New techniques for assessing audio and video quality in real-time interactive communications

Half day tutorial – Monday 10th September

Jim Mullin & Lucy Smallwood
Multimedia Communications Group
Department of Psychology
University of Glasgow
Glasgow G12 8QQ
Email: jim@mcg.gla.ac.uk, lucy@mcg.gla.ac.uk
http://www.mcg.gla.ac.uk/

Anna Watson & Gillian Wilson
Department of Computer Science
University College London
Gower Street
London WC1E 6BT
Email: g.wilson@cs.ucl.ac.uk
http://www.cs.ucl.ac.uk/

Acknowledgements:
This tutorial was produced as a result of a collaborative research project, Evaluation Taxonomy for Networked Multimedia Applications (ETNA), funded by the UK Engineering and Physical Sciences Research Council, grant numbers GR/M24462 & GR/M25599, awarded to Anne Anderson & Jim Mullin of the University of Glasgow and Angela Sasse of University College London.

Thanks to:
• Dr Angela Sasse who co-planned the tutorial with Jim Mullin and contributed to the notes;
• Professor Anne Anderson for material contributing to the notes;
• Mr Matthew Jackson of the University of Glasgow for material contributing to the notes.
# Table of Contents

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Timetable</td>
<td>4</td>
</tr>
<tr>
<td>SUMMARY</td>
<td>5</td>
</tr>
<tr>
<td>1. Networked multimedia and its usage</td>
<td>8</td>
</tr>
<tr>
<td>1.1 Context of use</td>
<td>8</td>
</tr>
<tr>
<td>1.2 Cost vs. benefit</td>
<td>8</td>
</tr>
<tr>
<td>2. Traditional methods of audio and video quality assessment</td>
<td>11</td>
</tr>
<tr>
<td>2.1 ITU Recommended Scales</td>
<td>11</td>
</tr>
<tr>
<td>2.1.1 Speech Quality Scales</td>
<td>11</td>
</tr>
<tr>
<td>2.1.2 Image Quality Scales</td>
<td>12</td>
</tr>
<tr>
<td>2.1.3 Audiovisual Quality Scales</td>
<td>12</td>
</tr>
<tr>
<td>3. Assessment of Task and Context</td>
<td>13</td>
</tr>
<tr>
<td>3.1 Task Characteristics</td>
<td>14</td>
</tr>
<tr>
<td>3.1.1 Telepresence vs. Teledata</td>
<td>14</td>
</tr>
<tr>
<td>3.1.2 Foreground vs. Background</td>
<td>14</td>
</tr>
<tr>
<td>3.1.2.1 Foreground Teledata (Video-as-Data/Audio-as-Data)</td>
<td>14</td>
</tr>
<tr>
<td>3.1.2.2 Background Teledata</td>
<td>15</td>
</tr>
<tr>
<td>3.1.2.3 Foreground Telepresence (Interpersonal Communication)</td>
<td>15</td>
</tr>
<tr>
<td>3.1.2.4 Background Telepresence (Interpersonal Awareness)</td>
<td>15</td>
</tr>
<tr>
<td>3.1.3 Media quality requirements by task type</td>
<td>15</td>
</tr>
<tr>
<td>3.1.3.1 Foreground teledata</td>
<td>15</td>
</tr>
<tr>
<td>3.1.3.2 Background teledata</td>
<td>16</td>
</tr>
<tr>
<td>3.1.3.3 Foreground telepresence</td>
<td>16</td>
</tr>
<tr>
<td>3.1.3.4 Background telepresence</td>
<td>17</td>
</tr>
<tr>
<td>3.1.4 Foreground Telepresence Task Categories</td>
<td>17</td>
</tr>
<tr>
<td>3.1.4.1 Interactive v. Non-interactive</td>
<td>17</td>
</tr>
<tr>
<td>3.1.4.2 Social vs. Cognitive tasks</td>
<td>18</td>
</tr>
<tr>
<td>3.1.4.3 Difficult/Complex tasks versus Simple/Routine tasks</td>
<td>19</td>
</tr>
<tr>
<td>3.1.4.4 Urgent Tasks</td>
<td>19</td>
</tr>
<tr>
<td>3.1.4.5 Emotive Tasks</td>
<td>19</td>
</tr>
<tr>
<td>3.2 User Characteristics</td>
<td>20</td>
</tr>
<tr>
<td>3.2.1 Sensory Impairments</td>
<td>20</td>
</tr>
<tr>
<td>3.2.2 Heterogeneity of Participants</td>
<td>20</td>
</tr>
<tr>
<td>3.2.2.1 Knowledge Background</td>
<td>20</td>
</tr>
<tr>
<td>3.2.2.2 Language Background</td>
<td>21</td>
</tr>
<tr>
<td>3.2.2.3 Personal Familiarity</td>
<td>21</td>
</tr>
<tr>
<td>3.2.3 Age of Users</td>
<td>22</td>
</tr>
<tr>
<td>3.2.4 Users’ experience with technology</td>
<td>22</td>
</tr>
<tr>
<td>3.3 Situation Characteristics</td>
<td>23</td>
</tr>
<tr>
<td>3.3.1 Geographical Remoteness of Participants</td>
<td>23</td>
</tr>
<tr>
<td>3.3.2 Amount of Background Noise</td>
<td>23</td>
</tr>
<tr>
<td>3.3.3 Number of Users</td>
<td>23</td>
</tr>
<tr>
