. Like MMC speech, MMC video is characterised by a large variety and range of impairments, which can change rapidly. This trait means that the single and double quality impairment tests are not suitable, since, as is reflected in the terminology of the scale (*imperceptible/perceptible*), they have been designed to determine whether individual *small* impairments are detectable. In the case of the DSIS in particular, the test has been shown to be more reliable for small rather than large impairments, meaning that it certainly is not a sensible method to use in assessing MMC video. The DSCQS is considered to be especially useful when it is not possible to provide test stimulus conditions that exhibit the full range of quality, which suggests this method would be better suited to the assessment of MMC video than the impairment scale. However, the problem with vocabulary still remains, as will be considered next.

With respect to use of the quality scale, the same criticism can be levelled as to its use with MMC speech: the vocabulary is unsuitable, and therefore it is expected that responses will be biased towards the bottom of the scale. Use of the DSCQS at least permits scoring between the categories (the subject places a mark anywhere on the rating line, which is then translated into a 0-100 score), but it is still the case that subjects shy away from using the high-end of the scale, and will often place ratings on the boundary of the good and excellent ratings (Aldridge et al., 1995). This is a common finding in rating scale studies.

The quality tests typically require the viewer to watch short sequences of approximately 10 seconds in duration, and then rate this material on the relevant scale. However, it is not the case that a 10-second video sequence would be representative of MMC video or long enough to experience the types of degradations common to MMC video (this type of problem has also been encountered in high quality video studies, and will be discussed in section 4.8).

Finally, the quality judgements are supposed to be made entirely on the basis of the picture quality. It should be questioned whether it makes sense to assess MMC video on its own (i.e. without audio) in this manner, given that it is commonly experienced at low-frame rates and small image size. At present, it would be true to say that the video image in MMC is not the focus of attention in the same way that the picture is when we watch television. It can be argued that the utility of the low frame rate video used in MMC arises mainly when it is used in conjunction with audio (and perhaps shared workspace), and so it is only in real task environments that it makes sense to evaluate the subjective quality of the video. It would be highly unusual, if not inconceivable, for users to be using low frame rate video as the sole means of communication across the Mbone at present[17]. For this reason, some of the

[16] These methods have also been applied to speech assessment (Handbook of Telephonometry).

[17] As desktop machines become increasingly powerful, and bandwidth reservation protocols take force, video frame rates in MMC will increase. It is already possible to communicate via video alone (e.g. using sign language), but for the purposes of this thesis the investigation is restricted to the average conditions experienced by the average user in the late nineties: 2-12 frames per second.

recommendations of the P.900 series, which specifically address audiovisual quality in multimedia services, should be more relevant.

4.6 Audiovisual quality in multimedia services

ITU-T Recommendation P.910 presents *non-interactive* subjective assessment methods for the evaluation of low and medium quality digital images (bit rates of up to 2 Mbit/s). The recommended test methods are the 5-point quality scale, the 5-point impairment scale (see Figure 8) and the pair comparison method, where viewers indicate which of two test stimuli they prefer.

The quality and impairment scales have been encountered already, and the queries raised about their validity for investigation of MMC quality apply here too (see section 4.4.2)[18]. The pair comparison method is recommended for use when several of the test items are nearly equal in quality, since it is high in discriminatory power. This alone renders it an unsuitable test for assessing MMC video.

Interactive test methods are covered by P.920 (ITU-T P.920). The methodology in this recommendation is based on conversation opinion tests. Conversational tasks more suited to the assessment of audiovisual quality than conventional conversational tasks (since they do not require attention to be focused elsewhere than the video medium) are proposed, but the rating scales that are to be applied after carrying out these tasks are the same as before: the 5-point quality scale is recommended for assessing the video quality, the audio quality and the overall audiovisual quality.[19]

4.7 Interim summary

In the previous sections the most widely used ITU-recommended subjective assessment methods for speech and video quality were described. As mentioned in section 2.3, these methods were originally developed to assess delivered quality in analog and CSN environments, which have very different characteristics compared to PSNs such as IP-based networks. One of the biggest problems in applying the existing methods to the assessment of MMC speech and video is due to the unstable nature of packet networks: quality can fluctuate rapidly and to a

[18] The recommendation does suggest that for the assessment of low-bit rate video codecs, rating scales with more than 5 grades can be used. The suggestion, however, does not extend to a different vocabulary on the scale, but rather a 9-grade scale where the 5 quality scale categories are used as labels for every second grade on the scale. Thus, the problem vocabulary remains.

[19] The only unfamiliar scale is one that addresses the effort needed to interrupt - a potentially valuable measurement when dealing with communication networks where delay plays a significant role. Again,

great degree. This means that assessing the quality of short segments of MMC speech or video may not be very meaningful, since it does not model what actually happens in the real world. If the test material is lengthened, the risk of measuring cognitive (memory) effects rather than perception of quality arises.

What is required is a means of measuring perceived quality in a dynamic fashion, as and when the changes occur. Dynamic rating methods are better suited to HCI evaluations, since they allow the perceptual effect of different quality levels to be registered and recorded instantly. The requirement for dynamic rating was identified in the research area of high quality digital video (MPEG-2), and has culminated in the development of a new assessment method, the Single Stimulus Continuous Quality Evaluation (SSCQE). It should be noted that the SSCQE was being developed at the same time as the research reported in this thesis, but for sake of completeness is reported now.

The next part of this chapter describes the method's development and applications.

4.8 The need for a continuous rating scale

Variable bit rate networks and new digital video channels cannot be best assessed via still pictures or short segments of 10 seconds duration (as recommended for the DSIS and DSCQS), for a number of different reasons. Firstly, a duration of 10 seconds may not be long enough to experience the quality variations that people would commonly experience in their home or office environments. However, the solution is not as simple as merely increasing the stimulus length in order that the full range can be encountered. The lengthened test time caused by showing two 30 second (or longer) segments - as opposed to only 10 second segments - render the use of the double stimulus methodologies problematic due to available time and subject fatigue. More importantly, however, Aldridge et al. (1995) found that, when the presentation length was increased from 10 to only 30 seconds, a human memory phenomenon, the *recency effect*, occurred. The recency effect is associated with working, or short term, memory whereby there is a recall advantage for the most recently presented material over what has been presented before. Obviously, this means that quality ratings attained from segments as long as/longer than 30 seconds will be biased by what has occurred in the most recent part of the display. Whether the segments are 10 seconds or longer, however, the method always implies that subjects will be registering an impression formed over time (de Ridder and Hamberg, 1997). Therefore, a new method has been sought and developed, entailing a new use of a single stimulus. This research was carried out, in the main, under the RACE MOSAIC (Methods for Optimisation and

this is a 5-point scale with the categories: No Effort, Minor Effort, Moderate Effort, Considerable Effort, or Extreme Effort.

Subjective Assessment in Image Communication) project. The resulting method is termed the SSCQE (Single Stimulus Continuous Quality Evaluation), and is intended for use in assessing the quality of digital TV pictures using long test sequences (which are more representative of real world viewing material). The SSCQE has since been included in the most recent version of recommendation ITU-R BT.500 "Methodology for the Subjective Assessment of the Quality of Television Pictures". The method is unique in this recommendation in that it advocates the inclusion of an audio channel while the video quality is being rated, bringing the assessment situation closer to that of the real world.

The SSCQE follows a similar approach to the direction the thesis research took (see the QUASS studies presented in sections 7.1 and 7.4). The similarity of the approaches reflects a growing awareness in the research community at large that a new method is required to evaluate time-varying quality from a subjective standpoint.

4.8.1 How SSCQE works

The SSCQE method involves continuous quality evaluation by the subject using a hardware slider with a linear range of 10 cm. It is usual for the range of the scale to be labelled with the ITU quality labels Excellent, Good, Fair, Poor and Bad, as per the DSCQS method. However, these labels can be altered to fit the requirements of the test (see next section). The position of the slider is registered at 500 ms intervals. Means and standard deviations of each point of vote are calculated over subjects and represented graphically, which permits a detailed examination of subjects' voting behaviour across time. Individual voting behaviour can also be illustrated graphically.

4.8.2 Applications of SSCQE

The ACTS (Advanced Communications Technologies and Services) project TAPESTRIES (The Application of Psychological Evaluation to Systems and Technologies in Remote Imaging and Entertainment Services) (1997) has used the SSCQE extensively to investigate the effects of a number of different dimensions on perceived video quality, including the use of audio. In these studies, the SSCQE has been adapted to fit the requirements of the test. For example, one study assessed subjects' perceived correlation between what they heard and what they saw in a multimedia learning environment where two types of video transmission were compared, Super High Definition TV and standard broadcast TV. The scale was adapted such that the 5 labels were: Very clearly visible, Clearly visible, Visible, Not clearly visible, and Not visible at all.

Another study used the SSCQE method in a videoconferencing environment, where two groups of 3 people played a variety of games across a connection where the effects of audio bandwidth, sound spatialisation, sound level, delay and video bit rate was examined. In this study the scale was bounded by the adjectives Excellent and Bad only.

These studies show that the SSCQE can be used for the assessment of attributes other than perceived picture quality. However, conclusions about the relative effects of variables such as bandwidth, delay and bit rate are hard to draw, stemming from the fact that the SSCQE was designed in order to measure the quality of a service in real world conditions rather than to draw out the fine-grained effects of certain variables (ACTS TAPESTRIES, 1997). Plus, there is undoubtedly a confounding element due to the performance of a task during rating, as was noted in the videoconferencing study where subjects were actually communicating over a link, rather than simply watching a video and monitoring the quality.

That there is a conflict between interactive task performance and using the SSCQE method is clear. But is the method otherwise effective and reliable?

4.9 Behavioural traits of continuous evaluation

A number of points about subject behaviour when using the SSCQE need to be noted. The primary question is obviously whether subjects are *able* to track the changes in picture quality as and when they occur. From both the ACTS TAPESTRIES project and the studies carried out by Aldridge et al. (1998), the answer is clearly 'yes', with a few provisos. Firstly, it can take a few seconds for subject responses to settle. Aldridge et al. (1998) noted in one study that this period can be as long as 15 seconds from the start of the test. Secondly, it has been observed that subjects are better at noticing *drops* in quality than improvements. The reaction to drops is much quicker than to improvements, and when an impairment is over, the slider is rarely returned to the original quality level. Aldridge et al. have hypothesised that viewers may readjust their standard or baseline after such an impairment, a type of adaptation that is also seen in analogue TV quality studies. Thirdly, there is a delay between the subject perceiving a change in quality and moving the slider. de Ridder and Hamberg (1997) have suggested that this is about one second in length, but in a previous paper suggested it was somewhat longer, approximately 2 seconds (Hamberg and de Ridder, 1995). It could be presumed that this delay may vary as a function of the material/task that a person is engaged in.

The concept of an anchor for such long presentations, where the whole point is to experience a range of conditions, is essentially untenable, and so anchors have not generally been provided or used in SSCQE studies. Aldridge et al. (1998) comment that an anchor does not seem to be required for use with the SSCQE since the presentations in the studies are literally thousands of seconds longer than the traditional 10 second presentations. They suggest that since a great range of material can be presented during this time, the "*technique may be self-anchoring to some extent.*"

Finally, it has been observed that there is a discrepancy between the ratings subjects give using the SSCQE, and an overall quality rating at the end of a session. For example, impairments occurring 20-30 seconds prior to the end of the presentation will be registered with the SSCQE, but seem to have little bearing in the overall quality rating collected at the end of the session. This discrepancy and its implications will be discussed in greater detail in the following section.

4.9.1 Investigating other potential effects

Aldridge et al. (1998) investigated a number of different factors that might impinge on performance when using SSCQE for prolonged intervals:

- fatigue;
- severity and duration of impairment ;
- drift;
- instructions to subjects when audio is present.

4.9.1.1 Fatigue

In order to determine whether subjects become fatigued during longer rating sessions using the slider, the authors compared the results from a 'long' (24 minutes) and a 'short' (10 minutes) group. Both groups watched MPEG-2 coded video at 4 Mbit/s except for 1 minute of 1Mbit/s, which either appeared in minute 6 (for the short group) or minute 20 (for the long group). It was hypothesised that if fatigue affects continuous responses, the long group's tracking of poor-quality video would be worse than the short group's since they would have become less responsive to variations in picture quality (through fatigue). However, it was found that the long group was actually *more* sensitive to quality variations than the short group, indicating that fatigue could not be playing a role. Indeed, the investigators concluded that "... *the use of long test sessions appears to improve rather than impair subjects' ability to react to and identify coding errors.*" Subjects filled out a 9-item questionnaire at the end of the session on using the slider as well, and this subjective data also showed that fatigue was not playing a role.

4.9.1.2 Severity and duration of impairment

The subjects were additionally asked to provide an overall quality rating at the end of the session. It was found that the 'long' group did not give significantly higher overall quality ratings than the 'short' group, despite the fact that only 1/24th of their material had been of poor quality compared to the short group, who experienced 1/10th as poor. This interesting finding was examined in more detail in a study which compared the overall rating of quality subjects gave at the end of a session with the continuous rating gathered by the SSCQE. Subjects were presented with three 30 second video clips, all coded at 4 Mbit/s, where one clip had 5 seconds of 1 Mbit/s quality inserted, and one had 10 seconds' worth. Using the DSCQS to extract an overall rating of the segments they found that the control condition (i.e. with no 1Mbit/s section) was rated best, but that there was no significant difference between 5 and 10 second conditions in terms of *overall* quality rating. However, results from the SSCQE showed that subjects do register the difference between 5 and 10 seconds of poor quality, which suggests that the duration of an impairment does not have a direct bearing on *overall* retrospective judgements. The authors report that this interpretation fits well with psychological studies of affective responses to positive and negative stimuli. They cite research that has shown that when subjects are presented with short (35 seconds) and long (113 seconds) video clips of a painful medical operation, and asked to give both continuous and overall ratings, overall ratings are influenced to a large extent by the peak intensity of the episode (i.e. the single most severe event), but the *duration* of the episode does not play a key role in forming perceptions.

The implication of this result is that in terms of *retrospective* picture quality perception, it may not be so important to recover quality as quickly as possible, but rather to protect against severe drops in quality in the first place.

4.9.1.3 Drift

Aldridge et al. (1998) were keen to determine whether responses drifted over time, owing to lack of awareness on the subject's part of the slider position. They compared the SSCQE results from two groups of subjects watching a film coded at 2 Mbit/s, a level at which impairment visibility varies considerably. One group watched the entire 24 minute segment at one sitting, and the second watched it in three 8-minute segments where the slider was returned to midpoint of the scale after each break. A divergence in the results between the groups was not found in general - a greater fluctuation for the second group was attributed to the resetting procedures during each break. It was concluded that drift does not occur, at least in sessions up to 24 minutes in duration: subjects remain aware of the position of their slider.

4.9.1.4 Instructions to subjects when audio is present

In the interests of maximising the similarities between the experimental set-up and real world, it was considered that subjects should be rating the quality of video while audio was also being played, as would happen in the real world. However, this raised the issue of what the subjects should be asked to assess: the picture quality only, or the overall (picture and sound) quality? Two groups of subjects were again compared in their ratings of 23 minutes of impaired video accompanied by unimpaired audio. One group was told to rate the picture quality, while the other was told to rate the overall quality. No statistical difference was found between the two.

This finding suggests that subjects take into account audio quality in addition to video quality whether consciously or not, and reveals the interactive relationship between audio and video.

4.9.2 Assessing perceived audio quality

Kokotopoulos (1997) reported a study that investigated the use of SSCQE in a multimedia system for distance learning. The study involved the use of MPEG-2 video and MPEG-1 audio. Unlike the other studies discussed above, Kokotopoulos varied the quality of the *audio* coding in one of his experiments, holding the video quality constant. The test sequence (2 mins 40s in duration) was subject to different bit rate compressions in order to cause the audio impairments. Watching the unimpaired video and listening to the soundtrack (with impairments) from a speaker directly below the monitor, it was found that subjects were able to rapidly track the drops in coding quality. As with the video studies, it was again observed that subjects are slower at responding to improvements than degradations in quality.

Kokotopoulos additionally asked subjects to provide an overall quality judgement of the test sequence. A comparison between the overall quality ratings for a *video*-impaired sequence and the audio-impaired sequence suggested that audio quality was of greater importance to the subjects than the video (at least in this distance learning scenario), since the mean rating for the audio-impaired sequence was less than half as good as that for the video-impaired sequence. Although it is hard make direct comparisons between audio and video degradations in terms of the overall impressions they make, this finding is in keeping with other research that has found that the audio channel is more important than the video channel in most multimedia situations (e.g. Sasse et al., 1994a).

4.9.3 Discussion

The SSCQE was developed originally for the assessment of digital video impairments. The reasoning was that digital video such as MPEG-2 afforded visible quality impairments that varied over time, and that a new assessment methodology was required to capture the perceived quality of the transmission. As de Ridder and Hamberg (1997) observed, small coding errors are far more likely to be noticed in short or static segments of video than in a continuous assessment approach where the subject's attention is focused on the main action, as it were. This makes the SSCQE a more realistic assessment tool when the goal is to model a home or office environment. However, in terms of applying it to IP *multicast* video, it is doubtful that the SSCQE is a useful technique, since the current quality (i.e. frame rate and size) of the image is of a much lower standard than MPEG video - the nature of the impairment in IP packet networks is very different to that of the bit errors seen in MPEG transmissions (see section 2.3.3). As has already been stated, what is of interest in MMC is whether the video quality is *good enough* for a certain task to be accomplished satisfactorily, not to what degree the coding impairments are visible. It is not clear that the SSCQE would enable this information to be gathered.

Investigations into the utility of the SSCQE have found that there is a discrepancy between the overall quality results registered at the end of a session, and the continuous opinions registered with the SSCQE. Whilst it could be argued that it is the overall end opinion that matters more (in that it is the overall opinion on which people are likely to base a judgement on whether to reuse a system or not), the SSCQE method affords a valuable means of assessing *where and when* drops in perceived quality occur, which would permit interpretation to be made of *why* a session may have been awarded an overall rating of *Poor*. This ability could prove especially valuable in terms of evaluating MMC *speech* quality - as previously stated, one of the key aims of the research presented in this thesis is to identify where the (perceptual) quality boundaries lie, such that resources are not squandered unnecessarily. That Kokotopoulos (1997) has successfully applied the SSCQE technique to the assessment of bandwidth impaired audio is therefore very encouraging.

The SSCQE is a new development and is not yet widely used, and in the meantime it is likely that category scales giving rise to MOS will continue to be used since they are quick and easy to apply, and results are apparently simple to understand. However, there remain serious issues that need to be raised with respect to the validity of these 5-point scales (and by extension to the terminology that is often used in the SSCQE).

4.10 The nature of the international interval scale

In evaluating speech and video quality, it has been the 5-point quality scale (a category interval scale) that is most widely used. The scale is easy to administer and score, and its recommendation by bodies such as the ITU has meant that its use has been accepted without question by many researchers, network providers and application developers. There is a growing number of researchers, however, who question whether trust in this scale is warranted, and the findings will be discussed here. These investigations have focused mainly on whether the quality scale is actually an interval scale, as represented by the labels on the categories. If the intervals on the scale are not equal in size, then it is doubtful whether the use of parametric statistics on the data gathered from quality assessments is strictly legitimate, since this would require a normal distribution (Jones and McManus, 1986). Investigations have also been carried out to substantiate the assumption that the scale labels have been adequately translated into different languages such that the scale is 'equal' in different countries of the world, so that quality results can be generalised across the world[20].

4.10.1 Internationally interval, or internationally ordinal?

Investigations of the interval (or otherwise) nature of the rating scales have generally been carried out using the graphic scaling method. Subjects are presented with a piece of paper on which there is a vertical line with the words *'Worst Imaginable'* at the bottom, and *'Best Imaginable'* at the top. On this line, they are required to place a mark where they feel a certain qualitative term would fit. Each word under test is ranked, and by measuring the distance of the marks from the bottom of the scale, the means and standards deviations for each term can be calculated. Using this method, Narita (1993) found that the Japanese ITU labels conform well to the model of an interval scale, although not perfectly. Whilst this is encouraging news for Japanese speakers, it is a different story for American English, Dutch, Swedish and Italian speakers.

Jones and McManus (1986) used this method to investigate whether the intervals represented by the labels are equal i.e. that the distance between *Good* and *Fair* is equal to the distance between *Poor* and *Bad.* They found that the scale terms were spaced almost as a 4-point, 3-interval scale as opposed to the 5-point, 4-interval scale they are supposed to represent. That is, the ITU terms constitute an ordinal rather than an interval scale. *Bad* and *Poor* were found to be perceived as very similar in meaning, whilst the perceptual distance to *Fair* was comparatively great. Since research in psychology has established that subjects tend to avoid the end points of scales, they question the usefulness of what appears essentially to be a "*3-point, 2-interval scale*".

Jones and McManus also carried out their study in Italy and in different regions of the US, and found not only differences between the meanings of the words between the countries, but even small regional differences across the US. The Italian ranking of the ITU terms produced a scale that has no mid-point. In the ranking of other terms, it is interesting to observe that a supposedly 'universal' word such as OK appears to mean different things to different nations: the Americans positioned OK around the centre of the scale, as roughly equivalent to Fair, whereas the Italians seemed to equate OK with Good. Other researchers have found similar results. Virtanen et al. (1995) investigated the placement of 37 Swedish quality descriptors on a vertical line. With respect to the ITU terms, it was again found that there was a flattened lower end (i.e. the terms equivalent to Bad and Poor were perceived as very similar), and there was a

[20] Although the ITU's Handbook of Telephonometry (CCITT, 1987) cautions that "*it may be difficult to translate the names of the individual categories into different languages whilst preserving the same inter-category relationship as in the original language*", it does not warn that the categories in the initial language may not have a very orthodox relationship...

large gap between Poor and Fair, such that Fair was actually above the midpoint of the scale. Teunissen (1996) investigated Dutch terms and found once more that the ITU terms do not divide the scale into equal intervals. As part of this thesis research, a similar study was carried out with British English speakers, and again it was found that Poor and Bad were perceived as similar, and that Fair was well above the midpoint of the scale (see section 5.9).

Teunissen asked whether it might be possible to construct a 10-term ordinal scale to investigate the perceived absolute quality of a display system. For this type of scale to be valid, it would be necessary for all subjects to place terms in the same order on a scale. Having selected the ten terms for his scale, Teunissen presented them to 85 subjects. He found that 74 of them put the Dutch words in the following order: awful, very bad, bad, not so bad, poor, reasonable, reasonably good, good, very good, excellent. Although Teunissen concluded that this margin of error was acceptable, it is interesting to note that all of the discrepancies involved people interchanging two terms from the majority order e.g. bad with not so bad, and very good with excellent. This type of result illustrates well that different people hold different semantic concepts and are therefore likely to use rating scales in different ways: individual differences and preferences make any type of qualitative research difficult.

4.10.3 Summary

The ITU-recommended quality scale is not the international interval scale it is purported to be. Nor is the quality scale internationally ordinal, since the positional rankings of the qualitative terms in different languages are not equal. Many studies have highlighted the differences between national concepts and meanings. However, there is another, more complex issue at hand, and that is the overall concept of quality. The 5-point quality scale treats quality as a single measurable dimension, despite much evidence to the contrary.

4.11 What is quality?

Virtanen et al. (1995) claim that quality is not a "*single monotone dimension*" - or at least the terms used to describe it are not. They investigated the semantic groups that qualitative terms fall into, and identified 6 different classes:

- Qualitative terms e.g. good, bad;
- Compound qualitative terms e.g. very good, very bad;
- Positional terms e.g. in the middle;
- Emotional terms e.g. terrible;
- Approval terms e.g. acceptable, tolerable;
- Comparative terms e.g. superior, inferior.

The authors determine, from this admittedly unscientific classification, that there are at least 4 types of quality scaling situations dependent on the task and context: qualitative/hedonic judgement; positioning in relation to a reference; emotional/communicative expression; and ‘people as judges’. It is clear, then, that there are many different ways of asking people to make qualitative judgements. The existence of so many semantic quality categories underscores the fact that many different variables can affect quality perception formation.

What can be said about the variables that contribute to speech and video quality perception?

4.11.1 Speech quality

Researchers from disciplines as diverse as hearing aid research and engineering have identified significant roles in speech quality for variables such as intelligibility, loudness, naturalness, listening effort, pleasantness of tone etc. (e.g. Kitawaki and Nagabuchi, 1988; Preminger and Van Tasell, 1995; ITU-T P.800). However, as Preminger and Van Tasell (1995) observe, *“Although a multidimensional view of speech quality has not been disputed, many researchers have taken a unidimensional approach to its investigation* [...] *When speech quality is treated as a unidimensional phenomenon, speech quality measurements are essentially judgements, and one or several of the individual quality dimensions may influence the listener’s preference.*” Kitawaki and Nagabuchi (1988) present this fact as an advantage since *“different impairment factors can be assessed simultaneously*”, but this approach does not allow researchers to determine which of the many factors that comprise quality carry most weight in perception formation.

Just as there is a unidimensional approach to measuring quality, within the networking community there is also a tendency to assume a unidimensional approach to improving quality: increasing bandwidth. For example, Jayant (1990) remarks that “... *the notion of quality as a function of speech bandwidth will become more pervasive, and subjective testing will lead to better quantification of the quality-bandwidth function.*” However, although increasing bandwidth would undoubtedly solve many quality issues, it should not be treated as a panacea. It may well be the case that many quality issues can be settled without resorting to increasing bandwidth, and since bandwidth is a valuable resource, exploring these possibilities is important, both for the HCI and networking communities (see section 1.2).

4.11.2 Video quality

Much as speech quality should not be assumed to be unidimensional, subjective video quality is formed through the influence of many different variables. Gili et al. (1991) investigated the perceived quality of 9 different digital codecs and identified seven key variables to be colour, brightness, background stability, speed in image reassembling, outline definition, 'dirty window', and the mosaic/blocking effect. As has been argued already (in section 4.5.2), however, for MMC video it is more important to investigate the interaction between speech and video, in the context of a task.

4.11.3 Speech and video quality in MMC

As illustrated in Figure 6 (Chapter 2), there are many factors that can impinge on perceived quality in MMC environments, ranging from the type of headsets a person is wearing to the lighting and background noise of the room he/she is sitting in. Clearly it is not possible to explore the impact of all of these variables, but in the context of the research that is presented in the remaining chapters of this thesis, these factors must be borne in mind, since the studies were often field-based rather than conducted in controlled laboratory settings.

4.12 Chapter summary

This chapter has presented a critical review of the most commonly used methods for assessing speech intelligibility and quality, and image and video quality. Although many of the methods discussed are well-established and easy to use, there are various issues that may render their application to the assessment of MMC speech and video quality inappropriate. To summarise:

- The characteristics of MMC degradations i.e. the size of the impairments, and also their unpredictable and fluctuating occurrence, suggest that the use of some intelligibility and quality assessment methods will be inappropriate. Many of the established methods recommend the use of relatively short testing stimuli. It is contended that the impact of MMC degradations will be better assessed, in a task context, over a longer period of time.
- The vocabulary on the widely used ITU quality scale appears unsuitable with respect to describing MMC speech and video, and in any case these terms have been demonstrated to be non-interval, calling into question results gained using this scale.
- The use of a single quality measurement scale implies that perceived quality is a unidimensional phenomenon, but in fact quality is multidimensional: it is the scales that are not. The different facets of quality that impact perceived quality in MMC need to be established and measured individually.

The issues therefore range from the size and duration of the test stimuli to more serious methodological problems. However, in lieu of any guidelines or established methodologies specifically addressing the subjective assessment of MMC speech and video, these existing methods should be explored in situ before being dismissed.

Based on the issues listed above, the research agenda is presented in the following section.

4.13 Research agenda

The early chapters of this thesis have described MMC technology and discussed the impact that this technology has on transmitted audio and video. Although addressing the technological impact of sending real-time data is of course important, researchers, network providers and application developers have an additional requirement to understand and measure the *subjective* impact of real-time MMC communication, since it is the end users' opinion that will determine the success of an application. There is great importance in measuring the subjective effects of different facets of MMC, such as speech encoding schemes, packet loss repair mechanisms, and the quality of the video image. How this should be done, however, has yet to be established.

4.13.1 Research aims

Little empirical research has been carried out to date on the assessment of MMC media quality. There are no established guidelines addressing either the conditions required for successful MMC task completion, or the assessment methods that should be used to derive these conditions.

The overall research objective is therefore to gather as much data pertaining to audio and video quality requirements in MMC as possible, using a variety of techniques. Exploring a range of techniques will enable the best assessment methods to be determined, such that the degree of substantive knowledge about quality requirements can be maximised.

This chapter has reported that quality is not a unidimensional phenomenon (section 4.11). Another research aim is therefore to identify and measure the most critical subjective quality dimensions for different MMC communication tasks. The ultimate aim is to be able to pinpoint actual *quantities* for the dimensions, i.e. to establish the critical quality boundaries (minimum and maximum quality thresholds) for a particular dimension, in the context of a particular task. Once a large set of empirical data has been collected, this approach will yield a taxonomy of quality boundaries for audio and video for a range of tasks. Application developers and service providers could apply this taxonomy to infer objective QoS requirements for particular applications.

The methods that are utilised in addressing these aims are described in the following section.

4.13.2 Research methods

In Chapter 3, the complexities of MMC communication were described, and Chapter 4 has considered established assessment methods with respect to MMC conditions. It has been argued that, given the novel type of degradations that characterise MMC communication, many existing methods may not be suitable for MMC quality assessment. However, the methods should not be dismissed on face value alone, and must be evaluated in the context of MMC studies.

The first research priority is therefore to investigate the use of established audio-visual assessment methods (i.e. speech intelligibility tests and quality rating scales) in measuring MMC communication. The performance of these methods must be evaluated and, if required, adaptations must be made to make the methods better fit the characteristics of MMC communication. If necessary, steps must be taken to develop new methods more suitable and meaningful with respect to MMC conditions.

It is a central tenet of the HCI discipline that evaluation should take place *in context*, and the research that is undertaken in this thesis conforms to this approach as far as possible (see section 2.8). The thesis research began with observations of a remote language teaching field trial, with specific questions arising from this being investigated explicitly in experimental set-ups where conditions can be better controlled. Advantage was taken of any ongoing MMC work at UCL to carry out further context-based MMC research. This included the testing and implementation of new speech repair schemes and an audio-video synchronisation mechanism. To a certain extent, therefore, the research follows an incremental approach, building up both substantive and methodological knowledge using a variety of different approaches and test-beds.

Since little previous empirical research into delivered MMC quality has been undertaken, the thesis research is necessarily of an exploratory and descriptive nature, occasionally without strict hypothesis testing. This means that it may not be possible to carry out statistical analyses on, or generalise from, all of the findings, but there is significant value in the research, in terms of contributing to general MMC knowledge and refining areas for future research.

A mixture of qualitative and quantitative data is gathered through small exploratory studies, and lab simulations of real-world conditions identified through extended field trials. Different user groups are investigated, as well as different tasks, meaning that a broader understanding of what matters in media quality assessment can be attained. Subjects and users in the studies are often asked to complete questionnaires and to participate in interviews and focus groups. This helps to extract the quality dimensions that are most important to the users.

In order to collect as much data as possible, a variety of different subjective evaluation approaches are employed: established assessment methods (speech intelligibility and quality scales), questionnaires, interviews, and performance observations (especially in the field trials). The established rating scale methods are indeed found to be lacking in addressing the conditions of MMC communication, leading to the development of novel quality rating methods. Through interviews and group discussions, and on the basis of comments made in questionnaires and during experiments, the quality descriptors that users use for different types of MMC degradations are extracted and associated with different quality dimensions. Suggestions for their application in future subjective measurement tools are put forward.

The overall methodological approach acknowledges that there are multiple factors that influence users' perception of multimedia speech and video, and by using multiple means of gathering data, the complexities and depths of perceived quality can be accessed. The knowledge gained can be employed in developing a HCI methodology for assessing and establishing MMC subjective quality requirements.

The next chapter (Chapter 5) documents the early stages of the research, where ITU-recommended scales were used in the assessment of MMC speech and video in firstly a field trial, and then in smaller controlled studies. The aim of the research was twofold: to assess the usefulness of the assessment methods used, and to begin to determine the critical levels of some of the subjective quality dimensions identified.

Chapter 5: Modelling and capturing Internet speech and video in early studies

This chapter first presents and discusses a long-term field trial undertaken to assess the feasibility of using MMC technology in a distance education application. The chapter then presents lab-based speech intelligibility and speech quality studies following on from observations made during the field trial. Both traditional 5-point and modified rating scales are used in these studies. The research highlights the importance of conducting research in the field, and both the advantages and drawbacks of using category scales and static, post-hoc ratings in attaining perceived quality ratings in MMC environments.

The motivation for, and the findings from, each of the studies can be divided into those which contribute to substantive knowledge, and those which advance methodological knowledge. It is these factors that will be the focus of the chapter.

5.1 The ReLaTe field trials

The key question addressed in this thesis is that of how to measure perceived quality in MMC environments. In order to do this properly, the necessary first step must be to observe the technology in context, in use by real users with explicit task goals, to see what variables have the greatest impact on perceived quality. The ReLaTe (Remote Language Teaching over SuperJANET[21]) project offered one such opportunity, in the context of small group language teaching. This section describes the issues involved in evaluating a novel communication system such as ReLaTe, and introduces the approaches used and the conclusions drawn. The section concludes with a discussion of issues that need to be investigated further, in controlled studies.

5.1.1 Overview of the ReLaTe project

ReLaTe was a joint project between UCL and Exeter University, with two goals:

- to provide a working demonstrator of a multicast-based conferencing system for remote language tuition;
- to assess the feasibility of using multicast technology to provide remote tutoring in a field trial with teachers and students.

It was hoped that the results of the trial would provide guidance for the development of networks and workstation technology fit for distance learning applications, and help in identifying corresponding changes required in distance education pedagogy.

[21] SuperJANET is the UK's national broadband network for the education and research community.

Language teaching was chosen as the project domain since it would place high demands on the quality of the audio. Students of a foreign language do not possess the native speaker's facility for compensating for poor audio quality, and lip synchronisation is commonly supposed to be required for at least some tasks, e.g. pronunciation. It was reasoned that if multicast audio can support language teaching, it should be good enough for most distance education tasks.

The technology was investigated in the context of small-group tutoring sessions. This was partly for practical reasons (number of workstations and students required), but also because groups with fewer students were expected to yield a higher degree of student involvement and interaction, which is more demanding for the technology than a one-to-many lecturing scenario would be.

A pilot trial took place in July and August 1995, with the main field trial taking place from October to December 1995, involving tutors and students from the Language Centres at UCL and Exeter University, over the SuperJANET IP network between the two sites. Four weekly sessions (Advanced French, French for Business, Portuguese for Beginners and Latin) ran through the term. Two or three students participated per tutorial. All the lessons were taught in two-hour sessions except for Latin (one hour). Each participant sat at a UNIX workstation (a Silicon Graphics Indy) equipped with a headset and a camera. The tools that were used were **vic**, **wb** and an early prototype of **RAT**. After the pilot trial the tools were combined into an integrated interface[22] (shown in Figure 9) which eliminated the screen real estate problem discussed in section 2.8. However, in hiding the 'unnecessary' windows from the users, the task of making quality observations was made harder for the evaluators (see section 5.1.3.2).

The sessions were observed and assessed from both a usability and pedagogical point of view. The usability findings are the focus of the following sections.

The ReLaTe project evaluation approach and results have been published in Watson and Sasse (1996b) and Watson and Sasse (1996a), which focuses at greater length on the pedagogical evaluation.

[22] In standard multicast conferences, users have to position and juggle windows for audio, video and shared workspace tools, plus a separate window for every video stream displayed. This allows experienced users to set up their own conferencing screens, but is a cumbersome and often confusing task for less experienced users.

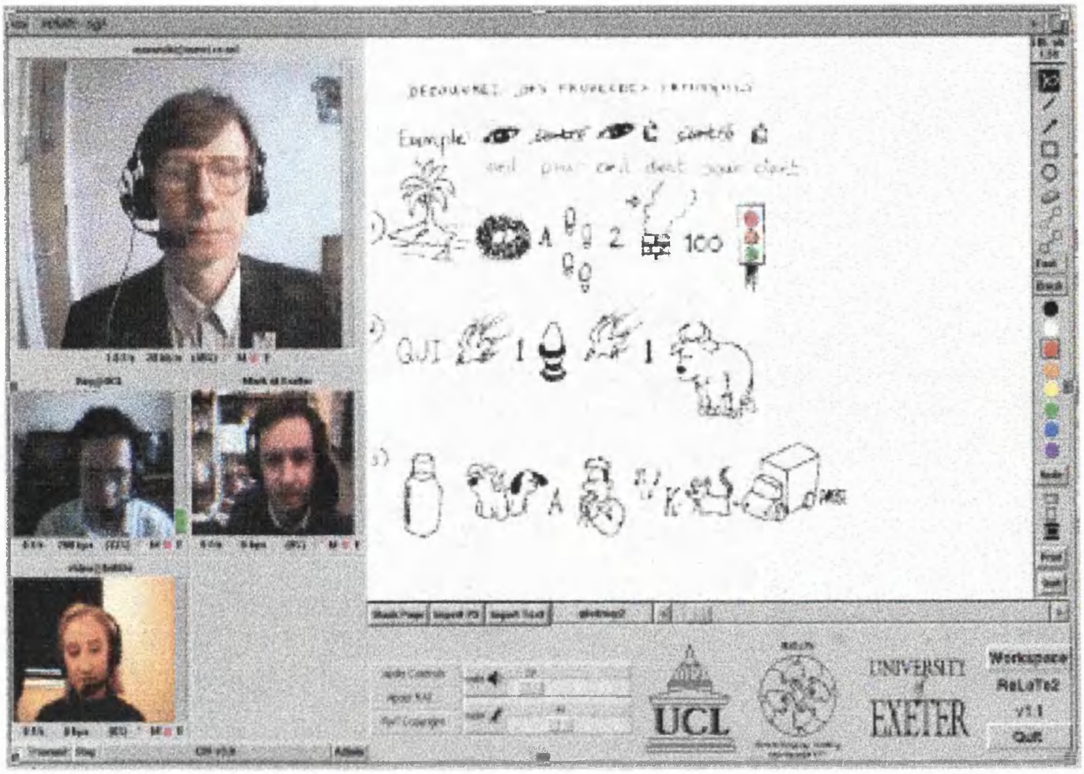


Figure 9: The ReLaTe integrated interface

5.1.2 Evaluation focus and methods

The ReLaTe study was of interest and importance to the thesis research because it would reveal which of the myriad variables have the greatest impact (in this particular task) on quality, from a subjective point of view. It also provided the first opportunity to start investigating how best to measure perceived quality of audio and video delivered in MMC environments. It allowed the issue of 'what is quality?' to be addressed.

Controlled experimental studies of videoconferencing systems have often measured task performance while aspects of the system quality (typically video) are varied, e.g. size of image, presence or absence of video. These types of studies emphasise participants' performance, rather than their perception of the system. However, the tasks that participants are asked to perform in these studies tend to be limited and artificial, the interactions tend to be one-to-one, and data collection methods are restricted. At the time of the ReLaTe project trials very little quantitative or qualitative analysis had been carried out using participants in a field study of a system in extended use over a real network such as the Mbone, although occasional observational studies such as those of the MICE seminars (Sasse et al., 1994a) had been reported.

It was accepted from the outset that *quantitative* evaluation of the ReLaTe system in use would be extremely difficult. A lab-based subjective quality evaluation would be based around known objective quality conditions, such that the relationship between the two can be properly established. However, a multicast conferencing field trial across a best-effort network such as the Mbone does not permit this degree of understanding of objective conditions. For example, the multicast route between UCL and Exeter involved a number of hops, and one of the routers was particularly prone to congestion problems. The level of connectivity between the two sites was therefore outside the control of the investigators, and it was not possible to predict what the objective quality would be for each language lesson[23]. Although technically possible, taking a recording of the objective conditions was not practical since an enormous amount of disk space would have been required. As a result, the focus of the evaluation was placed on the gathering of *qualitative* data. A variety of evaluation techniques were chosen so that the data captured was as rich as possible. Evaluation of the technological and pedagogical issues of the lessons therefore took place via

- observation (by trained observers sitting with the teacher/student and by language teachers acting as unobtrusive 'expert observers', from a separate workstation);
- questionnaires and rating scales administered after the lessons;
- a group discussion workshop with all the participants after the end of the trials.

5.1.2.1 Observation

The observations carried out were from two perspectives – usability of the system (by the study investigators) and language teaching pedagogy (by 'expert observers' – other language teachers). While the experts were looking in particular for educational aspects that could and could not be achieved, the usability investigators were focusing in particular on whether the delivered audio quality was good enough for the task, and on how the video channel was used. The discussion in Chapter 3 (section 3.4) has shown that there is much debate as to what, precisely, a visual channel adds to a conference (e.g. Whittaker, 1995). Given that the frame rate in the ReLaTe system was low (2-5 fps), the question of what, if anything, the video channel would be used for, was of special interest. It was also of interest to ascertain whether the size of the video images was large enough for use in a teaching application. The integrated interface allowed the participant to select one image to be CIF-sized while the others were QCIF-sized, by clicking on the name of the person above the image (see Figure 9). There was no restriction over how many times the images could be switched in this fashion.

5.1.2.2 Questionnaires and rating scales

Questionnaires were completed by the students after each lesson (see Appendix A). The questionnaire covered 4 areas: audio issues, video issues, the user interface and pedagogical issues.

[23] A certain lack of knowledge regarding objective conditions is likely to always be the case in field trials over a best-effort network, although there are now better techniques available to record the objective

Since one of the key aims of the project was to determine whether multicast audio was of sufficient calibre to support demanding applications, it was decided to include the 5-point listening quality and listening effort scales in the questionnaire to see whether an indication of the subjective opinion of the overall audio quality during each particular lesson could be gained. However, as discussed in the section 3.2.4, conditions over a wide-area network can change rapidly and unpredictably, such that a request to describe the quality for a two-hour lesson may be rendered meaningless.

5.1.2.3 Group discussion workshop

At the end of the project a group discussion involving all the students and the teachers was conducted. It was felt that bringing all the participants together would be of value since interviewing only students or only teachers might have resulted in biased responses. It was hoped that bringing the two groups together would also trigger some memories or impressions that might otherwise have been forgotten.

5.1.3 Observations and findings from group workshops and questionnaires

5.1.3.1 Overall findings

The main finding from the group workshop was a positive result for the project: both teachers and students pronounced the system a success. They found it enjoyable to use and appreciated the level of bonding that came about through all participants learning about the capabilities of the system together. Both teachers and students remarked on the high degree of concentration produced when using the system. The teachers and expert observers agreed that using the system produced as least as good learning as in a conventional FTF environment. One teacher reported finding that using the system did not impinge at all on what she could do with a class. She said it could be used to *replace* a FTF class.

Other positive aspects included an observation from the teachers that the students wanted to get to know each other, and took an initiative with the system that does not happen in a classroom, since physical presence can be taken for granted here. The teachers were impressed by the fact that it was not difficult to relate to the students through the system. They reported that using ReLaTe seemed to make lessons more amusing, and the teacher/student relationship was more informal because everyone was learning how to best deal with the system. The interaction between students and teachers was made 'more equal' by the equal access all had to the tools. For example, the whiteboard is normally only in the teacher's domain, but ReLaTe placed this tool with all participants. The point was made by one of the teachers that in a normal language

conditions as a conference occurs (see sections 7.2 and 8.1).

tutorial participants sit in a semi-circle and are unable to see the faces of all the other participants, but in Relate all participants can see all faces all of the time. The students agreed that this made them feel that they could not drift off, that they had to concentrate, because it would be immediately apparent to everyone else if they were not engaged – *"it's like Big Brother's watching you!"* The level of spoken behaviour was also felt to be increased, which is important since this is one of criteria language teachers use to assess the success of a session.

On the negative side, the participants reported that wearing heavy over-the-ear headphones made the lessons very tiring, since the lack of sidetone means that it is not possible to hear one's voice normally, and can lead to forcing of the voice. They also reported that any additional noise was stressful, such as somebody coughing.

All other negative aspects reported were to do with the audio quality, presented in the next section.

5.1.3.2 Audio findings

With respect to evaluating the audio channel, a number of problems were encountered. Firstly, because the participants wore headsets during the lessons, it was not possible for a usability observer to hear the sound quality being experienced on that specific workstation. Secondly, in the main body of the trials the integrated interface had been designed so that only the windows that were strictly necessary for the purpose of the language lesson were visible. Therefore it was not possible to view the **RAT** window with the objective packet loss statistics for the workstation in use (see Figure 4). For these two reasons, the observer had to rely solely on comments the user or expert observer made during or after the lesson with respect to the sound quality. The assessment of audio quality was also affected by the fact that the participants in a conference were all sitting in different environments, with different levels of background noise, different headsets and different workstations. Despite these hindrances, though, it is possible to make certain statements about the audio quality.

In the pilot trial an early version of **RAT** was used. This version required participants to push-to-talk, i.e. to place the cursor in the **RAT** main window and depress the left-hand mouse button. It was observed that participants found this very irritating, since it meant that they could not be writing in the whiteboard at the same time as speaking, reducing the ability of the teacher in particular to explain details while writing. In addition, it was also found to be disconcerting that the auditory backchannels were not open for the communication of encouragement and agreement while another person was speaking. These two facts led to an improved silence detection mechanism being implemented in **RAT** such that the microphones could be left open all the time in a full duplex manner.