<td>3.3.4 Distribution of Users</td>
<td>24</td>
</tr>
<tr>
<td>3.4 Applying the Taxonomy</td>
<td>25</td>
</tr>
<tr>
<td>4. Assessment Methods</td>
<td>26</td>
</tr>
<tr>
<td>4.1 Choice of methods</td>
<td>26</td>
</tr>
<tr>
<td>4.1.1 Context</td>
<td>26</td>
</tr>
<tr>
<td>4.1.2 Subjective and Objective – the differences</td>
<td>26</td>
</tr>
<tr>
<td>4.1.3 Structured v unstructured methods</td>
<td>26</td>
</tr>
<tr>
<td>4.1.4 Real vs. Contrived tasks</td>
<td>27</td>
</tr>
<tr>
<td>4.2 Task performance</td>
<td>28</td>
</tr>
<tr>
<td>4.2.1 Benchmarking</td>
<td>28</td>
</tr>
<tr>
<td>4.2.2 Output</td>
<td>28</td>
</tr>
<tr>
<td>4.2.3 Time</td>
<td>28</td>
</tr>
<tr>
<td>4.2.4 Errors</td>
<td>28</td>
</tr>
<tr>
<td>4.2.5 Costs of task performance methods</td>
<td>29</td>
</tr>
<tr>
<td>Section</td>
<td>Title</td>
</tr>
<tr>
<td>---------</td>
<td>-------</td>
</tr>
<tr>
<td>4.3</td>
<td>User satisfaction</td>
</tr>
<tr>
<td>4.3.1</td>
<td>Audio Quality</td>
</tr>
<tr>
<td>4.3.2</td>
<td>Video Quality</td>
</tr>
<tr>
<td>4.3.3</td>
<td>Complementary quality of audio and video</td>
</tr>
<tr>
<td>4.3.4</td>
<td>Post-hoc rating – quality and adequacy</td>
</tr>
<tr>
<td>4.3.5</td>
<td>Criticisms of the ITU scales</td>
</tr>
<tr>
<td>4.3.5.1</td>
<td>Real-time Internet Speech</td>
</tr>
<tr>
<td>4.3.5.2</td>
<td>Real-time Internet Video</td>
</tr>
<tr>
<td>4.3.5.3</td>
<td>The nature of the 5-point quality scale</td>
</tr>
<tr>
<td>4.3.5.4</td>
<td>Internationally interval, or internationally ordinal?</td>
</tr>
<tr>
<td>4.3.5.5</td>
<td>Summary</td>
</tr>
<tr>
<td>4.3.6</td>
<td>What is quality?</td>
</tr>
<tr>
<td>4.3.7</td>
<td>Unlabelled scale</td>
</tr>
<tr>
<td>4.3.8</td>
<td>Continuous Rating</td>
</tr>
<tr>
<td>4.3.9</td>
<td>Real-world interactive communication</td>
</tr>
<tr>
<td>4.3.10</td>
<td>Costs of Subjective Assessment Methods</td>
</tr>
<tr>
<td>4.4</td>
<td>Assessing User Cost</td>
</tr>
<tr>
<td>4.4.1</td>
<td>Physiological Measurements</td>
</tr>
<tr>
<td>4.4.2</td>
<td>The Physiology Setup at UCL</td>
</tr>
<tr>
<td>4.4.3</td>
<td>Some problems we have encountered</td>
</tr>
<tr>
<td>4.4.4</td>
<td>Some of our findings so far</td>
</tr>
<tr>
<td>4.4.4.1</td>
<td>Audio degradations due to hardware set-up and end-user behaviour</td>
</tr>
<tr>
<td>4.4.4.2</td>
<td>Differences in signal responses</td>
</tr>
<tr>
<td>4.4.4.3</td>
<td>Lack of correlation between subjective and physiological measures</td>
</tr>
<tr>
<td>4.4.5</td>
<td>How does physiology fit in with our other methods?</td>
</tr>
<tr>
<td>4.4.6</td>
<td>Methodology Guidelines</td>
</tr>
<tr>
<td>4.4.7</td>
<td>The future of physiology at UCL</td>
</tr>
<tr>
<td>4.4.8</td>
<td>Cost for physiology: capital outlay and time</td>
</tr>
<tr>
<td>4.5</td>
<td>Impact on User Behaviour</td>
</tr>
<tr>
<td>4.5.1</td>
<td>Eye tracking</td>
</tr>
<tr>
<td>4.5.1.1</td>
<td>The Eye-tracking setup in Glasgow</td>
</tr>
<tr>
<td>4.5.1.2</td>
<td>Some problems we encountered</td>
</tr>
<tr>
<td>4.5.1.3</td>
<td>Some of our findings so far</td>
</tr>
<tr>
<td>4.5.1.4</td>
<td>How does eye-tracking fit in with our other methods?</td>
</tr>
<tr>
<td>4.5.1.5</td>
<td>Costs for eye-tracking</td>
</tr>
<tr>
<td>4.5.1.6</td>
<td>Methodology Guidelines</td>
</tr>
<tr>
<td>4.5.2</td>
<td>Assessing impact on communication</td>
</tr>
<tr>
<td>4.5.2.1</td>
<td>Conversation analysis</td>
</tr>
<tr>
<td>4.5.2.2</td>
<td>Some conversation analysis findings</td>
</tr>
<tr>
<td>4.5.2.3</td>
<td>How to transcribe dialogues</td>
</tr>
<tr>
<td>4.5.2.4</td>
<td>Conversational games analysis</td>
</tr>
<tr>
<td>4.5.2.5</td>
<td>Some conversational games analysis findings</td>
</tr>
<tr>
<td>4.5.2.6</td>
<td>Costs</td>
</tr>
<tr>
<td>5.</td>
<td>Conclusions</td>
</tr>
<tr>
<td>6.</td>
<td>References</td>
</tr>
<tr>
<td>7.</td>
<td>Appendix A</td>
</tr>
</tbody>
</table>

Assessing context of use in video and audio conferencing | 59
The MARC Evaluation Checklist: Reference and Assessment | 59
Provisional Timetable

13.00  Introduction: interactive networked multimedia and its usage
       Characteristics of networked multimedia
       Context of use
       Cost vs. benefit

13.15  Background to audio and video quality assessment
       ITU MOS scales and degradation testing

13.30  Assessment context
       Assessment of task and context
       Task characteristics
       User characteristics
       Different usage, different assessment scenarios
       Taxonomy of networked communication tasks

13.45  Assessment Methods: Task Performance
       Benchmarking
       Output
       Time
       Errors

14.00  Assessment Methods: User Satisfaction
       Audio Quality
       Video Quality
       Complementary Audio and Video Quality

15.00  Assessment Methods: User Cost
       Introduction to Physiology
       The Physiological Set-up at UCL
       Problems we have encountered
       Methodology Guidelines

15.30  Assessment Methods: Impact on user behaviour
       Eye tracking introduction
       The eyetracking set-up at Glasgow
       Problems we have encountered
       Methodology guidelines

16.00  Assessment Methods: Impact on communication
       Dialogue analysis
       Content Analysis

16.15  Conclusions
SUMMARY
New techniques for assessing audio and video quality in real-time interactive communications

Jim Mullin & Lucy Smallwood
Multimedia Communications Group
Department of Psychology
University of Glasgow
Glasgow G12 8QQ
Email: jim@mcg.gla.ac.uk, lucy@mcg.gla.ac.uk