Despite these improvements, in the group discussion workshop both the teachers and students listed improved audio quality as their number one recommendation for change. The students remarked that this was particularly important for beginners since there is a degree of reliance on certain syllables which cannot be derived from context when just starting to learn a language. The objective statistics (which had been collected in the pilot trial) indicated that packet loss might be a prime factor in causing the poor audio quality. In the pre-trials using the separate tools, it had been possible to observe the audio packet loss rates that one person received from another participant by right-clicking on that participant's name in the **RAT** main window (Figure 4). By recording these figures at regular intervals, average objective loss statistics could be produced for a session. These had shown that loss levels were often in the region of 8-10%. Results taken at 5-minute intervals in one such lesson are presented in Table 4 below.

Time	**Packet loss reported**
1:10	8.49%
1:15	8.39%
1:20	8.13%
1:25	7.94%
1:30	8.08%
1:35	7.97%
1:40	8.73%
1:45	8.88%
1:50	8.72%
1:55	8.7%
2:00	8.58%
2:05	8.54%
2:10	8.68%
	Range:7.94 - 8.88 **Mean:** 8.45

Table 4: Audio loss levels reported during one hour-long session

Subjective quality opinions were gathered via questionnaires after individual lessons, and at the group workshop. Due to the small subject numbers involved in the study, it was hard to draw conclusions from the individual questionnaires. Table 5 shows the rating scale results for two students in a course which ran for 7 weeks. Looking at the audio quality results, there is a suggestion that perceived quality improves over time, but it is not possible to state whether this is due to better objective quality or adaptation. The effort scale results suggest that, if anything, it has become more effortful to understand the speech over the course (which may have been due to the level of the language lesson becoming harder).

	Audio Quality		Listening Effort (English)		Listening Effort (French)	
	S1	S2	S1	S2	S1	S2
Week 1	2	1	3	3	2	2
Week 2	3	3	4	2	3	3
Week 3	3	2	4	3	3	2
Week 4	NA	2	4	2	3	2
Week 5	4	4	4	NA	3	NA
Week 6	4	4	4	4	3	3
Week 7	4	3.5	4	3	3	4
Mean	3.33	2.79	3.86	2.83	2.86	2.67
Overall	4	3.5	5 & 4	NA & 4	4 & 2	NA & 2

Table 5: Audio scale rating results for two students (S1 and S2) over the course

In addition to the individual questionnaires completed after each lesson, the students also completed a post-course questionnaire at the group workshop. Seven of the students filled out a post-course questionnaire. Pertinent findings from this and the general discussion are presented below.

Five of the 7 respondents claimed that the audio quality varied between lessons, but 6 of the 7 managed to provide a rating for the overall quality of a typical session, nearly all good or fair (4, 4, 4, 3.5, 3 and 2). Opinion was split as to whether the audio quality varied within lessons (2 said yes, 3 said no, and 2 said sometimes).

The disparity between the volume levels of different participants was reported to cause annoyance, but the main audio problem was the short and unpredictable audio losses[24]. This was perceived to have a crucial impact on learning in a beginners' class (Portuguese), where the students were unfamiliar with the sound of the language and the teacher wanted to practice pronunciation of short sounds. However, as a number of participants commented, the poor audio quality required them to concentrate more, which was beneficial to their learning. Another student shrewdly observed that if the audio is of poor quality, the teacher will be more inclined to speak slowly...

5.1.3.3 Video findings

In designing the integrated user interface, a facility was incorporated for viewing one participant at CIF-sized, and the others at QCIF-sized. It was hypothesised that the participants might switch the CIF-sized image according to who was speaking, to maximise all available visual cues, but in fact most participants tended not to switch the large image, suggesting that there

[24] See section 8.2 for a fuller investigation of this problem.

were few extra cues to be picked up from the larger image size. Four out of the 7 respondents to the post-course questionnaire said they occasionally used the enlarging facility, but no more than 4 times in a lesson. Two of the 4 said that they would only change the image once (making the tutor the large picture). The other 3 respondents reported never using the facility. It is therefore clear that the participants do not change image size according to who is speaking at the time. Of course, this finding may be a factor of group size, and also due to the fact that the tutor may have done most of the talking. It would be interesting to observe whether behaviour differs with larger groups sizes or in a non-teaching environment.

Evaluating the video component revealed some interesting discrepancies between what the users thought they used the video for, and what they were observed using it for. In the post-course questionnaire the students were asked how often they monitored their own image during a lesson. Of the 7 respondents, 2 said that they monitored their own image often (11 or more times per lesson), 3 said that they monitored it occasionally (5-10 times per lesson), and 2 said rarely (1-4 times per lesson). Although it is true that one of the participants rarely looked at his video image (to the extent that he was rarely in shot at all), the evaluators had observed the rest of the participants all monitoring their own image far more than 11 times in a session, perhaps suggesting that users do not consciously register looking at their own images (see section 3.4 - Gale, 1991). The questionnaire answers suggested that the benefit of the video was mainly psychological, although the observers noted much and varied use of it. It is possible that the participants consciously dismissed the video since there was no synchronisation with the audio channel, but were unaware of how much use they made of the tool for indirect communication[25].

Although many of the participants could see the point of having a video image of the tutor and other students in the session, there was some ambivalence as to the usefulness of having their own video image on display ("*you don't need a mirror in the classroom*"). Only one of the students saw the importance of this from the teacher's point of view, commenting that her own video image was useful in that it helped the tutor to see when she was lost. It was commented that the position of the camera (on the top of the workstation) was not conducive to encouraging its use, since the participant's attention in the tutorial was usually focused on the computer screen. Eye contact was of course not possible due to the visual parallax problem (see section 3.3.6).

With respect to the low frame rates (2-5 frames per second) experienced in the trials, not everyone found them to be a disadvantage. One student commented: *"Tutor's lip movement was*

[25] Confirmatory evidence of this was suggested by an additional lesson that was run with a new synchronisation method in use at 6 frames per second, where subjects afterwards reported using the video far more than usual. They also reported that audio was easier to understand when video was faster.

delayed and didn't match speech. This actually encouraged me to listen harder to the French sounds - very useful". However, the lack of synchronisation with the audio stream may have been another reason for not selecting the larger image size more often.

A common problem observed was with participants moving partly out of camera shot. This was often related to space management: when the teacher used a textbook or students took notes on paper, they would lean sideways and out of the range of the camera.[26]

All the same, video was felt to be a valuable component, and participants were observed making use of the channel for many aspects of communication:

- **Trouble-shooting**: the video was used to check whether anyone was speaking but not being heard, indicating an audio problem.
- **Comprehension checking:** through nods and appropriate facial expressions.
- **Common reference**: props for the lesson were sometimes held in front of the camera to show what was being referred to.
- **Nonverbal gestures:** gestures pertinent to the target language could be demonstrated.
- **Psychological reassurance:** the lack of sidetone could make the system feel 'dead' - even with very low frame rates, participants appreciated being able to see that there were other people there.

In the group discussions at the end of the trials, much enthusiasm for the video channel was voiced by both the tutors and students. It was suggested that relationships might be beneficially affected by the fact that, unlike in a conventional classroom, all participants' faces were always visible to one another, but it is likely also that in an environment where communication was stressed (through being in a foreign language, and also with degraded audio quality), the video channel becomes more important, as concluded by Veinott et al. (1997 and 1999).

5.1.4 Interpretation of results

This section presents interpretations of the results presented above, and discusses how to set about confirming these interpretations.

Despite the positive result for the project as a whole, measuring perceived audio and video quality in the field trial was not straightforward. In particular, the participants reported that audio conditions varied both within and between lessons, meaning that it was hard to interpret the subjective quality ratings given on the rating scales. Since audio quality was identified by

[26] It is a testament to the absorbing nature of the classes that only the evaluators seemed to find this annoying!

all participants as the most critical aspect to improve, an understanding of delivered audio quality and its impact on the participant is imperative. The results from the logging of the objective statistics undertaken in the pilot trial suggest that packet loss plays an important role in perceived audio quality – loss rates of 10% and above were common. This factor needs to be investigated in a lab-based setting where more control can be exercised over the environment.

The rating scales that were used in the study, particularly the 5-point quality scale, did not allow any solid conclusions to be drawn about the *adequacy* of the delivered quality. In part this was due to the low subject numbers in the study, but it was also due to the fact that it can be presumed that the quality that the participants experienced varied over time. How, then, should the quality rating awarded at the end of such a session be interpreted? Does the 5-point scale allow enough flexibility to the user? It was witnessed that users often indicated between options on the scale, or placed *two* ratings on the scale, one for 'the beginning' and one for 'the end'. How should a rating that is anything other than *good* or *excellent* be interpreted? How can we know what is responsible for the quality that is perceived as *poor* or *bad*? Use of this rating scale in the context of MMC audio needs to be investigated in greater detail.

Overall, the perceived quality conclusions that could be drawn from the ReLaTe field trials suffered from a lack of control over various factors. It is difficult to prevent participants from behaving in a certain way, e.g. individual preferences for volume and lighting can lead to unsatisfactory sound and view for other participants, mainly because the participants do not understand the impact their behaviour could have on other participants. Due to real-life demands on the participants, it was often not possible to perform appropriate soundchecks/testing of levels before the lessons began. For these reasons, further research needs to be carried out in a more controlled environment so that the weight of these different factors can be explored and identified (see section 8.2).

On the basis of these points, an approach is outlined for subsequent work. The most likely main effects of poor subjective quality will be isolated and tested in controlled studies. In the first instance the effect that is investigated in detail will be audio packet loss, and different means of repairing this loss (e.g. sections 5.2, 5.3, 5.4 and 5.5). However, other factors are also investigated. For example, one of the problems highlighted in the ReLaTe project was the lack of synchronisation between the audio and video channels. Synchronisation between the audio and video channels in MMC is not straightforward because the two streams are sent separately and the packets arrive at different times. However, a new method for synchronising **RAT** playout with the video stream was developed in the autumn of 1996, and is reported in Kouvelas et al. (1996). The subjective testing of the synchronised audio and video that was attained is described in section 5.7. The material that is tested in the controlled studies is necessarily shorter than that that formed the basis of the opinions voiced in the ReLaTe project. Although this permits safer conclusions to be drawn, it always borne in mind that longer, more natural material should also be assessed in the research. This is why the research also takes into account experiences and results of further field trials (see sections 6.2 and 8.1). It is hoped that this combined, interleaving approach will lend itself to better identifying both the critical factors in MMC communication, and also the best means by which to assess the impact of these factors.

One of the key observations from the ReLaTe field trials was that the major impediment to successful conferencing was poor audio quality. It was assumed that the main cause of this poor audio quality was network packet loss: through informal logging of the packet loss statistics during lessons, a picture of the loss levels could be built up. Clearly, lab-based studies of user perception of different levels of packet loss would be valuable in determining more exactly the impact of this type of degradation. The impact of packet loss and different repair schemes was therefore investigated in a variety of different studies, starting with words and then building up to longer samples of speech.

5.2 Study 1: Compensating for packet loss in Internet audio

In ReLaTe it was found that the aspect most critical for a successful conference was the audio quality, and all too often the delivered audio quality was a source of dissatisfaction to the participants. It was hypothesised that the primary cause of this poor audio quality was packet loss due to network factors, since observation of the packet loss statistics had shown that loss was often in the region of 10% and greater. The impact of this loss on the intelligibility and perceived quality of the received speech needed to be understood, and steps towards this understanding were undertaken in a number of lab-based studies. The earliest of these studies investigated the effect of packet loss on single words. The study is presented here and the importance of the results is expounded.

The motivation for the study was twofold:

1. To investigate how to measure the effects of packet loss on speech intelligibility and quality from a perceptual point of view. Both speech intelligibility and perceived quality were investigated since it is believed that the two, although not mutually exclusive, are not identical – it is possible to get high speech intelligibility at the expense of perceived quality.
2. To use the identified methods to determine the efficacy of different techniques for repairing audio packet loss. The study formed part of the iterative development of the audio tool **RAT** (section 2.5.2), and at this early stage in the tool's design, it was important to gather information about the impact on user perception of different repair methods.

The experiment was designed to investigate the effect on speech intelligibility and speech quality of different levels of packet loss (0, 10, 15, 20, 30 and 40% loss[27]) where this loss is repaired by three different methods (silence substitution, packet repetition and LPC redundancy). The experiment also investigated the effects of different packet sizes (20, 40 and 80 ms).

The key experimental hypothesis was that speech intelligibility and quality would be worst where silence was used to repair loss, and that these would be enhanced where the loss was repaired using the techniques of packet repetition and LPC redundancy. It was also predicted

[27] These loss levels were chosen because - although Handley (1997) reported 2-5% loss as common - some receivers receive much more (see section 2.3.2). In this thesis research one aim was to understand what happens at high loss rates, particularly in view of the argument that as MMC popularity increases, so will traffic. It is simply not possible to guarantee that loss levels will remain at the level Handley reported.

that packet repetition would begin to fail as a method when loss rates were high and packet sizes large, since the speech characteristics would change within the lost packets of sound.

5.2.1 Material

A thorough investigation of the different types of speech intelligibility tests was carried out (see Table 3 in Chapter 4). Egan's (1948) Phonetically Balanced (PB) word lists were selected as the test material, since the words are all the same length (monosyllabic), and proportionally represent the sounds found in everyday English. There are 25 words in each list, and subjects are required to write down each word as it is heard.

The speech material was spoken by a male with Received Pronunciation[28] and recorded in 16-bit linear, 8kHz format and encoded using the DVI4 codec (IMA, 1992). A Sennheiser HMD25-1 headset was used for both the recording and the playback. Random packet loss of different constant percentages was generated on the recorded speech material using a software packet reflector[29]. The packet loss was repaired by one of three methods: silence substitution, packet repetition, or LPC redundancy. These repaired files were then stored for playback as test material. The order of the test material was randomised for the subjects.

The subjective quality of the speech was measured via the ITU-recommended 5-point quality scale. This scale was used to measure the perceived quality of the speech since it is the most widely used scale and the standard against which speech quality ratings are based in the research literature at large (see Figure 7, a).

5.2.2 Subjects

21 subjects participated. They were all native English speakers with good hearing. None had previous experience in Internet audio.

5.2.3 Procedure

A between-subjects design was employed – 3 groups of 7 subjects each listened to the 3 different repair schemes. (However, the design of the experiment was unbalanced in that not all loss rates were tested under all repair schemes.) The test files were played back through the software application Audio Tool on a Sun SPARC-10 workstation. The subjects were required to write down each word after it had been played. On completion of each list of 25 words, the

[28] Defined as a non-regional British accent (Ainsworth, 1976).

[29] An application level packet reflector and forwarder, written by Orion Hodson at UCL. Available online from http://www.cs.ucl.ac.uk/staff/O.Hodson/misc/reflector.tar.gz

subjects were asked to rate the *quality* of the speech as they perceived it on the ITU 5-point listening quality scale.

5.2.4 Results

Table 6 shows the means and standard deviations for the intelligibility results, whilst the results for the quality ratings are shown in Table 7. The complete data set can be found in Appendix B.

Loss Repair	Packet size	10%	15%	20%	30%	40%
Silence Substitution	20 ms	83.33 (5.89)	85.14 (3.02)	71.43 (14.32)	67.33 (7.34)	
	40 ms	83.43 (7.81)	84.57 (5.86)	71.43 (8.14)	69.33 (10.01)	
	80 ms	92 (5.66)	74.29 (13.03)	74.28 (5.59)	53.14 (9.99)	
Packet Repetition	20 ms	86.86 (3.02)	90.29 (6.05)	91.43 (3.6)	85.71 (3.15)	70.86 (15.27)
	40 ms	85.71 (5.09)	84.57 (5.85)	81.14 (9.14)	77.14 (7.9)	53.71 (11.97)
	80 ms	73.71 (7.61)	82.29 (6.87)	74.86 (8.23)	61.14 (10.76)	53.14 (14.74)
LPC Redundancy	20 ms			86.86 (9.72)	86.86 (8.55)	77.71 (6.05)
	40 ms			89.71 (8.9)	79.43 (7.81)	76.57 (6.94)
	80 ms			88 (6.53)	77.14 (6.82)	72.57 (7.81)

Table 6: Means and standard deviations for PB word list intelligibility results

Loss Repair	Packet size	10%	15%	20%	30%	40%
Silence Substitution	20 ms	2.57 (0.79)	2.57 (0.79)	2.14 (0.38)	2.14 (1.07)	
	40 ms	3 (0.82)	3.43 (0.53)	2 (0.58)	2 (0.58)	
	80 ms	3.29 (0.49)	2.71 (0.76)	2.43 (0.53)	1.57 (0.79)	
Packet Repetition	20 ms	3.43 (0.53)	3.43 (0.98)	3 (1)	2.86 (0.9)	2 (0)
	40 ms	3.71 (0.49)	2.86 (1.07)	2.57 (0.79)	2.14 (0.9)	1.57 (0.53)
	80 ms	2.43 (1.27)	1.86 (0.38)	2.14 (0.69)	1.43 (0.53)	1.71 (1.11)
LPC Redundancy	20 ms			3.14 (0.9)	3 (0.82)	2.71 (0.95)
	40 ms			3.57 (0.53)	3.14 (0.9)	2.57 (0.98)
	80 ms			3.29 (0.76)	3 (0.58)	2.29 (1.11)

Table 7: Means and standard deviations for PB word list quality ratings

These results are illustrated in Figure 10 and Figure 11 where SS stands for silence substitution, PR stands for packet repetition, and LPC for LPC redundancy.

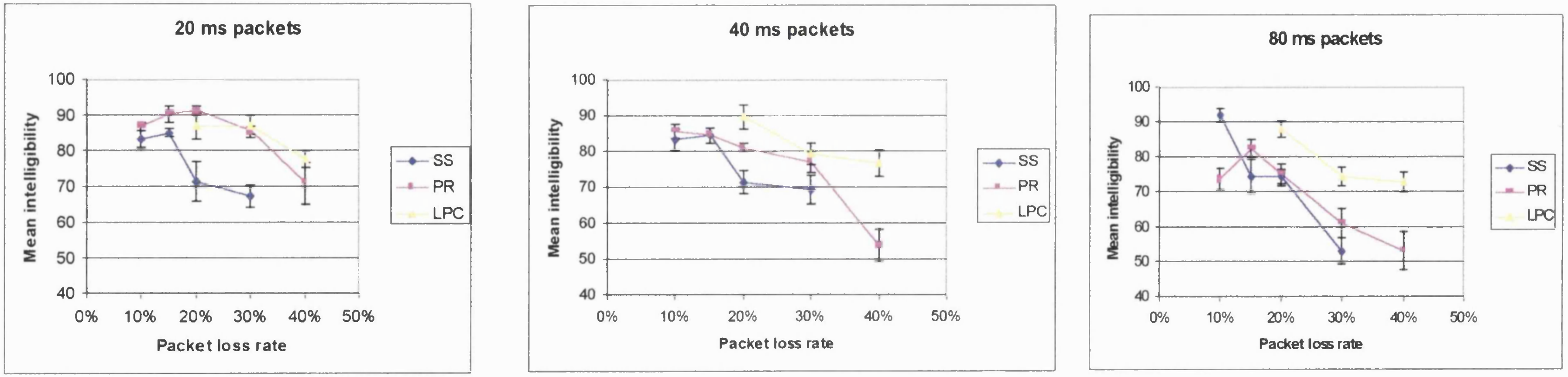


Figure 10: Intelligibility of packet loss speech repaired with SS, PR and LPC at 20, 40 and 80 ms packet sizes respectively

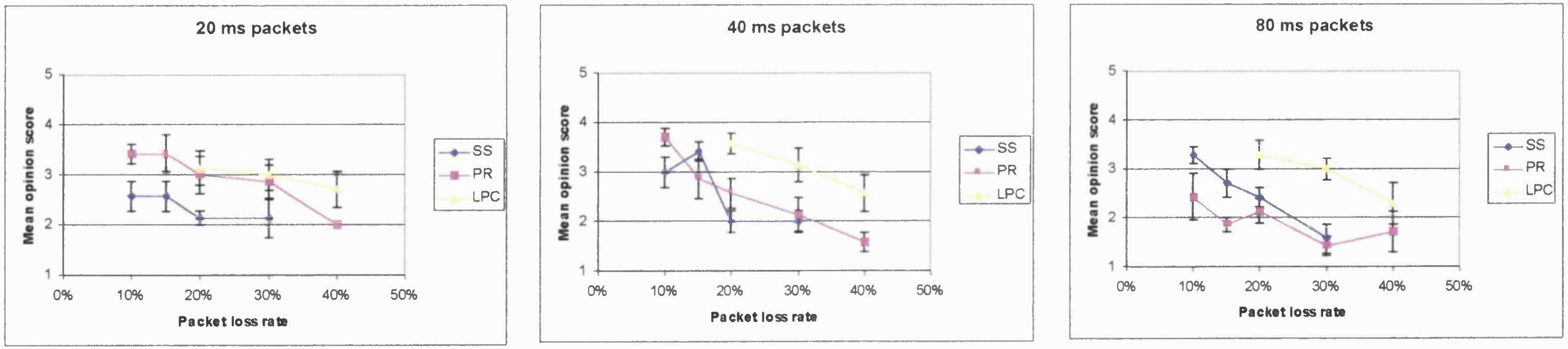


Figure 11: Mean opinion scores for packet loss speech repaired with SS, PR and LPC at 20, 40 and 80 ms packet sizes respectively

Since the study involved small subject numbers and was an incomplete design, caution must be paid in interpreting the results. However, the figures suggest that silence substitution generally produces the lowest speech intelligibility and perceived quality, at least when the packet sizes are as small as 20 ms. When packets are as large as 80 ms it can be seen that both intelligibility and perceived quality of speech repaired with packet repetition drops in comparison to the other phonemic restoration scheme, LPC redundancy. It appears from the figures that LPC redundancy produces the best intelligibility and quality in general. The figures also indicate that intelligibility and perceived quality do not always concur. For example, in the 80 ms, 15% loss condition, packet repetition produces better intelligibility than silence substitution, but the quality ratings show that silence substitution sounded better to the subjects.

Since the design of the experiment was unbalanced in that not all loss rates were tested under the same repair schemes, the results cannot be submitted to a 3x3x3 analysis of variance. However, it is possible to look more closely at the results for the 20 and 30% loss rates, across all repair schemes and packet sizes.

Figure 12 shows a plot of the interactions between the variables of repair scheme, packet size and loss rate, as measured by intelligibility scores. The figure illustrates the effect that phonemic restoration repair schemes such as packet repetition and LPC redundancy have on speech intelligibility compared to silence substitution – the resulting intelligibility is greater. However, it appears this advantage is not always present for packet repetition - when the packet size is increased to 80 ms, the advantage of using packet repetition over silence substitution decreases, an observation in agreement with the stated hypothesis that packet repetition will begin to fail when packet sizes are large due to the speech characteristics changing within the missing packet.

A 3x3x2 analysis of variance was conducted to determine whether the interactions were significant or not. The significant main effects and interactions are presented in Table 8, with the full analysis in Appendix B. There are significant main effects of all three factors, repair scheme, packet size and loss, but there is no 3-way interaction between these factors.

Main effects	Significance	Interactions	Significance
Repair	$p < 0.01$	Repair * packet size	$p < 0.01$
Packet size	$p < 0.01$	Repair * loss	
Loss	$p < 0.01$	Packet size * loss	$p < 0.01$
		Repair * packet size * loss	

Table 8: Main effects and interactions for intelligibility scores at 20 and 30% packet loss

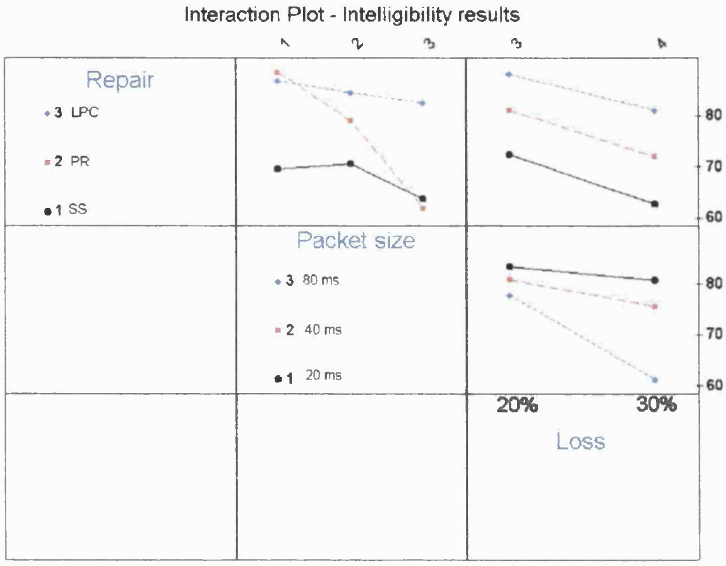


Figure 12: Interaction plot for intelligibility scores at 20 and 30% packet loss

The interaction between repair scheme and packet size is significant, with packet repetition becoming less effective as the packet size increases to 80 ms. The interaction between packet size and loss is also significant, again the 80 ms packet size producing lower intelligibility scores than others as loss rate increases.

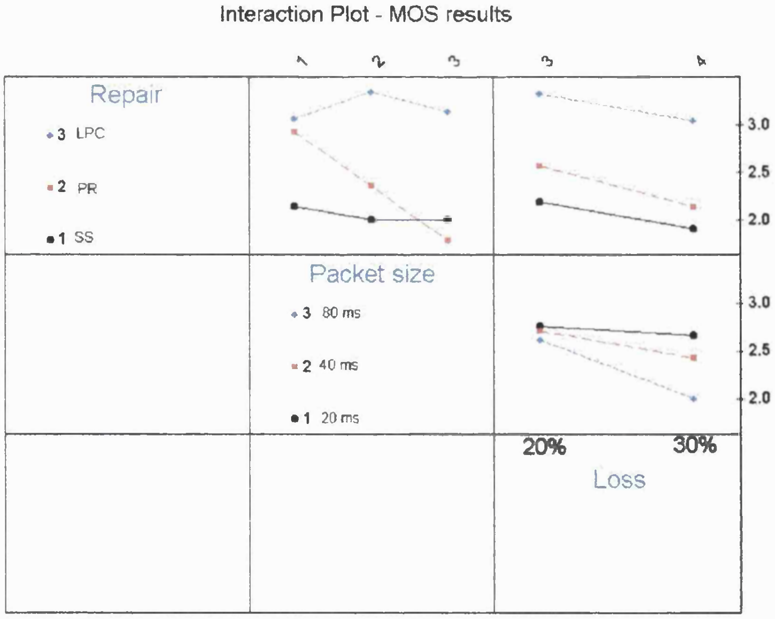


Figure 13: Interaction plot for quality ratings at 20 and 30% packet loss

A similar analysis was performed for the quality rating results for these conditions[30]. The interactions are plotted in Figure 13, and the significance of these are listed in Table 9. Looking at the interactions plot, it can be seen that the same general pattern as the intelligibility results is reflected. There are significant main effects of the three factors again, and a significant interaction between repair and packet size, packet repetition being perceived less favourably than LPC redundancy as packet size increases.

Main effects	Significance	Interactions	Significance
Repair	$p < 0.01$	Repair * packet size	$p < 0.05$
Packet size	$p = 0.05$	Repair * loss	
Loss	$p < 0.05$	Packet size * loss	
		Repair * packet size * loss	

Table 9: Interaction plot for quality ratings at 20 and 30% packet loss

5.2.4.1 Rating scale use

It was observed that use of the 5-point rating scale was depressed to a smaller range than that on offer. The range of responses given for each condition is presented in Table 10. It was found that excellent (5 on the rating scale) was only selected in one condition (15% loss, 20 ms packet size, repaired by packet repetition) whereas at the other end of the scale, bad was frequently invoked. This finding highlights the concern (expressed in section 4.4.2) that the vocabulary on the listening quality scale may not be suitable for use in quality judgements for speech over the Internet.

Loss Repair	Packet size	**10%**	**15%**	**20%**	**30%**	**40%**
Silence substitution	20 ms	2 - 4	2 - 4	2 - 3	1 - 4	▨
	40 ms	2 - 4	3 - 4	1 - 3	1 - 3	▨
	80 ms	3 - 4	2 - 4	2 - 3	1 - 3	▨
Packet repetition	20 ms	3 - 4	2 - 5	1 - 4	1 - 4	2
	40 ms	3 - 4	1 - 4	1 - 3	1 - 3	1 – 2
	80 ms	1 - 4	1 - 2	1 - 3	1 - 2	1 – 4
LPC redundancy	20 ms	▨	▨	2 - 4	2 - 4	2 – 4
	40 ms	▨	▨	3 - 4	2 - 4	1 – 4
	80 ms	▨	▨	2 - 4	2 - 4	1 – 4

Table 10: Minimum and maximum quality ratings awarded in each condition

[30] That there may be some question over the suitability of conducting parametric statistical analyses on data afforded by discontinuous – and non-interval – rating scales is acknowledged. However, since the practice is common throughout the research community, it is felt that the analyses here should be subject to the same treatment.

5.2.5 Discussion

The results from the intelligibility test broadly confirm the hypotheses outlined above: speech intelligibility is worst when silence substitution is used as a repair method, and better when packet repetition or LPC redundancy are used; and LPC produces better intelligibility than PR when packet loss rate increases and packet sizes are large. At 20 ms, LPC and PR produce greater intelligibility than SS. As packet size increases, LPC remains superior to SS, but PR does not. This is due to the fact that as packet size increases, it encompasses more of the changing speech signal, rendering a simple repetition of sthe previous packet inappropriate for enhancing intelligibility.

The quality ratings are in agreement with the intelligibility findings in that superior perceived quality can be achieved by using phonemic restoration schemes such as packet repetition and LPC redundancy rather than silence substitution. However, it can be seen from Figures 10 and 11 that the change in perceived quality does not always correspond to the change in intelligibility, highlighting the fact that intelligibility results cannot be taken as representative of perceived quality, or vice versa.

The substantive contribution of the study as a whole is clear: repairing packet loss speech with silence leads to inferior intelligibility and perceived quality when compared with phonemic restoration schemes such as packet repetition and LPC redundancy. This finding is of great significance since it is common practice in many internet audio tools to repair packet loss with silence to maintain the playout order. The damage to perception that this can cause has been illustrated clearly in this study. A clear recommendation to audio tool designers would therefore be to incorporate phonemic restoration schemes into their audio tools. Indeed, the findings from this study were instrumental in the implementation of packet loss repair methods in the early development of the audio tool **RAT** (Robust-Audio Tool), and were published in Hardman et al. (1995) and Watson and Sasse (1996b).

It was observed that the full range of the 5-point rating scale was not used in the study, with responses being concentrated towards the lower end of the scale. This suggests that the scale may not have been an appropriate one to use in an investigation of speech degradations of this type and level. Further investigations of the use of this scale need to be carried out.

In addition to these concerns over the rating scale, the question can be raised over the generalisability of the results from this study, since lists of monosyllabic words do not correspond to material encountered daily in the real world. The rest of the studies presented in this chapter therefore investigate longer bursts of continuous speech in further intelligibility and quality investigations. They also investigate only 40 ms packets, since this became the default setting in **RAT** after the completion of this study.

5.3 Study 2: Investigating the intelligibility of longer speech stimuli

5.3.1 Material

In Study 1 speech intelligibility was investigated using PB words as the test material. This material was selected since it did not require training of subjects, and could create no context effects. However, people do not generally listen to lists of monosyllabic words through choice, and so in the study reported in this section the material chosen for test was PB sentences published by the IEEE (1969). These take the form of short syntactically varied sentences with five key words, for example "*Nine rows* of *soldiers stood* in *line*" where the key words are italicised. Sentences such as these offer a good compromise in speech intelligibility investigations between the artificiality of single word tests and the contextual effects found in passage tests (see Table 3).

Five different percentages of packet loss were generated on PB sentences (0, 10, 20, 30, and 40% loss). The loss was either repaired by packet repetition (PR) or by LPC redundancy (LPC).

5.3.2 Subjects

Twenty-four subjects participated in the study. They were aged between 20 and 45, and all had normal hearing. None of them were experienced in MMC.

5.3.3 Procedure

A within-subjects design was employed. The subjects were presented 15 sentences, 4 of which were reference conditions, 11 of which were test. The order of the test sentences was randomised for each subject to prevent order effects. The subjects were asked to write down as much as possible of each sentence they heard.

5.3.4 Results

The results from this study were published in Watson and Sasse (1997).

Scoring of the sentences was carried out by marking the correctly perceived key words and working out a mean percentage intelligibility for each sentence from this. The means and standard deviations are presented in Table 11. The means and standard error bars are shown in Figure 14. The raw data can be found in Appendix C.

Loss rate Repair	0%	10%	20%	30%	40%
Packet repetition	96.67 (9.63)	87.5 (23.45)	74.17 (17.17)	83.33 (32.66)	30.83 (24.3)
LPC redundancy	96.67 (9.63)	98.33 (5.65)	79.17 (26.03)	85 (21.47)	27.5 (21.11)

Table 11: Means and standard deviations of sentence intelligibility

The results indicate that the LPC repair method offers superior intelligibility over PR, at least at lower packet loss rates: at 10% loss LPC repair produces significantly better intelligibility than PR ($t = 2.6$ with 23 d.f, $p < 0.02$). There is a significant drop in repaired speech intelligibility once packet loss exceeds 30%. (For PR, $t = 7.19$ with 23 d.f., significant at the 1% level; for LPC, $t = 9.49$ with 23 d.f, significant at the 1% level).

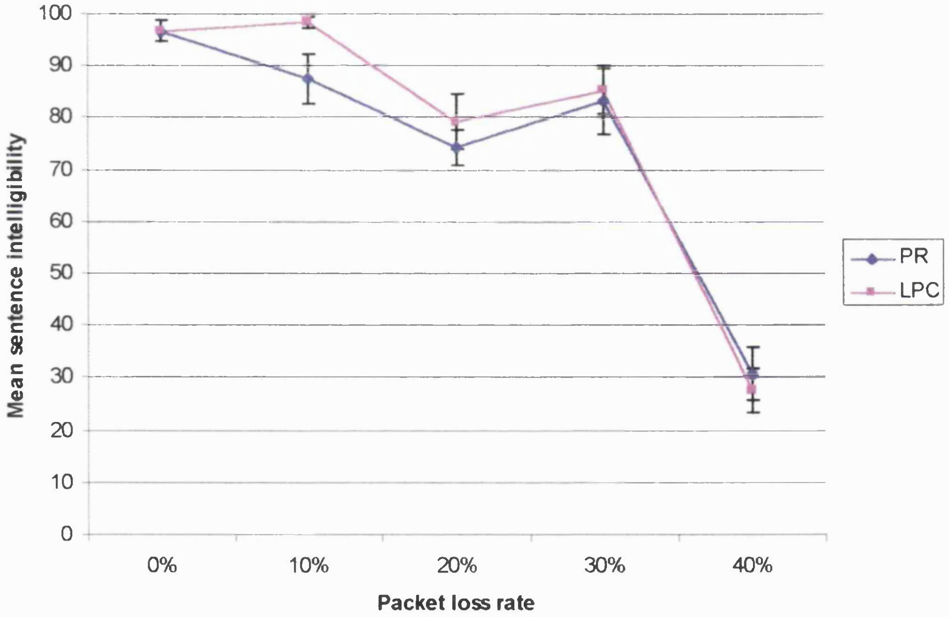


Figure 14: Mean sentence intelligibility with packet loss repaired with either packet repetition (PR) or LPC redundancy

5.3.5 Discussion

The substantive contribution of this study is that it confirms the intelligibility results of Study 1 with longer test stimuli than monosyllabic words, i.e. that with repair schemes such as PR and LPC, intelligibility drops significantly only once the loss rate exceeds 30%. It is also found that LPC is better than PR at low packet loss levels.

Although the real world-representative design of this speech intelligibility investigation was improved over that carried out in Study 1, the ability to generalise from results gained from sentence material is still lacking. It is not known what people think about the speech that they hear – does the intelligible speech sound acceptable to them? It is not safe to generalise from highly controlled experimental results such as these to the everyday fluctuating audio conditions and contextual speech that are commonly experienced on the Internet. Given these problems in generalising to the real world, and the assertion by researchers such as Preminger and Van Tasell (1995) that intelligibility is just one facet of overall quality, it can be argued that it is more meaningful to concentrate on perceived speech *quality*. The rest of this chapter reports studies that investigated the perceived quality of different sets of speech material.

As a result of the research reported in the intelligibility studies, **RAT** was developed with the capacity to compensate for some of the effects of packet loss, by offering a new selection of repair methods such as PR and LPC redundancy. Experimental results had shown that PR was a valuable means of attaining better intelligibility when loss rates were low and packet sizes less than 80ms, but that LPC redundancy offered superior performance at larger packet sizes (see section 5.2). However, there are overheads to implementing redundancy, so it was important to prove clear subjective benefits can be gained by using it.[31] It was of particular importance that the subjective testing should be carried out with longer passages of speech than the word lists that had been used in Study 1.

5.4 Study 3 Investigating the perceived quality of passages of speech

5.4.1 Material

Eleven 30-second passages were extracted from a magazine and recorded by a male speaker with received pronunciation. The passages had six levels of packet loss generated on them (0, 10, 20, 30, 40 and 50% loss) and these were repaired either by silence substitution or by LPC redundancy.

[31] These overheads are relatively small (13 or 14 bytes for a 20 ms segment of speech), but nevertheless any overhead in best-effort IP networks should be justified.

5.4.2 Subjects

A within-subjects design was employed. Ten native English speakers participated in the study. They were aged between 22 and 30, and all had normal hearing. None had previous experience of MMC.

5.4.3 Procedure

All the subjects heard all the passages of speech played out through Audio Tool on a SPARC 10 workstation through Sennheiser headsets. After each sample they rated the perceived quality on the 5-point quality scale.

5.4.4 Results

The mean opinion scores and standard deviations for each condition are given in Table 12. The means and standard error for each condition are shown in Figure 15. The raw data can be found in Appendix D.

Loss rate Repair	0%	10%	20%	30%	40%	50%
LPC redundancy	4.7 (0.48)	4.3 (0.48)	3.6 (0.7)	2.5 (1.18)	1.9 (0.74)	1.6 (0.7)
Silence substitution	4.7 (0.48)	2.5 (0.71)	1.7 (0.67)	1.4 (0.7)	1.2 (0.42)	1 (0)

Table 12: Means and standard deviations for perceived quality of passages

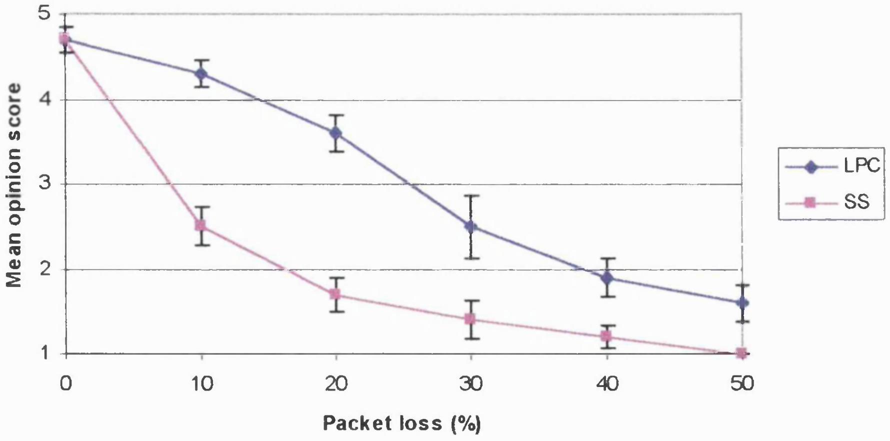


Figure 15: Mean perceived speech quality, with packet loss repaired with silence substitution (SS) or LPC redundancy

A two-way ANOVA was carried out on the data. A highly significant main effect of repair scheme was found (F 1, 108 = 71.1, $p < 0.01$) and also of loss rate (F 5, 108 = 75.82, $p < 0.01$). The interaction was significant (F 5, 108 = 6.21, $p <0.01$). A Tukey HSD test revealed that the significant differences between silence substitution and LPC occur at 10% loss (Qcrit = 5.37 at 1% level of probability, Qobt = 9), 20% loss (Qobt = 9.5) and 30% loss (Qobt = 5.5). Within silence substitution, there is a significant decrease in quality between no loss and 10% loss (Qobt = 11), and within LPC, there is a significant decrease in quality between 20 and 30% loss (Qobt = 5.5).

5.4.5 Discussion

In agreement with the previous study reported, a clear subjective preference for repaired speech was demonstrated, providing justification for the redundancy overhead.

In addition to illustrating the superiority of speech repaired with redundancy, the results support well the argument that the 5-point quality scale is an inappropriate scale to use in assessing perceived quality of Mbone speech (see section 4.4.2). For example, in the silence substituion condition, in the lowest packet loss condition (10%), the MOS drops to well below a rating of Fair, and continues to plummet throughout the remainder of the conditions. Of course, the results for the redundancy-repaired speech do occupy the full range of the scale, and therefore it could perhaps be argued that the scale *is* suitable for a new generation of Internet speech tools incorporating repair mechanisms. However, it is necessary to investigate use of the scale (and the perceived quality results arising) not only in passive listening environments, but also in *interactive* Mbone speech settings, since conversational behaviour between two or more participants is the norm in MMCs. This was the motivation for the study presented in the following section.

Unlike Study 1, the study presented here did not investigate PR as a repair mechanism. The results from another study investigating the perceived quality of passages of speech repaired with either LPC or PR are presented in section 6.1.

5.5 Study 4: Measuring the quality of a conversation

To investigate how subjects perceive the quality of an audio connection when actively engaged in a multimedia conferencing *task* (see section 2.8) rather than passively listening to Mbone speech, a study was carried out to investigate the perceived speech quality of a degraded connection between pairs of participants.

5.5.1 Material

The word game *Taboo* was selected as the study task as it provokes easy and rapid conversational exchange. The game involves one person providing descriptions of items without using certain other terms, and the other guessing what item is being defined.

Participants were connected across an experimental network using the tools RAT, vic and wb on SPARC 10 workstations. RAT was configured such that the audio packets played out at the receiving workstation effected a constant rate of packet loss. The packet loss rates were set at 0, 10, 20, 30 and 40% loss. The loss was repaired with packet repetition or LPC redundancy.

5.5.2 Subjects

Twenty-four subjects participated in the study, divided into twelve pairs. They were aged 20-45, had normal hearing, and were not experienced in MMC.

5.5.3 Procedure

Each pair of subjects played ten 2-minute games of Taboo across the network, each at a different level of loss and repair scheme. The order of conditions was randomised for each subject pair. After each game the participants were asked to rate the perceived quality of the connection on the 5-point quality scale.

Together with the results from Study 2, the results from this study were published in Watson and Sasse (1997).

5.5.4 Results

The mean quality ratings and standard deviations for each condition are given in Table 13, and means and error bars are shown in Figure 16. The full data for the study can be found in Appendix E.

Loss rate Repair	0%	10%	20%	30%	40%
LPC redundancy	3.88 (0.99)	3.54 (0.66)	3.13 (0.85)	2.46 (0.78)	2.08 (0.83)
Packet repetition	4.08 (0.88)	3.67 (0.87)	3.29 (0.55)	2.75 (0.9)	2.21 (0.72)

Table 13: Means and standard deviations for perceived quality of speech during a conversational task

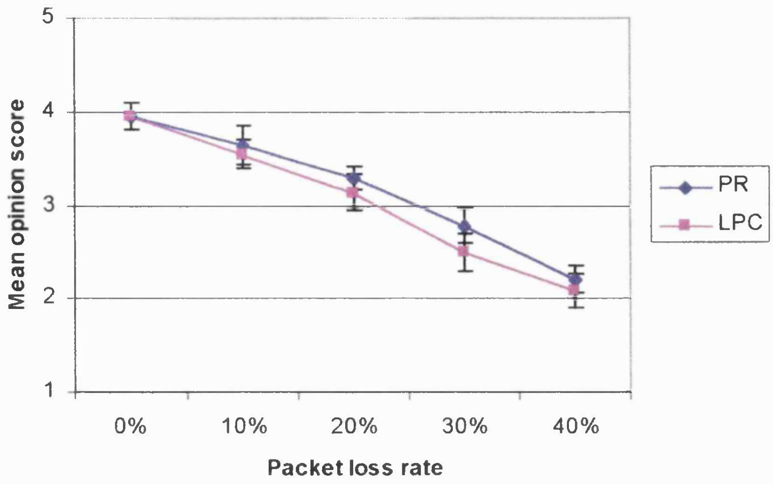


Figure 16: Mean perceived quality rating of connection during a conversational task, where packet loss is repaired with packet repetition (PR) or LPC redundancy

A two-way analysis of variance was carried out on the data. A small main effect of repair method was found (F 1, 230 = 4.17, $p < 0.05$), and a larger main effect of packet loss rate (F 4, 230 = 34.22, $p < 0.01$). There was no interaction between the two factors (F 4, 230 = 0.18, $p = 0.947$). Post hoc analyses (Tukey HSD) show that there is a steady decline in perceived quality with no significant effects found between adjacent levels of packet loss.