Anna Watson & Gillian Wilson
Department of Computer Science
University College London
Gower Street
London WC1E 6BT
Email: g.wilson@cs.ucl.ac.uk

ABSTRACT
The aim of the tutorial is to introduce participants to new methods for establishing audio-visual quality requirements for a range of real-time multimedia applications. Until recently, there were few HCI-specific methods for assessing audio and video quality and the usability of videoconferencing systems. The tutorial will survey methods commonly used in the telecommunications industry, and explain why these are usually not suitable for assessing usability of audio and video. The main part of the tutorial consists of description and demonstration of a range of new assessment methods that have been developed in research contexts over the past few years, and a discussion of their suitability for use in different evaluation contexts. As well as methods for measuring task performance and user satisfaction, we will cover innovative approaches to assessing the impact on user behaviour and user cost such as eye-tracking and physiological responses and the way in which users communicate with each other.

KEYWORDS: Multimedia, quality, CSCW, assessment methods.

INTRODUCTION
With the increasing popularity of networked multimedia applications such as media-on-demand and multimedia conferencing, application designers and service providers have to decide what level and type of audio and video should be provided. The awareness that such decisions have to be made with a view to user requirements and preferences has increased, but there is currently no well-founded methodology for determining what quality users both need and desire. Methods currently used in the telecommunications and broadcasting industries are not suitable because they have a different aim in that they seek only to determine whether users can spot degradations in short audio or video clips. These methods are not suitable for assessing usability of networked multimedia systems for several reasons:

Context of use
Usability should determine whether a level of quality is suitable to support the user in carrying out a particular task (such as interviewing via videoconferencing) rather than whether the user notices a degradation in passive viewing (where the task is "can you spot the degradation?"). Assessment needs to take into account users' tasks and the context in which it is used.

Cost vs. benefit
It is possible to provide real-time audio and video with no noticeable degradation, but the cost would be prohibitive. Low-cost multimedia applications are rapidly gaining in popularity, and mobile multimedia services are becoming available. Bandwidth of connections and processing power of devices is limited. The assessment question asked in this context is "what level of quality is good enough?", and
how limited resources should be utilised to provide maximum benefit to the user.

**Networked multimedia quality**

Networked multimedia quality is determined by a number of factors. Traditionally, assessment has focused on the effect of different compression techniques and coding errors. With networked multimedia, the effects of delay and different levels of bandwidth have been investigated; delivery over the Internet adds the effects of packet loss and jitter; delivery over mobile systems means that connections may be interrupted or lost. Audio and video quality can also be affected by peripherals (such as microphones and speakers) and user operation of applications and peripherals. Finally, audio quality can influence the assessment of video quality, and vice versa.

The tutorial introduces a framework for audio and video quality assessment based on HCI evaluation criteria (task performance, user satisfaction, and user cost), and explains how these criteria should be considered and balanced in any specific evaluation. In a work context, for instance, task performance is usually the most important consideration, whereas in a home entertainment situation, user satisfaction is the primary consideration.

We then present a catalogue of assessment methods that can be used to collect data for each of the three criteria. When choosing and applying a set of assessment methods, we have to consider which of the following provide relevant input to answering the assessment question.

- Type of data (subjective/objective, quantitative/qualitative)
- Data source (user report/observational)
- Granularity and timing of data collection (continuous/post-hoc)

**ASSESSMENT OF TASK AND CONTEXT**

The first consideration of a principled approach to evaluation is the recognition that any interaction involves both a user and a system working together in a particular physical and/or situational context to complete a task. Thus, any particular characteristics of user, task and physical and/or situational context that may generate particular requirements or preferences regarding the level and type of audio and video quality have to be identified, and replicated in the assessment as far as possible. It would, for instance, be very misleading to determine audio quality characteristics via experiments in a quiet lab situation with assessment being the only task, when the application is to be used in an environment with high background noise and many competing demands on the user.

**Task performance**

Task performance can be measured in a number of ways. We consider effectiveness (can the task be completed) and efficiency (over what duration, and with how many errors). When choosing indicators of task performance for a controlled experiment, we must be guided by how that task would be judged in the real world. This will provide benchmarks for assessing user performance at the different levels of quality, and thus identify the level at which necessary performance can be sustained, or optimum performance reached.