5.5.5 Discussion

There is a visible and steady depreciation in perceived quality as the packet loss rate increases, much as is seen in the listening quality results of the previous section. However, in contrast to the listening-only results for LPC (section 5.4), the range of the scale that is occupied is flattened. The relative depression of the 'no loss' rating in the conversation test, compared to ratings awarded in listening tests, can be attributed to the difference between listening passively to one male speaker and taking an active part in communicating with a different speaker/voice across a 'lossy' network. The degree of involvement in the task may also account for the apparent lack of perceived difference between the two repair schemes.

This finding is in agreement with Kitawaki and Itoh (1991), who found that MOS results varied according to the task being performed. They observed that people were more critical of quality in conversational tasks compared to listening-only tasks.

The loss rate in this study was kept constant throughout any one condition, and yet it was seen that the range of the rating scale used was effectively compressed to a three-point scale (good, fair, poor).

5.6 Combined findings

Figure 17 illustrates the combined findings for LPC and PR repair of words, passages and in conversations, when the packet size is 40 ms. It can be seen that the lowest perceived quality is achieved when the speech material is words and the repair method is packet repetition. The explanation for this is that with monosyllabic words, a packet could conceivably contain nearly the whole word, making the repetition of the preceding packet (silence) pointless, or it could contain a whole phoneme, the repetition of which would create the stutter effect described in section 2.5.2.5. LPC redundancy would cope better in this situation, containing a synthetic copy of the missing speech signal. The figure also illustrates the effect of context – differences between the repair methods become less noticeable as the context of the speech material increases and it becomes easier to interpret what the degraded signal must be. Because the difference between PR and LPC becomes negligible with longer samples of speech, PR was chosen as the default repair method in RAT, since LPC has a bandwidth overhead where PR has none.

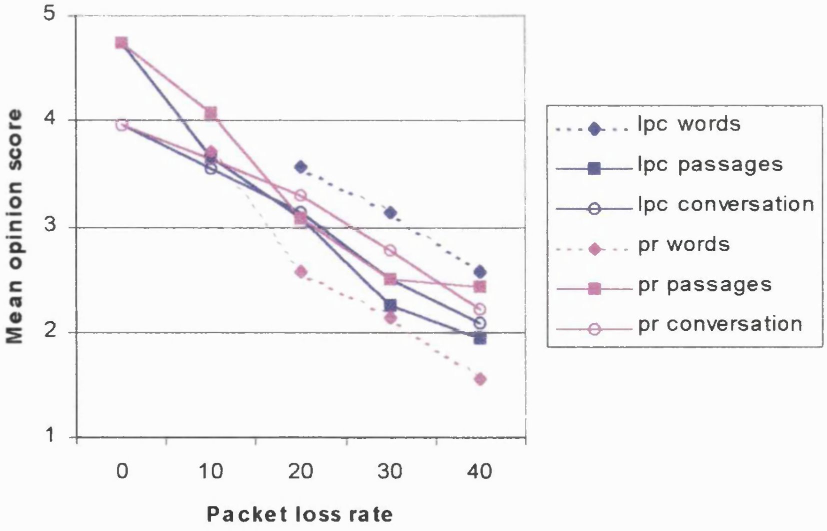


Figure 17: Combined perceived quality results from word, passage and conversation tests

So far, the lab-based studies that have been presented have focused on the audio channel. However, it was seen in the ReLaTe project (section 5.1) that one of the desired improvements was closer synchronisation between the audio and video channels. The next section presents findings from a study investigating perception of synchronisation at low frame rates.

5.7 Study 5: Assessing the perceived level of audio-video synchronisation

In real-time communication over the Internet, audio and video are sent in separate streams, and are not usually synchronised at the receiver. One reason for this is that the rate of video transmission is often low due to the large bandwidth demanded by the video channel, and the time required to encode and decode video data. Frame rates in the region of 2-8 fps could be considered too low for there to be benefits in synchronising the video with the audio stream, and the synchronisation of the two has not been previously attempted or supported in multicast technology. However, although it might be questioned why there could be any benefit to synchronising the audio and video streams at such low frame rates, Frowein et al. (1991) observed perceptual benefits at frame rates as low as 5 or 6. It is therefore worthwhile investigating the benefits afforded by synchronising Mbone audio and low frame rate video.

Informal and exploratory studies had revealed that frame rate and packet loss vary and fluctuate a great deal in normal video transmission using **vic**. This means that **vic** in its present instantiation is not a good testbed for video quality assessment studies: the **vic** settings do not seem to set any upper limit, and therefore it is not possible to control what the subject receives/sees. The experiment presented below, however, used a modified version of **vic** over a dedicated network.

The first method of synchronising multicast audio and video steams was developed in early 1996 at UCL. The method involves the two media streams communicating their required playout delay through a conference bus (Handley et al., 1995). The larger of these two delays (usually - but not necessarily - that of the video stream) is then adopted by the other stream, hence facilitating a synchronisation of the media playouts. A more detailed discussion of the technical details can be found in Kouvelas et al. (1996).

A subjective evaluation of the synchronised media resulting from the method was required. Two issues in particular had to be addressed: the method for indicating subjective opinion, and the audio-visual material to be transmitted during the test. An exploratory study was therefore undertaken.

5.7.1 Material

Creating a rating scale

A suitable method of ascertaining the perceived level of synchronisation by the subjects needed to be found. No suitable scales in the literature could be identified. In keeping with most existing image quality scales, it was decided to develop a novel 5-point scale. Due to the difficulty of identifying adequate descriptive terms that might fit the different levels of synchronicity, only the end points of the scale were labelled. The scale was accordingly bounded by the labels 'Synchronised' and 'Not synchronised', as shown in Figure 18.

Synchronised 5 4 3 2 1 **Not synchronised**

Figure 18: Rating scale for assessing degree of synchronisation

Assessment task

Since the evaluation was of the synchronisation method as it functioned in real time, there was no means of playing the subjects exactly the same conditions, for example by pre-recording the test material. This meant that the testing environment was constrained: a speaker must be able to provide each subject with as identical as possible conditions each time. It was felt that it would not be feasible for the speaker to read a passage of text, since this would entail the speaker not looking at the camera, making lip movements harder to see. Clapping and snapping of fingers were considered (as recommended by ITU-T P.920) but these tasks do not entail a judgement of *lip* synchronisation. Although somewhat artificial, it was decided that the test material would be of a speaker counting from 1 to 10 and back down again, since this was a task that could be repeated with ease by the speaker. The artificiality was not seen as too detrimental to the aims of the study, since the more predictable the stimulus, the easier it would be for the subjects to discern the level of synchronisation.

5.7.2 Subjects

Eight members of the Computer Science Department at UCL participated in the study. They were aged 24-35, with normal hearing and eyesight, and all had experience in MMC. They therefore all had roughly the same expectations of MMC quality and likely level of synchronisation.

5.7.3 Procedure

After each condition the subjects rated their perceived degree of synchronisation on the 5-point scale described above.

5.7.4 Results

The mean scores and standard deviations for each condition are presented in Table 14, and the mean results and standard error are presented in Figure 19. (The full data set for the study can be found in Appendix F.) The graph indicates that synchronising the audio and video streams at around 5 or 6 fps produces a noticeable improvement over streams that are not synchronised. Due to the small number of subjects in the study, no further analyses were carried out.

Fps Sync	2	5	6	8	12
Synchronised	2.38 (0.92)	3.5 (0.76)	3.88 (0.83)	4.25 (0.46)	4.38 (0.52)
Unsynchronised	1.63 (0.52)	1.75 (0.46)	1.88 (0.35)	2.25 (0.46)	2.25 (0.89)

Table 14: Mean scores and standard deviations for perceived degree of synchronisation

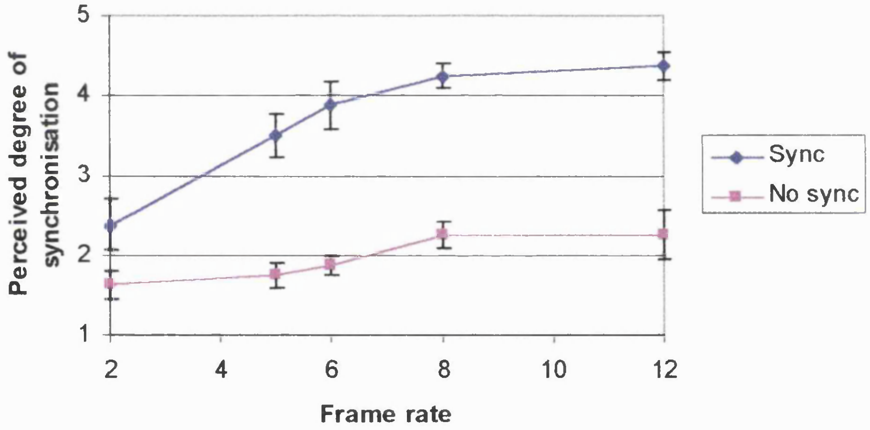


Figure 19: Perceived degree of audio-video synchronisation, where 1 represents 'not synchronised' and 5 'synchronised'

5.7.5 Discussion

The results support the findings of Frowein et al. (1991), who found a significant difference for speech reception between 5 and 6 frames per second, and suggest that perceptual benefits could exist were multicast audio and video to be synchronised as a matter of course. Unfortunately, there is a serious drawback to the implementation of the synchronisation method in everyday MMC use. Most platforms that are connected to the Internet (e.g. UNIX workstations, Windows 95/98 PCs) have (different) time-sharing operating systems that do not adequately support real-time applications. It is the case that, as the load on a workstation increases, the event timings become less accurate, which has serious implications for the wide implementation of the synchronisation mechanism. For the foreseeable future, synchronisation between the audio and video streams in multicast multiway conferences will be less than perfect (see Kouvelas et al., 1996 for a fuller discussion).

The 5-point scale worked effectively within the confines of the study, although a tendency to indicate between the options on a scale was observed for a few of the subjects.

The test task was adequate for the demands of the study, but its artificiality makes it hard to generalise to 'normal' speech and video in a genuine interactive conferencing environment.[32] This problem is addressed in section 6.2, by focusing on the *adequacy* of the video for the task at hand.

5.8 Interim summary

In Chapter 4, it was contended that the vocabulary on the 5-point rating scale would not be suitable in assessments of MMC speech quality, since the magnitude and type of degradations likely to be experienced would be likely to restrict the effective range of the scale. Some support for this supposition has been illustrated by the speech quality studies thus far. It was seen in section 5.2.4.1, for example, that the excellent response was only chosen once in the study, and in the interactive setting presented in section 5.5.4, the range of the scale that was used was depressed. Chapter 4 presented another reason as to why the listening quality scale is not a valid assessment method: the labels on the scale do not represent equal category intervals. However, the research that had established this fact has not, to the knowledge of the author, been repeated with British English speakers. The final section of this chapter details a graphic scaling study carried out using British English speakers, which supports the findings of graphic scaling studies carried out with other nationalities.

5.9 Study 6: Graphic scaling of quality terms

The literature review had revealed that no graphical scaling study involving the ITU terms had been carried out with British English speakers, and it was therefore of interest to repeat this study with speakers from this make-up. Although it is unlikely that there would be major differences between American English and British English speakers, this should not be taken for granted.

5.9.1 Material

In addition to the ITU quality terms, the placement of *different* terms than those used by previous researchers (Teunissen, 1996; Narita, 1993; Jones and McManus, 1986), as well as the ITU terms, was investigated. The reason for this is that it is not the quality of the audio or video delivered in MMC *per se* that is of ultimate interest, but rather whether that quality is ***good enough*** in order for the task at hand to be completed satisfactorily. As Virtanen et al. (1995) identified, qualitative terms can fall into various semantic categories (section 4.11), and the 'approval' terms may prove a more useful scaling terminology for MMC quality than other communication technologies.

Twenty-one terms were selected for presentation. In addition to the 5 ITU quality terms, 5 terms were selected from previous research (fine, ok, passable, marginal and very bad). Many high quality terms (e.g. ideal, superior) that had been used in previous studies were omitted in favour of terms that addressed the concept of acceptability or 'good enough', since the ultimate goal is to develop a MMC rating scale that allowed people to indicate clearly at which point quality becomes unacceptable for the task at hand. The 11 novel terms that were investigated were: satisfactory, unsatisfactory, acceptable, unacceptable, adequate, inadequate, tolerable, intolerable, sufficient, good enough and very good. Teunissen (1996) suggested that people may prefer a scale to be symmetrical, hence the majority of these options were antonyms. Very good was included to balance very bad.

5.9.2 Subjects

Twenty-six subjects participated in the study. They were aged between 22 and 28, and were all British English speakers.

5.9.3 Procedure

The subjects were presented with all 21 terms at once, and asked to place them in order on a line 200 mm in length bounded by the labels 'Best Imaginable' at the top and 'Worst Imaginable' at the bottom (after Jones and McManus, 1986; Teunissen, 1996).

5.9.4 Results

The position of each term on the graphic scale was calculated by measuring the distance, in millimetres, from the bottom of the scale. The mean position for each term was calculated and is presented with the standard deviations in Table 6. The ITU term results are presented in bold italic font. Figure 20 shows the mean positions as they appeared on the 200 mm line.

[32] However, the synchronisation method was employed in a later ReLaTe session, to positive approval from

5.9.5 Discussion

It is clear from the results that, as has been found in studies in other languages and American English, the ITU terms do not divide the scale up into equal intervals. These findings reveal that the perceptual distance between poor and bad is not great, likewise excellent and good, and also that fair does not represent a clear midpoint. There is a difference between the results attained from the American English speakers reported by Jones and McManus (1986) and the British English speakers in this study. While Jones and McManus found that the terms excellent, good and fair were equally spaced in the top half of the scale, the results here position all of these terms in the top *third* of the scale (see Figure 20). However, with the large standard deviations observed in this study, a larger number of subjects would be required to see whether this difference between American and British English is reliable (Jones and McManus tested 49 subjects, more than twice the number tested in this study).

Term	Mean rating (SD)
Excellent	***197.7 (3)***
Very good	186.8 (7.5)
Good	***167.7 (14.1)***
Fine	156.3 (19.9)
Fair	***137.8 (17)***
Satisfactory	135.7 (23.9)
Good enough	133.9 (27.5)
Acceptable	129.7 (24.8)
Adequate	125.3 (21.4)
OK	123 (20.1)
Sufficient	122 (21.2)
Passable	107.9 (23.8)
Tolerable	104.8 (22.3)
Marginal	89.2 (26.6)
Poor	***54.6 (25.3)***
Inadequate	54.3 (24)
Unsatisfactory	47.9 (26)
Bad	***41.2 (25.4)***
Unacceptable	21.5 (20.7)
Very bad	16.7 (12.7)
Intolerable	10.2 (12.6)

Table 15: Means and standard deviations for position of terms on graphic scale

the participants.

The positioning of the terms that addressed the issue of acceptability revealed that some terms were almost synonymous e.g. acceptable, adequate, ok and sufficient. Passable and tolerable were closest to the mid-point of the scale, with marginal hovering just beneath the mid-point. At the lower end of the scale poor and inadequate seem synonymous, with relatively little difference between them and unsatisfactory and bad. The lowest ratings were awarded to unacceptable, very bad and intolerable. There is support for Teunissen's comment that very good and very bad might produce a more symmetrical scale, but there is little evidence that the other antonymic pairings reflect evenly across a mid-point of the scale.

If the 200 mm line is divided up into five equal intervals in the manner of the continuous quality scale, and the terms that are closest to the mid-points of those intervals are extracted, what would arise is very different from the ITU quality scale. The five labels of the new scale would be very good, fair, tolerable, poor, and unacceptable.

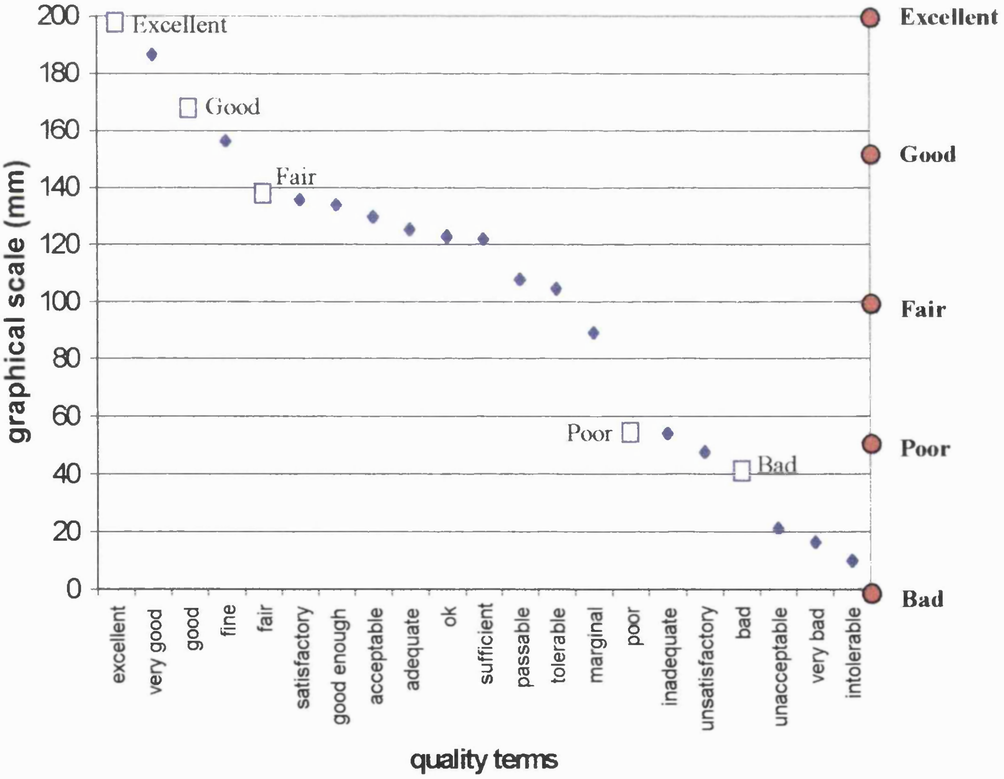


Figure 20: Mean positions for quality terms placed on 200 mm line. ITU terms are indicated by an unfilled square. The right-hand axis shows the theoretical positions of the ITU terms on the 5-point scale.

The key conclusion to be drawn from this study is *that the ITU scale is invalid and should not be used as an interval category scale.* That it continues to be used all over the world by telecommunications companies, in face of evidence that it is not a reliable method, is alarming at best. Better, more valid methods need to be developed and applied *post haste.*

5.10 Chapter conclusions

The results of the research presented so far can be divided into substantive and methodological findings.

5.10.1 Substantive results

- The ReLaTe field trial demonstrated that MMC technology can be used to support distance language learning, but highlighted the necessity of good audio quality. It was found that measuring the objective (network) conditions in the field is extremely difficult.
- It is clear from both the speech intelligibility and speech quality studies that perception benefits from a packet loss repair method that substitutes something other than silence in the place of the missing packet. Although the audio tool RAT offers different forms of packet repair, the other most common MBone audio tool, **vat** (Jacobson, 1992), offers no such facility. Designers need to be made aware of the large perceptual benefits that can be gained at relatively low cost.
- There are subjective benefits to having a video channel, and if this video channel can be synchronised with the audio, the benefits to users may be found at relatively low frame rates. (However, this would have to be qualified with research into more meaningful tasks than the one presented in Study 5.)

5.10.2 Methodological results

- Intelligibility testing should be carried out with sentence rather than word material, in order to increase the generalisability of results to the real world, but assessing speech intelligibility is not in itself sufficient: speech quality must also be assessed.
- The ITU-recommended quality scale is not suitable for use in MMC testing for the following reasons:
  - The labels (excellent, good, fair, poor, bad) do not lend themselves well to describing the conditions of MMC degradations.
  - Audio and visual quality can fluctuate drastically throughout the duration of a MMC session. The scale affords only a post hoc 'summary vote' of perceived quality.
  - The scale assumes that quality is a single dimension, whereas in fact perceived quality can be affected by many different variables.

- Subjects do not use the full range of the scale, and tend to indicate between options (see below).
- And most importantly, the labels *do not represent equal intervals.*

Based on these findings, and an analysis of the literature (section 4.10), it is argued that the labels on the scale should be *removed*, and the ability of subjects to rate on a polar scale should be explored. If quality is not unidimensional, then no one set of labels are likely to capture what it is that renders the quality poor to a user. It would be better to explore the use of a polar scale, and if this scale can be used consistently by users, then asking them to describe *why* a rating was given will reveal more about user perception than a qualitative term will. Further support for exploring a polar scale comes from the fact that, in all of the studies discussed so far, a tendency for subjects to indicate *between* options on a scale was witnessed, even when they were instructed not to do this. Some claimed dissatisfaction with having to circle only one quality rating, since even in the controlled studies the speech quality of the test file could range quite significantly (although the overall loss rate in these files was kept constant, the loss *pattern* within the file was random). A polar scale will allow this tendency to be explored further, by allowing subjects to rate in whatever manner they wish, by providing them with a rating scale that has no formal divisions. (Although the ITU-recommended continuous quality scale (DSCQS) permits the user to score in-between intervals, the intervals are labelled with the vocabulary that has already been argued to be inappropriate for MMC.)

The development and use of a new scale that had no labels on it at all, a polar continuous quality scale, is the focus of the next chapter, Chapter 6.

Chapter 6: Investigating a polar continuous quality scale

This chapter presents studies exploring the use of a novel, polar continuous scale. The reason for developing a scale such as this was to investigate whether a polar scale can be used as well as/instead of the ITU quality scale. If the answer is yes, then many opportunities for investigating and establishing the multidimensionality of perceived quality arise.

6.1 Study 7: Investigating a new quality scale

A radical approach was taken to eliminating the drawbacks (namely the vocabulary and the non-interval nature) of the 5-point quality scale by investigating whether a *polar continuous scale* might allow listeners to indicate their perceived quality rating in a less restrictive, and hopefully more indicative, manner than the traditional 5-point scale. Effectively, a visual analogue scale was created, where only the end-points had labels, and these labels took the form of '+' and '-' symbols.

The primary aim of the study was to demonstrate the workability of a polar scale, by comparing perceived quality of passages of speech repaired with either packet repetition or LPC redundancy, using both the 5-point quality scale, and the novel polar scale. If the polar scale can be used with ease by subjects and produces consistent results, then it can be claimed to be a reliable assessment method.

Results from this study, together with the studies reported in sections 5.3 and 5.5, were published in Watson and Sasse (1997).

6.1.1 Material

The polar scale is an example of non-categorical continuous scaling (ITU-R BT. 500-8). It is a vertical line 20cm long, with no labels other than a '+' sign at the top and a '-' at the bottom to indicate the polarity of the scale (see Figure 21). In addition to this scale, the 5-point quality scale was also used.

Eleven 30-second long passages were extracted from a newspaper and recorded by a male speaker with received pronunciation using a Sennheiser headset. As in previous studies, 6 levels of loss were generated on the files (0, 10, 20, 30, 40 and 50%) which were then repaired with either packet repetition or LPC redundancy.

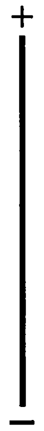


Figure 21: The polar continuous scale (not to scale)

6.1.2 Subjects

Twenty-four native English speakers participated in the study. Their ages ranged from 20-45 and their hearing was normal. None had previous experience of MMC.

6.1.3 Procedure

The design was counterbalanced, such that half of the subjects rated the speech samples on the 5-point scale in the first half of the experiment, and in the second half of the experiment they rated on the polar scale, and the other half of the subjects completed this in reverse order.

When using the polar scale, the subjects were instructed to place a cross on the line at the point they felt represented the quality of the speech to which they were listening. In order to anchor the scale as far as possible the subjects were told both orally and on the rating form that the '+' limit of the scale should be taken to represent the best quality they could imagine, and conversely the '-' limit represented the worst quality they could imagine.

Clearly one of the biggest concerns in using the polar scale must be whether users are consistent in their use of it, or whether its lack of guidance as to where a rating should be put means that users indicate their rating in a haphazard and unreliable manner. In order to address this concern, the 24 subjects in the study rated each sample twice on the 5-point scale and twice on the polar scale. The order of the conditions was randomised to prevent order effects.

6.1.4 Results

Scoring of the 5-point scale gave rise to the Mean Opinion Scores (MOS). The polar scale results were translated into a 100-point % scale. Figure 22 shows the mean results and standard error for ratings on both the polar scale and 5-point scale, divided in the order in which the scale was experienced, and showing the difference in ratings between the first and second times the passage was heard.

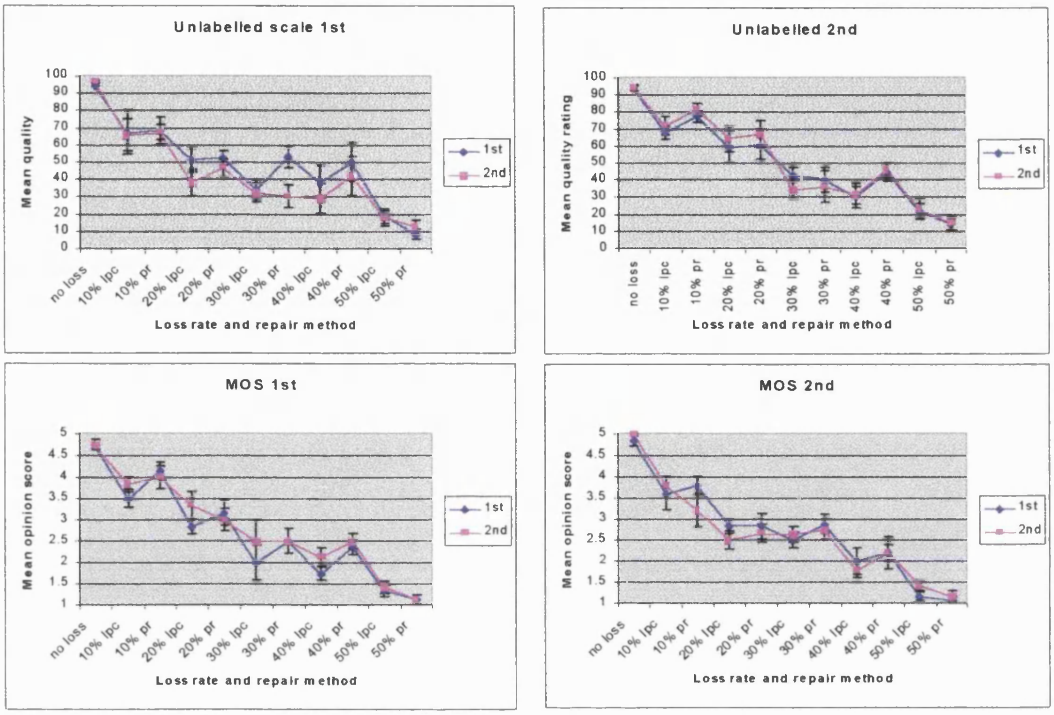


Figure 22: Mean ratings awarded for speech passages with loss repaired with packet repetition (PR) or LPC redundancy. The top two graphs show the consistency of rating on the polar scale by those who experienced the scale before the 5-point scale (on the left), and those who experienced it after (on the right). The bottom two graphs show the consistency of rating using the 5-point scale.

Single factor analyses of variance were carried out on the polar scale results for each condition. Firstly, it was established that there are no significant differences between those who experienced the polar scale first and those who experienced the 5-point scale first, by comparing the scores awarded the first time the stimulus was encountered. These results are presented in Appendix G. Since there are no significant differences between those who rated on the 5-point scale first and those who rated on the polar scale, the polar scale results were combined in order to investigate whether there is consistency between subjects' first and second rating on the scale. A Pearson correlation between first and second rating on the polar scale was calculated to be 0.936, $p < 0.01$, indicating that the subjects use the scale consistently and reliably.

It is clear that the two scales follow the same trend. For example, the polar scale results replicate the result that at 40% loss PR is preferred (for MOS, F(1,50) = 7.4 at the 1% level of significance; for the polar scale, F(1,50) = 7.56 at the 5% level of significance), whereas at 50% loss LPC is favoured (for MOS, F(1,94) = 6.15 at the 5% level; for the polar scale, F(1,94) = 7.14 at the 1% level). This is encouraging in the case of the polar scale, since the subjects set their own rating criteria.

Use of the polar scale also reduced the tendency of subjects to avoid the end points of the scale. The polar scale was popular with many of the subjects, who commented that they preferred it, since it allowed them more flexibility in making quality judgements.

The conclusion of the study is that subjects vote consistently on the polar scale, and there is little evidence of shaping by the 5-point scale.

6.1.5 Discussion

The main result from the study is that the feasibility of using a polar scale to indicate perceived quality has been demonstrated. This is an important methodological finding since it means that the scale could be used to assess specific *dimensions* of quality without having to label different degrees of the dimension on the rating scale. (For example, the scale is used in section 6.2 to assess the perceived *adequacy* of the delivered quality.)

The results from the study further illustrate the complexity of the relationship between speech intelligibility (section 5.3) and perceived quality (section 5.4). The results demonstrate that it is not safe to assume that more intelligible speech will receive a higher quality rating than less intelligible speech. Results have shown that LPC redundancy produces greater intelligibility than packet repetition at lower loss rates, yet in both passive listening (this section) and interactive conversations (section 5.5), packet repetition is awarded at least as good as, or a higher quality rating. Further research needs to be carried out to corroborate and extend these findings across different speech coding schemes and packet loss repair methods.

The polar scale has been shown to be as consistent as the 5-point scale in terms of users' ratings. This finding is extremely encouraging, firstly because it means that it is not necessary to use labels such as the ITU terms on such a scale, and secondly because a continuous scale such as this can be applied in many different settings.

Since the polar scale has proved to be a reliable assessment method for measuring the subjective quality of speech, might the technique also be employable in investigating the perceived quality of video? As discussed in section 4.5.2, the ITU-recommended scales for video quality assessment are not considered to be useful in terms of measuring low frame rate Mbone video quality. What is required is a method by which the *adequacy* of the video quality can be assessed for a certain task. Overall adequacy might be reflected well via a polar scale.

6.2 Study 8: Distance learning course on networks and communications

An opportunity to measure perceived video quality in a real-world environment arose when an eight-week long course on networks was conducted between UCL and the University of Westminster between October and December 1997. Thirty final-year undergraduate students were involved at UCL, and one tutor at Westminster. The tutor was experienced in MMC and had previously taught distance learning courses using the technology. Since the tutor was interested in the impact of video image size and frame rate on student learning behaviour (in terms of dialogue content), the opportunity was available to measure the subjective impact of changes in video quality across a period of time in a real world setting. The link between UCL and Westminster was of high bandwidth, which permitted the assumption that packet loss would be negligible between the two sites, allowing a reasonable degree of 'control' over the objective network conditions (in contrast to the situation of the ReLaTe project reported in section 5.1).

6.2.1 Assessment plan

In keeping with the objectives of the course tutor, it was agreed that the objective video quality would be improved halfway through the course, i.e. after 4 weeks of a lower quality. It was determined that for the first half of the course the video would be run at 2 frames per second (fps) using QCIF image sizes, and in the second half of the course the frame rate would be increased to 8 fps and the image to CIF sized. The resolution quality was maintained at 10 for the duration of the course. In order to cope with the increased frame rate in the latter half of the course, the video bandwidth was increased from 128 kbit/s (the default setting in **vic**) to 256 kbit/s.

The 30 students were divided into 6 groups of 5. Each group received a one-hour tutorial every week for a total of 8 weeks. Each student in a group was assigned to a particular workstation in the Computer Science department at UCL for the purpose of participating in the tutorials.

Two different techniques were employed in assessing perceived media quality: rating scales and group discussions.

6.2.2 Rating scales

The polar continuous quality scale was used to assess the video channel. Instead of asking the students to rate the *quality* of the video channel *per se*, they were asked to rate the *overall adequacy* of the video pictures for the purpose of the tutorial, where the bottom of the scale represented 'completely inadequate', and the top represented 'completely adequate'. The ratings were scored as in the previous study (section 6.1), by translating the marks into a 100-point scale.

The scales were administered first 2 weeks into the course (i.e. the middle of the first video condition), and 6 weeks into the course (i.e. the middle of the second video condition).

6.2.3 Group discussions

An overall group discussion was held halfway through the course and addressed audio and video quality issues. During this discussion the students were asked to complete a short paper-based task which asked them to select terms that best described the audio and video quality that they had experienced in their tutorials. The results of this task are presented and discussed in section 8.3.2.

6.2.4 Video rating results

Thirteen students filled in rating scales on both occasions, permitting a comparison between the ratings given under the 'old' and 'new' video qualities. Mixed results were obtained from the data. From the figures it does not seem that there was a consistent trend within the subjects: 4 of the students gave greatly improved ratings, 3 gave markedly lower scores, and 6 gave roughly the same rating both times (between 1 and 5% difference only). A Wilcoxon large-sample test confirmed that there was no effect of video quality ($z = 1.08$).

The result that some of the subjects rated the video the same or worse in the latter half of the course than they had in the first half of the course was somewhat unexpected. The explanation for this may lie in the fact that increasing the image size made packet losses and the lack of synchronisation more apparent to the students. It is also the case that some of the students found it hard to position all the CIF-sized video images on the screen such that they did not obscure the whiteboard or **RAT**. This screen real estate problem has been mentioned previously in sections 2.8 and 5.1.1. Some support for this interpretation was found in the results to a questionnaire administered by the course tutor. Of the 19 anonymous responses, 6 students said that they preferred the QCIF-sized images. The reasons given were that it required less juggling of screen space ($n = 5$) or was 'less intense' (n

= 1). However, 12 students preferred the CIF-sized images since they felt that this enabled them to communicate better since they could see more facial reactions (Hearnshaw, 1999).

Rating 1	Rating 2	Type of change
0	56.5	Increase ↑
44	84	Increase ↑
5	33	Increase ↑
63.5	97	Increase ↑
79	35	Decrease ↓
90	75	Decrease ↓
85	35	Decrease ↓
83	84	Static —
57	54	Static —
91	92	Static —
80.5	83.5	Static —
77	80	Static —
87	82	Static —

Table 16: Video rating results for weeks 2 and 6, where 0 represents a rating of 'completely inadequate' and 100 represents a rating of 'completely adequate'.

It is also possible that no effect of the video quality improvement was found because the improvement was not great enough. That is, 8 fps may still have been too low for image improvements to be seen as beneficial to overall adequacy. It is also true to say that the video channel was not critical in this course since the students already knew each other well, and the focus of the tutorials was the whiteboard where the students collaborated on providing answers to course material questions. It is likely that had the study been looking at tasks where shared video data was key (as in Whittaker's 1995 "video-as-data" model) or where affective information was critical, such as remote counselling, the effects of improved video quality would have been more tangible.

6.2.5 Issues arising from the general discussion

After the first four weeks of the course, the students were invited to participate in a discussion and feedback session about the way the course was proceeding so far. Twenty-seven students attended. The discussion topics dealt with the usability of the desktop learning environment in general, covering the ease of use of the whiteboard, audio and video tools. Some of the points that emerged were useful and revealing. For example, it became clear that the students were only willing to 'play around' with the volume controls at the very beginning of a session, and any late arrivals were not appreciated in so far as this required either another adjustment of volumes or degraded quality from that person for the rest of that session[33]. This is a good argument for effective *automatic gain control* to be incorporated into **RAT**, such that users do not have to constantly readjust their incoming volume according to whoever is speaking. Related to the volume issue, some students commented that the silence suppression mechanism was sometimes too efficient, cutting out a person as his/her speech naturally tailed off at the end of a sentence, leaving the listeners unsure as to whether the person had really finished what he/she had to say.

Although the students appreciated that headsets were necessary, some reported that having to wear them was oppressive and made them feel cut off from the outside world. Others complained that people accidentally moving their microphones, or deliberately moving them in order to cough or sneeze, again resulted in problems with the audio. Although they were aware that using the push-to-talk function would negate the need to move the microphone, this function was not popular as it was felt that it reduced the naturalness of the communication.

With respect to the video channel (which, at the time of this discussion had only been at 2 fps and QCIF-sized), the majority of the students agreed that they would not want to do without it (only two said they would rather not have the video channel), but that it was the whiteboard that was the focus of their visual attention. They claimed that they did not use the video images a great deal, but that the video channel:

- looked nice;
- made them feel 'more connected' to each other;
- allowed them to reassure themselves that the others were 'still there' if the audio got bad;
- told them when another person had arrived in a tutorial.

It also emerged that the students would have felt more comfortable with an integrated user interface such that they did not have to position the tool windows themselves. A related issue was the way in which many names, not just the current speaker, would light up in the **RAT** window. It was explained that this was a function of the silence suppression mechanism not working effectively, but the students argued that a more sensible means of indicating the present speaker would be to have the video image of a speaker changing in some way to indicate that he/she was now talking. (This is indeed a design feature that was incorporated into the ReLaTe integrated interface, but due

[33] Volume issues such as these were investigated explicitly in a study reported in section 8.2.

to the larger group size and video manipulations that were necessary in this course, the integrated interface could not be used.)

6.2.6 Summary and critique

The video rating scale data showed that there was no perceived change in overall adequacy of the video channel for the purpose of the tutorials once the quality had been improved from 2 to 8 fps and the video image size enlarged.

The students were invited to leave additional comments alongside the rating scales, and what emerged from this is that they seemed to like the bigger, faster video channel, but in terms of it looking better, rather than being more useful. The comments seemed to reflect an appreciation factor, rather than a 'usefulness' one, an observation that seems to be borne out by the fact that no difference in 'overall adequacy' was found with the improved video quality.

Whether due to bursts of packet loss or sudden drops in frame rate (or even changes in audio quality), it was observed that some of the students changed their minds when rating the video adequacy. Although they were instructed to wait until the end of the lesson in order to score it, some subjects clearly did not, placing one mark, scoring it out and then substituting another. It can be hypothesised that the fluctuation of the quality made them revise their opinions as time went on. This behaviour is extremely interesting and suggests that some kind of dynamic rating tool would be very useful in terms of trying to understand at what point in a video stream quality degradations become objectionable and affect perceived adequacy (see section 7.4).

6.3 Chapter conclusions

The studies in this chapter contribute further to the substantive and methodological findings listed in section 5.10.

6.3.1 Substantive results

- The relationship between speech intelligibility and speech quality should not be taken as straightforward - greater speech intelligibility was gained with LPC redundancy as a speech repair mechanism (section 5.3), but in terms of perceived speech quality, speech repaired with packet repetition achieves at least as good quality ratings (section 6.1).

- In an interactive learning environment, where the video link is not critical to the task being performed, varying the image size or frame rate does not appear to have either an adverse or beneficial effect, for the task and the levels investigated.

6.3.2 Methodological results

- The polar quality scale produces reliable and consistent subjective quality results (when compared to the traditional 5-point quality scale) and therefore is a promising tool for the investigation of different quality dimensions (see next point).
- The polar scale can clearly be adapted for different purposes e.g. *overall adequacy,* and can be used for both audio and video quality rating.

This chapter has presented studies utilising a new polar scale which avoids the problems associated with the 5-point quality scale. The scale can be adapted for use with different subjective dimensions, such as perceived adequacy. However, a key problem remains: that of interpreting post hoc quality ratings. It has been observed that even within a short period of time, subjects can change their minds as to how adequate the delivered media quality is. What is required is a means of measuring perceived quality *as and when* conditions fluctuate. Building on the concept of the polar scale, Chapter 7 presents a method for gathering dynamic quality ratings, and then takes the approach a crucial step forward by handing *control* of the quality to the user.

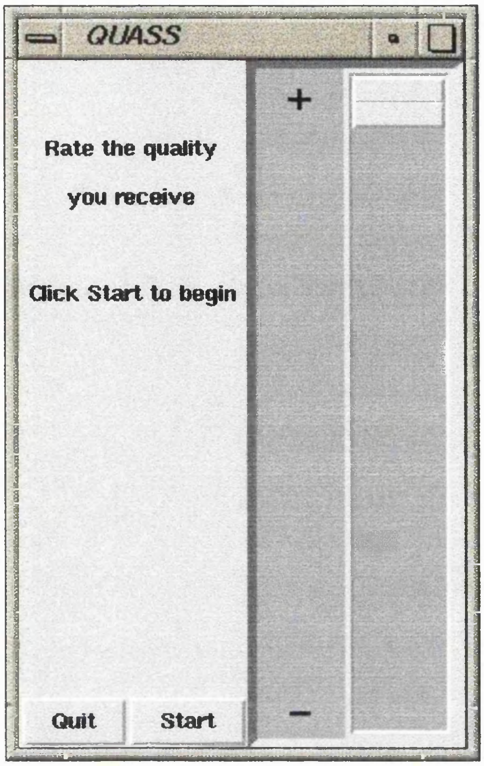
Chapter 7: Investigating real-time continuous quality rating

The previous chapter presented and discussed research exploring the use of a novel polar continuous scale, with encouraging results. But the inability to measure the effects of time-varying quality, a key characteristic of MMC, is a problem that remained. This chapter therefore presents the logical next step, a new method for assessing the effects of time-varying quality, as and when it occurs. An experiment using a specially designed software tool, QUASS (Quality ASsessment Slider) is described, where it is shown that users are able to monitor time-varying changes in audio quality. However, the results do not allow conclusions to be drawn about the quality that is *required* by participants in *interactive* conferencing environments. Therefore the final part of the chapter documents the findings from a study in which QUASS was modified such that movement of the slider allowed a participant to *control* the audio quality that he/she received from another participant. Results from this interactive, contextual environment have more real-world relevance than those from previous studies. The chapter ends with a discussion of the pros and cons of the new method.

7.1 Study 9: Investigating perceived quality using QUASS

On the basis of the studies reported in the previous chapter demonstrating the reliability of the polar scale, a software tool was implemented which allows subjects to move a virtual slider bar (controlled by the mouse) up and down a polar continuous scale (Bouch et al., 1998). The position of the slider bar is translated onto a 0-100 scale, and the figure is recorded in a results file every second, which allows a direct mapping of objective quality with perceived quality ratings. The interface to the tool, known as QUASS (QUality ASsessment Slider), is shown in Figure 23.

Although QUASS is similar to the SSCQE method presented in section 4.8, it differs in two important respects. Firstly, the tool was developed as a logical extension of the polar continuous scale, and therefore the scale is not divided up into quality 'regions', but bounded at the upper and lower limits by a '+' and '-'. The SSCQE method, on the other hand, advocates the use of the ITU quality scale terms. Secondly, the tool is a software rather than hardware tool. As such the tool can be incorporated into any desktop conferencing environment with minimal disruption. Manipulation of the slider is via the mouse, so no additional equipment is required by the end user.



Rate the quality

you receive

Click Start to begin

Figure 23: The QUASS interface

7.1.1 Study overview

An experiment was designed to examine subject behaviour when using QUASS to rate perceived audio quality. The object of the experiment was to determine whether subjects are able to perceive and rate quality changes in audio continuously. Although Kotokopolous (1997) had shown that subjects were able to detect and monitor changes in high quality audio compression schemes, the study reported here investigated comparatively low quality audio in which the changes in quality were due to packet loss. In addition, the audio was played out through headsets, minimising additional cues. All previous studies using the SSCQE methodology have played sound out through speakers, thereby maximising the perceptual cues for the listener.

7.1.2 Material

Four two-minute extracts were taken from a talking book on CD[34]. The extracts of speech had loss imposed on them as in previous studies (see section 5.2), using three different loss patterns. The packet loss repair method was silence substitution (since the object of the study was to investigate rating behaviour, it was necessary to ensure that the objective deficits were noticeable). Different

[34] "The Hitchhikers' Guide to the Galaxy" by Douglas Adams, 1979.

loss patterns for the four files were generated to avoid habituation effects and any influence of quality expectancies[35]. Although the patterns varied, the time period for each loss level was constant (17 seconds). This relatively long interval was to give the subjects a sufficiently long period in which to adjust to a level of loss before it changed. Within each loss level, the pattern of the loss was random. Two of the test files followed a loss pattern in which the loss increased and then decreased in regular intervals (loss pattern 1), and the other two followed less regular loss patterns (loss patterns 2 and 3). The three loss patterns are shown in Table 17.

Seconds	Loss pattern 1 (% loss)	Loss pattern 2 (% loss)	Loss pattern 3 (% loss)
0-16	0	0	10
17-33	5	5	15
34-50	10	0	10
51-67	15	10	20
68-84	10	15	25
85-101	5	10	20
102 - end	0	20	30

Table 17: Packet loss patterns imposed on the speech files

7.1.3 Subjects

Twenty-four subjects participated in the study. They were all native English speakers, aged between 21 and 28, with normal hearing. None of them had experience in MMC.

7.1.4 Procedure

A within-subjects design was followed. The subjects rated the quality of the four passages continuously using QUASS. In a pilot trial, the starting position for the slider was in the middle of the interface, but it was observed that the subjects were inclined never to raise the bar above this midpoint when listening to the speech files, despite the fact that the first part of the stimulus was a no loss speech burst. It was therefore decided to place the slider at the top, at the start of the study proper, since three of the four test files began with no loss.