**User satisfaction**

Currently, subjective methods of rating quality dominate audio and video quality assessment. Typically, users are asked to rate the quality on a 5-point scale of bad-poor-fair-good-excellent; ratings are then averaged across users and/or repeated assessments to produce a Mean Opinion Score (MOS). Whilst subjective ratings are important because they detect when users are actually bothered by some aspect of the quality they experience, we advocate:

- more specific assessment questions (asking users to rate the adequacy and quality for the particular task they carried out), and
- use of a 100 point scale with positive and negative poles, since the 5-point MOS scale has been shown to be unreliable.

In longer-term field trials, logs allowing a correlation between the frequency and duration of
use and the quality delivered can provide some objective measure of user satisfaction.

**User cost**

As stated above, measures of user satisfaction will indicate that users are bothered by some aspect of the quality they experience. We cannot, however, assume that users can reliably identify and report all effects of audio and video quality they experience, especially after short periods of use: effects may be too subtle to be noticed, or attributed to other aspects of the technology or task.

User cost can be measured subjectively through mood scales, and objectively via users' physiological responses, which are indicative of perceptual strain. When a user is presented with insufficient audio and video quality in a multimedia conference, he/she has to expend extra effort at the perceptual level. If they struggle to decode the information, this should induce a response of discomfort/stress, even if the user remains capable of performing the main task. We have used Heart Rate (HR), Blood Volume Pulse (BVP) and Galvanic Skin Response (GSR) to detect perceptual strain in response to low video quality and audio degradations.

**Impact on user behaviour**

What people do during an interaction via computer mediated communication can be difficult to measure. Our recent research has led us to consider what people are looking at during interaction. Advances in eye tracking technology have produced remote eye tracking devices, which no longer in themselves have an adverse impact on the communication processes. While our research in this area is in its infancy, we believe that, in conjunction with other methods, eye tracking can provide insight into explanations for user behaviour.

**Impact on communication**

In computer mediated communication, a key question in assessment is often whether the technology and the levels of quality impact on the process of the communication. We have had success employing measures of task performance such as efficiency and effectiveness of the conversational dialogues. Other methods can be used to look at the content of the conversation in order to consider the more subtle implications of the technology.
1. NETWORKED MULTIMEDIA AND ITS USAGE

The aim of the tutorial is to introduce the participants to a principled HCI approach to assessing audio and video quality in real-time interactive multimedia environments, and a range of novel methods emerging from the research community. The aim of any assessment must be to determine the minimum levels of audio and video quality required for the intended users to carry out their communication tasks effectively and without undue cost to themselves. There are existing methods used in the telecommunications and broadcasting industry, but their aim is different in that they seek to determine whether users can spot degradations in short audio or video clips. These methods are not suitable for assessing usability of networked multimedia systems for a number of reasons:

- The degradation effects on networked multimedia are of a different nature than in broadcasting or telecommunications, and short clips are not suitable for replicating these effects.
- Asking users to determine whether degradations are noticeable or not in a passive viewing environment is not a suitable usability criterion for networked multimedia communication: many users want to use lower quality networked multimedia communications because broadcast quality would be unattainably expensive.
- Audio quality has an effect on the perception of video quality, and vice versa.

We will briefly survey the use of networked multimedia communications in different contexts of use today, and who the users are, followed by a taxonomy of communication tasks. The main part of the tutorial consists of description and demonstration of a range of new assessment methods that have been developed in research contexts over the past few years. We will show how they can be applied, and how data generated can be analysed and interpreted (using example data we have collected in previous assessments), what resources and skills are required, and what the pitfalls of these methods are. There will be a demonstration of physiological monitoring, and video recordings of eyetracking recording and analysis.

With the increasing popularity of networked multimedia applications such as media-on-demand and multimedia conferencing, application designers and service providers have to decide what level and type of audio and video should be provided. The awareness that such decisions have to be made with a view to user requirements and preferences has increased, but there is currently no well-founded methodology for determining what quality users both need and desire. Methods currently used in the telecommunications and broadcasting industries are not suitable because they have a different aim in that they seek only to determine whether users can spot degradations in short audio or video clips. These methods are not suitable for assessing usability of networked multimedia systems for several reasons:

1.1 Context of use

Usability should determine whether a level of quality is suitable to support the user in carrying out a particular task (such as interviewing via videoconferencing) rather than whether the user notices a degradation in passive viewing (where the task is "can you spot the degradation?"). Assessment needs to take into account users' tasks and the context in which it is used.