The subjects listened to the first file twice. This first playout served as an anchor condition such that the subjects understood the range of qualities they would experience in the experiment, and so they were not required to rate the quality. During the second playout of the file they used QUASS to register their opinions of the quality. The other three passages were played only once, and

[35] See Bouch and Sasse (2000), and also footnote 15, section 4.4.2.

quality rating was performed simultaneously. All subjects received the same order of conditions to ensure that everyone had experienced exactly the same levels of loss when they made their quality ratings. The order of conditions was therefore:

- Condition 1: loss pattern 1
- Condition 2: loss pattern 2
- Condition 3: loss pattern 1
- Condition 4: loss pattern 3

7.1.5 Results

The position of the slider was recorded every second for each subject, and the average position across subjects was calculated for every second. The results for the regular loss pattern (loss pattern 1) are plotted in Figure 24, and those for the irregular loss patterns are shown in Figure 25 and Figure 26. The objective loss for the conditions is shown as the lower curve on these graphs.

The results clearly show that listeners are able to rate speech quality continuously. The graphs in Figure 24 illustrate this particularly clearly in that the slider position curve reflects that of the objective loss rate. The results from condition 1 are especially smooth, which is undoubtedly due to the fact that the subjects had heard this speech file once already before rating, but the condition 3 results also illustrate clearly that the quality changes are registered and have an impact on the perceived quality. The results for conditions 2 and 4, shown in Figure 25 and Figure 26 respectively, also reflect the objective loss pattern, although the effect is harder to determine for Condition 4, where the objective quality started off already degraded and became far worse. Pearson correlations were calculated between each subject's rating and the objective loss levels for each condition (see appendix H). In each condition strong negative correlations between the loss rate and the ratings were found, indicating that the subjects detected and responded in suitable manner to decrease that they were hearing.

The results from this condition suggest that once the packet loss exceeds 15% quality will be rated very negatively. However, the results for condition 2 register a relatively higher rating (around 30 on the scale) for 15 and 20% loss than the ratings found in condition 4 (around 20 on the scale). This finding may arise directly from the low quality that condition 4 started with (10% loss). Aldridge et al. (1998) noted that subjects tend to respond less quickly and with less magnitude to good quality than they respond to bad. In condition 4 it is therefore likely that once the slider has been lowered in response to the bad quality at the start, it would not be raised by a great degree again, meaning that the available rating range of the scale becomes compressed, and poor quality is consequently rated even lower.

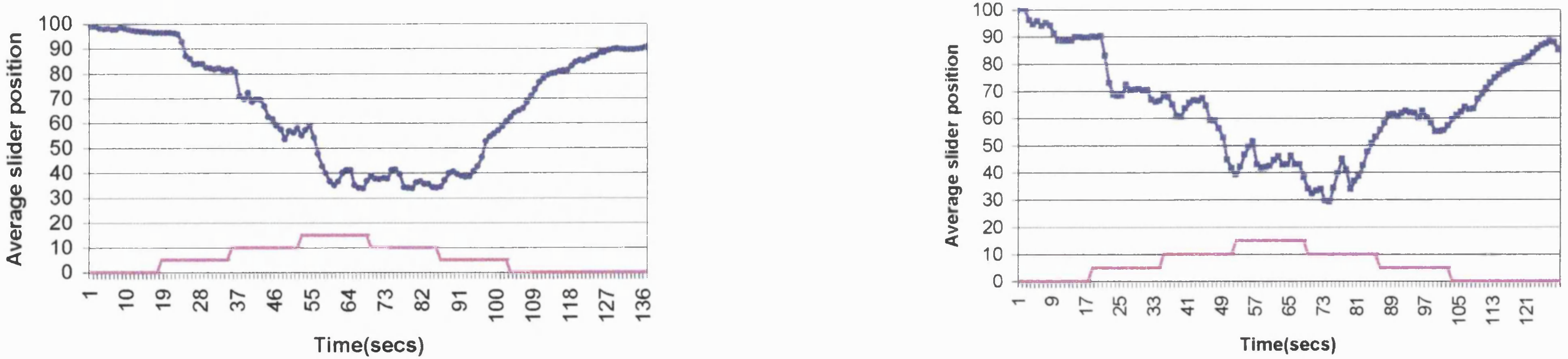


Figure 24: Mean slider position every second for conditions 1 (left graph) and 3 (right graph). Objective packet loss rate is shown by the lower curve on each graph.

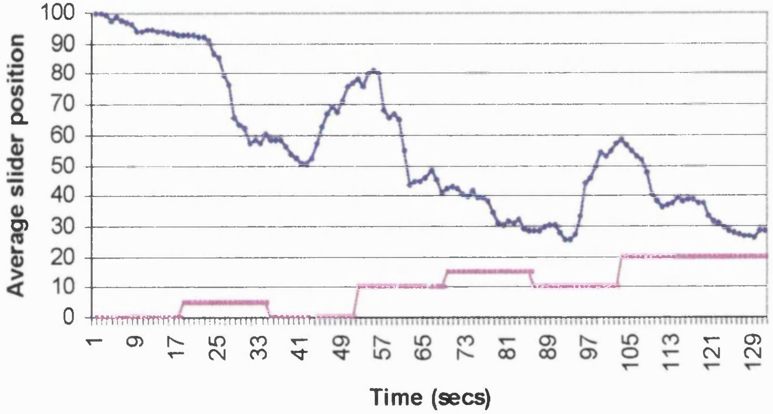


Figure 25: Mean slider position every second for condition 2.

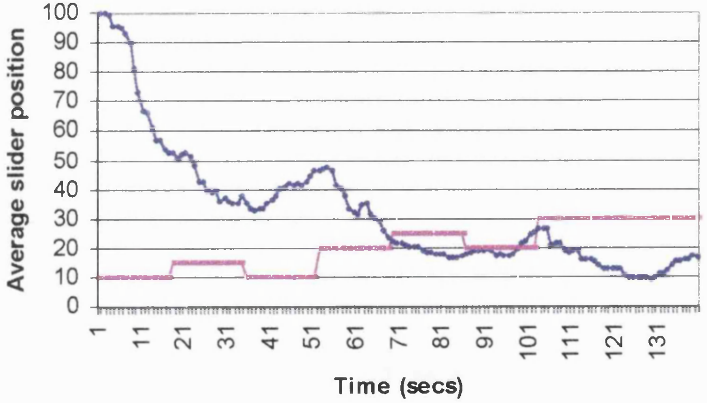


Figure 26: Mean slider position every second for condition 4.

The results show that there is a delay of roughly 3-4 seconds between the objective quality changes and their reflection in the quality ratings. This figure is slightly longer than that reported in studies using the SSCQE method (1-2 seconds). Two reasons can be put forward to explain this. Firstly, the nature of the impairment, packet loss, is different to that of bit errors that were the focus of the SSCQE studies. Secondly, random packet loss within the 17 second period means that a relatively steady level of loss may result, or a burst of severe loss: it may be that the listener is unsure whether the packet loss perceived is a severe but temporary blip, or an actual overall change in quality.

The standard deviations across the subject responses were quite large, revealing that individual rating behaviour varies quite significantly. Figure 27 shows the results for condition 2 with the standard deviations imposed. The standard deviations shown here are fairly representative of all the conditions. However, the correlation results show that the subjects were, in the vast majority of cases, rating in direct response to the objective loss conditions (see Appendix H).

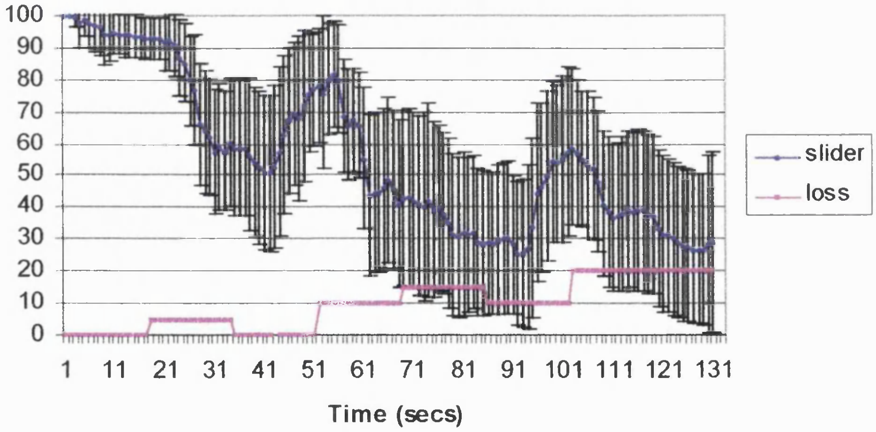


Figure 27: Standard deviations for condition 2

7.1.6 Discussion of rating study

The results of the study clearly illustrate that subjects are able to rate the quality of speech continuously, supporting the findings of Aldridge et al. (1998), but also that they can do this *without requiring quality labels* on the rating scale. This result is extremely encouraging because it shows that relative quality judgements from each individual can be extracted without the presupposition of QoS categories: it is no longer necessary to constrain subjects by imposing a quality criterion on them.

The study shows that when the packet loss rate reaches 15%, perceived quality ratings dip below the midpoint of the scale. It can be hypothesised that the midpoint of the scale represents the boundary between acceptable and unacceptable quality to a subject, and so it appears that 15% loss is the level at which sound quality is rejected by listeners. (This conclusion is supported by the next study - see section 7.4.5.2.)

There is a considerable degree of variation in the rating behaviour exhibited by the subjects, however, and it is likely that this is due to the non-critical nature of the task. Since it was not critical that the subjects understand the speech files, nor that they refrain from being as judgmental as they like, some subjects may have moved the slider lower than strictly reflected their opinion, or indeed left it higher.

The next section discusses what happens when QUASS is used in a real-world application, where the task is critical and interactive.

7.2 Using QUASS in real world conferences

The results from the first QUASS study demonstrated that participants listening in a MMC environment are able to rate sound quality continuously as they listen, but would they be able to do this in an *interactive* task setting? For example, it can be inferred from the results of the previous study (Study 9) that listeners to a lecture delivered across the Mbone would be able to register their approval or disapproval of the sound quality, but would they be able to do the same in an interactive MMC application where their attention is being divided between a number of sub-tasks? To explore this question further a version of QUASS was incorporated into the desktop conferencing interface for a series of real world conferences under the auspices of the PIPVIC (Piloting IP-based VideoConferencing) (Sasse et al., 1998) and ReLaTe projects.

The basic QUASS functionality and method remained virtually the same: the participants in the conference were instructed to rate the perceived quality using the slider, and the slider position was recorded to a file, together with a time stamp. However, the slider position was only recorded when it was moved, since it was anticipated that conference participants would not move the slider as often as in the experimental set-up since they were engaging in another task. The objective loss statistics were gathered using a tool that gathers packet reception statistics from the different sources in a conference[36], such that a match could be made between objective and subjective conditions in the analysis stage. However, a number of problems were revealed.

Firstly, it was difficult to match the objective statistics with the timestamps of the subjective ratings, since the objective statistics were not gathered at equal intervals, and represent more of a 'sampling' of the loss statistics, averaged over several seconds. Fine-grained matching was therefore impossible. Even more important was the observation that the rating task is very intrusive for real-world conference participants. There is often little free space on a screen to incorporate the QUASS interface, and when people are engaged in a task such as working on shared workspace, they have little time, or a free hand, to rate the quality. Essentially, people are unable to attend to moving the slider to record their opinion of the quality when they are performing other tasks in a conference: the cognitive load is too great[37].

7.3 Interim summary

The studies reported in the thesis thus far have investigated how best to measure the perceptual effects of different qualities of audio and video in both lab settings and field trials. It has been argued that a continuous measurement tool is required to fit the special characteristics of MMC audio and video, and such a tool, QUASS, has been developed. QUASS has allowed an examination of the subjective effects of quality fluctuations to be conducted, but it is clear that continuous assessment can only be effective as a measurement tool in a task where the participant is not engaged in any competing activity, so that the participant's undivided attention can be given to moving the slider. QUASS is likely to be most effective as a quality measurement tool in passive MMC situations, such as listening to seminars and lectures over a network, watching recorded conferences, or in home entertainment scenarios, such as viewing movies delivered from servers (see Chapter 9). This approach, however, is unlikely to lead to hard and fast rules about what quality is actually *required* by users in order to carry out an interactive MMC task.

A method of establishing *when* quality is not good enough to accomplish the task at hand is needed. In theory, this could be accomplished via a means of handing *control* of the quality to the user, whilst allowing the experimenter to retain the ability to determine cause and effect through manipulating the objective quality. To this end, QUASS was extended to function in a second

[36] **Crtpdump**, written by Colin Perkins at UCL. The software is available online at http://www.cs.ucl.ac.uk/staff/C.Perkins/software/crtpdump/

mode, whereby movement of the slider controls the objective quality delivered. The rationale for this approach was that full attention could be paid to the task at hand, and only when the quality became so poor as to be intrusive to the accomplishment of the task, would the user have to direct attention to the slider.

The next section details a study where QUASS was used in this fashion.

7.4 Study 10: Controlling received quality[38] using QUASS

An experiment was designed whereby participants conversed across a network with another person (the co-experimenter). One-way packet loss between the co-experimenter and the subject was artificially generated, and the subject was able to control the degree of the loss he/she received by manipulating the QUASS tool.

There were two main aims to the experiment:

- To see whether people were able to perform a conversational task and control the quality simultaneously.
- To determine if there is a common limit to the loss that people will tolerate in an interactive task of this nature.

Two subsidiary aims came about through the addition of a video channel for half of the subjects, and a 'budget' on the QUASS interface that visibly decreased in response to quality increase in half of the conditions (see Figure 28):

- To determine whether the addition of a visual channel affects the audio quality required.
- To see whether the addition of a visible budget affects the audio quality required.

[37] A confounding element due to the performance of a task during rating was also reported by the ACTS TAPESTRIES project, in a videoconferencing study where subjects were actually communicating over a link, rather than simply watching a video and monitoring the quality (section 4.8.2).

[38] The quality dimension under investigation in this study was packet loss, but the methodology could be used to investigate other quality dimensions.

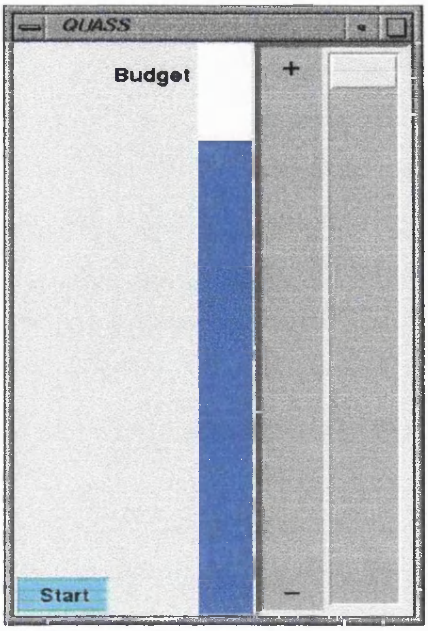


Figure 28: QUASS interface with budget

7.4.1 Material

7.4.1.1 QUASS configuration

For the experiment QUASS was configured such that the slider range corresponded to packet loss settings of 0-50%. The scale of the slider was set at 0-50 such that the position of the slider was inversely related to the level of loss. For example, a slider position of 40 set the packet loss at 10%. In the non-budget condition, QUASS was configured to automatically decrease its slider position and the corresponding quality received by the participant by 0.1% per second[39], since there would be little motivation for the subjects to decrease the quality themselves. No automatic decrease in quality was configured for the budget condition. In this condition, the budget was linked to the position of the slider such that the higher the slider position (i.e. the lower the loss), the faster the budget level would decrease, and the lower the slider position, the slower the budget level would decrease. This device was intended to give preliminary insights into pricing mechanisms and how they might relate to quality required by the end user (see Bouch and Sasse, 1999).

The position of the slider and the objective loss level was recorded every second for every subject, along with the amount of budget consumed (if applicable).

[39] This figure represented a trade-off between a just noticeable decrement and a larger decrement that might lead users to believe they had no control over the interface, and hence confounding the experiment.

7.4.1.2 Set-up

The set-up for the study is illustrated in Figure 29. **RAT** was started up on computer A and connected to computer B. On computer B, **RAT** was started up and connected to computer C, which hosted the reflector. Computer C forwarded audio packets from computer B to computer A. The user interface (QUASS) was on computer A and movement of the slider affected the behaviour of the reflector (on computer C). The subject sat at computer A, and the co-experimenter sat at computer B. Movement of the slider on computer A determined how the audio packets were forwarded from C to A i.e. when the slider was positioned at the top, this corresponded to a request for zero packet loss, and so the audio packets were forwarded from B to A with no additional loss. When the slider was lowered, the reflector forwarded audio with a set level of packet loss. Since speech from A to B did not pass through the reflector, the co-experimenter never received degraded audio from the subject.

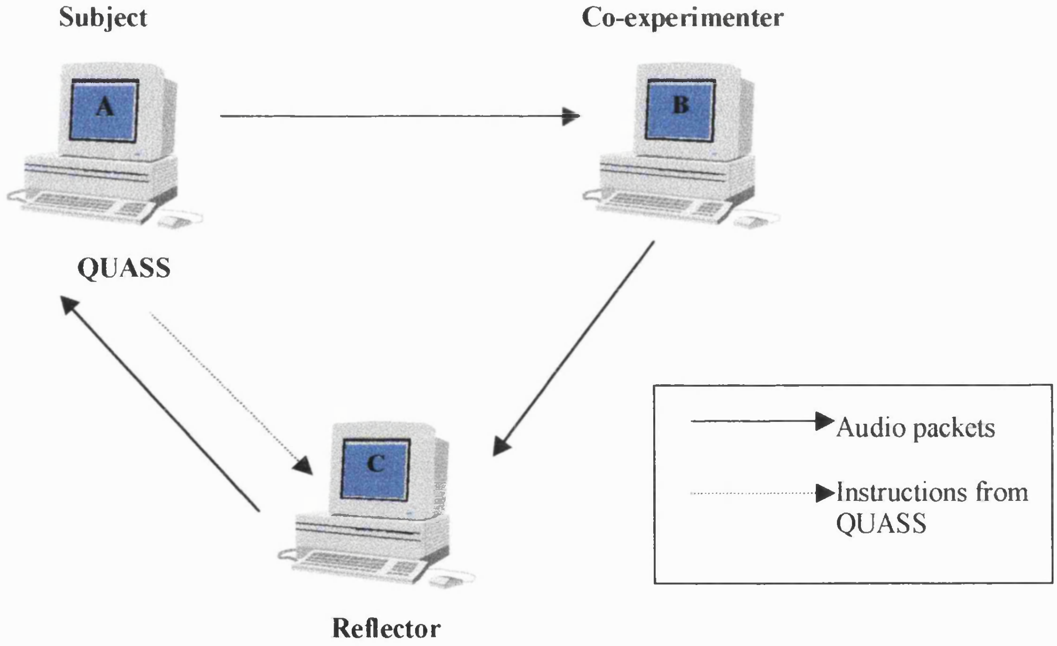


Figure 29: Experimental set-up

7.4.1.3 Task

The experimental task consisted of the subject and co-experimenter taking turns to define words using Taboo cards. This task was chosen as it involved conversational turn-taking, and was felt to be more interesting and fun than some other conversational tasks. The fact that it was relatively 'hands-free' was also important, since it meant that subjects were able to leave one hand on the mouse to manipulate the slider. The task had also been observed (see section 5.5) to be absorbing, which was seen as positive in that it would detract attention from the slider and its functionality.

7.4.2 Subjects

Twenty-six subjects participated in the study, all with normal hearing and aged between 21 and 30. None had previous experience of MMC. The subjects were divided into two groups, one of which communicated using the audio channel only (Group A), and the other who had a video channel available as well (Group V). (For these subjects the **vic** default settings were used, and the image was displayed at QCIF size.)

7.4.3 Procedure

All the subjects played Taboo in two 5-minute conditions: the first without the budget, and the second with the budget.

Following the conditions there was a brief questionnaire and semi-structured interview in which the subjects were asked to describe the quality they had experienced, and answer questions about the task and conditions in general.

7.4.3.1 Instructions to subjects

The subjects were informed that by moving the slider they could improve (or reduce) the quality. They were asked to move the slider to a position that was adequate to perform the task, and to only change the quality if they felt it was really necessary. The first condition incorporated the auto-decrement of the slider position, and so the subjects were told to be aware that if they left the slider still, it would start to drop and decrease the quality slightly over time. In the other condition, the budget interface was included, at which point the subjects were told that it would decrease in relation to the position of the slider, and that they should try to conserve as much of it as possible. However, in order not to prejudice behaviour in the first non-budget condition, the subjects were not told in advance that they would be participating in a condition with a budget.

7.4.4 Results

The key issues that were under investigation in this experiment were:

- Are people able to perform an interactive task and control received quality at the same time?
- Is there a loss level that can be identified as critical i.e. a level beyond which people find the quality intolerable?

- Does the presence of a video channel affect the audio quality demanded?
- How does a visible budget affect quality 'expenditure'?

As the following sections will reveal, some of these issues are easier to resolve than others.

Two types of data were collected: the slider positions, and comments from the semi-structured interview. The slider results form the major part of the analysis, and reference is made to the interview comments when relevant.

7.4.4.1 Analysing the slider data

Questions to be addressed included how should the movements of the slider be analysed, and is it possible to interpret these movements to mean that the subject wanted better quality? At what point do subjects increase the quality: when do they 'feel' (as evidenced by moving the slider) that an improvement in quality is required in order to accomplish the task effectively?

Separate results files for each subject in each condition were generated. The data from one subject was discarded, leaving 13 sets of data for Group A but only 12 for Group V.

Analysing the slider results is a complex undertaking. Averaging the overall results for the different participants could be misleading, since subjects' behaviour with respect to the slider varied quite widely: different participants started moving the slider at different points in time, and to different degrees. (The slider movements for each subject are plotted against time and represented graphically in Appendix I.) However, since it was important to see if there were any trends in subject behaviour, it was decided to look at individual and mean results from the point at which the subjects first *increased* the slider position i.e. asked for the objective loss to be decreased.[40] This is because this could be taken as the moment at which the subject first *required* a quality increase.

The maximum, minimum and mean loss rates for each subject are given in Table 18 (Group A) and Table 19 (Group V). The first figure in each column refers to the result for the no budget (NB) condition, and the second figure refers to the condition in which a budget was shown (B). The group means and range of results between different subjects is provided at the bottom of each column.

[40] This figure should not be interpreted to mean that this was the first time that the subject had moved the slider. Indeed, many subjects had *lowered* the slider before this time i.e. increased the loss. The degree to which the loss was lowered varied however, which helps explain the wide discrepancy between the times for

The range of loss levels selected during the conditions is high, indicating highly idiosyncratic subject behaviour. This indicates that group mean results must be interpreted with care.

Looking at the group means, clear differences between the no budget and budget conditions can be observed:

- The loss level that is rejected is noticeably greater in the budget conditions than in the no budget conditions for both groups. In Group A, the slider was first increased *earlier* in the budget condition (second 33.92) than in the no budget condition (second 54.17), which seems counterintuitive: if subjects are trying to conserve their budget, then they should presumably not raise the quality (i.e. spend the budget) too early on. However, when the level of loss that is rejected is considered, an explanation for this behaviour can be proffered: the loss level that is rejected in the budget condition (25.77%) is markedly higher than the loss level that is rejected in the no budget condition (12.45%). The subjects have obviously decreased the slider to a far greater degree in the initial stage of the budget condition than in the no budget condition (presumably to see the effect on the budget reservoir), to the point where an improvement in quality is required in order to continue with the conversation. This effect can also be observed in Group V, where the loss level rejected in the budget condition (12.63%) is much lower than in the budget condition (29.42%). The slider is not increased earlier in the no budget condition for this group, however.

- The mean maximum loss level that is tolerated in the period subsequent to the first increase of the slider is higher in the budget conditions (28% and 28.75%) than in the no budget conditions (20.32% and 24.52%), due, it can be assumed, to the fact that the subjects are being more cautious about the quality they request. However, this difference is not as great as might be expected, and this may be explained by the different strategies that the subjects employ to conserve budget (see section 7.4.6.1).

- There is a negligible difference between the minimum loss tolerated in the subsequent period between groups A and V, but there is a constant difference between the budget and no budget conditions of approximately 6%, again indicating that subjects are prepared to put up with more loss when there is a visible budget.

the slider first being increased, since different subjects were experiencing different levels of loss at the same point in time, through their own actions. This is illustrated clearly in the 'Loss level rejected' column.

- There does not appear to be great deal of difference between the groups and conditions as to the mean range of loss that is tolerated, although again the range of individual results that make up this figure is very large, especially in Group V (45.7 and 46).

The observations made from the average loss tolerated figures suggest that the presence of a budget makes a difference to the level of audio quality that will be tolerated, but that the presence or absence of a video channel does not affect the level of audio quality requested. These two factors are considered in more detail in the following sections.

Table 18: Individual subject behaviour in the audio-only group, Group A, in the no budget (NB) and budget (B) conditions.

Subject number	Slider first increased (second)		Loss level rejected (%)		Loss level decreased to (%)		Max loss tolerated (in subsequent period) (%)		Minimum loss tolerated (%)		Range		Average loss tolerated (%)	
	NB	B	NB	B	NB	B	NB	B	NB	B	NB	B	NB	B
1	45	39	4.4	15	3	14	7.5	24	1	3	6.5	21	3.5	12.44
2	26	30	8.1	14	6	13	12.5	31	1	13	11.5	18	6.82	22.22
3	29	13	12.1	43	9	22	9	24	0	0	9	24	3.51	4.38
4	39	43	15.3	27	11	22	19.4	22	9	12	10.4	10	13.35	16.19
5	34	63	3.3	11	2	7	12.8	20	2	7	10.8	13	6.93	10.9
6	44	40	5.8	21	4	20	8.1	20	1	0	7.1	20	3.18	3.71
7	24	25	10.1	28	4	20	20.8	43	4	7	16.8	36	15.27	22.84
8	15	10	10.1	39	9	34	29.7	34	9	30	20.7	4	21.4	30.96
9	26	26	23	28	17	25	23	27	3	15	20	12	12.34	19.91
10	75	80	21.5	27	19	26	25.3	26	17	23	8.3	3	21.97	24.37
11	63	11	12.1	25	11	23	45	36	11	17	34	19	18.14	23.9
12	31	17	12.3	33	11	31	29.9	34	11	12	18.9	22	22.3	25.24
13	199	44	23.7	24	21	23	21.2	23	2	13	19.2	10	7.63	17.53
Mean	54.17	33.92	12.45	25.77	9.77	21.54	20.32	28	5.46	11.69	14.86	16.23	12.03	18.05
Range	184	70	20.4	32	19	27	37.5	23	18	31	14.2	33		

Table 19: Individual subject behaviour in the audio and video group, Group V, in the no budget (NB) and budget (B) conditions.

Subject number	Slider first increased (second)		Loss level rejected (%)		Loss level decreased to (%)		Max loss tolerated (in subsequent period) (%)		Minimum loss tolerated (%)		Range		Average loss tolerated (%)	
	NB	B	NB	B	NB	B	NB	B	NB	B	NB	B	NB	B
14	42	57	13.7	27	12	26	18.1	26	12	15	6.1	11	14.69	19.77
15	138	11	13.8	25	12	22	18.1	23	11	5	7.1	18	14.44	17.12
16	58	9	18.2	46	16	43	29.2	43	14	28	15.2	15	21.28	33.47
17	19	116	9.1	18	8	16	19.1	17	4	9	15.1	8	14.57	14.76
18	51	47	5	28	4	25	16.8	25	3	14	13.8	11	10.11	19
19	28	52	6.1	24	5	23	18.3	36	3	16	15.3	20	10.85	25.56
20	21	33	17.9	24	16	23	22.6	24	9	6	13.6	18	15.82	22.29
21	27	12	11.4	28	7	22	24.6	22	5	18	19.6	4	10.67	20.13
22	64	35	22.2	23	18	19	23.2	19	8	13	15.2	6	14.4	15.52
23	49	31	31.6	33	24	30	50	30	0	15	50	15	5.92	18.75
24	7	27	0.6	50	0	34	50	50	0	0	50	50	13.59	21.43
25	21	55	2	27	0	23	4.3	30	0	0	4.3	30	1.52	11.3
Mean	43.75	40.42	12.63	29.42	10.17	25.5	24.52	28.75	5.7	11.58	18.77	17.17	12.32	19.92
Range	131	107	31	32	25	27	45.7	33	14	29	45.7	46		

7.4.4.2 The effect of video

The average loss tolerated in the two treatment groups was compared and subjected to an independent samples t-test. It was found that there was no significant difference between groups A and V in terms of the average loss tolerated in both the no budget (t= 0.11, p=0.46) and budget condition (t=0.66, p=0.26). The average results for each group are shown in Figure 30.

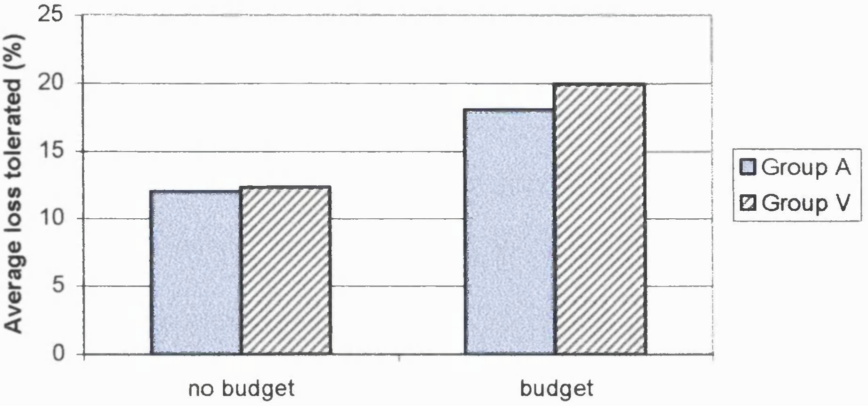


Figure 30: Average loss tolerated by Group A (no video) and Group V (video), in the 'no budget' and 'budget' conditions.

Various reasons can be postulated as to why the video channel had no impact on the level of audio quality requested. It is possible that the task ensured that the subjects' visual attention was directed at the Taboo cards and slider, not the video image, rendering the video channel effectively redundant, so that no communicative benefits could be gained. On the other hand, it is possible that the frame rates that the subjects were receiving were not high enough to transmit communicative cues, or that the delay may have caused the cues to be irrelevant or misleading. Given the findings from Study 8 (section 6.2), this latter possibility seems plausible.

In the interview at the end of the study, the subjects were asked whether they felt, in the case of having a video channel, that the video had had an effect on the quality of the audio that they had requested, and in the case where they only had an audio channel, whether having a video channel would have made a difference to the audio quality they would have required. Most of group A (9 out of 13) thought that video would have made a difference to the audio quality that they would have required, whilst only 2 members of group V definitely felt that the presence of the video channel had made a difference. What is illustrated here is a discrepancy between expectancy and experience: group A has no way of knowing what the video quality would actually be like, whereas the experience of group V allows them to make a different type of judgement. The evidence from the quantitative analysis shows that group V was right, in the main: the presence of a video channel (at least as it was presented in this study) does *not* impact on the level of audio quality requested.

On the whole, the fact that there was no negative impact of video can be seen as a positive finding, since it reveals that low-quality video quality does not impact on the audio quality required.

Since there was no effect of group, it was decided to look at all 25 subjects as one group in investigating the effect of the budget.

7.4.4.3 The effect of the budget

Looking at the effect of the budget on individual loss tolerance (Figure 31), it is clear that subjects become much more tolerant to poor audio quality when there is a visible budget. This effect is highly significant (t = 7.62 with 24 d.f, $p < 0.01$).

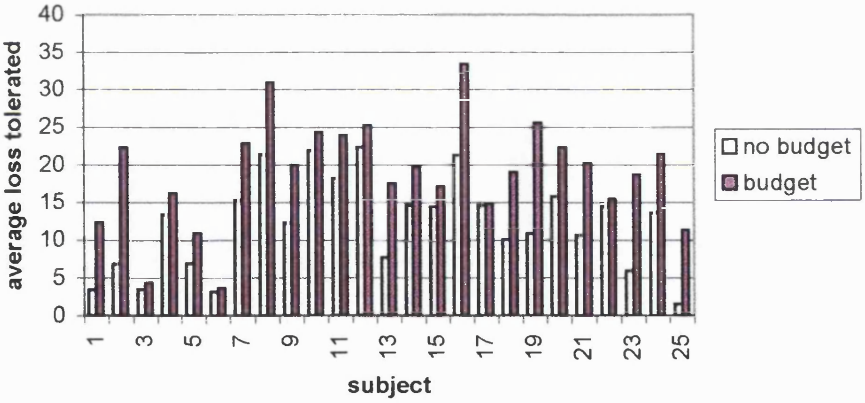


Figure 31: The effect of the budget on individual rating behaviour

The effect of the budget is clear. The graph also clearly illustrates that individuals vary widely in the amount of audio loss they will tolerate. Some subjects (e.g. S3 and S6) are very intolerant to audio loss, and the budget incentive does not appear to change their behaviour, whereas others, such as S8 and S16 are far more tolerant of audio loss (approximately 21% in each case) without the budget, and their average tolerance increases by approximately 10% when the budget is present. Due to these differences in behaviour, it is likely to be useful to study individual behaviour across time during the studies.

7.4.4.4 Individual behaviour: rejecting loss

One of the prime questions of interest in the study was whether there was a common critical level at which loss was rejected. The average loss tolerated figures (see last column in Table 18 and Table 19) do not allow conclusions of this nature to be drawn, since the range of individual averages is very large: the group A non-budget range is 3.18 to 22.3 (a difference of 19.12%), and for the budget condition the range is 3.71- 30.96 (a difference of 27.25%). In group V, the range for the non-budget condition is 1.52 - 21.28 (a difference of 19.76%), and for the budget condition the range is 11.3 - 33.47 (a difference of 22.17%).

It was decided to look at the behaviour of each subject individually and record at which point the slider was raised, i.e. a loss level was rejected. These points have been extracted and graphed for each subject (see Appendix H). The means of these points for each subject are shown in Table 20, together with the overall mean across subjects for each condition. There is a clear difference between the non-budget and budget conditions for each group, but a comparison between the mean differences for each group (7.41 and 7.01) is unlikely to be significant (further statistical analyses are not warranted since the data points in the table are already mean results, with sometimes large standard deviations). Clearly all participants tolerate worse quality when a budget is involved, and the presence or absence of video channel makes no difference.

The tabulated results permit the conclusion to be drawn that somewhere in the region of 13-15% loss is unacceptable to participants in multimedia conferences where there is no packet repair mechanism employed. Very few people (only 5 out of the 25 subjects) will tolerate loss greater than 20%.

Group A (audio only)			Group V (audio and video)		
Subject	**Non-budget**	**Budget**	**Subject**	**Non-budget**	**Budget**
1	4.28	13.42	14	15.47	21.5
2	8.22	23.08	15	14.32	16.43
3	5.28	14.83	16	22.98	36.33
4	15.35	19.17	17	14.75	15.33
5	7.41	13.57	18	10.69	23.33
6	3.82	9.31	19	12.37	23.11
7	16.46	22.16	20	17.73	16.82
8	19.92	33.33	21	13.16	22.75
9	14.46	22.3	22	16.85	19.33
10	23.19	25.75	23	9.69	23.18
11	22.25	27.13	24	31.47	27.18
12	21.88	27.21	25	2.03	20.38
13	11.24	18.84			
Mean	**13.37**	**20.78**	**Mean**	**15.13**	**22.14**
Mean Difference	**7.41**		**Mean difference**	**7.01**	

Table 20: Average loss rejected by individuals in Group A (no video) and Group V (video), with condition (no budget and budget) and group means.

A number of observations can be made from the individual subject's graphs shown in Appendix H. Firstly, it is immediately clear that there are a number of different sets of behaviour with respect to the task. Three different types of behaviour can be observed: 'exploration', 'adaptation' and 'conservation'.

- **Exploration** can be categorised as a high level of loss rejected early on in the condition, followed by a period of much lower loss toleration. Since the auto-decrement (in the non-budget condition) is neither fast nor great, a loss rejection of greater than 10% before 50 seconds of the condition has passed can be seen as a significant 'exploration' of the slider functionality. If the subject had not increased the slider by this point the objective loss would only be 5% loss. Exploratory behaviour can be clearly observed in the graphs of about half of the subjects for the non-budget conditions (e.g. subjects 3, 8 and 9), and nearly all of the subjects in the budget condition (e.g. 1, 7 and 15). That it is so prevalent in the budget condition is hardly surprising, since the subjects must understand the relationship between the quality and the budget in order to conserve the latter.

- **Adaptation** is defined as a visible trend towards higher loss rates being rejected as the condition continues. This behaviour is seen more often in the non-budget than the budget condition: about one third of the subjects exhibit the behaviour in the non-budget condition (e.g. subjects 8, 12 and 16), whereas fewer subjects exhibit it in the budget condition (subjects 2, 19, 15). It is, in fact, more usual to see a trend towards *lower* loss rates being rejected in the budget condition, which is likely to be caused by the subjects having learned that the budget does not plummet rapidly in response to higher quality requests. Since the subjects have an abundance of budget left towards the end of the condition, they feel less inclined to guard it jealously, and therefore increase the quality. There are therefore two types of adaptation occurring: that in the non-budget condition towards greater loss tolerance as the subjects become more used to the audio quality, and that in the budget condition towards spending less cautiously. This latter is related to the development of conservation strategies, the final behaviour.

- **Conservation** strategies can be observed most clearly in the budget condition (e.g. subjects 2, 3, 9, 11 and 12). For many of the subjects, the intervals of the loss rejected peaks are consistent with the points at which a changeover in the conversation would occur, i.e. the point when subject has to start listening, and therefore has a requirement for better quality audio. (The number of these changeovers varied between players and conditions, but in general it is true to say that 4-6 words were defined by each player in one 5-minute condition.) Two of the most extreme cases can be seen in subjects 7 and 24 where it is clear that the subjects maximise the packet loss when talking (i.e. conserve budget), and decrease the packet loss when listening. Surprisingly (considering the lack of 'financial' incentive) similar behavioural strategies are found in the non-budget condition. Although there was no budget to conserve in this condition, the subjects were under instruction to keep the quality at a level where they could just perform the task, and it can be seen that some subjects used the same strategy as in the budget condition to deal with this instruction i.e. they raised or lowered the slider according to whether they were talking or listening (see subjects 7, 9, 11, 22, 24).

The existence of these strategies, and some of the other behaviours, is supported by comments made by the subjects during the semi-structured interviews. These comments are displayed underneath the individual slider results in Appendix H.

7.4.5 Discussion of results

The main questions for which answers were sought were:

- Are people able to perform a conversational task and control received quality simultaneously?
- Is there a common limit to the loss that people will tolerate in completing the task?
- Does the addition of a visual channel affect the audio quality requested?
- Does the addition of a visible budget affect the audio quality requested?

7.4.5.1 Controlling the conversation and the slider

In as much as only one subject never moved the slider (the subject whose results were discarded), a straightforward answer to the first question would be yes, subjects are able to perform both tasks. However, from the interview data it is clear that there was a high level of task interference: many of the subjects found it impossible to operate the slider while they were providing definitions. Some of the subjects were explicitly aware that moving the slider only really mattered when they were listening anyway, but the inability to do both tasks at the same time may have been one reason why extreme conservation strategies were witnessed relatively infrequently.

However, Taboo had been picked as a task precisely because it *was* absorbing. This was seen as positive in that it would detract from the slider and its functionality. It was, after all, an aim that the quality should only be improved when it was really required, and if the task was absorbing, then it could be assumed that any increasing of the quality was because the subject had become aware that the quality was not good enough for the task to be continued satisfactorily. So from this point of view, comments that the definition task interfered with the slider 'task' are positive.

7.4.5.2 The limit to loss tolerance

Still, the data do not lend themselves easily to an interpretation of a certain loss level that is critical. The range of loss levels rejected and tolerated for individual subjects is very large, and the interpretation of the figures must be tempered by a consideration of any optimisation strategies that subjects may have been engaged in whilst carrying out the task. The best conclusion that can be drawn is that *most* subjects find a figure of 13-15% packet loss objectionable, but when there is an 'incentive' (such as the budget) to tolerate more loss, *most* subjects will tolerate around 20% loss. Very few subjects will regularly tolerate loss greater than 20%. Whether this conclusion can be extended to different multimedia conferencing tasks, and using different packet loss repair mechanisms, remains to be seen.

The importance of the fact that subjects do not request the highest level of quality in the non-budget condition should not be overlooked, since it supports one of the central arguments of the thesis that participants in conferences do not *require* the maximum quality that can be delivered. Bandwidth can therefore undoubtedly be saved in certain task contexts.

7.4.5.3 The video non-effect

It was shown that there was no difference between the audio quality demanded by subjects, whether they had an additional communication channel (video) available to them or not. This could be explained by the task not being complex enough to require the use of video, or that the quality of the video was not sufficiently high to afford additional communicative cues. However, it should be viewed as a positive fact that this low video quality did not actually have a negative effect. Supporting findings of the distance learning course (section 6.2), the result here shows that low video quality is at least not detrimental to communication.

7.4.5.4 The effective budget

The effect of the budget was marked. Most subjects demonstrated that they were able to make do with far lower quality when they were asked to 'pay' for what they used. Some subjects developed cunning strategies to optimise their use of this budget. That such changes in behaviour could be provoked by a budget that really did not decrease rapidly (or indeed where there was *no* budget - see section 7.4.4.4) is interesting, and raises the prospect of all sorts of unethical practices for pricing on the Internet!

7.4.6 Experimental critique

The method employed in the experiment was successful, and offered the opportunity to capture much more realistic quality requirement data than had previously been possible. However, extreme individual differences render the drawing of firm conclusions problematic. In concluding that somewhere between 13-15% (cross-group average was 14.21%) loss may be the critical level for many subjects, it was necessary to average results which ranged from 3.82 to 31.47% loss.

Given the occurrence of such differences between individuals, it would be sensible, in any future study, to employ within-group treatments (as was done with the budget), in order to better determine the effect of changes made. For example, a future study could have the same subjects performing a task with video and without, allowing comparisons to be made on an individual basis. A possible different measure of performance could be how many definitions are given and received in the two conditions, to see whether a video channel might enhance task performance.

Future investigations of the effect of a budget should employ a faster rate of decrement - in this study it was observed (and reported by some of the subjects in the interview), that once the speed of budget drop had been perceived as relatively slow, some subjects stopped worrying about it and altering their behaviour in response to it. This again could explain why some subjects exhibited strong behavioural strategies, and others did not. A future investigation of the non-budget condition should investigate the effects of a faster auto-decrement. In the experiment reported here, the decrement was only 0.1 per second, meaning that after 1 minute of playing the game (without moving the slider) the loss would only be at 6%. Since many subjects were able to tolerate significantly larger levels of loss than this, a faster decrement might force subjects to raise quality more often, lending more support to any conclusions drawn about loss levels found objectionable by those individuals.

Future work should seek to find out whether the findings observed here hold true in other multimedia conferencing set-ups, in particular for tasks where shared workspace is utilised. Experiences of real-world conference participants using QUASS as a rating device (section 7.2) showed that the continual rating task was too intrusive, but QUASS used as a quality *controller* in a conference using a variety of different media may be less intrusive, since it should only become the focus of attention when quality becomes poor enough to be distracting. In this case, offering QUASS as a solution to the problem of poor quality would presumably be seen as a welcome rather than annoying distraction.

Future work issues will be returned to in the final chapter (Chapter 9).

7.5 Chapter conclusions

The research documented in this chapter has investigated the applications of a software quality slider. The results from the studies can be divided up into substantive and methodological contributions.

7.5.1 Substantive results

- With silence substitution as the repair mechanism, passive listeners object to a packet loss rate of 15% (see section 7.1.5). In an interactive MMC application, most participants will tolerate up to between 13 and 15% loss before requesting better quality.
- Participants in conversational MMC tasks can perform the task satisfactorily with less than perfect quality: participants do not request the best quality that they can get.

- Where a 'budget' is included as an incentive to tolerate lower quality, most participants will tolerate up to 20% packet loss.
- Video does not appear to affect the level of audio requested when the video is at 8 fps.

7.5.2 Methodological results

- The application of the polar scale to a continuous quality rating method has been proved viable.
- MMC participants are able to rate perceived quality continuously using QUASS, but only if they are passively, as opposed to actively, engaged in a task.
- MMC participants can use QUASS to *control* the objective level they receive of a quality dimension, but only when they are not actively performing another task.

The thesis research has culminated in the development of a new methodology by which perceived media quality can be measured continuously (using QUASS in its first mode), and the level of different quality components can be controlled by a user (using QUASS in its second mode). However, the different quality components to be measured and controlled need to be identified through *asking the user.* What are the elements of delivered audio and video that most affect perceived quality, from the end user's point of view? Chapter 8 describes findings from a large-scale field trial which suggested that the quality factors that concerned users most were *not* caused by network congestion, but rather were due to end-user behaviour and the hardware in use. The observations made in this field trial were then investigated further in a controlled experiment. In this experiment the participants were asked to *describe* the degradations they perceived, in addition to providing quality ratings. Taking these descriptions in conjunction with comments gathered in some of the previous studies, it is possible to build an initial categorisation of different audio and video quality dimensions, as experienced by the end user.

Chapter 8: What components have the greatest impact on perceived MMC quality?

This chapter presents the results of research designed to bridge the gap between objective and subjective measurements in networked communication, and also to ascertain which are the most important aspects of 'quality' to the end user (audio quality in particular). The results of a large-scale field trial (in which methods of objective and subjective data matching were explored) informed the design of a lab-based experiment in which the impact of different types of speech degradation was investigated in detail. The results from this experiment, taken with observations made in previous studies reported in this thesis, allow for certain conclusions to be drawn about the importance of hardware and end-user behaviour, and also about the way end users describe different types of degradation.

8.1 The PIPVIC field trials

The UKERNA-funded PIPVIC (Piloting IP-based VIdeoConferencing) project sought to gain a greater understanding of the issues that may be encountered in running a large-scale IP videoconferencing service, and to determine whether a best-effort network service can provide the reliability and quality required to support teaching and learning activities on this scale. The project was unique in its sheer size and diversity (13 UK universities and 150 participants were involved over a period of 9 months), and collecting meaningful subjective *and* objective performance data presented a challenge (see Watson and Sasse, 2000a).

8.1.1 Issues in evaluating IP-based videoconferences

As discussed in Chapter 2, collecting evaluation data in IP videoconferencing is problematic, due to issues such as rapidly changing network conditions, leading to packet loss and fluctuations in delivered audio and video quality during the course of one conferencing session. It has been demonstrated in lab-based studies that perceived quality is affected negatively by packet loss (see Chapter 5). However, in the real world there are other sources of 'objective' degradation, such as background noise, or problems caused by the hardware used, such as 'leaky' headsets or inadequate lighting. The impact of these real-world objective factors is difficult to assess, especially since network conditions, hardware and set-up will all be different for different end-sites.

As reported in Chapter 4, conventional audio-video subjective quality measurements are taken at the end of a session, but when participants report that quality was "bad at times" over an hour-long session, to which times are they referring? How can it be established *when* quality is not good enough, and how can it be known *what* is responsible for the perception of poor quality? As discussed in section 7.2, taking continuous measurements using QUASS is not practical due to task interference. In the field trial described here, steps were taken to address this issue by time-matching network statistics and subjective opinions. When registered opinion is poor, but network conditions are good, a better understanding of the effects of other objective factors can begin to be established.

8.1.2 Objective and subjective data collection

8.1.2.1 Objective data collection

RAT was modified such that *reception reports* gathered from participants in a conference could be logged. RTP (Real-Time Transport Protocol) reception reports are generated once every 2-5 seconds, and can be used to produce an overall picture of the level of loss experienced at different end sites in a conference. This development afforded a means of recording (objectively) how a participant's audio stream is received at other sites. The logged statistics could then be matched in time with the web-based opinion ratings discussed below, since there would be digital timestamps on both.

Scripts were used by the field trial participants to launch **RAT**, meaning that the settings used were the same across all participants. The settings used in the project were 40 ms packets encoded in DVI with packet repetition as the packet loss repair method.

8.1.2.2 Subjective data collection

Subjective evaluation data was collected from some 150 participants in a range of activities encompassing small group working, tutorials, seminars and lectures, in various topics including Russian, Sociology, Art History and Business. The conferencing activities were mainly, but not exclusively, desktop rather than room-based. Over 500 hours of MMC experiences were evaluated in total.

Subjective data was gathered through paper-based questionnaires at the end of each particular course, addressing perceived audio *quality*, and the *adequacy* of the quality for the task at hand; group workshops with the tutors and students; and through web-based opinion scales completed at set points during a conference. The perceived quality and adequacy, and web-based scales all permitted a closer match between subjective ratings and objective conditions to be derived. The scales that were used were modified versions of the polar scale explored in the previous chapter. The scale was developed such that the end-points are bounded by the numbers 1 and 100 (see Figure 32). This development was necessary to gather data via a web-based form, since a polar scale was impractical to implement in HTML. Instead a decision was taken to ask participants to provide a figure between 1 and 100 that they felt represented the audio quality or adequacy. In order to keep the data as consistent as possible, the paper-based scales were numbered between 1 and 100, in intervals of 10.

As discussed in section 4.11, the research in this thesis is conducted in agreement with the point made by Preminger and Van Tasell (1995), who state that speech quality should not be treated as a unidimensional phenomenon, since one or many different dimensions may affect the listener's opinion. This is why there are no descriptive labels other than at the end points on the 100-point scale. Instead subjects are asked to give a numerical rating and, if this rating is less than very good, to describe *how the sample sounded* – explain why a certain rating was awarded. Knoche et al. (1999) argue that use of the traditional 5-point scale leaves the experimenter ignorant of the subject's perspective and rationale for positioning on the scale: by asking the user to explain why a rating is awarded on the 100-point scale, a deeper insight into factors that affect perceived quality can be gained, with a long-term view to producing a series of diagnostic scales along different quality dimensions (see section 8.3.4).

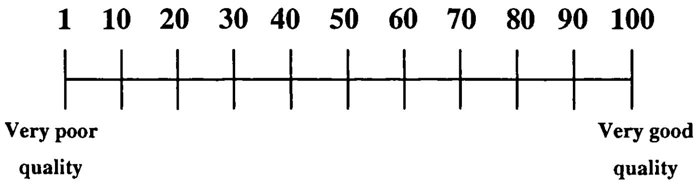


Figure 32: The 100-point quality rating scale

8.1.3 Results

The outcome from the project's network monitoring activities showed that, in general, the level of packet loss on the SuperJANET multicast service, and participants' local networks, was low during the field trial. One such picture is shown in Figure 33, where reception reports reflecting the level of loss received by a participant in Glasgow from a participant in London are shown.

It was found that packet loss on the audio stream was relatively rare, and in general occurred at a level of 5% loss or below. Higher levels of packet loss tended to appear in very short-lived bursts, but could reach levels of 20% or greater.

Despite these encouraging objective statistics, however, participants reported speech quality problems in one out of three multimedia conferencing activities, where sufficient bandwidth was available. The most commonly reported factors were: missing words or incomplete sentences;

variation in volume between participants; and variation in quality between participants. These problems, and their likely causes, are summarised in Table 21.

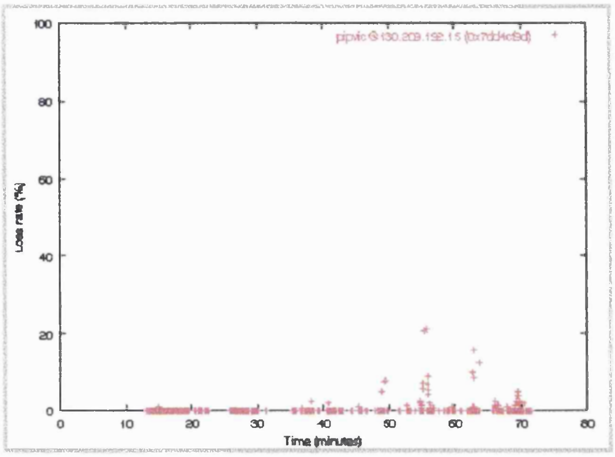


Figure 33: Loss reported from UCL by a Glasgow conference participant over a one-hour project meeting

Problem	Likely causes
Missing words or incomplete sentences	Packet loss; silence suppression clipping beginnings and endings of words; machine 'glitching'
Variation in volume between participants	Insufficient volume settings; poor headset quality
Variation in quality between participants	High background noise; open microphone; poor headset quality

Table 21: Key problems reported by users in the PIPVIC field trials

Missing words would normally be attributed to the occurrence of packet loss, but through consultation of the objective reception reports, and cross-matching with the time that the web rating forms were completed, it is clear that in many cases missing words are *not* due to network packet loss. The likely cause is therefore either hardware (e.g. a machine becoming temporarily overloaded and 'glitching' as a result) or software (e.g. faulty silence suppression in **RAT**).

The variation in volume and quality between participants are again not network factors, but due to hardware (e.g. inadequate headset or soundcard) or user behaviour (e.g. moving microphone away to cough, then not replacing it near the mouth; insufficient raising of the volume bar on the audio tool).

8.1.4 Discussion

The results from this field trial show that the more mundane aspects of MMC, such as discrepancy in volume between participants, or the quality of the headset in use, can have greater effects on participants than network conditions. The impact of these factors would not have been apparent in controlled, lab-based simulations of MMC, but the combined approach of objective and subjective measurement undertaken in the PIPVIC field trials has revealed their importance. The findings can now be used to inform the design of meaningful controlled studies to assess the relative weights of these factors. One such study is described in the following section.

8.2 Study 11: Investigating the effects of other factors on perceived quality

An experiment was designed to compare the relative impact of different types of degradation on subjective quality ratings of interactive speech transmitted over a packet-switched network. The experiment was inspired by the observations made during the long-term large-scale PIPVIC field trials described in the previous section, that in most cases, unsatisfactory perceived speech quality was caused by end-system hardware, equipment set-up or user behaviour. The experiment presented here investigates the effects of volume differences, echo and 'bad' microphones as well as different levels of repaired packet loss.

The experiment and its findings was presented at ACM Multimedia 2000 (Watson and Sasse, 2000b).

8.2.1 Rationale for conditions chosen

The user assessment of the conferencing sessions in the PIPVIC trials showed that three factors were most often reported as problematic: missing words or incomplete sentences; variation in volume between participants; and variation in quality between participants (see Table 21). Although missing words were frequently cited as a problem by the participants, the outcome from the project's network monitoring activities showed that, in general, the level of packet loss on the SuperJANET multicast service, and participants' local networks, was low during the trial: generally 5% loss or below. Higher levels of packet loss tended to appear in very short-lived bursts, but could reach levels of 20% or greater.

In view of these findings, a controlled experiment investigating the impact of different factors on subjective opinion was designed.

8.2.1.1 Codec, packet size and packet loss repair method

In selecting the codec and packet size to investigate in the experiment, it was decided to use the RAT version 3 default settings that had been available to participants in the field trials, even though in recent months these defaults have been changed and are now likely to produce better perception[41]. Therefore the experimental speech material that was generated was coded in DVI, using 40 ms packets, and repaired with the receiver-based packet loss repair method packet repetition.

8.2.1.2 Packet loss rates

As a result of the PIPVIC findings reported above, for the experiment 5% was selected as a lower level of packet loss, which is representative of the level of packet loss users are likely to experience on the SuperJANET multicast service today. The figure of 20% was chosen as a higher level of packet loss for two reasons. Firstly, it is the upper level of loss that was witnessed in the PIPVIC field trial. Secondly, the studies presented in Chapter 5 showed that 20% loss will impair perceived quality, but not damage intelligibility – it was an aim of this study that intelligibility not be impaired. For these two reasons, 20% loss would act as a reference point in the planned study.

8.2.1.3 'Bad' microphone

A poor-quality microphone was chosen as a condition because during the field trials users had reported and complained about 'tinny' or 'hummy' microphones. The selection of a 'bad' microphone is, of course, somewhat subjective. In addition, a microphone that produces 'bad' audio when used with one soundcard will not necessarily be a 'bad' microphone for another, making the matter more complicated. However, the effect of microphone distortion was still felt to be worthwhile investigating, since so many subjective comments refer to how the voice sounds, and whether it is pleasant to listen to (see Preminger and Van Tasell, 1995). The microphone chosen for the experiment was an Altai A087F.

8.2.1.4 Volume differences

Many users in the field trials complained of extreme volume differences between participants in multi-way conferences. Although it is possible to alter the incoming volume from a particular participant by adjusting the incoming volume slider in **RAT**, when the next participant speaks at either a softer or louder volume, users find themselves constantly having to adjust the slider, and getting frustrated. For the experiment it was decided to investigate the subjective effects of one speaker at 'normal' volume, and the other at 'too loud', and also the impact of 'normal' and 'too

[41] At the time of writing, the defaults in RAT version 4 are 20 ms packets, with pattern matching as the preferred method of receiver-based packet loss repair.

quiet'. Again, it is recognised that determining what is 'too loud' or 'too quiet' is a subjecti decision to be taken by the experimenter, but by piloting the experiment with both network audi experts and novices it was possible to determine levels that were commonly agreed to be 'too lou or 'too quiet'.

8.2.1.5 Echo

Echo, or feedback, commonly occurs in multicast conferences when people are working individual offices and using a speaker and open microphone, and forget to mute their micropho when not speaking, or when 'leaky' headsets are used (i.e. the headphones leak sound into t microphone). Although the echo effect is primarily annoying to the speaker, it is also distracting other listeners.

8.2.2 Material

A two-person dialogue was created from recordings of multicast project meetings, with names a locations changed from the original recordings. This script was acted out by two male acto without regional accents. The actors sat at Sun Ultra workstations at different locations on the sa local network for the duration of the recording. The recording was made at 16 bit linear quality a recorded via the record facility in **RAT**. Silence suppression was left on and both micropho were kept open during the recording. The actors wore identical Canford DMH12OU headset Once the whole dialogue had been recorded, the actors read certain parts of the dialogue agai whilst the experimenter manipulated aspects such as the volume, the headset in use, and t feedback of one of the speakers. The resulting recordings were then split into 2-minute files a coded into DVI, at 8kHz sampling rate, and 40 ms packets. Packet loss and repair (pack repetition) were generated on the files where required, using the software program **test_repair**[42]. The conditions that were generated were:

- **reference**: a no-degradation reference condition;
- **5% loss**: 5% packet loss generated on both voices, and repaired with packet repetition;
- **20% loss**: 20% packet loss generated on both voices, and repaired with packet repetition;
- **echo**: one person using an open microphone and speaker rather than headset, such that t other person generates echo/feedback ;
- **quiet**: one voice recorded at a low volume, the other at a normal volume;
- **loud**: one voice recorded at a high volume, the other at a normal volume;
- **bad mic**: one person using a poor quality microphone.

Three Internet audio experts agreed that the conditions were identifiable as containing the degradations to be tested, and also that the intelligibility of the recorded speech was not affected by the impairments[43]. A pilot study of the recorded samples with 6 subjects (all first-time users of Internet audio) confirmed the expert assessment.

8.2.3 Subjects

Twenty-four subjects (12 men and 12 women) participated in the study. They all had good hearing and were aged between 18 and 28. None of them had previous experience in Internet audio or videoconferencing.

8.2.4 Procedure

A within-subjects design was employed. The subjects each listened to the seven 2-minute test files twice (to determine the consistency of subjects' scores on the 100-point scale). The files were played out through the program Audio Tool[44] on a Sun Ultra workstation. Each subject listened to the files wearing a Canford DMH12OU headset. The test files were preceded by a 1-minute volume-test file. Following the volume-test file the subjects were asked whether or not they wanted the playout volume to be adjusted. This volume was then used to play out the subsequent test files. The subjects were also instructed that the volume test file should be taken as indicative of the best quality they would encounter in the following test files. The order of the test files was randomised, with one exception: the **reference** (no degradation) condition was always heard first and eighth. The 7 conditions were therefore all heard once before they were repeated in a different order.

After each test file the subject was asked to provide a quality rating, for the file as a whole, from the 1-100 scale where 1 represents *Very Poor Quality* and 100 represents *Very Good Quality* (see Figure 32). The subject was then asked to explain why that rating was awarded i.e. how the speakers had sounded to him/her. These answers were tape-recorded.

8.2.5 Results

The results from the subjective rating and qualitative responses confirm that the impact of volume discrepancies and voice feedback affect perceived quality more adversely than the levels of packet loss typically experienced in the PIPVIC field trials.

[42] **test_repair** is a component verification program included in the RAT version 4 application.

[43] It was an explicit aim of the study that the intelligibility of the speech should not be affected. As argued in section 4.11.1, intelligibility and perceived quality are not the same thing - it is possible to get high intelligibility with speech that receives very poor quality ratings e.g. with synthetic speech, but not vice versa.

[44] Audio Tool is an OpenWindows DeskSet application for recording, playing and simple editing of audio data.

8.2.5.1 Quantitative results

The mean results and standard error for the perceived quality ratings are shown in Figure 34. The full data set is presented in Appendix J. The graph suggests that a 'normal' level of packet loss (5%) when repaired with packet repetition has little impact on perceived quality when compared to the **reference** (no degradation) condition. As expected, **20% loss** repaired with packet repetition has a profound effect on perceived quality, but it appears a loud-normal volume discrepancy, and an echo effect also affect perceived quality adversely. Are these apparent differences statistically significant?

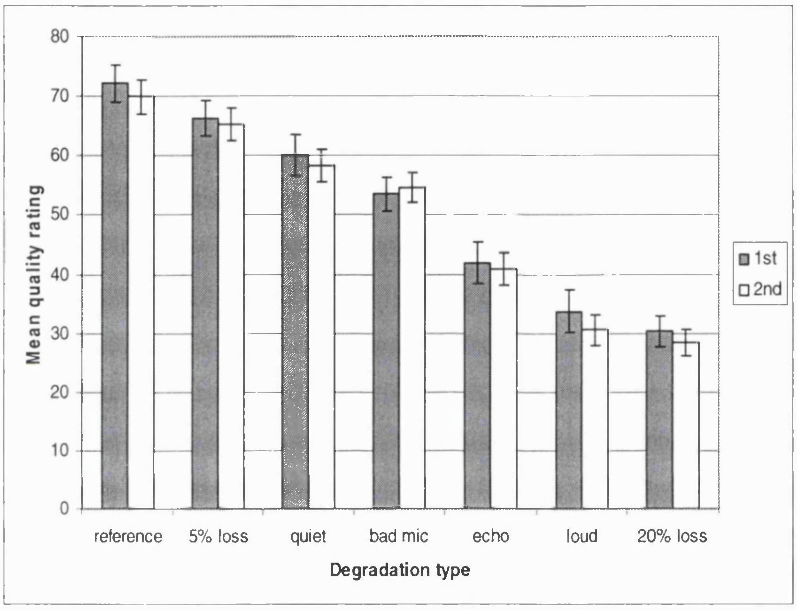


Figure 34: Mean quality rating awarded for different degradation types, on 1st and 2nd occasion of hearing

Analyses of variance were carried out on the data. A two-factor with replication ANOVA at the 1% level of probability revealed that there is a highly significant effect of condition (F 6, 322 = 62.25, p < 0.01), and that there is no significant difference between the quality ratings awarded on 1st presentation and those awarded the 2nd time of hearing (F 1, 322 = 0.799).

Since there is no significant difference between the 1st and 2nd presentation ratings, the mean response for each person can be taken. These results are presented in Table 22. An analysis of variance on these combined means again confirms that there is a highly significant main effect of condition at the 1% level of probability (F 6, 161 = 36.598, p < 0.01). Post hoc analyses (Tukey HSD) allow further statements to be made as to where these significant differences lie. There is no significant difference between the **reference** condition and the **5% loss** condition (Qcrit = 4.88, Qobt = 1.97) or the **quiet** condition (Qobt = 4.36). The differences between the **reference** condition and all other conditions are significant. The **5% loss** condition is not significantly different from the **quiet** condition (Qobt = 2.39), but it is rated significantly higher than the **echo** (Qobt = 9), **loud** (Qobt = 12.41) and **20% loss** (Qobt = 13.43) conditions at the 1% probability level, and higher than the **bad mic** condition at the 5% level (Qcrit = 4.17, Qobt = 4.33). Although **20% loss** gives the worst performance according to the graph, the difference between this condition and the **echo** and **loud** conditions is not significant at the 1% level (Qobt = 4.43 and 1.02 respectively).

	ref	**5%**	**quiet**	**bad**	**echo**	**loud**	**20%**
1	77.5	67.5	62.5	50	37.5	27.5	30
2	77.5	78.5	49.5	67.5	35	51	29.5
3	81.5	81	81.5	71	30	18.5	17
4	83.5	73.5	51	52.5	60	31.5	32.5
5	72.5	60	27.5	40	20	10	25
6	60	50	65	52.5	17.5	25	9
7	58.5	45	59	44	30	10.5	19
8	60	60	47.5	50	42.5	24	27.5
9	77.5	50	60	62.5	40	17.5	34
10	90	90	50	57.5	42.5	27.5	32.5
11	35	50	40	35	25	20	15
12	87	81.5	76	66.5	60	66	47.5
13	72	73.5	80	57	51	34	27.5
14	65	59	67.5	52.5	37.5	27.5	37.5
15	67.5	62.5	56	51.5	52.5	50	40
16	82.5	65	60	65	75	30	35
17	90	80	79	66.5	50	27.5	21.5
18	80	72.5	67.5	70	57.5	52.5	45
19	72.5	62.5	52.5	37.5	30	35	40
20	60	60	57.5	40	32.5	42.5	20
21	77.5	72.5	62.5	55	45	30	30
22	75.5	77	79	72.5	57.5	61	42.5
23	60.5	68.5	54	47.5	41.5	31	33
24	42.5	38	37.5	32.5	22.5	21	14
Mean	**71.08**	**65.75**	**59.27**	**54.02**	**41.35**	**32.12**	**29.35**

Table 22: Combined subjective rating means for 1st and 2nd presentation of the conditions

8.2.5.2 Qualitative results

In addition to providing a rating on the 100 point scale, subjects were asked to describe why they had awarded each rating. The primary aim of this part of the study was to search for common descriptive terms used by non-expert users to describe different types of degradations to aid in the building of diagnostic scales, as discussed in section 9.3. The descriptions also functioned as a check on the experimental conditions by making it possible to check that users had perceived and reacted to the effect intended.

As might be expected, subjects were able to clearly identify and describe the problems in the **quiet**, **loud** and **echo** conditions. From the answers given it was found that the **quiet** condition was rated relatively highly because the subjects found it not *too* quiet or annoying to listen to, unlike the **loud** and **echo** conditions. In the **loud** condition, subjects complained of the increased level of noise in general e.g. the speaker's breathing could be heard.

For the **bad mic** condition three main types of description were found: 'distant' or 'far away', 'muffled', and descriptions likening the speaker to being 'on the telephone', or 'walkie-talkie', or 'in a box'.

In the **5% loss** condition the terms that appeared most frequently were 'fuzzy' and 'buzzy', (mentioned by 13 of the subjects) with 'metallic', 'robotic' and 'electronic' appearing slightly less often (7 times) than might have been anticipated. This fuzziness/buzziness is due to the speech waveform changing in the missing packet, and not being catered for well enough in the repeated packet.

In the **20% loss** condition, the descriptive terms used most often were words that suggested the mechanical nature of the sound: 'robotic', 'metallic', 'digital', 'electronic' (mentioned by 15 of the subjects), in addition to terms such as 'broken up' and 'cutting out' (10 times). Compared to the **5% loss** condition, 'fuzzy' and 'buzzy' were generated infrequently - just twice each. Interestingly, 5 subjects described the impairment as 'echo', and 10 of the subjects described major volume variations in the file.

The frequency with which the subjects ascribed volume differences (in the **20% loss** condition especially, but also in the **5% loss** and **bad mic** conditions) as a problem was surprising. Since the original recordings did not have volume differences, and because subjects were not consistent in attributing the problem to the first or second speaker, the conclusion must be that users do not always reliably identify the cause of a degradation. This has implications for the type of support that users require, as will be argued in the following sections.

8.2.6 Discussion

The results of the experiment have shown that the typical level of packet loss encountered in the PIPVIC trials (which was generally below 5%), when repaired with a method such as packet repetition, does not affect users' subjective ratings adversely when compared to a no-loss condition, whereas non-network factors such as volume discrepancies between speakers, poor quality microphones, and echo or feedback do. It is not the case that the users do not *notice* the degradation in the **5% loss** condition (since their descriptions of the files are different from those of the **reference** condition), but rather that it has less impact on perceived quality than other types of degradation. This finding highlights the difference between the results that would be obtained from a study investigating merely whether degradations were noticeable or not, and the richness of an HCI approach, which enables the prioritisation of different types of degradation from a user's point of view.

Importantly, the study demonstrated that users will rate the different conditions *consistently* on a 100-point rating scale. This means that the scale is reliable (like the polar scale it is based upon – see section 6.1) and can be used with confidence. Participants in the study also described the repeated degradations in a similar manner to the first time they heard them, indicating that they formed a consistent impression of the cause of the degradation. However, it was observed that, although their ratings and descriptions may be consistent, users often attribute impairments inaccurately, suggesting there is a need for a diagnostic tool to aid users in correctly identifying the source of different impairments, and then enable them to take appropriate steps to correct them. This issue will be returned to in Chapter 9.

8.3 Further vocabulary findings

Support for the vocabulary categorisations identified in section 8.2.5.2 comes from additional qualitative data collection carried out in some of the studies presented in the earlier chapters, namely the two main QUASS studies (sections 7.1 and 7.4) and the distance learning course on networks and communication (section 6.2).

8.3.1 Findings from the first QUASS study

As a subsidiary part of the first QUASS study (section 7.1), the 24 subjects were asked to select terms from a list that they felt described the audio quality they had experienced when that quality was at its *worst*. Under the circumstances of the experiment, it was known precisely what that worst objective quality had been – 30% loss repaired by silence substitution.

The subjects were instructed to select as many terms as they liked from a list of 19 descriptive words. The word lists were composed of 17 descriptive terms that had been generated by participants in the ReLaTe questionnaires, and also in the conversation study (section 5.5), and two extra descriptive words that are commonly used by multicast researchers to describe packet loss speech: lossy and bubbly. In addition, the subjects were instructed to write down any terms they felt were appropriate, but were not on the list.

The number of times each term was selected by the 24 subjects was totalled, and the figures are presented in Table 23.

Term	No. of times selected
Broken/breaking up	22
Irregular	15
Jumping	15
Cut up	13
Cracking/cracked	11
Crackling	11
Choppy	10
Metallic	8
Disconnected	7
Fluctuating	7
Clipped	5
Stuttering	5
Echoing	4
Clicking	3
Lossy	3
Muffled	3
Fading	2
Bubbly	1
Staccato	1

Table 23: Words and number of times selected to describe audio quality 'at worst' by 24 subjects

The terms that were generated in addition to those on the list were: distorted (generated twice), incoherent, slurring, noisy, interfered with, discontinuous, tinny, hazy, non-continuous, spliced, disjointed, and fragmented.

All but two of the subjects chose broken/breaking up to describe the worst quality. Irregular, jumping and cut up were selected by more than half of the subjects. Bubbly, clicking, fading, lossy, muffled and staccato were each selected only 3 or fewer times. In contrast to the study presented in section 8.2, the terms *metallic, echoing* and *muffled* were not selected often. It can be hypothesised that this is because the conditions did not provoke the requisite degradations, e.g. the 30% packet loss was not repaired with packet repetition (silence substitution was the method used) so the speech should not have sounded metallic. Likewise, the listening material had been recorded from a CD, so there would have been no echo in the material, nor would the material be affected by a poor quality microphone such as to create a muffled sound.

The term lossy, which is often favoured by MMC researchers, was selected only three times, suggesting novice users and expert users do not speak the same language when it comes to describing impairments. The term bubbly, another term popular with MMC researchers, was only selected once by the subjects. That bubbly was not a popular choice is perhaps less surprising since even experts would not describe 30% packet loss using this term, but again the overall evidence suggests that care must be taken to develop a 'common vocabulary' between experts and users of multicast conferencing technology.

8.3.2 Findings from distance learning course

In section 6.2, observations from the distance learning field trial were presented. At the midpoint of the course, 23 students participated in a group workshop, during which they were asked to select words from a long list of terms describing possible audio and video degradations[45]. The students were asked to select as many of the words as they felt described the audio and video degradations they had experienced, and additionally instructed to write down any suitable terms they could think of that were missing from the list. As with the study presented in the previous section, the choice lists had been generated through analysis of the written responses to quality questionnaires from the conversation study (section 5.5) and the ReLaTe project (section 5.1) (where video, as well as audio, terms were identified). In addition to these terms, terms from likely semantic groups in a thesaurus were included (for example, video terms were taken from the semantic groups of 'nonuniformity', 'break' and 'jerky'). Clearly, there was an element of subjectivity in determining which terms from the thesaurus should be included in the overall list, but the final list of possible responses numbered 50+ options for both audio and video, so a comprehensive range was covered.

The descriptive terms and how often they were selected are presented in Table 24 and Table 25 for audio and video respectively.

8.3.2.1 Audio results

From the table it can be seen that terms commonly used by MMC researchers, such as choppy and clipped, were only selected twice each. Although three descriptive terms were identified as missing by the students (disjointed, cut off and fuzzy), bubbly and lossy, terms that MMC researchers commonly use, were *not* generated by this user population. These findings clearly support the hypothesis of the study: the vocabulary of different user groups cannot be assumed to match.

In contrast with the results from the QUASS study (section 8.3.1), the terms choppy and clipped were chosen infrequently. This can be explained by the absence of a great deal of packet loss, whereas the opposite was true in the QUASS experiment.

[45] N.B. The object of the study was to find words that describe the *problems*, not the overall quality.

Table 24: Frequency of audio term selection

Broken up	18	**Rumbling**	2
Echoing	16	**Spluttering**	2
Crackling	11	**Whispered**	2
Disconnected	9	**Brassy**	1
Distant	9	**Chunky**	1
Irregular	9	**Clicking**	1
Cut up	8	**Cluttered**	1
Stuttering	8	**Grating**	1
Cracking/ed	7	**Murmuring**	1
Fading	7	**Muttered**	1
Jumping	7	**Pinched**	1
Fluctuating	6	**Rattling**	1
Muffled	6	**Ruptured**	1
Staccato	6	**Rusty**	1
Blasting	5	**Scratched**	1
Breaking	5	**Scrunched**	1
Clashing	5	**Shattered**	1
Faint	5	**Shrill**	1
Rough	5	**Warped**	1
Stammering	5	**Whistling**	1
Squeaky	4	**Booming**	0
Buzzing	3	**Carved up**	0
Humming	3	**Clattering**	0
Popping	3	**Crunching**	0
Blaring	2	**Grumbling**	0
Choppy	2	**Jagged**	0
Clipped	2	**Ragged**	0
Damaged	2	**Ripped**	0
Deadened	2	**Scraping**	0
Harsh	2		
Hissing	2		
Metallic	2		

Table 25: Frequency of video term selection

Frozen	14	**Fitful**	2
Inconsistent	13	**Fractured**	2
Jerkiness	13	**Jittery**	2
Variable	12	**Lumpy**	2
Breaking up	11	**Mosaic**	2
Disjointed	9	**Changeableness**	1
Patchy	8	**Jumbled**	1
Unpredictable	8	**Motley**	1
Fuzzy	8	**Patchwork**	1
Shaky	7	**Piecemeal**	1
Sporadic	7	**Splintered**	1
Erratic	7	**Disordered**	0
Bumpy	7	**Divided**	0
Blurred	7	**Gaining**	0
Fluctuating	6	**Haphazard**	0
Fragmented	6	**Jaggedness**	0
Intermittent	6	**Jiggly**	0
Losing	6	**Ruptured**	0
Unsystematic	6		
Fits and starts	5		
Uneven	5		
Flickering	4		
Inconstant	4		
Segmented	4		
Bitty	3		
Choppy	3		
Disintegrating	3		
Irregular	3		
Jigsaw-like	3		
Jolty	3		
Shuddery	3		
Confused	2		

Since the packet loss in the course was minimal, the silence suppression mechanism in the audio tool is probably responsible for the prevalence of broken up as a selected term. This mechanism can have the tendency to 'suppress silence' when in fact people are just speaking more quietly (which often happens as people draw to the end of their speech bursts).

8.3.2.2 Video findings

As with the audio results, it was found that the word MMC researchers most commonly use, blocky, was not spontaneously produced (strobe and delayed were), and the most similar seeming terms, mosaic and patchwork, were not prime choices. Of the ten most popular terms, five come from the semantic grouping of 'nonuniformity' i.e. inconsistent, jerkiness, patchy, unpredictable and variable. Two cluster under the grouping of 'break': breaking up, disjointed. Frozen, the most popular choice, seems to be a dimension of its own.

There is a relatively broader spread of terms with the video results than with the audio results (no one term was chosen by over 50% of the participants), perhaps indicative of the relative importance and emphasis that people give to audio compared with video.

8.3.2.3 Discussion

The prevalence of the terms *broken up* and *crackling* for audio terms is interesting since it is known that there was little network congestion during the course. The fact that the term *metallic* was selected only twice backs this up, since if the packet loss had been network induced, the loss would have been repaired with packet loss, causing the 'metallic' sound reported in section 8.2.5.2. The breaking up of the signal is therefore likely to be due to local workstation overload, which can cause packet loss, rendering the speech broken up and/or crackling. Echoing was selected frequently, suggesting that there was poor insulation between the microphone and the earphones on some of the headsets in use - the relative popularity of distant also suggests that the microphones were of poor quality.

Given that there should have been negligible packet loss during the tutorials, it is hard to infer what objective conditions the subjects were applying the terms *to*. Some of the terms that were chosen clearly pertain to hardware and environmental problems, rather than the networked communication per se. For example, echoing was selected by 16 of the participants. Echo is caused by a speaker's voice being received and fed back, either through headphones that do not provide a sufficient barrier between the earphones and microphone, or through users listening through speakers and forgetting to mute their microphones when not speaking. Muffled can be attributed to poor quality microphones, or people not placing their microphones close enough to their mouths, whilst faint may be similarly attributed or due to the perennial problem of people not setting their speaking volume to the correct level. (These interpretations were confirmed by the study presented in section 8.2.)

With respect to the video terms, the popularity of terms such as frozen, jerkiness, breaking up and disjointed suggests that the irregular block updates that are characteristic of H.261 video transmission, coupled with the low frame rates set in the study, had a negative impact on perceived quality.

These two studies have revealed the terms that non-expert users would commonly use to describe audio and video quality degradations, but how do users describe the quality at the other end of the scale i.e. when there are no degradations? The next study, carried out as part of the second QUASS study, investigated descriptions of both good and bad quality. The subjects were asked to generate their own descriptions, as they were in the study reported in section 8.2.

8.3.3 Findings from the second QUASS study

As part of the second QUASS study (section 7.4), the subjects were asked to imagine describing the quality *range* of the speech that they had experienced to someone who had no idea what speech over the Internet sounds like, the aim being to make them understand what to expect.

The subjects were asked to provide as many descriptive adjectives as possible, i.e. not just terms such as 'good' or 'bad'. They were asked to describe what the audio sounds like when there is no problem hearing what the other person is saying ('at best'), and then what it sounds like when it is hard to make out what the other person is saying ('at worst'). The subjects who had had a video channel (group V) were additionally asked to complete the same task for the best and worst video quality they had experienced.

The descriptive terms that were generated are shown in Table 26 and Table 27, with the frequency of generation shown in brackets.

From the prevalence of the term clear, the clarity of both the audio and video is obviously one of the most important quality dimensions to the users. The results also confirm the negative descriptors of the previous studies i.e. broken up and crackling describe high packet loss repaired with silence substitution.

Positive audio terms	Negative audio terms	
Clear (15)	Broken (up) (13)	Clipped (1)
Continuous (4)	Crackling/crackly (9)	Interrupted (1)
Smooth (3)	Metallic (5)	Jerky (1)
Good (2)	Choppy (4)	Impaired (1)
Unbroken (2)	Interference (4)	Unintelligible (1)
Full-bodied (1)	Tinny (3)	Halting (1)
Rounded (1)	Fuzzy (3)	Inconsistent (1)
Crisp (1)	Unclear (3)	Hollow (1)
Whole (1)	Cut up (2)	Thin (1)
Intelligible (2)	Jumpy (2)	Distant (1)
Sharp (1)	Stutters/stuttery (2)	Discontinuous (1)
Connected (1)	Delayed (2)	Glitches (1)
Flowing (1)	Distorted (2)	Echo (1)
Unobtrusive (1)	Intermittent (2)	Mangled (1)
Normal (1)	Muffled (2)	

Table 26: Descriptive audio terms generated by the subjects

Positive video terms	Negative video terms	
Clear (3)	Slow (4)	Intermittent (1)
Smooth (2)	Broken up (4)	Ill-defined (1)
Swift (1)	Blocky (2)	Blurred (1)
Bright (1)	Fuzzy (2)	Interrupted (1)
Fast (1)	Delayed (2)	Twitchy (1)
Transparent (1)	Stilted (2)	Jerky (1)
Solid (1)	Disjointed (2)	Jumbled (1)
	Jigsaw (1)	Low resolution (1)

Table 27: Descriptive video terms generated by the subjects

8.3.4 Discussion

The studies have provided support for the findings of the experiment reported in section 8.2. The descriptive terms that non-expert users use with respect to MMC quality degradations are starting to be identified. What is interesting to note is that these terms are not necessarily those which MMC researchers commonly use, and also that the terms seem to cover a number of semantic areas, which can be attributed to the impact of different quality factors and dimensions. Different types of degradation fall under different semantic labels (corresponding to the different components of quality identified by researchers such as Preminger and Van Tasell, 1995, and Virtanen et al., 1995).

On the basis of all of these results, it is possible to hypothesise a number of terms that are indicative of different factors that impact on perceived quality. For example, with respect to the audio channel:

- Broken/breaking up, and also clipped are indicative of silence suppression cutting in when the speaker is too quiet or tails off at the end of a speech burst.
- Choppy, clipped and cracking/crackling are all indicative of packet loss.
- Echoing indicates hardware problems, i.e. there is feedback from the headset or speaker to the microphone.
- Distant indicates that the speaker's microphone is too far away from the mouth, or possibly that the volume level is too low.
- Metallic refers to the speech repair mechanism - if there is a large amount of packet loss and the speech repair mechanism employs a synthetic algorithm, the speech may sound metallic.
- Jumping possibly indicates the presence of jitter.

With respect to the video channel:

- Frozen indicates that the frame update rate was not fast enough to create an impression of movement, or that momentary bursts of packet loss effectively rendered the image motionless.
- Broken up indicates packet loss, where the image is not updated smoothly.
- Blocky and possibly patchy indicate a high level of packet loss, since packets contain a small block of the image. When packets get lost, different blocks of the image will not appear to 'fit' the image.
- Delayed indicates that the degree of mismatch between the audio and video streams is apparent.
- Inconsistent, variable and jerkiness indicate irregular frame rate update, and possibly irregular block updates within the frames.
- Fuzzy and blurred indicates that the resolution of the image is not good enough.

When quality is good, it is the *clarity* of the both the image and the sound that is valued most highly, and the *continuity* of the speech stream.

To verify these hypothesised factor descriptors, a fuller investigation needs to be undertaken. A straightforward method of doing this would be to record speech and video material and then degrade it in a specific fashion e.g. by imposing packet loss, echo or delay. Subjects could then either be given a subset of terms to choose from, or asked to generate words that describe what they hear or see. This approach should allow the multidimensionality of quality to be further understood. Clearly, the degree to which these factors matter will depend on different user groups and tasks, so once the descriptors of the factors have been verified, the relative impact of each would need to be assessed with different user groups performing different MMC tasks.

In the meantime, the words that have been identified as common descriptors of quality degradations have been incorporated into a web-based evaluation form for participants in small group tutorials under the PIPVIC (Sasse et al., 1998) project (see Appendix K). Ultimately, this data could be matched with the objective statistics from a session to understand better the relationship between the level of packet loss and what the resulting audio and video sounds and looks like to the listener.

Chapter 9: Conclusions and recommendations

The purpose of this chapter is to present and reflect upon the conclusions and implications of the research conducted in this thesis, to detail how HCI researchers and others can benefit from the findings, and to propose future research topics and directions.

The research presented in this thesis is novel in that it investigates networked desktop videoconferencing across a range of different user groups and applications in both the field and in laboratory experiments. The incremental findings have allowed a vastly improved understanding of the factors that impact perceived quality in videoconferencing environments to be built up. A new subjective assessment methodology has been outlined and shown to be reliable, from which both network designers and HCI practitioners stand to benefit.

9.1 The research problem restated

Videoconferencing across networks is growing in popularity, both in research communities and with commercial service providers, as it offers the opportunity for work and leisure collaboration across great distances. Since the research reported in this thesis started, Internet usage and tool development have both grown enormously (see http://livinginternet.com/ for a comprehensive overview of this growth). Interest from remote gaming and the home entertainment industries, together with the distance education, military and academic research communities will ensure that uptake of digital networked conferencing and communication continues.

However, growth is not enough: the technology must be proved viable from the *end user's* point of view, meeting a desired level of task performance and user satisfaction, and without causing too much cost, in terms of fatigue, irritation or stress. At present, there is no established method for delivering this proof. There is, therefore, an urgent requirement for reliable techniques to measure the subjective quality of the audio and video delivered in the applications developed, and to link them to the objective QoS factors which can be applied to network services. The ultimate aim is to use these techniques to define the level of quality required to complete a particular videoconferencing task successfully, and the point of diminishing return, i.e. the threshold beyond which increasing objective quality (and hence bandwidth) does not increase user performance or satisfaction. This aim can only be achieved by carrying out HCI research into the *subjective* effects of objective videoconferencing conditions, i.e. assessing perceived quality (in the context of a specific task).

The thesis has summarised a growing body of evidence which indicates that existing speech and video quality rating scales - which were developed to assess quality for very different types of networks and applications (see section 2.3) - may be widely used, but results obtained with them are imprecise at best, and may be thoroughly misleading at worst. These scales should not be used to determine subjective quality required by networked multimedia applications developed today, or used to infer bandwidth or other QoS requirements for network services.

Instead, the thesis has outlined a new approach to assessing the speech and video quality delivered in networked videoconferencing. The approach acknowledges that there are multiple factors that influence users' perception of multimedia speech and video, and the research has begun to establish which audio-visual factors have the largest impact on overall perceived quality. Principal factors, including packet loss, repair scheme, volume differences and echo, have been investigated using a variety of established and new methods. The new methods offer a better means of capturing and measuring the multidimensionality of quality, and assessing the impact of different levels of these dimensions on subjective quality. For the first time it will be possible, using the new methods developed in the course of the research, to establish the critical quality boundaries (minimum and maximum quality thresholds) for a particular dimension in the context of a particular task, and hence build up a taxonomy of audio-visual requirements[46].

9.2 Contributions of the thesis

The laboratory-based research that was undertaken in this thesis was informed by observations made in and from MMC field trials. The field trials were fundamental to the lab-based research since they enabled the identification of problem aspects which could then be explored and verified in experiments.

The application of existing qualitative assessment methods to this new form of communication was investigated, and it was found that many of these methods are lacking in suitability for the characteristics of MMC speech and video. A large number of studies were conducted addressing the areas of speech intelligibility and perceived speech quality, and perceived video quality. The main findings for these areas have been divided into substantive and methodological contributions, and are summarised in the following sections.

[46] Steps towards establishing a taxonomy of requirements for different tasks are being furthered in the EPSRC-funded project ETNA (Evaluation Taxonomy for Networked multimedia Applications) between UCL and Glasgow University. The project aims to produce a set of audio-visual guidelines for different real-time

9.2.1 Speech intelligibility and quality

9.2.1.1 Substantive contributions

The results from the speech intelligibility and speech quality studies reported in Chapter 5 showed that, under conditions of packet loss in the audio channel, speech perception benefits from a packet loss repair method that substitutes noise rather than silence in the place of the missing packet. It is possible to attain a high level of speech intelligibility (up to 30% packet loss) by employing a packet loss repair scheme such as packet repetition or LPC redundancy. The findings from this research were instrumental in the iterative development of the audio tool **RAT** (Hardman et al., 1995; 1998) - as a direct result of the research, **RAT** offers various different forms of packet repair, in contrast to the other most common Mbone audio tool, **vat** (Jacobson, 1992), which offers no such facility.

The ReLaTe (section 5.1), distance learning (section 6.2) and PIPVIC (section 8.1) field trials all demonstrated that MMC technology can be used to support distance learning effectively. However, in these field trials the importance of reliable audio quality was highlighted.

In a continuous speech quality rating task and an interactive MMC conversational task (sections 7.1 and 7.4), it was found that 13-15% loss is the critical level at which users will object to, or request better, quality (where the packet repair mechanism is silence substitution). The finding that maximum quality is not requested (for this conversational MMC task) is somewhat counter-intuitive and should be investigated further (see section 9.3), since such a trait has important implications for proposed 'charging for quality' mechanisms as the networking community moves towards bandwidth reservation protocols (see section 1.2). When there is a 'budget' incentive provided in order to conserve resources, most participants will tolerate up to 20% packet loss. However, differences in individual preferences are large, which again has implications for proposed charging mechanisms.

9.2.1.2 Methodological contributions

In Chapter 5, it was argued that intelligibility testing should be carried out with sentence rather than word material, in order to approximate real world situations more closely (in terms of longer speech bursts), but it has also been argued that assessing speech intelligibility is not in itself sufficient: speech quality must also be assessed.

multimedia tasks and applications undertaken by different user groups. Applications developers and service

Subjective speech quality has most commonly been assessed via category rating scales such as the ITU 5-point quality scale, giving rise to the Mean Opinion Score (MOS). This quality scale was used in early investigations into speech quality in this thesis research (see sections 5.2, 5.4 and 5.5), but these studies investigated packet loss (and repair) at constant loss rates over *short* speech samples, rather than the longer, more variable conditions which characterise real-world MMC environments. Using the scale in real-world conditions (section 5.1) led to concern over the post hoc nature of the quality rating: when quality fluctuated, which part of the conference carried most weight in forming the overall quality perception? In addition to these points, concern was expressed over the labels of the categories used on the scale: Excellent, Good, Fair, Poor, Bad are not appropriate for the level and type of degradation experienced in speech over the Internet, since the quality encountered will rarely be described as Excellent. The scale has also been criticised by other researchers for the following reasons:

1. Although treated as such, the scale is not an interval scale as represented by its 5 qualitative labels (Excellent, Good, Fair, Poor, Bad) (Jones and McManus, 1986; Teunissen, 1996; Virtanen et al., 1995). Fair, for example, is not indicative of a midpoint to most people.
2. Use of the 5-point scale leaves the experimenter ignorant of the subject's perspective and rationale for positioning on the scale (Knoche et al., 1999 - see section 8.1.2.2).
3. Quality is a *multi*dimensional phenomenon (Preminger and Van Tasell, 1995; Virtanen et al., 1995), and means are required by which the dimensions that have the largest effects can be identified.

The first development in addressing these problems was to explore the use of a polar continuous quality scale (section 6.1). The scale was shown to be a reliable means of measuring perceived quality, since users are consistent in their use of it, and the rating trend follows the same slope as that attained with the MOS. In addition, subjects liked using the polar scale for the flexibility it allowed. The major benefit of the scale is that it can be adapted easily to address different quality dimensions and purposes e.g. *overall adequacy* (section 6.2).

Although this scale dispensed with the problem of the quality labels, it remained difficult to interpret whether any particular points of the fluctuating material presented in real-world conditions were more influential than others, since it was an overall quality rating that was gathered at the end of a certain time period. Asking subjects to give a quality rating at the end of a conference, the length of which can be anywhere between 30 minutes and two hours, means that the rating must

providers will be able to apply the taxonomy to infer objective QoS requirements for particular applications.

necessarily be of a cumulative nature: subjects must take into account what may have been a large range of qualities throughout the session. Since research into video quality has shown that factors such as the severity and recency of an impairment (Aldridge et al., 1998; Seferidis et al., 1992) can influence perceptions, it could be dangerous to make system design decisions on the basis of these overall quality judgements: people may form an overly negative (or positive!) impression depending on when and where quality impairments are experienced. The need for a method of assessing perceived quality in a continuous manner was therefore identified, to avoid the problems with one-off rating at the end of a session, and to allow investigators to identify precisely *where* quality becomes objectionable and unacceptable.

The development of a continuous subjective quality measurement device, QUASS, enables the correspondence between the objective and subjective quality of different variables to be investigated (section 7.1). Although similar in concept to the new ITU-R recommended method SSCQE (ITU-R BT.500-8), QUASS differs in that it is a software tool, and does not employ the quality scale category labels that have proved to be invalid. Since it is a software tool it can be modified easily to provide additional functionality, which was explored in the second QUASS study (section 7.4), where movement of the slider actually *controlled* the objective quality level (in this case packet loss) that the participant received. It was found that subjects are able to rate perceived quality continuously using QUASS, but in agreement with the ACTS TAPESTRIES (1997) project findings, the application of the continuous rating method did not work well in real-world environments. However, the modification of QUASS to enable the conference participant to control a received quality dimension is an important step toward future real-world applications, especially in light of the findings that individual quality preferences can range widely, particularly when pricing via a 'budget' is involved (see section 7.4.4.4).

Although the QUASS method has only been applied to audio quality dimensions so far, there is no reason why it should not be applied to video quality dimensions in future research. (In fact, QUASS would be particularly well suited for use in studies investigating quality perception of new home entertainment applications such as videos delivered from a server – see section 9.3). Other video findings are discussed in section 9.2.2.

Since QUASS did not work well in interactive tasks, it was necessary to adapt the polar scale for use in real conferences. This was achieved by asking conference participants to complete a web-based rating form at set point in the conference. Since the polar scale could not be easily represented in HTML, the scale was numbered, between 1 and 100, and participants were asked to select a number that they felt represented their opinion of the quality. Like the polar scale, it was found that a 1-100 scale was easy to use and produced consistent results from the users (see section 8.2). Using the scale on a web form meant that time-stamped opinions of the quality could be gathered, and these could then be matched with objective statistics (level of packet loss received) logged during the conference. When these results were analysed, it was found that many of the audio problems that were cited were not caused by network conditions, rather by end-user behaviour and hardware used.

These findings led to the study reported in section 8.2, where non-network impairments were compared to network impairments. The experiment clearly demonstrated that the perceived quality of network audio is not primarily affected by the level of packet loss observed in the PIPVIC field trial (provided that a packet loss repair method such as packet repetition is in use). Volume discrepancies, poor quality microphones and echo have a greater impact on the user, meaning that it is possible to have perfect transmission from a network viewpoint, but still have poor quality audio from a user's viewpoint. The solutions envisioned are mainly a case of raising and improving user awareness, both of what the problem is, and how to solve it. These can be low-cost solutions – a huge amount of people-support should not be required once audio tools are better set up to support non-expert users (see section 9.3).

9.2.2 Perceived video quality

Although taking a secondary role to audio quality issues in this thesis, perceived video quality was also investigated to a certain extent in the research. In section 4.5.2, it was argued that assessing the perceived quality of MMC video at its present level of sophistication (2-12 fps) is essentially meaningless unless the video is accompanied by audio in the context of a task. Studies carried out under this format were reported in sections 5.1, 5.7, 6.2 and 7.4.

9.2.2.1 Substantive contributions

In agreement with previous research (e.g. Tang and Isaacs, 1993; Gale, 1991), the studies reported in this thesis found that there are clear (subjective) advantages to including a video channel in real-world videoconferencing environments, even at the low quality levels common today. Participants greatly appreciate being able to see their conversational partners. Subjects in both the ReLaTe field trial (section 5.1) and the distance learning course (section 6.2) valued the video channel, and if manipulations of the channel quality (image size and frame rate) were not proven to have a beneficial impact on perceived quality, neither were they shown to be a negative influence. The impact of the video channel, in these distance learning environments, seems to lie more in the sense of shared presence it creates, allowing participants to put a face to a name, and its use as a means of common reference (whether physical or emotional).

In a lab-based conversational MMC task (section 7.4), the presence of a low-quality video channel did not affect the level of audio requested. (As with the remote learning field trials though, the impact of video quality is likely to be more important where the video channel is more critical to the task.)

The audio-video synchronisation study (section 5.7) used a modified version of **vic**, which delivered a set frame rate, either synchronised with the audio stream or not. The findings from this study suggest that low frame rate synchronisation with the audio channel can offer perceptual benefits from 6 fps upwards. This result is very encouraging, for two reasons. Firstly, because it is unlikely that full-motion video will be transmitted in MMC in the foreseeable future, and secondly because conference participants are willing and inclined to accept lower quality when being 'charged' (section 7.4).

9.2.2.2 Methodological contributions

In assessing the video channel, it was not straightforward to know what to ask/look for. It was assumed that the video channel would not be the main subjective focus of most conferences, and due to the low quality it did not make sense to ask subjects to rate the overall image quality. This led to the application of the polar scale in searching for an 'overall adequacy' rating for the video (section 6.2). As with the audio studies, the problem with this was not knowing which part of the stimulus carried most weight in forming the perception. (Employing QUASS in determining the impact of video quality factors would be better, but first it is necessary to gain a measure of control over the objective video quality variables: this cannot be done using the video tool **vic** as it now stands.) In addition, the quality of the *audio* undoubtedly had an effect on perception of the video channel. Aldridge et al. (1998) reported that there was no difference in results when telling people to rate image quality or overall quality when an audio channel was included, but those studies involved high quality audio. It is far more likely (and was borne out by subjects' comments) that when the audio channel is itself of comparatively low quality, judgements of the video channel will be negatively affected.

9.2.3 Other substantive and methodological contributions

9.2.3.1 Conducting research in the field

The interleaved data-gathering approach that is advocated in this thesis involves both the field and the laboratory. It is only by conducting research in the field that MMC applications can be properly assessed, since it is only in this environment that the factors that affect perceived quality can be identified. The ReLaTe field trial (section 5.1) demonstrated that the technology is feasible for an application as demanding and complex as language teaching, but highlighted the fact that poor audio quality was the most disruptive factor to end-users. This led to a series of experiments investigating the impact of different packet loss repair schemes on different levels of packet loss. The findings from the ReLaTe pilot trial also led to the development of an easy-to-use integrated interface which has been used in many subsequent MMC applications. The communications and networks course (section 6.2) also demonstrated the capacity of MMC to meet complex educational demands.

From conducting research in the field, a number of points became clear. Firstly, in the ReLaTe project (section 5.1) it was seen that users *adapt* to typical MMC quality levels occurs over time. However, this should not be interpreted as a reason to avoid improving quality, since it is more likely that a user will be put off by initially encountering poor quality than be willing to 'stick it out' until adaptation occurs (unless the benefit to the user is really obvious).

Concern was voiced over the post-hoc quality ratings that are gathered at the end of long conferencing sessions, but it was found that participants in real-world conferences cannot effectively rate perceived quality continuously whilst performing an interactive MMC task (section 7.3). However, based on encouraging findings using QUASS in a passive listening study (section 7.1), the technique has potential to be applied in passive real-world MMC tasks such as remote lectures and home entertainment (e.g. movies), rather than interactive field trials. An effective compromise between these two options was explored in the PIPVIC project, where participants were asked to complete ratings on a web form at 3 set points during a conference. By matching the time stamps on the received web forms with the objective network logs, it became possible to get a clearer picture of the impact the network profile had (or indeed did *not* have) on end-user perception.

The use of this method in this field trial underlined the importance of the hardware used in conferencing. For example, different types of headsets can affect the quality perceived by a participant in terms of the feedback that may be caused through sound 'leaking' from the microphone. Heavy over-the-ear headphones may guard against background noise becoming intrusive, but in extended conferences these can become uncomfortable for the wearer, and the inability to hear one's own voice can become distinctly unnerving (see section 5.1.3.1). Different types of workstation also have an impact on perceived quality. The processing power of the machine, the operating system it is running, its type of soundcard and the camera used will all have an impact on the delivered audio and visual quality. Observations of this type informed the design of the experiment reported in section 8.2.

9.2.3.2 Investigating the vocabulary of quality

A subsidiary issue throughout the research was that of the vocabulary that participants use to describe the quality that they hear/see. Given the difficulties of hearing/seeing exactly what participants experience, especially in field trials, it was necessary to try to gain an understanding of what exactly is *meant* (in terms of objective quality) by the words they use to describe their perceptions. There are two main reasons why this is important:

1. Observations from field trials and experiments suggest that users of MMC audio and video do not necessarily use the same vocabulary as technical 'experts' to describe the degraded audio and video that they experience. If this is the case, then there is a danger that misconceptions can and will arise between novice users and designers of multimedia applications as the two groups converse in design and evaluation processes. This potential gulf needs to be investigated, and corrected if necessary.
2. The multidimensional nature of quality has been asserted throughout this thesis, but how might these different dimensions be described? The relationship between the objective quality dimension and the way in which the subjective experience is described needs to be determined.

Therefore investigations were undertaken to identify the groupings of terms that describe what the stimulus actually sounds or looks like, so that a link between different types of degradations and the quality perceptions they create can be established. Collating a wider range of suitable descriptive terms will also be useful in ultimately designing new quality rating methods and scales. Once the key quality dimensions and the related vocabulary have been identified, subjects could, for example, rank the collected terms using the graphic scaling method to generate more meaningful labelled rating scales for MMC speech and video quality dimensions.

In Chapter 8 it was shown that novice MMC users do not use the same descriptive vocabulary as 'experts', and also that different terms can be associated with different types of degradations e.g. packet loss, echo, irregular image update. Different quality dimensions have been hypothesised, and these should be verified through further research. This methodological approach has been implemented in the guise of an evaluation form to be completed by participants at the end of a conference (see Appendix K). Web-based assessment forms such as this would make it far easier to gather immediate subjective quality impressions from distributed conference participants.

The substantive findings provide recommendations mainly for tool designers and network providers, whilst the methodological findings provide recommendations for HCI researchers. These will be presented in section 9.2.5.

9.2.4 Shortcomings of the thesis research

It is recognised that the research undertaken could have been improved in a number of ways.

One shortcoming of the studies reported in this thesis is that there was no common task to make it easier to generalise results across studies. It would have been sensible to develop/identify a key task that could be used in future experiments. Steps have been taken to redress this in on-going user cost studies (measuring physiological responses to different quality levels) at UCL, where the key task is that of remote interviewing (see Wilson and Sasse, 2000a).

It was stated in section 5.2 the reasoning behind the levels of packet loss that were explored in the early lab-based studies, but with respect to generalising results, it would have been good to have been able to investigate more realistic packet loss patterns. However, having a set, as opposed to variable, level of loss at least allowed the establishment of benchmark levels of perception.

It would have been interesting, and arguably more valid (since this became the default setting in **RAT**), to perform the QUASS studies using packet repetition as a repair mechanism rather than silence substitution. However, it was principally the *method* that was under investigation. Likewise time contraints did not allow investigation of other repair methods/codecs.

Field study attendance was not as high as one would have liked, in particular in the distance learning course (section 6.2). This made it hard to draw firm conclusions, due to a lack of numbers. It is hard to see how this problem can be overcome in future non-compulsory course environments – perhaps the provision of incentives for participation should be explored (bearing in mind always the chance of inducing a Hawthorne effect...).

Also in field trials, there is a clear need to develop better methods of logging hardware/settings in use. This needs to be undertaken at all end sites, preferably automatically, since even one participant forgetting to record their details makes it hard to draw overall conclusions about the ensuing reports of quality. The field studies in this thesis sometimes suffered from a lack of detail of this kind.

Finally, the criticism could be levelled that the results from the packet loss studies reported in this thesis will become irrelevant with the incremental increase in bandwidth that is occurring. For example, with the deployment of ADSL giving users 1.5 Mb bandwidth at home, quality as defined by bandwidth will no longer be an issue, and packet loss should become far less common. However, the negation of packet loss as a problem will serve only to increase the importance of the other quality aspects brought to light in Chapter 8. As Doerry (1995) observes, "*the development of the current crop of computer-mediated environments has largely been driven by and oriented around the technical challenges posed by distributed interaction. By focusing on issues like bandwidth, frame rate, colour depth, and sampling rate, these projects make the tacit assumption that 'more is better'; that higher bandwidth and better resolution inevitably lead to a more robust and efficient interaction. Clearly, this approach places form before function, ignoring functional utility of the environment in favour of abstract parameters.*"

On the other hand, although bandwidth to the home and office may become less of a problem, bandwidth to the mobile user is unlikely to increase at such a rate. With the new generation of WAP phones being launched, it is likely that a whole catalogue of new perceived quality issues will be brought to the fore.

9.2.5 Recommendations to HCI researchers and network designers

It was stated in section 1.5 that the findings of the thesis should benefit both network and tool developers and HCI evaluators wishing to assess the subjective quality of speech and video delivered in MMC environments. Key recommendations to both groups, based on the research findings of this thesis, are presented below.

Recommendations to developers:

- Use **RAT** not **vat** – all the studies reported show the benefits of repairing audio packet loss with sound, not silence. Only **RAT** offers this facility.
- Perceived speech quality cannot be predicted solely on the basis on speech intelligibility. It is possible to achieve high intelligibility at the expense of low quality. Since user

satisfaction will ultimately determine the success of an application, taking into account perceived *quality* results is vital.

- 13-15% audio packet loss repaired with silence substitution can be tolerated in interactive conversations when all speakers are native. This figure is likely to be higher when a phonemic restoration scheme such as packet repetition is used. (However, the *user cost*, in terms of physiological stress for tolerating these loss levels this has not yet been established - see section 9.3)
- Conference participants do not always request the highest quality, even when they can have it without paying – so designers may not need to supply it.
- There are *subjective* benefits to video, but it may not be necessary to expend many resources on its quality. For certain tasks where the video link is not critical to the goal of the activity, there may be no advantage to increasing frame rates above about 5 fps (see section 6.2). In addition is has been found that low quality video does not affect the level of audio requested in conversational tasks. However, this interpretation is tempered by recent findings (Wilson and Sasse, 2000a) that although subjectively users may not notice the difference in frame rates, objectively (as measured by blood volume pulse, heart rate and galvanic skin response) they do. This finding underlines the importance of considering all three HCI aspects: task performance, user satisfaction, and user cost (see section 2.1.2).
- Individual preferences can and should be catered for in design (but not at the expense of help for beginners...)
- Users and designers will not necessarily speak the same language when describing quality impairments – the onus is on designers to learn the language of the users.
- Designers should seek to produce tool diagnostics to help the users help themselves. The thesis has demonstrated clearly that subjective quality in MMC environments is not just a bandwidth issue: many factors have an impact on perceived quality, meaning that perceived quality may be improved without resorting to increasing bandwidth. For example, by ensuring that the silence suppression in an audio tool does not cut out the ends of speech bursts, or that headsets do not leak sound into the microphone, perceived quality could be greatly improved. Since many users may not be aware of the problems they are causing other participants in a conference, or know how to fix the cause, designers of future network tools should consider incorporating diagnostics in audio and video tools (see section 9.3).

Recommendations to HCI researchers evaluating existing systems or prototypes:

- Do not use the ITU 5-point quality scales to measure perceived quality, since they are invalid and do not capture the multi-dimensionality of quality (see section 4.11). These scales have been designed with the aim of determining whether a degradation is detectable or not, rather than determining whether a task can be performed successfully or not under certain conditions[47]. Instead, use either the 1-100 or polar scales – both are used consistently by subjects.
- In addition to attaining quality ratings using these scales, HCI researchers should always *ask the user* to explain why a rating was awarded, leading to a far richer data collection. Participants in both field and lab studies should be encouraged to describe the quality degradations experienced in their own words, or be provided with rating scales labelled with descriptive terms (if these have been validated against different types of degradations).
- Use the field to inform the lab – through observation and user reports in the field, the most critical factors of the MMC environment can be identified, then tested under more controlled settings in the lab.
- If possible, gather time-stamped impressions in the field (as in the PIPVIC field trial) which can be correlated with objective network statistics. It is possible to perform much (and valuable) post hoc quantitative analysis if both subjective and objective data are collected over long periods.
- Set the task: many of the studies presented in this thesis were of an exploratory (and therefore somewhat artificial) nature, but it is important for future researchers to consider the nature of the experimental task that is set. Since it can be assumed that different applications will have different subjective quality requirements, it is important that the assessment task models the type of interaction for which the application is being designed.
- It is strongly recommended that evaluation should be undertaken using a *holistic, context-based approach.* Researchers should forego 'basic' tests such as speech intelligibility tests and instead ground research in *adequacy of quality* for task. The assessment should address the *adequacy* of the quality delivered in order for the task to be performed satisfactorily. This measure can be investigated explicitly (by asking the user, or employing 'expert' judgements, such as was used in the ReLaTe project), or implicitly by investigating

[47] There is in fact a means by which degradation detection may again become a useful technique within the research area of establishing required audio and video quality guidelines. Bouch and Sasse (2000) have identified how important stable quality is to users. As a result, QoS researchers are now debating how to manage quality in a way that is more acceptable to users (especially paying ones). One proposed method is by smoothing out abrupt changes i.e. increasing or lowering quality gradually so that the difference cannot be perceived (assuming that the quality is basically good enough for the task). In order to establish that the quality change really is not visible, degradation detection studies may be called for (see section 4.4.1 for a description).

required quality via QUASS (where it is used to control the delivered level of a quality dimension).

- Record as many details about the set-up and user group as possible. Olson and Olson (1997) have argued that there is a need to develop a framework against which past and future video-mediated communication findings can be evaluated. Apparent contradictions in the literature may be partly explained by inadequate detailing of set-up and conditions. By providing explicit information of this nature, and undertaking multidimensional research of the type advocated in this thesis, an important and valuable taxonomy for audio-visual quality requirements across different tasks and networks can be derived. Given the findings reported in section 8.2, this requirement seems especially pertinent.
- Use QUASS. The development of a dynamic rating mechanism, QUASS, has shown that MMC participants are able to rate quality continuously in passive MMC situations, and also that MMC participants can *control* the level of quality they require in interactive settings. This enables the impact of different conditions to be meaningfully assessed
- If possible, measure the *user cost* in terms of physiological stress to difference levels of quality (see next section).

9.3 Future work

The methods that have been developed in the course of this research open the door for many future studies.

In Study 11 (section 8.2) it was shown that end-user and hardware degradations can have a more negative effect on perception than network problems. There are a number of further steps that should be taken in this line of research. Firstly, it is important to quantify the exact levels of degradation imposed, in order that the study can be replicated accurately by other interested researchers. Secondly, the experimental conditions presented in the study should be recreated in an *interactive* task environment, where people engage in active conversations as opposed to passive listening. It can be hypothesised that the effects of the factors investigated in the initial study will be altered in this interactive setting. It can be predicted that the effect of echo, for example, will have an even more negative effect when a subject is trying to engage in a conversation with another person, but keeps hearing his own voice fed back to him. Another important aspect to investigate will be the effects of the interaction *between* different impairment types, for example one person with a bad microphone conversing with someone speaking too loudly. Again, presenting this scenario as an interactive experiment is likely to lead to different results than would be obtained in a passive listening environment.

The potential for future work using the QUASS methodology is good. A first step would be to compare the quality requirements from different speech encoding methods and repair schemes to determine which combination offers best performance from a network and user perspective. However, investigations need not be restricted to audio factors. In this thesis QUASS has only been used in studies investigating the effects of audio packet loss, but the tool could be adapted with ease to investigate the effects of other factors on perceived quality, e.g. video frame rate, lighting, delay and background noise, as well as the effects of different digitising techniques such as MPEG. In this manner a model of the impact and weight that different factors have on perceived quality, in different MMC tasks, could be built up. An effective use of QUASS could be in assessing new home entertainment applications, e.g. movies or news broadcasts delivered from a video server, since the tool has been demonstrated to be particularly useful in passive task settings (section 7.1), and continuous rating has been shown to be both reliable and sensitive in long sessions (see section 4.9.1.1).

QUASS can also be used in viewing *recorded* conferences. By marrying the subjective rating with objective statistics (packet reception reports gathered using **rtpdump**[48], or received frame rates logged by the video tool), the impact of different factors can again be investigated. As was discussed in section 7.3, continuous rating by the end user of the quality delivered in real-time interactive conferences using QUASS causes task interference. The development of web-based forms (which become time-stamped when submitted) to be completed at set intervals in a conference has allowed a closer match between objective conditions and subjective opinions to be forged. Another concept that should be explored is to provide a participant in a real conference with a 'dissatisfaction' button on the interface which, when pressed, will record the objective statistics for that moment, plus the statistics for a certain time interval before the button was pressed. This will allow a picture to be built up of when and why quality becomes objectionable to a participant. Of course, this method will only allow correlation to be made with packet loss (since packet reception reports are the information recorded by **rtpdump**) and frame rates, but by recording the conference and playing it back to the same participant, he/she should be able to cast light as to why the button was pressed, if not for these reasons.

The second QUASS study (section 7.4) demonstrated two interesting findings: that conference participants do not necessarily request the highest quality, and where a pricing mechanism is

[48] Software written by Henning Shulzrinne, available from
http://www.cs.columbia.edu/~hgs/rtp/rtpdump.html

involved, participants often request a quality level that in other studies has been found to be less than acceptable. These findings should be explored in longer conferences, with different tasks and user groups. It is not clear how long people would be willing to put up with poor quality, nor what the *user cost*, in terms of physiological stress, might be.

A fully comprehensive HCI approach should take into account user cost as well as user satisfaction (Wilson and Sasse, 2000a). New research at UCL is investigating the effect of different media quality on user stress, as measured by blood volume pulse, heart rate, and galvanic skin response. By placing sensors on the fingers of subjects, it is possible to monitor their physiological responses to different types of impairment, in order to assess the relationship between expressed opinion and user cost. Data of this type was gathered in the study reported in section 8.2. The results were intriguing because a discrepancy between subjective and objective results was found: it was found that users are more adversely affected by the **bad mic** condition than the subjective rating results would suggest (Wilson and Sasse, 2000b). In a previous study by Wilson and Sasse (2000a) looking at the effect of video frame rate on both subjective ratings and user cost, it was found that viewers did not notice (subjectively) the change in frame rate from 5 to 25 fps, but that their physiological readings registered the change: 5 fps was significantly more stressful than 25 fps. These types of findings emphasise the importance of carrying out more research of this nature in the future, combining subjective ratings with measurements of user cost – it is clear that subjective results alone do not provide a wholly accurate picture.

Further refinement and application of the descriptive vocabulary results should be pursued. Semantic groupings according to different types of degradation (e.g. packet loss, echo, delay and volume problems) can be further explored in controlled settings, with the goal of developing new labelled scales to assess the subjective perception of different quality dimensions represented by different objective conditions.

By further analysing how people describe different types of degradation, it should also be possible to provide improved fault diagnosis to novice users of MMC technology. For example, a help menu on an audio tool should provide a list of problems described in terms that users most commonly generate, such as 'fuzzy' and 'buzzy' which, as has been shown, are related to a specific type of packet loss repair (packet repetition). The user could search down this list for the terms that describe his or her problem, then follow the solution suggested (e.g. change the receiver-based packet loss repair method to another, such as pattern matching).

There is perhaps less that a user can do about someone else's bad microphone, other than tell them that they sound 'muffled', 'distant', or like they're 'on the phone'. One solution would be to reflect the user's audio as heard by other participants, since at present the user cannot hear what he/she sounds like. It is proposed that developers design a tool to perform an expert system style diagnostic of a user's speech stream and point to likely causes of problems. Users could be required to record sample sentences – as in a voice recognition package for word processing, for example – and only be allowed onto the network once the quality of the sample files is matched or recognized as providing satisfactory quality.

The key problems highlighted in Study 11 also provide a strong case for the inclusion of aspects such as automatic gain control and reliable echo suppression in Internet audio tools. These are already present in the most recent version of **RAT** (version 4), but they are optional settings – users need guidance on when to apply them.

As discussed above, physiological responses do not necessarily agree with subjective responses - perhaps users are not always capable of making sensible quality level decisions on their own. (This suggests that physiological measurements will be particularly useful and valuable in determining required and acceptable quality levels in long-term field trials.) It is important to establish a relationship between the price that users are willing to pay, the subjective quality delivered, the adequacy of that quality for task performance *and* the user cost. Future work should therefore continue to gather physiological data, to gain a better understanding of the user cost of different types of degradations, and the relationship between user cost, subjective opinion, and task performance. Physiological data should be gathered in lab-based studies, but also in the field: as the research presented in this thesis has illustrated, it is only by conducting research in both arenas that the impact of different quality degradations on the user can be accurately determined.

References

ACTS TAPESTRIES (1997): Acceptability studies in selected areas of audio-visual communications. ACTS Project AC055, Deliverable R/003/b2.

Ainsworth, W.A. (1976): Mechanisms of Speech Recognition. Pergamon Press.

Aldridge, R., Davidoff, J., Ghanbari, M., Hands, D. and Pearson, D. (1995): Measurement of scene-dependent quality variations in digitally coded television pictures. IEE Proceedings - Vision, Image and Signal Processing, 142(3), 149-154.

Aldridge, R.P., Hands, D.S., Pearson, D.E. and Lodge, N.K. (1998): Continuous assessment of digitally-coded television pictures. IEE Proceedings - Vision, Image and Signal Processing, 145(2), 116-123.

Anderson, A.H., Mullin, J., Newlands, A., Doherty-Sneddon, G. and Fleming, A. (1994): Video-mediated communication in CSCW: Effects on communication and collaboration. Presented at Workshop on VMC at CSCW '94, Oct. 20-25, Chapel Hill, NC.

Anderson, A.H., Bard, E.G., Sotillo, C., Newlands, A. & Doherty-Sneddon, G. (1997a): Limited visual control of the intelligibility of the speech in face-to-face dialogue. Perception and Psychophysics, 59(4), 580-592.

Anderson, A.H., O'Malley, C., Doherty-Sneddon, G., Langton, S., Newlands, A., Mullin, J., Fleming, A. and Van der Velden, J. (1997b): The impact of VMC on collaborative problem solving: An analysis of task performance, communicative process, and user satisfaction. In Video-Mediated Communication, ed. K.E. Finn, A.J. Sellen and S. Wilbur, LEA, NJ, 133-155.

ARC report: On retient mieux ce que l'on entend bien. Report from the Association de Recherche en Communication et Formation Langues, Paris. (No date on publication).

Argyle, M. (1990): Bodily Communication. London, Routledge.

Barber, P. and Laws, J.V. (1994): Image quality and communication. In Multimedia Technologies and Future Applications, ed. R.I. Damper, W. Hall and J.W. Richards, Pentech Press, London, 163-178.

Beerends, J.G. and Stemerdink, J.A. (1994): A perceptual speech-quality measure based on a psychoacoustic sound representation. Journal of the Audio Engineering Society, 42(3), 115-123.

Bellotti, V. (1988): Implications of Current Design Practice for the Use of HCI Techniques. In People and Computers IV. Proceedings of the Fourth Conference of the British Computer Society HCI Specialist Group, ed. R.Winder and D.M. Jones, Univ. of Manchester, 5-9 Sept. 1988. Cambridge: CUP.

Blokland, A. and Anderson, A.H. (1998): Effect of low frame-rate video on intelligibility of speech. Speech Communication, 26 (1-2), 97-103.

Bolot, J., Crepin, H. & Vega Garcia, A. (1995): Analysis of audio packet loss on the Internet. Proceedings of NOSSDAV, 163-174, Durham, NH.

Bouch, A. (1997): Redesigning the RAT user interface. MSc thesis, Department of Computer Science, University College London, University of London.

Bouch, A. and Sasse, M.A. (1999): Network quality of service: What do users need? Proceedings of the 4th International Distributed Conference (IDC '99), 21-23 Sept., Madrid, Spain.

Bouch, A. and Sasse, M.A. (2000): The case for predictable media quality in networked multimedia applications. In Proceedings of MMCN'2000, San Jose, CA, January 24-27, 2000, 188-195

Bouch, A, Watson, A. and Sasse, M.A. (1998): QUASS - A tool for measuring the subjective quality of real-time multimedia audio and video. Poster presented at HCI '98, 1-4 September 1998, Sheffield, England.

Boyle, E., Anderson, A. and Newlands, A. (1994): The effects of visibility on dialogue and performance in a cooperative problem-solving task. Language and Speech, 37 (1), 1-20.

Brown, G., Anderson, A.H., Yule, G. and Shillcock, R. (1984): Teaching Talk, Cambridge University Press.

Bruce, V. (1996): The role of the face in communication: implications for videophone design. Interacting with Computers, 8(2), 166-176.

Carroll, J. M. & Campbell, R. L. (1986): Softening up Hard Science: reply to Newell and Card. Human Computer Interaction, 2, 227-249.

Carroll, J.M. and Campbell, R.L. (1989): Artifacts as psychological theories: the case of human-computer interaction. Behaviour and Information Technology, 8, 247-256.

CCITT (1987): Handbook on Telephonometry, CCITT.

Clark, L. (1997): vic usability study. Internal Note 97/1, Dept. of Computer Science, University College London.

Clark, L. and Sasse, M.A. (1997): Conceptual design reconsidered - The case of the Internet Session Directory Tool. Proceedings of HCI'97, Bristol, August 12-15, 67-84. Springer.

Crowcroft, J., Handley, M.J. and Wakeman, I. (1999): Internetworking Multimedia. UCL Press.

Deering, S.E. (1988): Multicast routing in internetworks and extended LANs. SIGCOMM Symposium on Communications Architectures and Protocols, (Stanford, California), 55-64, ACM, Aug. 1988.

Dix, A., Finlay, J., Abowd, G. and Beale, R. (1998): Human-Computer Interaction. Prentice Hall.

Dodd, B. (1977): The role of vision in the perception of speech. Perception, 6, 31-40.

Doerry, E. (1995): Evaluating distributed environments based on communicative efficacy. CHI '95 Conference Companion, 47-48.

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

Fluckiger, F. (1995): Understanding Networked Multimedia, Prentice Hall.

Frowein, H.W, Smoorenburg, G.F., Pyters, L. & Schinkel, D. (1991): Improved speech recognition through videotelephony: experiments with the hard of hearing. IEEE Journal on Selected Areas in Communication, 9, 611-616.

Gale, S. (1991): Adding audio and video to an office environment. In Studies in Computer Supported Cooperative Work, ed. J.M. Bowers and S.D. Benford, North-Holland, 49-62.

Gili Manzanaro, J., Janez Escalada, L., Hernandez Lioreda, M., Szymanski, M. (1991): Subjective image quality assessment and prediction in digital videocommunications. COST 212 HUFIS Report.

Goodman, D.J., McDermott, B.J. and Nakatani, L.H. (1976): Subjective evaluation of PCM coded speech. Bell System Technical Journal, 55 (8), 1087-1109.

Gruber, J.G. and Strawczynski, L. (1985): Subjective effects of variable delay and speech clipping in dynamically managed voice systems. IEEE Transactions on Communications, 33(8), 801-808.

Hamberg, R. and de Ridder, H. (1995): Continuous assessment of image quality. SID 95 Digest, 121-124.

Handley, M.J. (1997): An Examination of MBone Performance. USC/ISI Research Report: ISI/RR-97-450.

Handley, M.J. and Crowcroft, J. (1997): Network Text Editor (NTE): A scalable shared text editor for the MBone. Proceedings of ACM Sigcomm 97, Cannes, France.

Handley, M.J., Kirstein, P.T. & Sasse, M.A. (1993): Multimedia Integrated Conferencing for European Researchers (MICE): Piloting Activities and the Conference Management and Multiplexing Centre. Computer Networks and ISDN Systems, 26, 275-290.

Handley, M.J., Wakeman, I. and Crowcroft, J. (1995): The conference control protocol (CCCP): A scalable base for building conference control applications. Proceedings of SIGCOMM, Cambridge, Massachusetts, 275-287.

Hardman, V., Sasse, M.A, Handley, M.J. and Watson, A. (1995): Reliable audio for use over the Internet. Proceedings of INET '95, 27-30 June 1995, Honolulu, Hawaii, 171-178.

Hardman, V., Sasse, M.A., and Kouvelas, I. (1998): Successful multi-party audio communication over the internet. Communications of the ACM, 41 (5), 74-80.

Hearnshaw, D. (1999): Desktop conferencing for tutorial support. PhD Thesis, Dept. of Computer Science, University College London, University of London.

Heath, C. and Luff, P. (1991): Disembodied conduct: Communication through video in a multi-media office environment. Proceedings of CHI '91, 99-103.

Hollier, M.P. and Cosier, G. (1996): Assessing human perception. BT Technol. Journal, 14(1), 4-13.

House, A.S., Williams, C.E., Hecker, M.H.L. and Kryter, K.D. (1965): Articulation-testing methods: consonantal differentiation with a closed-response set. Journal of the Acoustical Society of America, 37(1), 158-166.

IEEE Recommended Practice for Speech Quality Measurements, IEEE Transactions on Audio and Electroacoustics, Vol. AU-17(3), 1969, 225-246.

IMA (1992): Recommended practices for enhancing digital audio compatibility in multimedia systems (version 3.00). Technical Report, Interactive Multimedia Association, Annapolis, Maryland.

ITU-R BT.500-8 Methodology for the subjective assessment of the quality of television pictures. Available from http://www.itu.int/publications/itu-t/itutrec.htm

ITU-T G.711 Pulse code modulation (PCM) of voice frequencies. Available from http://www.itu.int/publications/itu-t/itutrec.htm

ITU-T G.723 Speech coders. Available from http://www.itu.int/publications/itu-t/itutrec.htm

ITU-T P.800 Methods for subjective determination of transmission quality. Available from http://www.itu.int/publications/itu-t/itutrec.htm

ITU-T P.910 Subjective video quality assessment methods for multimedia applications. Available from http://www.itu.int/publications/itu-t/itutrec.htm

ITU-T P.920 Interactive test methods for audiovisual communications. Available from http://www.itu.int/publications/itu-t/itutrec.htm

Jacobson, V. (1992): vat manual pages, Lawrence Berkeley Laboratory, USA. Software available from http://www-nrg.ee.lbl.gov/vat/

Jacobson, V. (1993): wb README file, Lawrence Berkeley Laboratory, USA. Software available from http://www-nrg.ee.lbl.gov/wb/

Jardetzky, P.W, Sreenan, C.J. and Needham, R.M. (1995): Storage and synchronisation for distributed continuous media. Multimedia Systems, 3, 151-161.

Jayant, N.S. (1990): High-quality coding of telephone speech and wideband audio. IEEE Communications Magazine, Jan. 1990, 10-20.

Jones, B.L. and McManus, P.R. (1986): Graphic scaling of qualitative terms. SMPTE Journal, November 1986, 1166-1171.

Jordan, T.J and Sergeant, P.C (1998): Effects of facial image size on visual and audio-visual speech recognition. In Hearing by Eye II, ed. R. Campbell, B. Dodd and D.Burnham, Psychology Press Ltd.

Kent, R.D. and Read, C. (1992): The Acoustic Analysis of Speech. San Diego, California: Singular.

Kies, J.K., Williges, R.C. and Rosson, M.B. (1996): Controlled laboratory experimentation and field study evaluation of video conferencing for distance learning applications. Virginia Tech HCIL-96-02. Available from http://hci.ise.vt.edu/~hcil/

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

Kitawaki, N. and Nagabuchi, H. (1988): Quality assessment of speech coding and speech synthesis systems. IEEE Communications Magazine, October 1988, 36-44.

Knoche, H., De Meer, H.G. and Kirsh, D. (1999): Utility curves: Mean opinion scores considered biased. Proceedings of the 7th International Workshop on Quality of Service (IWQoS '99), June 1-4, UCL.

Kohler, W. (1930): Gestalt Psychology. London: Bell.

Kokotopoulos, A. (1997): Subjective assessment of a multimedia system for distance learning. Lecture Notes in Computer Science, 1242, 395-408.

Kouvelas, I., Hardman V.J. and Watson, A. (1996): Lip synchronisation for use over the Internet: analysis and implementation, Proceedings of GLOBECOM '96.

Kryter, K.D. (1972): Speech communication. In Van, Cott & Kinkade (Eds), Human Engineering Guide to Equipment Design.

Long, J. and Dowell, J. (1989): Conceptions of the discipline of HCI: Craft, applied science, and engineering. Proceedings of the Fifth Conference of the British Computer Society, Human-Computer Interaction Specialist Group. In People and Computers V, ed. A. Sutcliffe and L. Macaulay, Cambridge University Press.

Luce, P.A., Feustel, T.C., and Pisoni, D.B. (1983): Capacity demands in short-term memory for synthetic and natural speech. Human Factors, 25, 17-32.

Macedonia, M. & Brutzman, D. P. (1994): Mbone provides Audio and Video Across the Internet. IEE Computer, April 1994, pp.30-36.

Mack, R. and Montaniz, F. (1994): Observing, predicting and analysing usability problems. In Usability Inspection Methods, ed. J. Nielsen and R.L. Mack, Wiley, 295-339.

Mack, R. and Nielsen, J. (1994): Executive summary. In Usability Inspection Methods, ed. J. Nielsen and R.L. Mack, Wiley, 1-23.

Massaro, D.W. and Cohen, M.M. (1993): Perceiving asynchronous bimodal speech in consonant-vowel and vowel syllables. Speech Communication, 13, 127-134.

McCanne, S. and Jacobson, V. (1995): vic: A flexible framework for packet video. Proceedings of ACM Multimedia '95, San Francisco.

McGurk, H. and MacDonald, J.W. (1976): Hearing lips and seeing voices. Nature, 264, 126-130.

Miller, G.A. and Isard, S. (1963): Some perceptual consequences of linguistic rules. Journal of Verbal Learning and Behaviour, 2, 217-228.

Miller, G.A. and Licklider, J.C.R. (1950): The intelligibility of interrupted speech. Journal of the Acoustical Society of America, 22(2), 167-173.

Miller, G.A., Heise, G.A. and Lichten, W. (1951): The intelligibility of speech as a function of the context of the text materials. Journal of Experimental Psychology, 41, 329-335.

Monk, A.F. and Watts, L. (1995): A poor quality video link affects speech but not gaze. Proceedings of CHI'95, 274-275.

Moran, T. (1981): The Command Language Grammar: A Representation for the User Interface of Interactive Computer Systems. International Journal of Man-Machine Studies, 15, 3-50.

Mouly, M. and Pautet, M-B. (1993): The GSM system for mobile communications. Lassay-les-Chateaux, France: Europe Media Duplication.

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

Nagabuchi, H. and Kitawaki, N. (1992): Evaluation of coded speech quality degraded by cell loss in ATM networks. Electronics and Communications in Japan, Part 3, 75 (9), 14-24.

Narita, N. (1993): Graphic scaling and validity of Japanese descriptive terms used in subjective-evaluation tests. SMPTE Journal, July 1993, 616-622.

Negroponte, N. (1995): Being Digital. Hodder & Stoughton.

Newell, A. and Card, S.K. (1985): The prospects for psychological science in human-computer interaction. Human-Computer Interaction, 1, 209-242.

Newell, A. and Card, S.K. (1986): Straightening out Softening up: a response to Carroll and Campbell. Human Computer Interaction, 2, 251-267.

Nielsen, J. (1994): Heuristic evaluation. In Usability Inspection Methods, ed. J. Nielsen and R.L. Mack, Wiley, 25-62.

Nye, P.W. and Gaitenby, J.H. (1974): The intelligibility of synthetic monosyllabic words in short, syntactically normal sentences. Haskins Laboratories: Status Report on Speech Research, 1974.

Olson, G.M. and Olson, J.S. (1997): Making sense of the findings: Common vocabulary leads to the synthesis necessary for theory building. In Video-Mediated Communication, ed. K.E. Finn, A.J. Sellen and S. Wilbur, LEA, NJ, 75-91.

O'Malley, C., Langton, S., Anderson, A., Doherty-Sneddon, G., and Bruce, V. (1996): Comparison of face-to-face and video-mediated interaction. Interacting with Computers, 8(2), 177-192.

Ostberg, O., Lindstrom, B., Renhall, P. (1989): Contribution of display size to speech intelligibility in videophone systems. International Journal of Human-Computer Interaction, 1(1), 149-159.

Payne, S. J. & Green, T. R. G. (1986): Task-Action Grammars: a Model of the Mental Representation of Task Languages. Human-Computer Interaction, 2, 93-134.

Perkins, C., Hodson, O. and Hardman, V. (1998): A survey of packet loss recovery techniques for streaming media. IEEE Network, 12(5), 40-48.

Podolsky, M., Romer, C. and McCanne, S. (1998): Simulation of FEC-based error control for packet audio on the Internet. Proceedings of IEEE INFOCOM '98 - The Conference of Computer Communications, 505-515.

Pollack, I. and Pickett, J.M. (1963): Intelligibility of excerpts from conversation. Language and Speech, 6, 165-171.

Postel, J. (1981): Internet Protocol. RFC 791. Available from http://www.rfc-editor.org/

Postel, J. (1981b): Transmission Control Protocol. RFC 793. Available from http://www.rfc-editor.org/

Preece, J., Sharp, H., Benyon, D., Holland, S. and Carey, D. (1994): Human-Computer Interaction. Addison-Wiley.

Preminger, J.E. and Van Tasell, D.J. (1995): Quantifying the relationship between speech quality and speech intelligibility. Journal of Speech and Hearing Research, 38, 714-725.

RACE ISSUE Usability Guidelines (1992): Vol. 1-4, RACE ISSUE Project 1065. HUSAT Research Institute, Loughborough University of Technology, UK.

RACE TELEMED (1992): Final report on recommendations for screen display settings for a videolink from a human performance perspective, and the concordance of results arising from different tasks. RACE Project R 1086, Deliverable 29-3.

Ralston, J.V., Pisoni, D.B., Lively., S.E., Greene., B.G. and Mullenix, J.W. (1991): Comprehension of synthetic speech produced by rule: word monitoring and sentence-by-sentence listening times. Human Factors, 33, 471-491.

Reeves, B. and Nass, C. (1996): The Media Equation. Cambridge University Press/CSLI Publications.

de Ridder, H. and Hamberg, R. (1997): Continuous assessment of image quality. SMPTE Journal, February 1997, 123-128.

Risberg, A. and Lubker, J.L. (1978): Prosody and speechreading. Stockholm: Department of Speech Communication and Music Acoustics, Royal Institute of Technology (STL-QPSR 4/1978).

Roy, R.R. (1994): Networking constraints in multimedia conferencing and the role of ATM networks. AT & T Technical Journal, July/Aug 1994, 97-108.

Sasse, M.A. (1997): Eliciting and describing users' models of computer systems. PhD thesis in Computer Science at the Faculty of Science of the University of Birmingham.

Sasse, M.A., Bilting, U., Schulz, C-D. & Turletti, T. (1994a): Remote Seminars through Multimedia Conferencing: Experiences from the MICE project. Proceedings of INET'94/JENC5.

Sasse, M.A., Clark, L. and Perkins, C. (1998): Piloting IP multicast conferencing over SuperJANET: The PIPVIC project, In Proceedings of the UKERNA Networkshop, Aberdeen.

Sasse, M.A., Handley, M.J and Ismail, N.M (1994b): Coping with Complexity and Interference: Design Issues in Multimedia Conferencing Systems. In D. Rosenberg & C.S. Hutchison [Eds.]: Design Issues in CSCW, Berlin: Springer-Verlag,179-195.

Seferidis, V., Ghanbari, M. and Pearson, D.E. (1992): Forgiveness effect in subjective assessment of packet video. *Electronics Letters*, 28(1), 2013-2014.

Segui, J. (1984): The syllable: a basic unit in speech processing. In H. Bouma & D. Bouwhuis (Eds.) Attention and Performance X, 165-181.

Sekuler, R. and Blake, R. (1994): Perception. McGraw-Hill Inc.

Sellen, A.J. (1992): Speech patterns in video-mediated conversations. Proceedings of CHI 92, 49-59.

Shannon, C. (1948): A mathematical theory of communication. Bell System Technical Journal, 27, 379-423.

Short, J., Williams, E., and Christie, B. (1976): The Social Psychology of Telecommunications. Wiley.

Sumby, W.H. and Pollack, I. (1954): Visual contribution to speech intelligibility in noise. Journal of the Acoustical Society of America, 26, 212-215.

Summerfield (1992): Lipreading and audio-visual speech perception. Philosophical Transactions of the Royal Society of London, B335, 71-78.

Suzuki, J. and Taka, M. (1989): Missing packet recovery techniques for low bit rate coded speech. IEEE Journal on Selected Areas in Communications, 7(5), 707-717.

Tang, J.C. and Isaacs, E.A. (1993): Why do users like video?: Studies of multimedia collaboration. Computer-Supported Cooperative Work, 1, 163-196.

Teunissen, K. (1996): The validity of CCIR quality indicators along a graphical scale. *SMPTE Journal*, March 1996, 144-149.

Veinott, E.S., Olson, J., Olson, G.M. and Fu, X. (1997): Video matters! When communication ability is stressed, video helps. Proceedings of CHI'97, Atlanta, GA, March 22-27, 315-316.

Veinott, E.S., Olson, J., Olson, G.M. and Fu, X. (1999): Video helps remote work: Speakers who need to negotiate common ground benefit from seeing each other. Proceedings of CHI'99, Pittsburg, PA, May 15-20, 302-309.

Virtanen, M.T., Gleiss, N. and Goldstein, M. (1995): On the use of evaluative category scales in telecommunications. Proc. Human Factors in Telecommunications '95, 253-260.

Vitkovitch, M. and Barber, P. (1994): Effect of video frame rate of subjects' ability to shadow one of two competing messages. Journal of Speech and Hearing Research, 37, 1204-1210.

Voiers, W.D. (1977): Diagnostic evaluation of speech intellibility. In Speech and Intelligibility and Speaker Recognition, ed. M.E. Hawley, Benchmark Papers in Acoustics, Dowden, Hutchinson & Ross Inc., Stroudsberg, PA, 374-386.

Warren, R.M. (1970): Perceptual restoration of missing speech sounds. Science, 167, 392-393.

Warren, R.M. (1982): Auditory Perception: A New Synthesis. Pergamon Press Inc.

Warren, R.M., Obusek, C.J., Fermer, R.M., and Warrren, R.P. (1969): Auditory sequence: Confusions of patterns other than speech or music. Science, 164, 586-587.

Wasem, O.J., Goodman, D.J., Dvorak, C.A. and Page, H.G. (1988): The effect of waveform substitution on the quality of PCM packet communications. IEEE Transactions on Acoustics, Speech and Signal Processing, 36(3), 342-348.

Waterworth. J.A. and Thomas, C.M. (1985): Why is synthetic speech harder to remember than natural speech? Human Factors in Computing Systems II, 1985, Ch. 30, 201-206.

Watson, A. and Sasse, M.A. (1996a): Assessing the usability and effectiveness of a remote language teaching system, Proceedings of ED-MEDIA '96 - World Conference on Educational Multimedia and Hypermedia, June 17-22 1996, Boston, Mass., 685-690.

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

Watson, A. and Sasse, M.A. (1997): 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, 15-16 September 1997, Aberdeen, Scotland, 189-194.

Watson, A. and Sasse, M.A. (1998): Measuring Perceived Quality of Speech and Video in Multimedia Conferencing Applications. Proceedings of ACM Multimedia '98, 12-16 September 1998, Bristol, England, 55-60.

Watson, A. and Sasse, M.A. (2000a): Distance Education via IP Videoconferencing: Results from a National Pilot Project. CHI 2000 Extended Abstracts, 113-114.

Watson, A. and Sasse, M.A. (2000b): The Good, the Bad and the Muffled: the Impact of Different Degradations on Internet Speech. Proceedings of ACM Multimedia 2000, Oct. 30- Nov. 3, Marina Del Rey, CA; 269-276.

Whitefield, A., Wilson, F. and Dowell, J. (1991): A framework for human factors evaluation. Behaviour and Information Technology, 10(1), 65-79.

Whittaker, S. (1995): Rethinking video as a technology for interpersonal communications: theory and design implications. International Journal of Human-Computer Studies, 42 (5), 501-529.

Whittaker, S. and O'Conaill, B. (1997): The role of vision in face-to-face and mediated communication. In Video-Mediated Communication, ed. K.E. Finn, A.J. Sellen and S. Wilbur, LEA, NJ, 23-49.

Wilson, G. and Sasse, M.A. (2000a): 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 HCI 2000: People and Computers XIV - Usability or Else! Springer, 327-339, September 5th - 8th, Sunderland, UK.

Wilson, G. and Sasse, M.A. (2000b): Investigating the impact of audio degradations on users: Subjective vs. objective assessment methods. Proceedings of OZCHI 2000, 4-8 December, Sydney, Australia, 135-142.

Zhang, L., Deering, S., Estrin, D., Shenker, S. and Zappala, D. (1993): RSVP: A new resource ReSerVation Protocol. IEEE Network Magazine, 7(5), 8-18.

Appendix A: Sample ReLaTe questionnaire

NAME
AGE
LEVEL OF LANGUAGE ABILITY

DATE
TIME OF LESSON

PLEASE INDICATE WHETHER YOU HAVE EXPERIENCE WITH USING COMPUTERS AND A MOUSE.

____ yes
____ no

Please indicate your responses by ticking the option that you feel best describes your agreement with the statements below.

Q1. How did the audio information sound during the lesson?

(a) How would you rate the overall quality of the audio during the lesson?
____ excellent
____ good
____ fair
____ poor
____ bad

Comments:-

(b) In general, how much effort did you feel was required on your part on listening to the tutor when he/she was speaking ENGLISH?
____ complete relaxation possible, no effort required
____ attention necessary, no appreciable effort required
____ moderate effort required
____ considerable effort required
____ no meaning understood with any feasible effort

(c) In general, how much effort did you feel was required on your part on listening to the tutor when he/she was speaking FRENCH?
____ complete relaxation possible, no effort required
____ attention necessary, no appreciable effort required
____ moderate effort required
____ considerable effort required
____ no meaning understood with any feasible effort

Q2. How did the video image assist in communication?

(a) The video image of the tutor helped me understand his/her speech
____ strongly agree
____ agree
____ undecided
____ disagree
____ strongly disagree

(b) I found the video image of the tutor reassuring, but not particularly useful
____ strongly agree
____ agree
____ undecided
____ disagree
____ strongly disagree

Q3. How easy was the ReLaTe system interface to use? (where 'interface' refers to the on-screen facilities e.g. whiteboard and audio control)

(a) I found the ReLaTe system in general easy to use
____ strongly agree
____ agree
____ undecided
____ disagree
____ strongly disagree

(b) I found the screen to be well laid out
____ strongly agree
____ agree
____ undecided
____ disagree
____ strongly disagree

(c) I found that having to manipulate the interface interfered with the smooth running of the lesson
____ strongly agree
____ agree
____ undecided
____ disagree
____ strongly disagree
(dependent to a greater or lesser extent on whether the student has to push to talk or not)

(d) I found all aspects of the interface equally hard/easy to use
____ strongly agree
____ agree
____ undecided
____ disagree
____ strongly disagree

If you disagreed with the above statement, please record which aspect you found most difficult.

Q4. How did the audio information sound during the lesson?

(a) How would you rate the overall quality of the audio during the lesson?
____ excellent
____ good
____ fair
____ poor
____ bad

(b) In general, how much effort did you feel was required on your part on listening to the tutor when he/she was speaking ENGLISH?
____ complete relaxation possible, no effort required
____ attention necessary, no appreciable effort required
____ moderate effort required
____ considerable effort required
____ no meaning understood with any feasible effort

(c) In general, how much effort did you feel was required on your part on listening to the tutor when he/she was speaking THE TARGET LANGUAGE?
____ complete relaxation possible, no effort required
____ attention necessary, no appreciable effort required
____ moderate effort required
____ considerable effort required
____ no meaning understood with any feasible effort

Q5. How did the video image assist in communication?

(a) The video image of the tutor helped me understand his/her speech
____ strongly agree
____ agree
____ undecided
____ disagree
____ strongly disagree

(b) I found the video image of the tutor reassuring, but not particularly useful
____ strongly agree
____ agree
____ undecided
____ disagree
____ strongly disagree

Q6. How did the remote lesson compare to a normal face-to-face lesson?

(a) Compared to a normal face-to-face lesson, the extent to which I spoke in English during this lesson was
____ greater
____ the same
____ lesser

If the answer was 'greater than' or 'less than', please indicate why you think this was so:

(b) From a language learning point of view, I found the session overall
____ excellent
____ good
____ fair
____ poor
____ bad

Additional Comments - please feel free to note any additional points, good or bad: your input is very valuable.

Appendix B: Raw data from Study 1, Compensating for packet loss in internet audio

Intelligibility ratings

SS, 20ms	10%	15%	20%	30%	LPC 20ms	20%	30%	40%
	88	88	72	64		68	88	68
	80	84	44	80		92	80	76
	84	88	80	64		84	76	80
	88	84	80	72		88	100	76
	92	80	72	60		88	84	76
	80	84	64	64		100	96	80
	76	88	88			88	84	88

SS, 40ms	10%	15%	20%	30%	LPC 40ms	20%	30%	40%
	92	88	76	56		96	68	60
	84	80	84	80		72	84	68
	72	84	68	80		100	76	88
	76	80	60	72		88	88	76
	92	80	72	60		92	80	80
	80	96	64	68		92	88	84
	88	84	76			88	72	80

SS, 80ms	10%	15%	20%	30%	LPC 80ms	20%	30%	40%
	96	80	84	68		84	80	80
	88	88	72	60		80	68	64
	100	72	76	56		84	84	72
	96	64	68	52		96	76	76
	88	56	68	36		92	84	80
	92	92	76	48		96	68	76
	84	68	76	52		84	80	60

PR 20ms	10%	15%	20%	30%	40%
	92	84	88	88	84
	84	96	92	80	96
	88	96	88	88	72
	88	92	88	88	56
	84	92	96	84	68
	88	92	96	88	68
	84	80	92	84	52

PR 40ms	10%	15%	20%	30%	40%
	84	92	80	84	40
	84	80	80	80	48
	80	92	84	88	52
	92	84	76	80	60
	80	84	84	68	68
	92	76	84	72	40
	88	84	80	68	68

PR 80ms	10%	15%	20%	30%	40%
	72	76	76	72	64
	80	80	64	76	68
	68	92	84	56	48
	84	76	76	56	68
	80	80	64	52	44
	68	92	84	68	28
	64	80	76	48	52

Perceived quality ratings

20 ms ss	10%	15%	20%	30%	20 ms lpc	20%	30%	40%
	2	2	2	4		3	3	2
	2	4	2	2		4	4	2
	4	2	2	2		4	4	4
	2	2	3	3		3	2	2
	3	3	2	1		2	2	3
	2	3	2	1		4	3	4
	3	2	2	2		2	3	2

40 ms ss	10%	15%	20%	30%	40 ms lpc	20%	30%	40%
	3	3	2	2		4	2	1
	4	3	2	3		4	4	3
	4	4	1	2		4	4	3
	3	3	2	2		3	2	2
	3	4	3	1		3	4	4
	2	3	2	2		3	3	3
	2	4	2	2		4	3	2

80 ms ss	10%	15%	20%	30%	80 ms lpc	20%	30%	40%
	3	3	2	3		3	3	2
	3	2	3	2		3	3	4
	4	4	3	1		4	3	3
	4	2	2	2		3	3	3
	3	3	2	1		4	4	1
	3	3	2	1		2	2	1
	3	2	3	1		4	3	2

20 ms pr	10%	15%	20%	30%	40%
	3	3	4	3	2
	4	5	1	1	2
	3	2	3	4	2
	4	3	3	3	2
	4	4	4	3	2
	3	4	3	3	2
	3	3	3	3	2

40 ms pr	10%	15%	20%	30%	40%
	4	3	3	1	2
	4	2	3	1	2
	3	1	1	3	1
	4	4	3	3	1
	4	3	3	2	2
	3	4	3	2	1
	4	3	2	3	2

80 ms pr	10%	15%	20%	30%	40%
	2	2	2	2	2
	4	2	3	1	1
	1	2	1	1	4
	3	1	3	1	1
	4	2	2	2	1
	1	2	2	2	1
	2	2	2	1	2

```
Analysis of Variance for SS,PR,LPC, 20-30%, intelligibility ratings

Source                   DF     Seq SS     Adj SS     Adj MS       F      P
Repair                    2    5828.95    5795.48    2897.74   36.35  0.000
Packet s                  2    3321.72    3254.91    1627.45   20.41  0.000
Loss                      1    2225.53    2199.86    2199.86   27.59  0.000
Repair*Packet s           4    2278.78    2265.87     566.47    7.11  0.000
Repair*Loss               2      35.20      30.01      15.01    0.19  0.829
Packet s*Loss             2    1051.67    1057.91     528.95    6.63  0.002
Repair*Packet s*Loss      4     346.44     346.44      86.61    1.09  0.367
Error                   106    8450.67    8450.67      79.72
Total                   123   23538.97
```

```
Analysis of Variance for SS, PR, LPC, 20-30%, quality ratings

Source                   DF     Seq SS     Adj SS     Adj MS       F      P
Repair                    2    29.3492    29.3492    14.6746   25.56  0.000
Packet s                  2     3.5397     3.5397     1.7698    3.08  0.050
Loss                      1     3.5000     3.5000     3.5000    6.10  0.015
Repair*Packet s           4     6.4127     6.4127     1.6032    2.79  0.030
Repair*Loss               2     0.1429     0.1429     0.0714    0.12  0.883
Packet s*Loss             2     1.4762     1.4762     0.7381    1.29  0.281
Repair*Packet s*Loss      4     0.9524     0.9524     0.2381    0.41  0.798
Error                   108    62.0000    62.0000     0.5741
Total                   125   107.3730
```

Appendix C: Raw data for Study 2, Investigating the intelligibility of longer speech stimuli

	0	10PR	10LPC	20PR	20LPC	30PR	30LPC	40PR	40LPC
1	100	100	100	80	100	100	100	60	0
2	100	60	100	80	100	100	80	0	20
3	100	100	100	80	100	100	80	20	20
4	100	100	100	80	40	100	100	20	20
5	80	0	80	80	60	20	100	20	0
6	100	80	100	40	60	100	80	20	20
7	100	60	100	80	20	100	80	40	40
8	100	100	100	80	100	100	80	20	20
9	100	100	100	80	100	100	100	60	20
10	100	100	100	80	60	100	80	20	60
11	100	100	100	80	80	100	100	40	40
12	100	100	100	80	100	100	80	0	20
13	100	80	100	80	100	80	80	60	60
14	100	80	80	80	60	0	100	20	20
15	100	100	100	80	80	100	60	40	20
16	100	100	100	80	100	100	80	60	20
17	100	100	100	100	60	100	100	0	20
18	100	80	100	80	80	100	100	20	20
19	100	100	100	60	100	100	100	40	20
20	60	60	100	40	20	20	0	20	20
21	100	100	100	80	80	60	80	40	20
22	80	100	100	20	100	20	100	0	20
23	100	100	100	80	100	100	80	100	40
24	100	100	100	80	100	100	100	20	100
Mean	96.66667	87.5	98.33333	74.16667	79.16667	83.33333	85	30.83333	27.5
stdev	9.630868	23.45208	5.646597	17.17345	26.02953	32.65986	21.46787	24.30185	21.11048
st error	1.9658	4.79	1.15	3.51	5.31	6.67	4.38	4.96	4.31

10PR	10LPC
100	100
60	100
100	100
100	100
0	80
80	100
60	100
100	100
100	100
100	100
100	100
100	100
80	100
80	80
100	100
100	100
100	100
80	100
100	100
60	100
100	100
100	100
100	100
100	100

t-Test: Paired Two Sample for Means

	Variable 1	*Variable 2*
Mean	87.5	98.33333
Variance	550	31.88406
Observations	24	24
Pearson Correlatior	0.623818	
Hypothesized Mear	0	
df	23	
t Stat	-2.6	
P(T<=t) one-tail	0.008005	
t Critical one-tail	1.71387	
P(T<=t) two-tail	0.01601	
t Critical two-tail	2.068655	

30PR	40PR
100	60
100	0
100	20
100	20
20	20
100	20
100	40
100	20
100	60
100	20
100	40
100	0
80	60
0	20
100	40
100	60
100	0
100	20
100	40
20	20
60	40
20	0
100	100
100	20

t-Test: Paired Two Sample for Means

	Variable 1	*Variable 2*
Mean	83.33333	30.83333
Variance	1066.667	590.5797
Observations	24	24
Pearson Correlatior	0.237378	
Hypothesized Mear	0	
df	23	
t Stat	7.187638	
P(T<=t) one-tail	1.28E-07	
t Critical one-tail	1.71387	
P(T<=t) two-tail	2.56E-07	
t Critical two-tail	2.068655	

30LPC	40LPC
100	0
80	20
80	20
100	20
100	0
80	20
80	40
80	20
100	20
80	60
100	40
80	20
80	60
100	20
60	20
80	20
100	20
100	20
100	20
0	20
80	20
100	20
80	40
100	100

t-Test: Paired Two Sample for Means

	Variable 1	*Variable 2*
Mean	85	27.5
Variance	460.8696	445.6522
Observations	24	24
Pearson Correlati(	0.028781	
Hypothesized Mea	0	
df	23	
t Stat	9.493468	
P(T<=t) one-tail	1.01E-09	
t Critical one-tail	1.71387	
P(T<=t) two-tail	2.02E-09	
t Critical two-tail	2.068655	

Appendix D: Raw data for Study 3, Investigating the perceived quality of passages of speech

	LPC 0	10	20	30	40	50
	5	4	4	2	1	1
	5	4	4	2	1	2
	4	4	4	2	1	1
	5	4	4	2	2	2
	5	4	4	5	2	1
	5	4	4	3	3	3
	5	5	3	4	3	2
	4	5	3	2	2	1
	5	5	2	1	2	1
	4	4	4	2	2	2
Mean	4.7	4.3	3.6	2.5	1.9	1.6
st dev	0.483046	0.483046	0.699206	1.178511	0.737865	0.699206
se	0.15	0.15	0.22	0.37	0.23	0.22

	SS 0	10	20	30	40	50
	5	2	2	1	1	1
	5	2	3	2	2	1
	4	2	2	1	1	1
	5	2	2	3	1	1
	5	4	1	2	1	1
	5	3	1	1	1	1
	5	3	2	1	1	1
	4	2	2	1	1	1
	5	3	1	1	2	1
	4	2	1	1	1	1
Mean	4.7	2.5	1.7	1.4	1.2	1
st dev		0.707107	0.674949	0.699206	0.421637	0
se		0.22	0.21	0.22	0.13	0

Two-way ANOVA: Score versus Repair, Loss

```
Analysis of Variance for Score
Source        DF        SS        MS        F        P
Repair         1    31.008    31.008    71.10    0.000
Loss           5   165.342    33.068    75.83    0.000
Interaction    5    13.542     2.708     6.21    0.000
Error        108    47.100     0.436
Total        119   256.992
```

Appendix E: Raw data for Study 4, Measuring the quality of a conversation

	no loss	no loss	10lpc	10pr	20lpc	20pr	30lpc	30pr	40lpc	40pr
	1	2	2	2	2	2	1	1	1	1
	2	3	3	3	2	3	2	2	1	1
	3	3	3	3	2	3	2	2	1	1
	3	3	3	3	2	3	2	2	1	2
	3	3	3	3	2	3	2	2	1	2
	3	3	3	3	2	3	2	2	1	2
	4	4	3	3	2	3	2	2	2	2
	4	4	3	3	3	3	2	2	2	2
	4	4	3	3	3	3	2	2	2	2
	4	4	3	3	3	3	2	2	2	2
	4	4	3	3	3	3	2	3	2	2
	4	4	4	3	3	3	2	3	2	2
	4	4	4	4	3	3	2	3	2	2
	4	4	4	4	3	3	2	3	2	2
	4	4	4	4	4	3	2	3	2	2
	4	5	4	4	4	3	3	3	2	2
	4	5	4	4	4	4	3	3	2	2
	4	5	4	4	4	4	3	3	3	3
	5	5	4	4	4	4	3	3	3	3
	5	5	4	5	4	4	3	3	3	3
	5	5	4	5	4	4	3	4	3	3
	5	5	4	5	4	4	4	4	3	3
	5	5	4	5	4	4	4	4	3	3
	5	5	5	5	4	4	4	5	4	4
Mean	3.875	4.083333	3.541667	3.666667	3.125	3.291667	2.458333	2.75	2.083333	2.208333
St Dev	0.9918143	0.880547	0.658005	0.868115	0.850192	0.550033	0.779028	0.896854	0.829702	0.72106

Two-way ANOVA: Score versus Repair, Loss

```
Analysis of Variance for Score
Source          DF        SS         MS        F        P
Repair           1     2.604      2.604     4.17    0.042
Loss             4    85.525     21.381    34.22    0.000
Interaction      4     0.458      0.115     0.18    0.947
Error          230   143.708      0.625
Total          239   232.296
```

Appendix F: Raw data for Study 5, Assessing the perceived level of audio-video synchronisation

sync	2	5	6	8	12
	2	4	4	4	4
	2	3	4	4	5
	3	4	5	4	5
	3	4	3	4	5
	4	4	3	5	4
	2	2	5	5	4
	2	4	4	4	4
	1	3	3	4	4
Mean	2.375	3.5	3.875	4.25	4.375
St dev	0.916125	0.755929	0.834523	0.46291	0.517549
St err	0.32	0.27	0.029	0.16	0.18

	2	5	6	8	12
	1	1	2	3	4
	1	2	1	2	2
	2	2	2	2	1
	2	2	2	2	2
	2	2	2	3	3
	2	1	2	2	2
	2	2	2	2	2
	1	2	2	2	2
Mean	1.625	1.75	1.875	2.25	2.25
St dev	0.517549	0.46291	0.353553	0.46291	0.886405
St err	0.18	0.16	0.12	0.16	0.31

Appendix G: Raw data for Study 7, Investigating a new quality scale

no loss	unlab 1st	unlab 2nd
	94	99
	91	91
	98	100
	96	98
	94	87
	99	98
	96	98
	97	96
	91	94
	95	79
	96	97
	89	92

Anova: Single Factor no loss

SUMMARY

Groups	Count	Sum	Average	Variance
unlab 1st	12	1136	94.66667	9.151515
unlab 2nd	12	1129	94.08333	37.17424

ANOVA

Source of Varia	SS	df	MS	F	P-value	F crit
Between G	2.041667	1	2.041667	0.088144	0.769335	4.300944
Within Gro	509.5833	22	23.16288			
Total	511.625	23				

10lpc	unlab 1st	unlab 2nd
	85	58
	48	59
	26	67
	90	76
	87	82
		64

Anova: Single Factor 10lpc

SUMMARY

Groups	Count	Sum	Average	Variance
unlab 1st	5	336	67.2	823.7
unlab 2nd	6	406	67.66667	91.46667

ANOVA

Source of Varia	SS	df	MS	F	P-value	F crit
Between G	0.593939	1	0.593939	0.001425	0.970715	5.117357
Within Gro	3752.133	9	416.9037			
Total	3752.727	10				

10pr	unlab 1st	unlab 2nd
	75	85
	47	68
	52	77
	86	65
	82	85
		84

Anova: Single Factor 10pr

SUMMARY

Groups	Count	Sum	Average	Variance
unlab 1st	5	342	68.4	316.3
unlab 2nd	6	464	77.33333	80.26667

ANOVA

Source of Varia	SS	df	MS	F	P-value	F crit
Between G	217.6485	1	217.6485	1.175396	0.306483	5.117357
Within Gro	1666.533	9	185.1704			
Total	1884.182	10				

20lpc	unlab 1st	unlab 2nd
	60	51
	77	83
	45	31
	39	84
	30	66
	55	42

Anova: Single Factor 20lpc

SUMMARY

Groups	Count	Sum	Average	Variance
unlab 1st	6	306	51	278.8
unlab 2nd	6	357	59.5	477.1

ANOVA

Source of Varia	SS	df	MS	F	P-value	F crit
Between G	216.75	1	216.75	0.573489	0.466345	4.964591
Within Gro	3779.5	10	377.95			
Total	3996.25	11				

20pr

unlab 1st	unlab 2nd
55	51
72	83
51	53
39	81
46	70
51	26

Anova: Single Factor

SUMMARY

Groups	Count	Sum	Average	Variance
unlab 1st	6	314	52.33333	123.0667
unlab 2nd	6	364	60.66667	470.6667

ANOVA

Source of Varia	SS	df	MS	F	P-value	F crit
Between G	208.3333	1	208.3333	0.701774	0.421759	4.964591
Within Gro	2968.667	10	296.8667			
Total	3177	11				

30lpc

unlab 1st	unlab 2nd
41	46
50	42
19	32
24	52
39	
34	
10	
55	

Anova: Single Factor

SUMMARY

Groups	Count	Sum	Average	Variance
unlab 1st	8	272	34	238.8571
unlab 2nd	4	172	43	70.66667

ANOVA

Source of Varia	SS	df	MS	F	P-value	F crit
Between G	216	1	216	1.146497	0.309446	4.964591
Within Gro	1884	10	188.4			
Total	2100	11				

30pr

unlab 1st	unlab 2nd
58	53
55	25
53	30
85	52
25	
50	
40	
58	

Anova: Single Factor

SUMMARY

Groups	Count	Sum	Average	Variance
unlab 1st	8	424	53	291.4286
unlab 2nd	4	160	40	212.6667

ANOVA

Source of Varia	SS	df	MS	F	P-value	F crit
Between G	450.6667	1	450.6667	1.682848	0.223675	4.964591
Within Gro	2678	10	267.8			
Total	3128.667	11				

40lpc

unlab 1st	unlab 2nd
65	27
7	34
45	28
40	35
35	56
	4
	48
	8

Anova: Single Factor

SUMMARY

Groups	Count	Sum	Average	Variance
unlab 1st	5	192	38.4	437.8
unlab 2nd	8	240	30	316.2857

ANOVA

Source of Varia	SS	df	MS	F	P-value	F crit
Between G	217.1077	1	217.1077	0.602286	0.45407	4.844338
Within Gro	3965.2	11	360.4727			
Total	4182.308	12				

40pr

unlab 1st	unlab 2nd
73	41
10	32
72	59
49	33
46	64
	23
	59
	47

Anova: Single Factor

SUMMARY

Groups	Count	Sum	Average	Variance
unlab 1st	5	250	50	657.5
unlab 2nd	8	358	44.75	224.2143

ANOVA

Source of Varia	SS	df	MS	F	P-value	F crit
Between G	84.80769	1	84.80769	0.222142	0.646622	4.844338
Within Gro	4199.5	11	381.7727			

50lpc	unlab 1st	unlab 2nd
	8	19
	51	22
	3	23
	18	10
	2	25
	38	39
	17	42
	5	6
	13	44
	20	8
	28	10
	25	7

Anova: Single Factor

SUMMARY

Groups	*Count*	*Sum*	*Average*	*Variance*
unlab 1st	12	228	19	218.7273
unlab 2nd	12	255	21.25	195.4773

ANOVA

Source of Varia	*SS*	*df*	*MS*	*F*	*P-value*	*F crit*
Between G	30.375	1	30.375	0.146667	0.705417	4.300944
Within Gro	4556.25	22	207.1023			
Total	4586.625	23				

50pr	unlab 1st	unlab 2nd
	2	21
	38	6
	0	23
	3	4
	0	31
	15	12
	6	34
	1	0
	8	29
	13	7
	4	2
	8	1

Anova: Single Factor

SUMMARY

Groups	*Count*	*Sum*	*Average*	*Variance*
unlab 1st	12	98	8.166667	111.9697
unlab 2nd	12	170	14.16667	160.8788

ANOVA

Source of Varia	*SS*	*df*	*MS*	*F*	*P-value*	*F crit*
Between G	216	1	216	1.583296	0.221477	4.300944
Within Gro	3001.333	22	136.4242			
Total	3217.333	23				

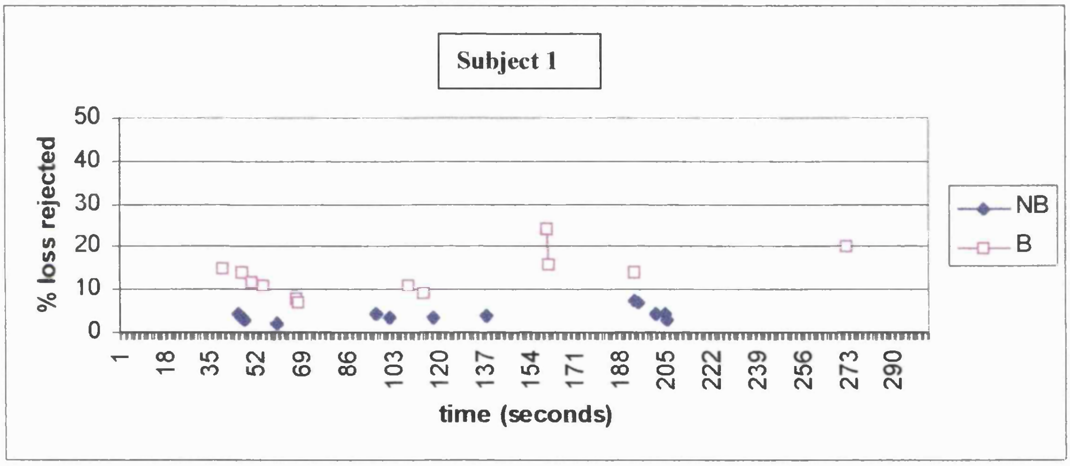
Appendix H: Individual subject rating correlations with loss rate in the first QUASS study

Condition 1	Condition 2	Condition 3	Condition 4
S1 -0.266 0.002	S1 0.029 0.743	S1 0.018 0.839	S1 -0.365 0.000
S2 -0.649 0.000	S2 -0.731 0.000	S2 -0.692 0.000	S2 -0.640 0.000
S3 -0.683 0.000	S3 -0.469 0.000	S3 -0.754 0.000	S3 -0.823 0.000
S4 -0.692 0.000	S4 -0.726 0.000	S4 -0.660 0.000	S4 -0.574 0.000
S5 -0.756 0.000	S5 -0.418 0.000	S5 -0.650 0.000	S5 -0.778 0.000
S6 -0.872 0.000	S6 -0.647 0.000	S6 -0.838 0.000	S6 -0.684 0.000
S7 -0.371 0.000	S7 -0.515 0.000	S7 -0.554 0.000	S7 -0.579 0.000
S8 -0.801 0.000	S8 -0.679 0.000	S8 -0.775 0.000	S8 -0.699 0.000
S9 -0.552 0.000	S9 -0.772 0.000	S9 -0.482 0.000	S9 -0.627 0.000
S10 -0.536 0.000	S10 -0.831 0.000	S10 -0.375 0.000	S10 -0.841 0.000
S11 0.029 0.740	S11 -0.712 0.000	S11 -0.604 0.000	S11 -0.812 0.000
S12 -0.495 0.000	S12 -0.704 0.000	S12 -0.759 0.000	S12 -0.766 0.000
S13 -0.573 0.000	S13 -0.632 0.000	S13 -0.626 0.000	S13 -0.659 0.000
S14 -0.667 0.000	S14 -0.604 0.000	S14 -0.918 0.000	S14 -0.692 0.000
S15 -0.696 0.000	S15 -0.456 0.000	S15 -0.335 0.000	S15 -0.502 0.000
S16 -0.809 0.000	S16 -0.594 0.000	S16 -0.752 0.000	S16 -0.642 0.000
S17 -0.727 0.000	S17 -0.525 0.000	S17 -0.559 0.000	S17 -0.748 0.000
S18 -0.646 0.000	S18 -0.646 0.000	S18 -0.357 0.000	S18 -0.851 0.000
S19 -0.857 0.000	S19 -0.466 0.000	S19 -0.725 0.000	S19 -0.681 0.000
S20 -0.770 0.000	S20 -0.527 0.000	S20 -0.538 0.000	S20 -0.690 0.000
S21 -0.706 0.000	S21 -0.698 0.000	S21 -0.250 0.004	S21 -0.721 0.000
S22 -0.692 0.000	S22 -0.647 0.000	S22 -0.835 0.000	S22 -0.575 0.000
S23 -0.774 0.000	S23 -0.728 0.000	S23 -0.751 0.000	S23 -0.662 0.000
S24 -0.284 0.001	S24 -0.763 0.000	S24 -0.789 0.000	S24 -0.595 0.000

Appendix I: Individual rating behaviour in the second QUASS study

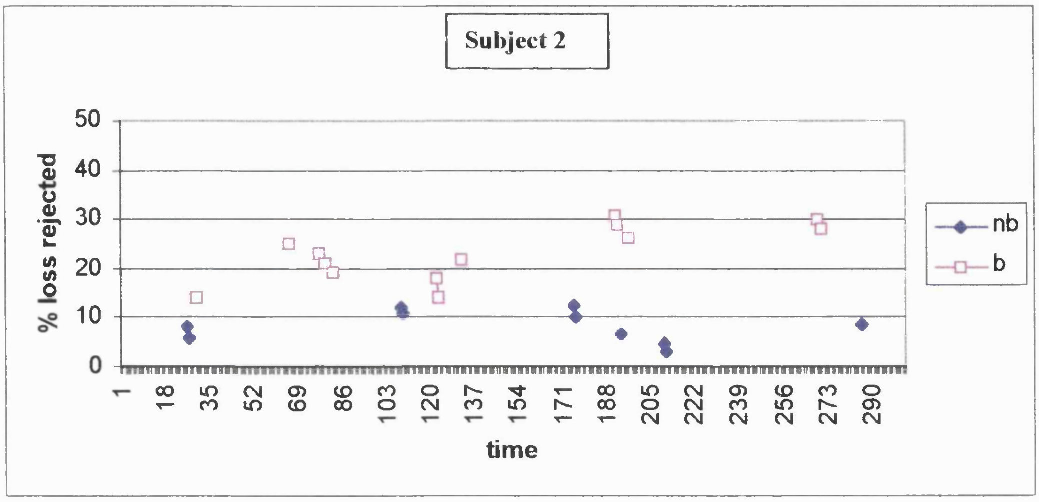
This appendix contains the individual rating behaviour for subjects in the non-budget (NB) and budget (B) conditions of the second QUASS study (section 7.4). The plots on the graph represent the point at which the slider was *increased* i.e. that specific loss level was rejected. Under each graph are comments made by that subject in the semi-structured interview after the QUASS experiment. The first answer relates to a specific question asked in the interview: did the task (giving and guessing word definitions) interfere with the subject's awareness of moving the slider? The second 'answer' was not explicitly sought in the semi-structured interview, but is comprised of revealing comments made during the interview about any behavioural strategy the subject may have been engaging in during the conditions.

Answers to the task interference question are prefaced by **TI**, and comments revealing behavioural strategies are prefaced by **BS**.



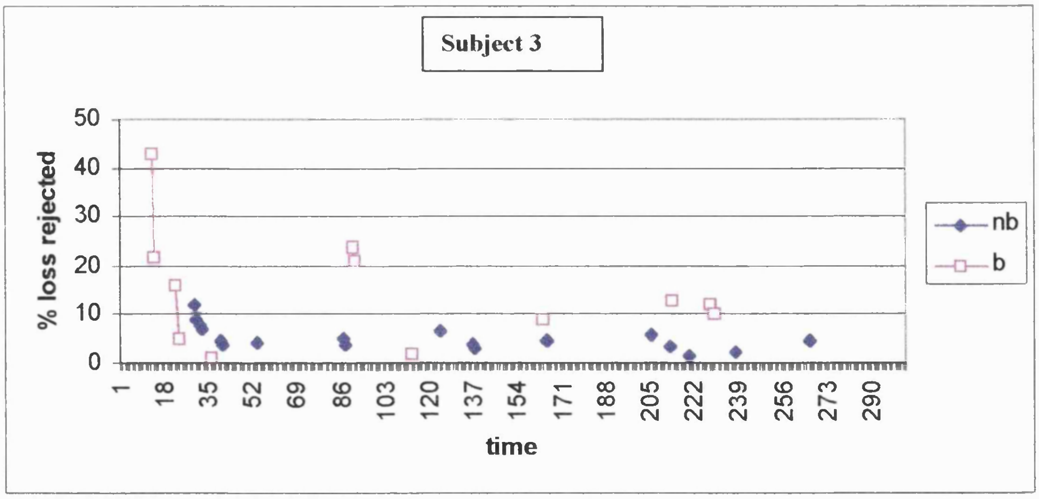
TI: Yes

BS: So, you develop strategies, for example, if you're intent on listening to something you put the quality up and you're willing to sacrifice budget for that but it it's not important that you hear somebody, for example when you're explaining something, then you put the quality low."



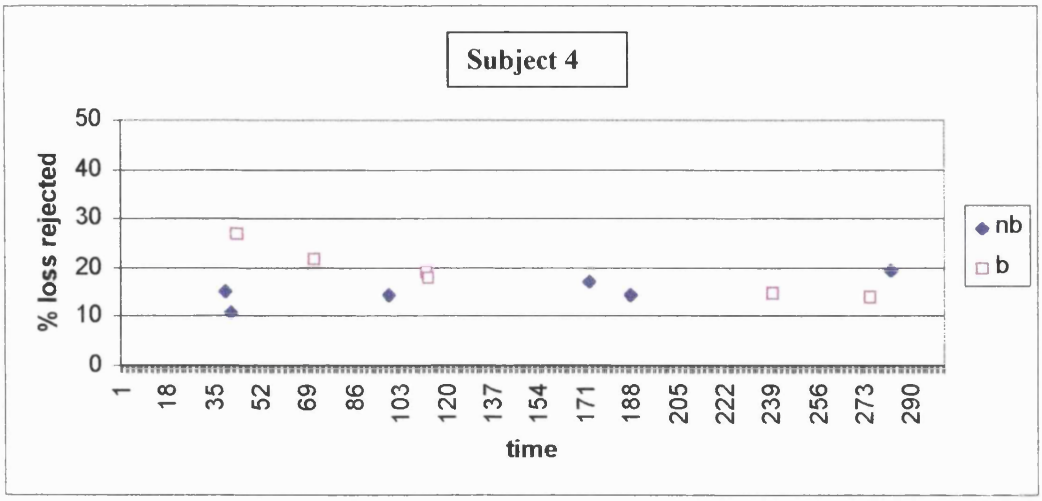
TI: No.

BS: Oh yeah, I was willing to accept a lower quality, I think because I was aware that I was losing a budget. I reduced it to a point where I found I could do it, which is substantially below what I had before [check]. Effectively what was happening, originally in the first case, without the budget, I wanted a higher level of sound and I didn't really need it.



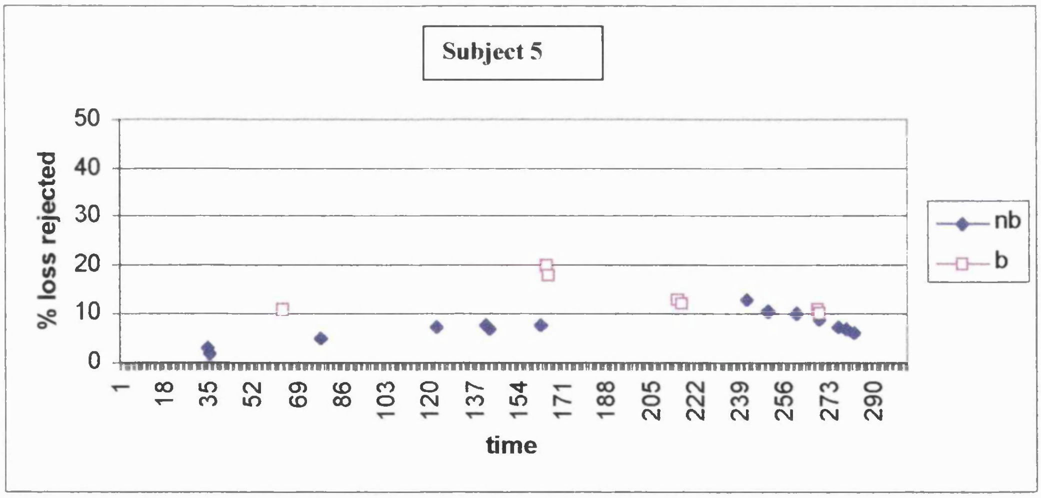
TI: It didn't really interfere.

BS: I had it almost all the time at the top anyway, I didn't seem to run out of budget. If I pick a difficult one for her to answer then I adjust the quality, it takes two minutes. I can turn down the quality when I'm explaining something.



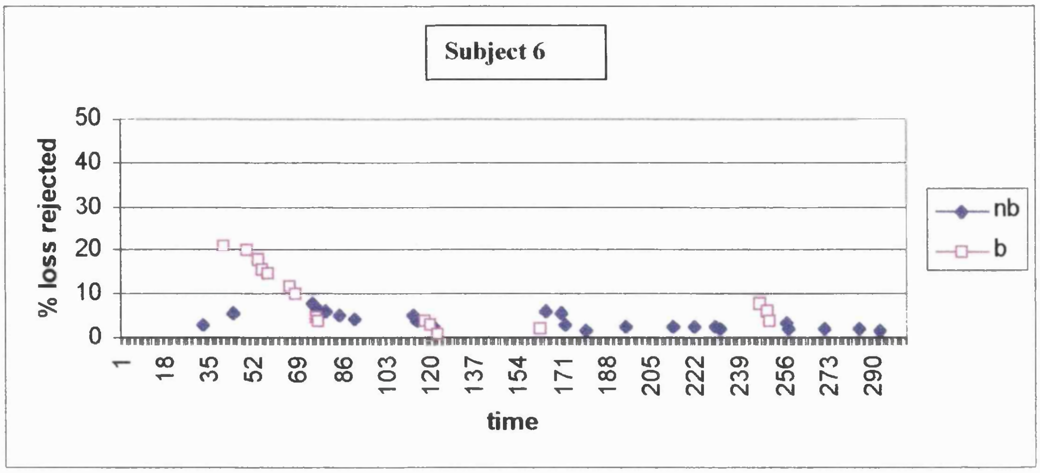
TI: No

BS: [no comments relevant]



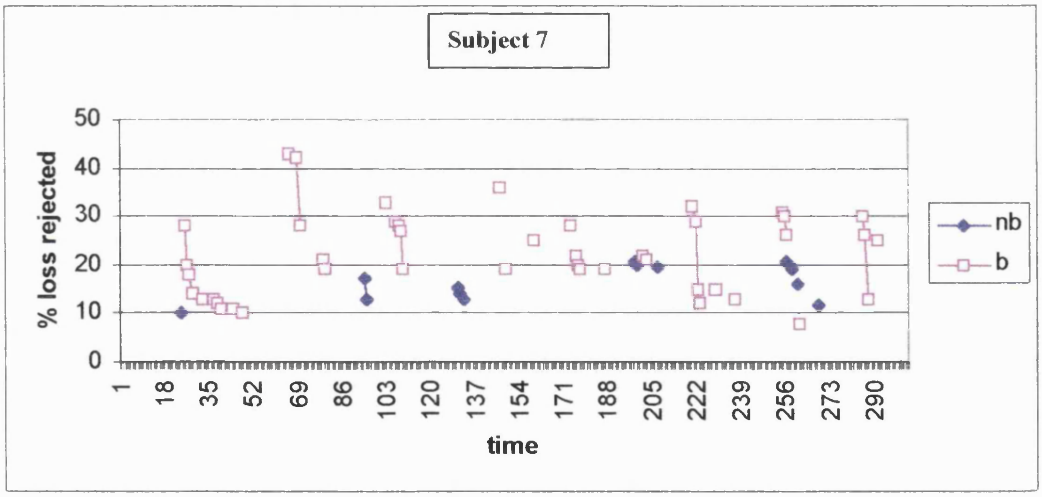
TI: No, not at all

BS: I think I was being very very generous because I had no expense so I left it at the top as long as I could. I think in the second condition the quality seemed to deteriorate much much quicker, so I had to maintain the quality myself. I noticed at the very beginning, when you first started the task, the cost bar was dropping quite quickly so you think 'I'm going to run out of money very quickly' so I dropped it considerably, re-evaluated the situation according to how much I could tolerate with respect to the task so I tried to be quite mean. But I noticed when there was still 2 minutes left to go that there was still quite a bit of money left so I splashed out a bit more but not enough for it too accelerate really quickly

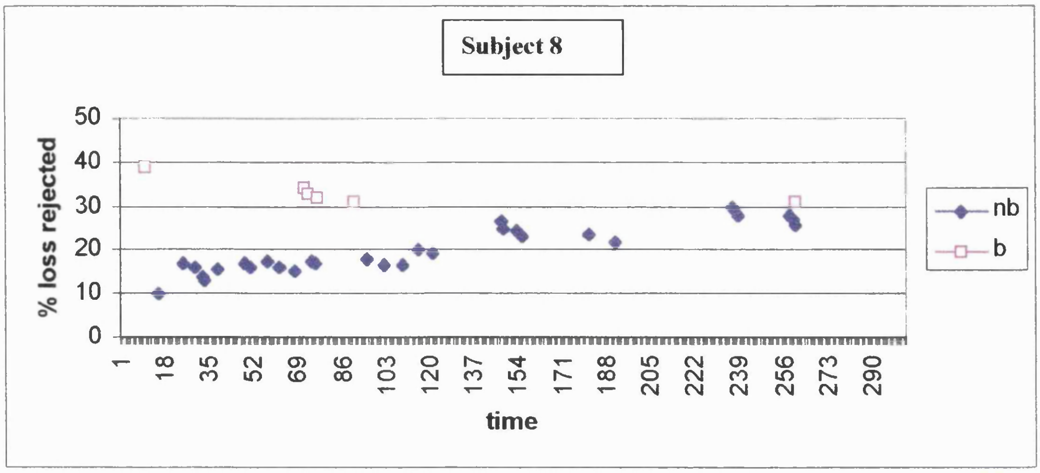


TI: Yeah it was - obviously I couldn't do it at all when I was talking - I could only do it when [co-experimenter] was talking.

BS: [no relevant comment]

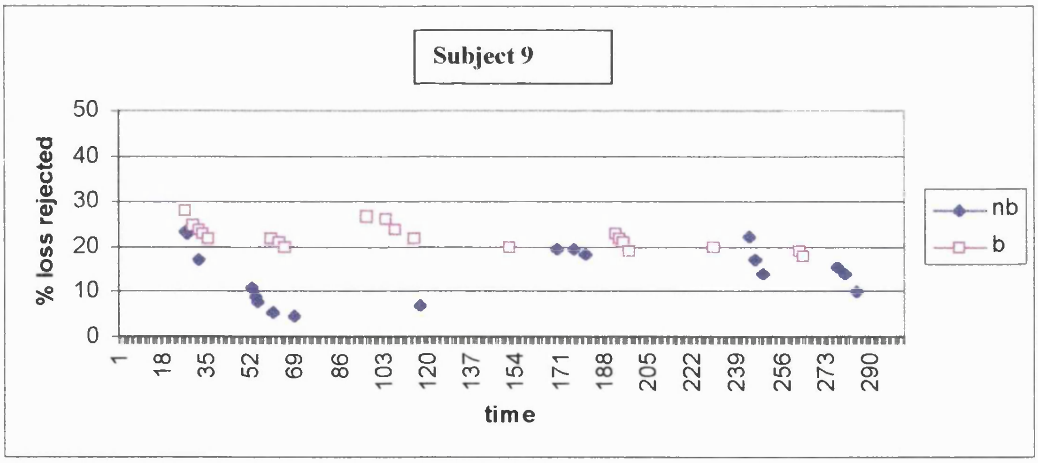


TI & BS: It was actually easier when you had the budget thing - it was easier to forget paying attention to it when there wasn't a budget there. With the budget you think oh yes, I must pay attention to it, because in the second lot, you've got this thing going down and you turn it down a bit but then you think 'oh my god I can't hear them', turn it up a bit, but hang on how much can I go up so I don't lose too much… So you're trying to pay attention to that and then play the game and what you end up doing is making trade-offs between does it matter if the quality is not good when I'm speaking but then it's got to be good when I'm listening so you do that trade-off and you end up… To be able to do the game properly you whack it up when it's on yours and whack it down… so you work out little strategies.



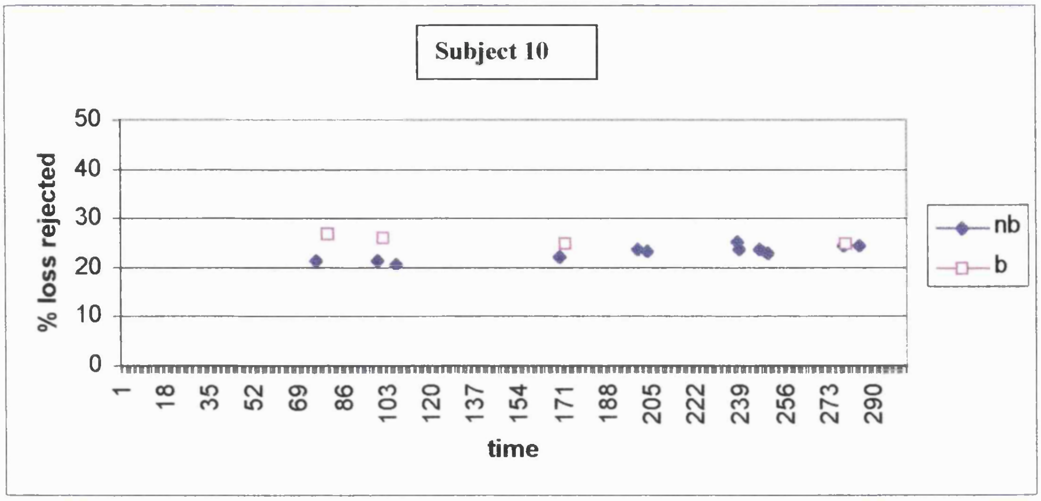
TI: I think it can be a little distracting in that you tend to play around with your slider most of the time perhaps when you've just finished a game (that is when you finish one of the sessions where someone has to guess). It's worse when you're trying to explain it than when you're trying to guess: when you're trying to explain it you're not actually moving the slider that much.

BS: I tried to adjust it to the point where I was just able to complete the task.
Budget didn't really make a difference because, given the amount of time I was spending on it, the budget didn't start dropping massively, it was a constraint but it wasn't a very binding constraint. If it had dropped
more I could have started paying attention to it.



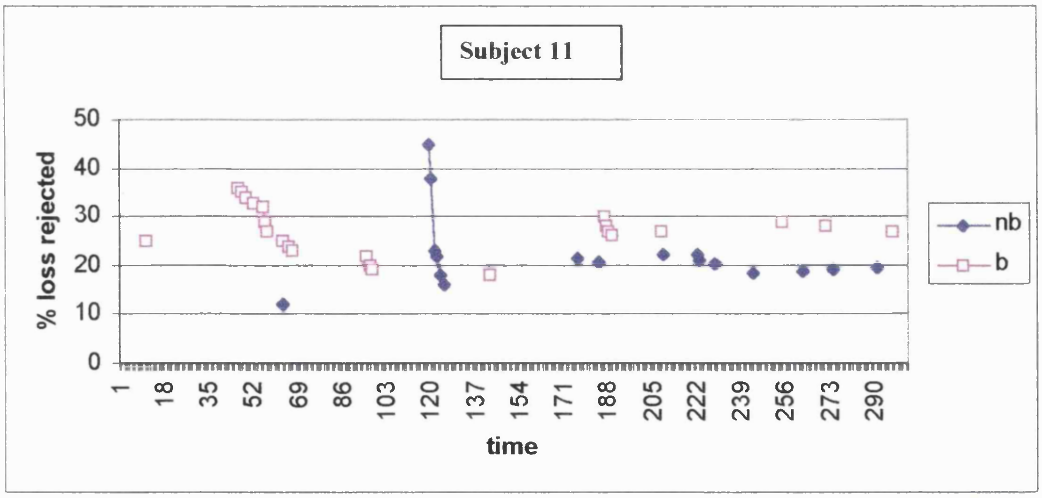
TI: It was a little bit, because if you're thinking about the game you're not thinking about the dial unless the quality drops to a level where you can no longer understand.. you have to think about it

BS: I had the quality turned up a little higher when I wasn't thinking about the budget…If you're aware that the budget might run out and you won't be able to talk anymore you might attempt to extend your budget by turning the quality down. For instance you might turn the quality right down when you're speaking and turn it up when the other person is speaking.



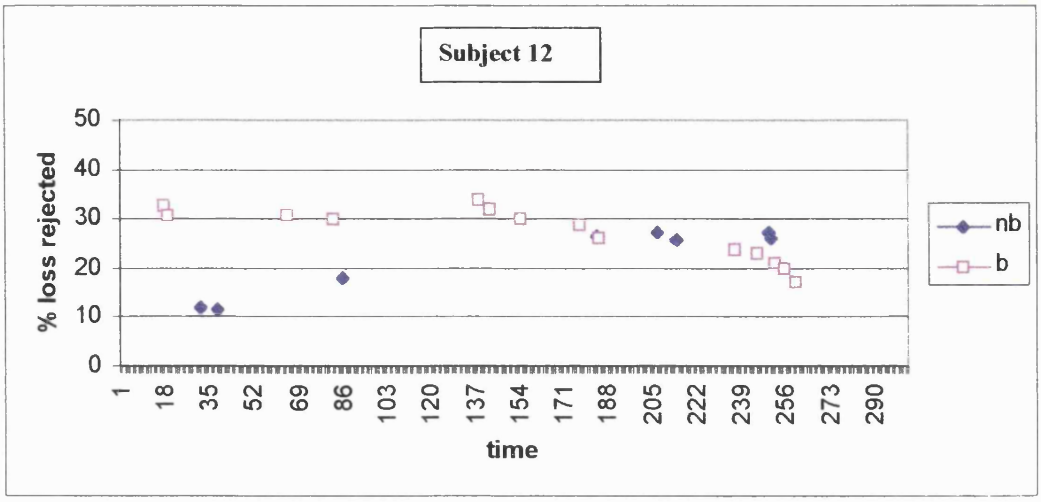
TI: No, not particularly

BS: I put it a little lower, not hideously lower… just kept an eye on the budget to make sure it wasn't plummeting down



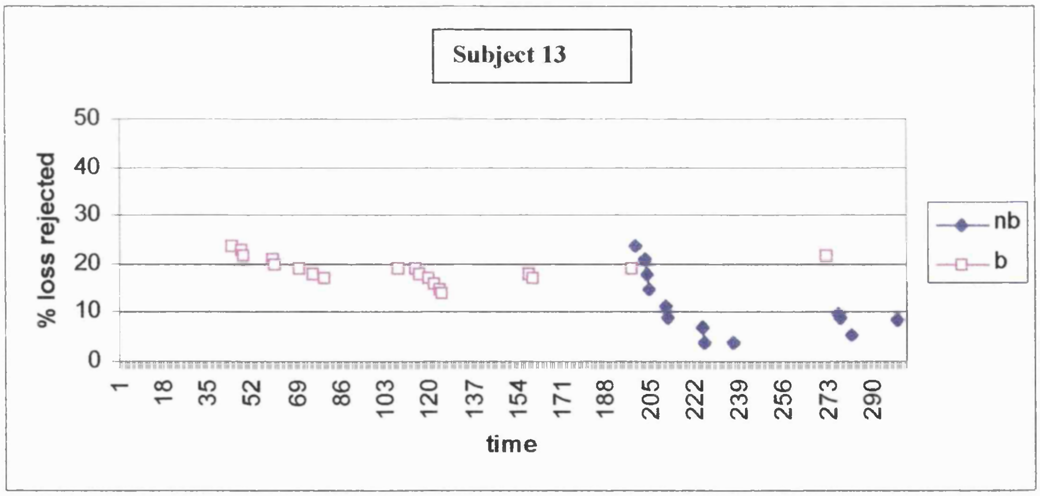
TI: No, that was cool

BS: [no relevant comment]



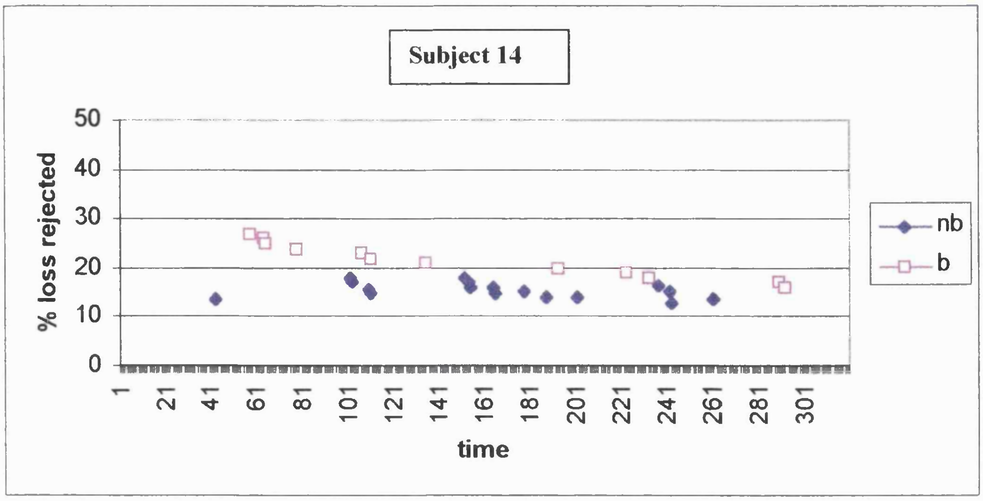
TI: No, that was ok. Sometimes maybe you forgot a bit to bring it down or bring it up

BS: Every now and then you remembered that you had a budget, but once you'd placed the cursor where you could hear best without losing information, the sound was more important



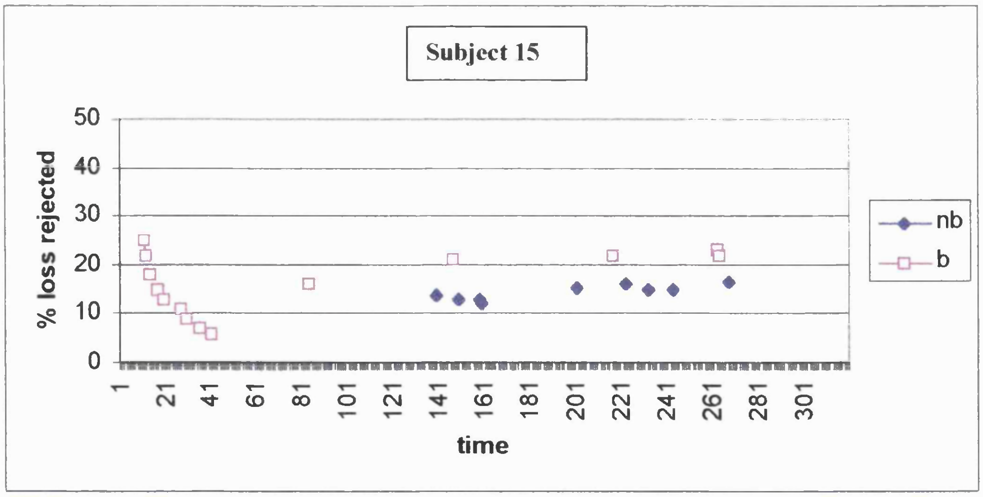
TI: it takes a minute or two to get used to the whole concept, and after it's ok.

BS: [no relevant comment]



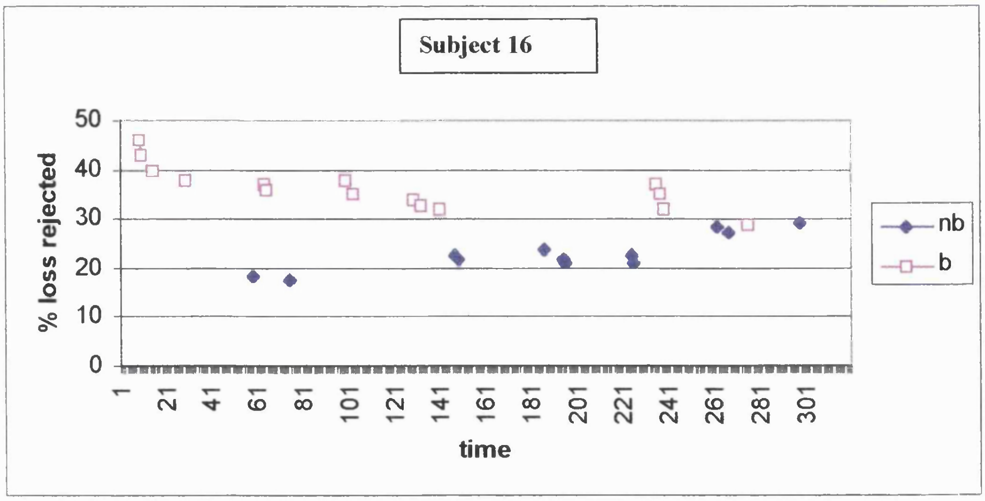
TI: Yes

BS: I was more conscious of keeping it to the bare minimum of quality so that the budget wouldn't run out. I got a really gradual scale of increasing the level. I only increased it as I needed to… really really slowly to increase the quality.



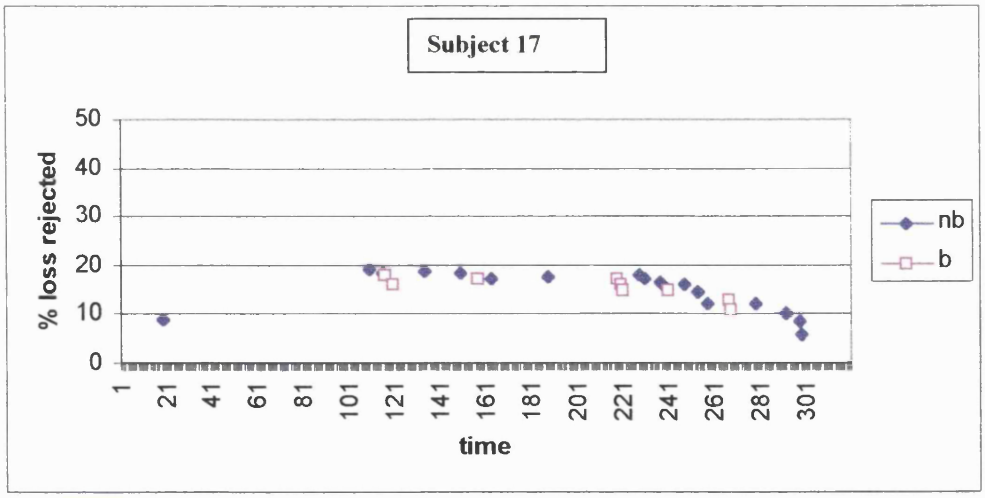
TI: It was difficult when I was reading cards, yeah, but when I was listening to [co-experimenter] reading then I could operate the slider then.

BS: I wanted to keep it so that there was a little bit of budget there left at the end



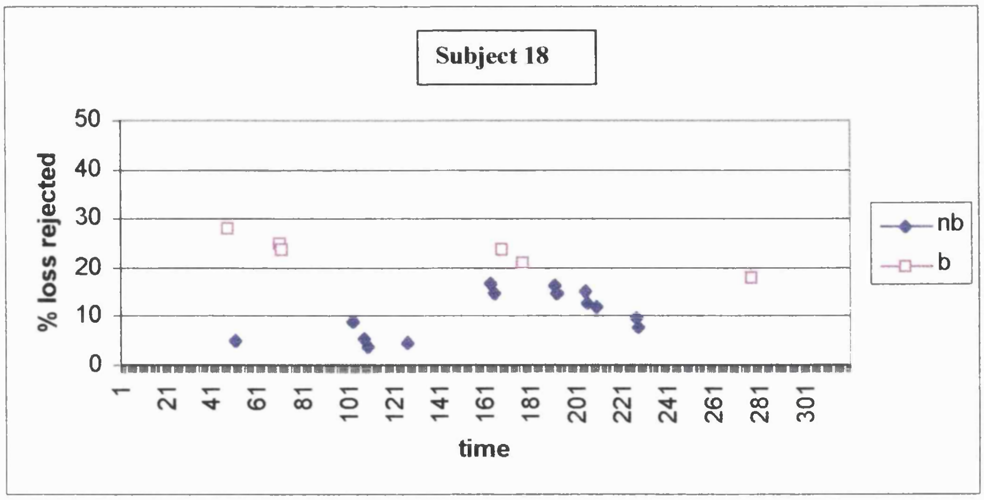
TI: It did interfere a bit because I got well into the Taboo game and forgot about the slider - tended to just leave it there instead of adjusting it all the time. Occasionally when it started cutting out badly I had to adjust it - you can only leave it alone for so long.

BS: I think it [the budget] did - I think it made me very tight. I put the slider right down. I think it made me put up with more.



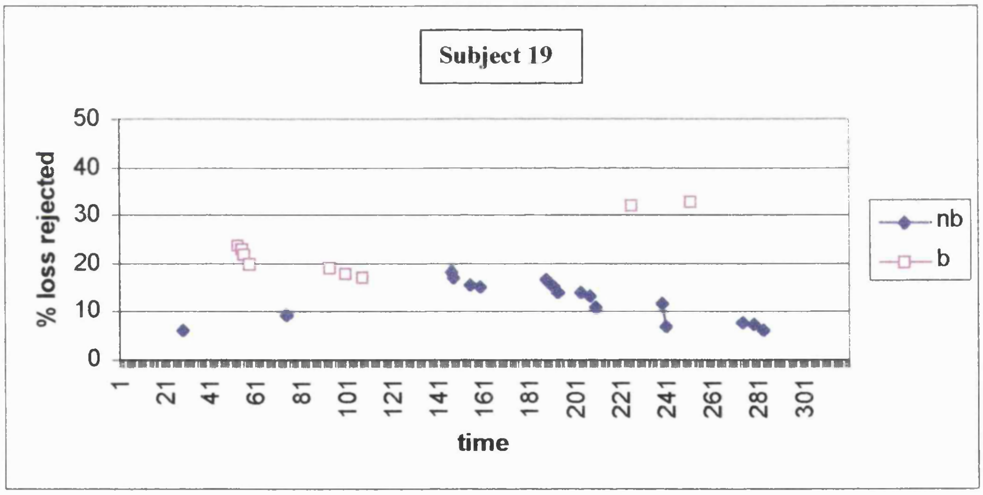
TI: The task was easy to do. You could only really concentrate on adjusting the slider when you were listening to what they were saying. You could only adjust from what you were hearing, not from what you were saying.

BS: [no relevant comment]



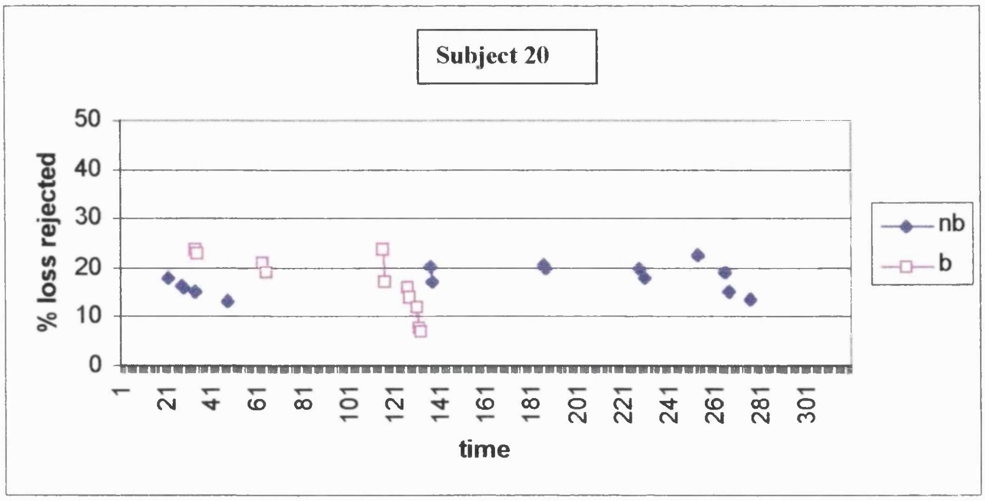
TI: Yeah, because tend to concentrate on the conversation a bit more.

BS: [no relevant comment]



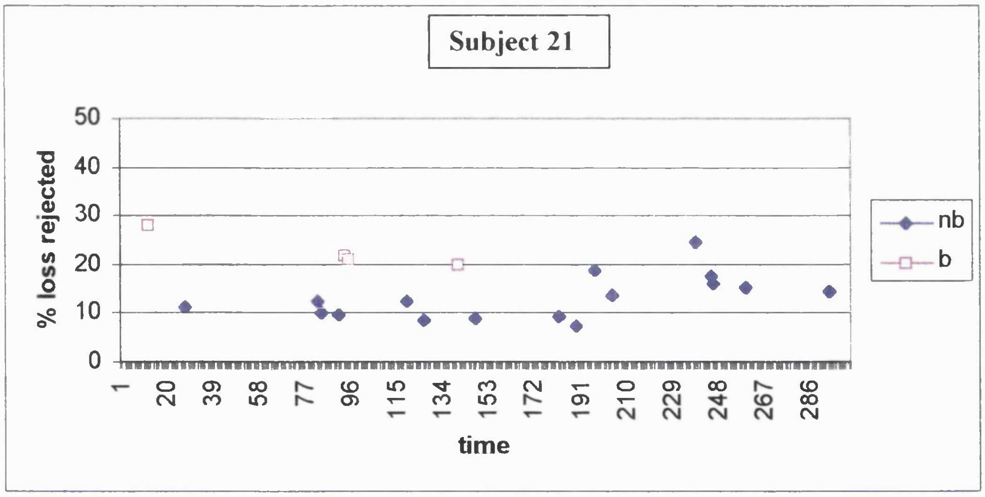
TI: It was harder when it was your go. It was quite easy when you were able to listen and then adjust it, but I think you're capable of doing both at the same time.

BS: [no relevant comment]



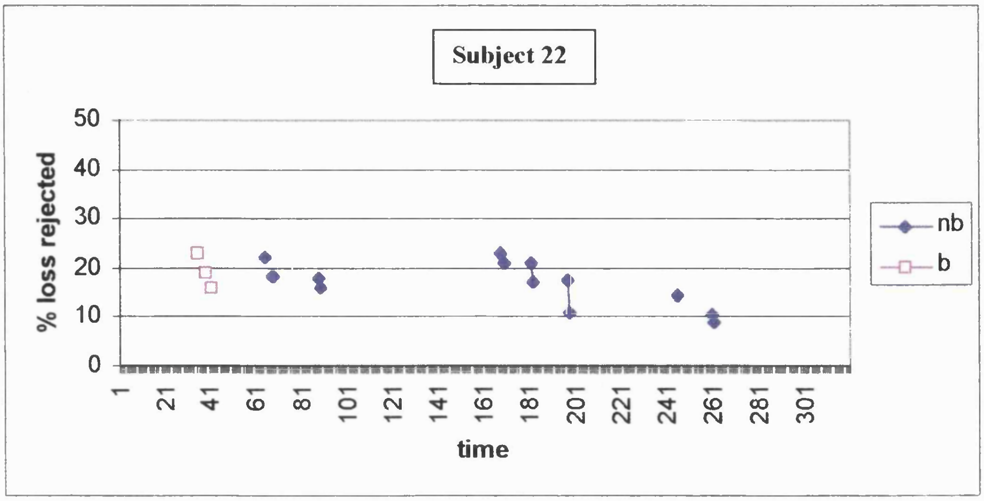
TI: Yeah, it was difficult to do when [co-experimenter] was trying to guess, like when I was speaking.. but of course I did when I was listening to [co-experimenter]. It's pointless trying to do it when you're speaking anyway because it makes no difference to your quality.

BS: [no relevant comment]



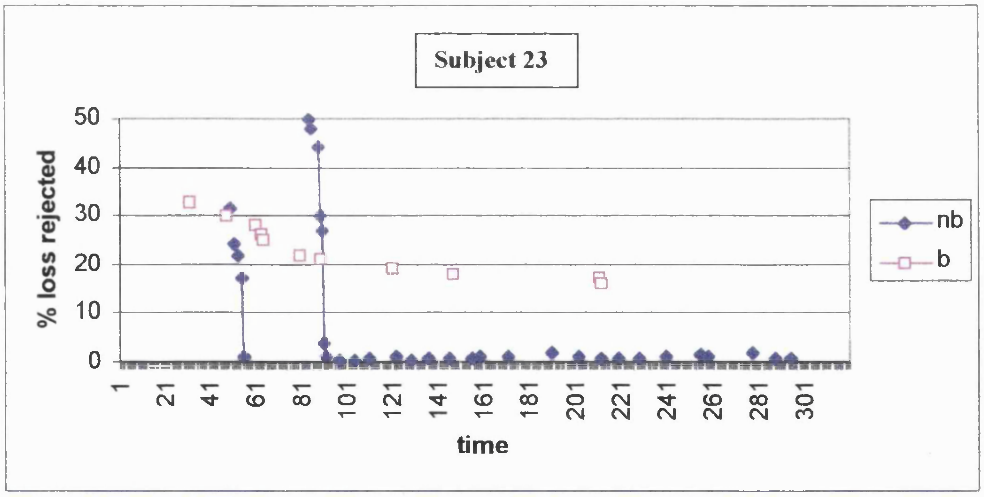
TI: I'd sort of forget that I was supposed to play with the slider so I'd find a comfortable place for it where I could hear everything and then I wouldn't really touch it that often .. so I wasn't constantly fiddling with it
but I don't know, it could all have been a trick so in fact the slider in the beginning quality goes down so I raised it

BS: I saw how fast the budget was going down … I kept it considerably higher because I knew I wouldn't run out.



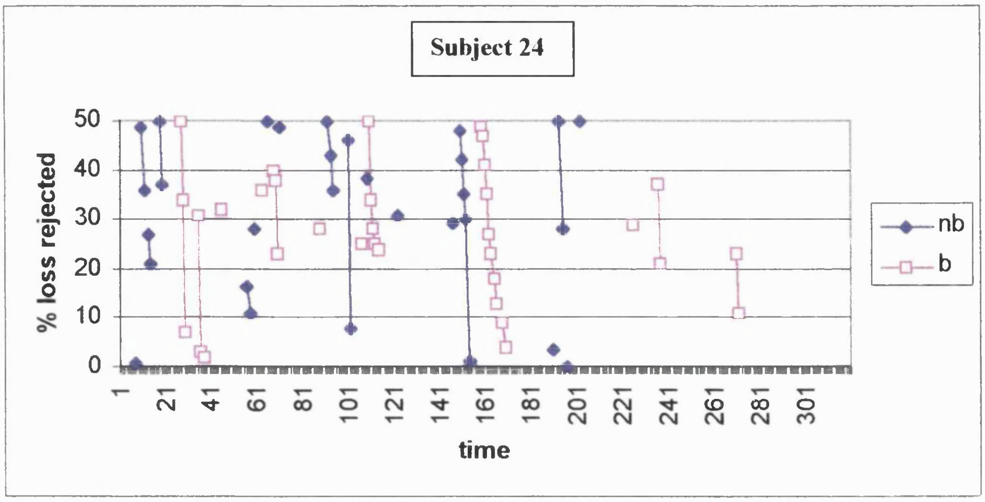
TI: Not too bad.

BS: It [the budget] did to an extent but I kept it at a fairly low level throughout to see what happens so that it was understandable all the way through.



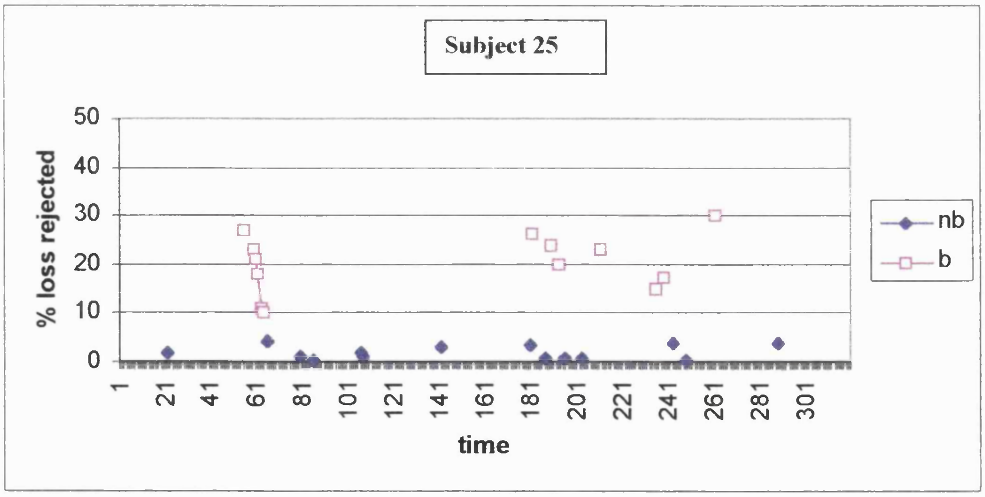
TI: No

BS: As soon as I had a budget I tried to keep it where I could just understand it, where I could just hear her and not spend money - rather than keep it at the optimum quality I just went for comprehension.



TI: It's virtually impossible to give clues and play [with the slider], but you can guess and play at the same time.

BS: Of course yeah - I found myself finding the minimum at which it would still be audible [w/budget].



TI: No, I'm fairly coordinated.

BS: [no relevant comment]

Appendix J: Raw data for Study 11, Investigating the effects of other factors on perceived quality

ref 1	ref 2	5% 1	5% 2	20% 1	20% 2	badmic 1	badmic 2	loud 1	loud 2	quiet 1	quiet 2	echo 1	echo 2
85	70	80	55	35	25	50	50	30	25	55	70	40	35
70	85	80	77	30	29	70	65	50	52	35	64	35	35
85	78	80	82	17	17	70	72	20	17	83	80	30	30
80	87	78	69	45	20	65	40	35	28	42	60	65	55
65	80	60	60	15	35	30	50	10	10	25	30	10	30
60	60	50	50	5	13	45	60	20	30	70	60	15	20
55	62	55	35	18	20	50	38	10	11	68	50	35	25
65	55	60	60	40	15	50	50	30	18	50	45	45	40
90	65	45	55	38	30	55	70	20	15	60	60	35	45
90	90	90	90	30	35	55	60	25	30	50	50	45	40
30	40	50	50	15	15	35	35	20	20	40	40	20	30
90	84	80	83	52	43	66	67	75	57	80	72	65	55
72	72	67	80	25	30	49	65	33	35	75	85	60	42
70	60	53	65	40	35	50	55	30	25	65	70	40	35
65	70	65	60	40	40	50	53	50	50	55	57	50	55
90	75	65	65	50	20	70	60	30	30	70	50	70	80
90	90	80	80	25	18	68	65	15	40	80	78	50	50
80	80	75	70	40	50	75	65	55	50	75	60	65	50
75	70	60	65	40	40	45	30	40	30	55	50	30	30
60	60	65	55	20	20	35	45	45	40	55	60	35	30
80	75	85	60	30	30	60	50	40	20	70	55	50	40
75	76	74	80	40	45	70	75	77	45	88	70	60	55
65	56	66	71	29	37	40	55	29	33	68	40	38	45
45	40	28	48	8	20	30	35	20	22	30	45	15	30

Anova: Two-Factor With Replicat 0.01

SUMMARY	reference	0.05	0.2	badmic	loud	quiet	echo	Total
1st								
Count	24	24	24	24	24	24	24	168
Sum	1732	1591	727	1283	809	1444	1003	8589
Average	72.16667	66.29167	30.29167	53.45833	33.708333	60.16667	41.79167	51.125
Variance	236.5797	217.5199	166.0417	186.6938	319.08514	295.1884	291.3895	463.5591
2nd								
Count	24	24	24	24	24	24	24	168
Sum	1680	1565	682	1310	733	1401	982	8353
Average	70	65.20833	28.41667	54.58333	30.541667	58.375	40.91667	49.72024
Variance	194.5217	181.3895	114.6014	159.6449	177.47645	188.2446	175.3841	401.029
Total								
Count	48	48	48	48	48	48	48	
Sum	3412	3156	1409	2593	1542	2845	1985	
Average	71.08333	65.75	29.35417	54.02083	32.125	59.27083	41.35417	
Variance	212.1631	195.5106	138.2336	169.8081	245.55851	237.3932	228.6166	

ANOVA

Source of Varia	*SS*	*df*	*MS*	*F*	*P-value*	*F crit*
Sample	165.7619	1	165.7619	0.799193	0.3720023	6.714231
Columns	77469.65	6	12911.61	62.25118	5.49E-51	2.858314
Interaction	130.0714	6	21.67857	0.10452	0.9958719	2.858314
Within	66786.5	322	207.4115			
Total	144552	335				

Anova: Single Factor 0.01

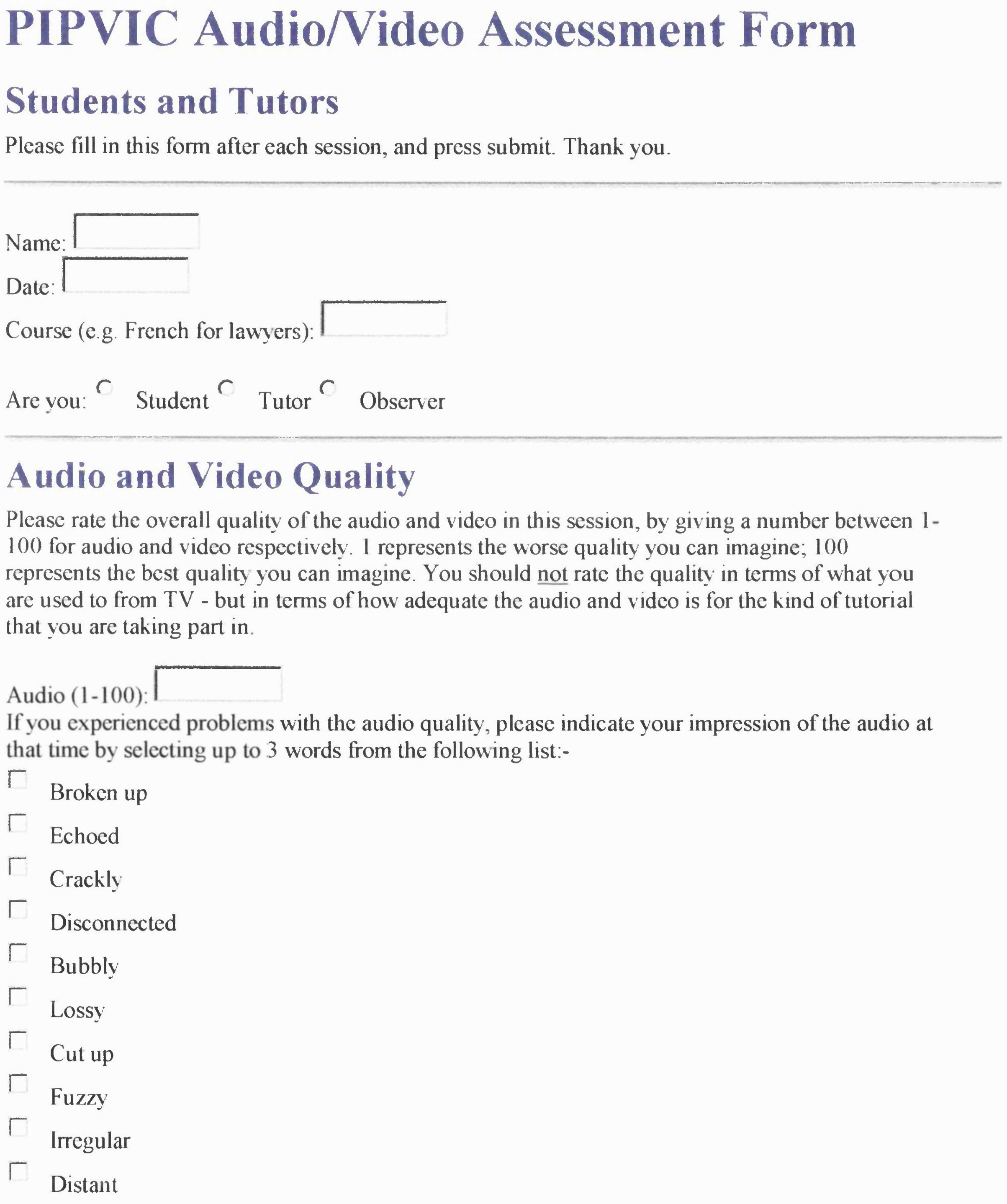
SUMMARY

Groups	Count	Sum	Average	Variance
ref	24	1706	71.08333	192.971
0.05	24	1578	65.75	169.3043
0.2	24	704.5	29.35417	105.1409
badmic	24	1296.5	54.02083	142.2713
loud	24	771	32.125	218.788
quiet	24	1422.5	59.27083	195.7822
echo	24	992.5	41.35417	210.5104

ANOVA

Source of Varia	SS	df	MS	F	P-value	F crit
Between G	38734.83	6	6455.805	36.59848	9.53E-28	2.91584
Within Gro	28399.67	161	176.3954			
Total	67134.49	167				

Appendix K: Example of a Web-based assessment form used in the PIPVIC field trials



PIPVIC Audio/Video Assessment Form

Students and Tutors

Please fill in this form after each session, and press submit. Thank you.

Name: □

Date: □

Course (e.g. French for lawyers): □

Are you: ○ Student ○ Tutor ○ Observer

Audio and Video Quality

Please rate the overall quality of the audio and video in this session, by giving a number between 1-100 for audio and video respectively. 1 represents the worse quality you can imagine; 100 represents the best quality you can imagine. You should not rate the quality in terms of what you are used to from TV - but in terms of how adequate the audio and video is for the kind of tutorial that you are taking part in.

Audio (1-100): □

If you experienced problems with the audio quality, please indicate your impression of the audio at that time by selecting up to 3 words from the following list:-

□ Broken up

□ Echoed

□ Crackly

□ Disconnected

□ Bubbly

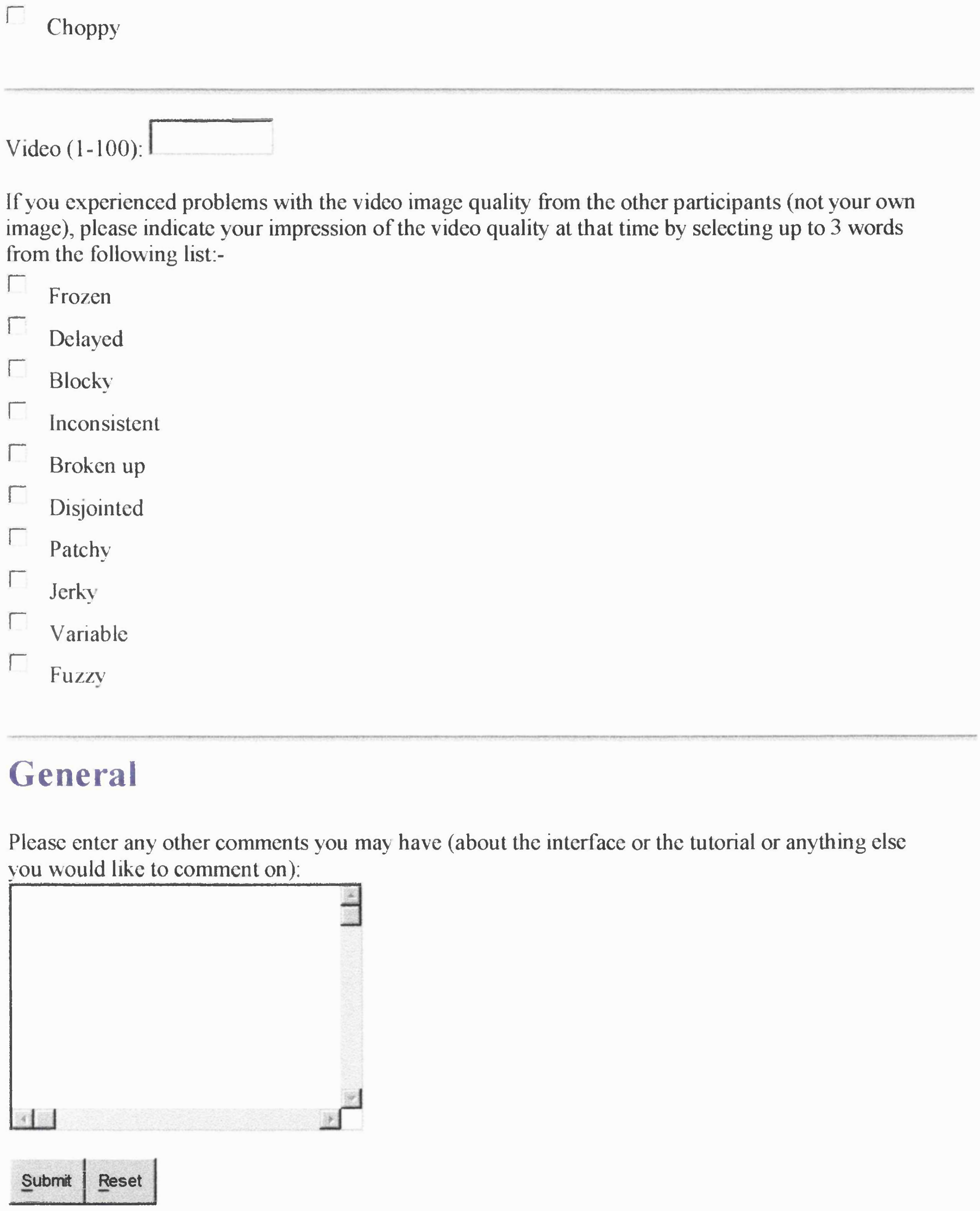
□ Lossy

□ Cut up

□ Fuzzy

□ Irregular

□ Distant



□ Choppy

Video (1-100): □

If you experienced problems with the video image quality from the other participants (not your own image), please indicate your impression of the video quality at that time by selecting up to 3 words from the following list:-

□ Frozen

□ Delayed

□ Blocky

□ Inconsistent

□ Broken up

□ Disjointed

□ Patchy

□ Jerky

□ Variable

□ Fuzzy

General

Please enter any other comments you may have (about the interface or the tutorial or anything else you would like to comment on):