1.2 Cost vs. benefit

It is possible to provide real-time audio and video with no noticeable degradation, but the cost would be prohibitive. Low-cost multimedia applications are rapidly gaining in popularity, and mobile multimedia services are becoming available. Bandwidth of connections and processing power of devices is limited. The assessment question asked in this context is "what level of quality is good enough?" and how limited resources should
be utilised to provide maximum benefit to the user.

Characteristics of networked multimedia
In networked digital multimedia, two or more multimedia end-systems communicate with each other over a network. The end systems can be dedicated communication equipment (such as videophones and videoconferencing terminals), general-purpose workstations such as PCs and laptops, and mobile devices such as PDAs and mobile phones.

The data transmitted may be:
- audio only (e.g. IP telephony)
- video only (e.g. web traffic camera)
- audio and videoconferencing (e.g. videotelephony)
- audio and/or video combined with other data (text, still images, animations - e.g. multimedia conferencing with a shared text editor – see Fig 1.1 for an example)

In CSCW terms, digital multimedia can support synchronous (real-time, such as conferencing) or asynchronous (multimedia message) applications. Whilst the assessment methods presented in this tutorial have been developed working with technology for supporting real-time interactions, most of them can be used or adapted for assessment of asynchronous tasks such as media streaming of music, or entertainment videos. Digital multimedia is also increasingly embedded in Web applications (e.g. a video walkthrough of a house, with audio commentary, on an Estate Agent’s Website).

Networked multimedia quality is determined by a number of factors. The quality at which networked multimedia can be delivered is largely determined by the bandwidth of the networked connection, and the processing power of the end-system (speed at which audio and video can be encoded and decoded). The quality that can be achieved increases with bandwidth and processing power, but so does the cost.

All network transmission is subject to some delay - for transmitting data this is usually be-
However, delivery of multimedia data can incur additional delays due to:
- time it takes to encode and/or decode the data
- insufficient bandwidth available on the network

In circuit-switched networks (such as ISDN), a set amount of bandwidth for the duration of the session is allocated. In packet-switched networks (such as the Internet), effects such as packet loss and jitter may impact the quality of audio and video received. Delivery over mobile systems may mean connections can be interrupted for short periods (e.g. leading to a "frozen video image"), or lost entirely.

Whilst the "impossibly low" quality of Internet multimedia was the subject of gloomy predictions 5 years ago ("it will never take off", ISDN salesmen used to say), businesses and consumers have voted with their feet. According to recent statistics, multimedia now accounts for 70% of Internet traffic, supporting everything from live webcasts (Bill Gates, Madonna) to Gaelic tutorials shared between Scottish Highlanders and their friends in Nova Scotia.

Application designers and network providers need to make decisions about the basic level of network connectivity and power required for the end-system. They also need to decide how the application using the networked multimedia should deal with network problems (e.g. lower the video quality, or stop video transmission).

Finally, audio and video quality can also be affected by peripherals (such as microphones and speakers) and user operation of applications and peripherals. This is often overlooked, but a usability assessment needs to consider the combined effect of all components involved in delivering multimedia to the user.

The research we advocate utilises a range of analytic techniques which we have devised to investigate the impacts of cognitive technologies. All of these techniques have provided insights into the impacts of technologies on users and the ways in which users adapt (or fail to adapt) to new ways of interacting. The implications of these studies for the design and implementation of effective multimedia technologies will be discussed.
2. TRADITIONAL METHODS OF AUDIO AND VIDEO QUALITY ASSESSMENT

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 International Telecommunication Union (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.

2.1 ITU Recommended Scales

ITU-T and ITU-R recommendations address subjective assessment of speech transmission over telephone networks and image quality over television systems, respectively. A series of ITU-T recommendations also address the subjective assessment of multimedia applications. The recommended scales fall into four main categories.

2.1.1 Speech Quality Scales

For the assessment of speech quality, the recommended rating scale for both listening-only and conversation tests is a 5-point category scale commonly known as the quality scale (ITU-T P.800). Listening-only tests can also be assessed via the listening effort scale. In conversation tests, a binary difficulty scale follows the (connection) quality scale. These scales are shown in Figure 2.1 (a-c).

<table>
<thead>
<tr>
<th>Quality of the speech/ connection</th>
<th>Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>Excellent</td>
<td>5</td>
</tr>
<tr>
<td>Good</td>
<td>4</td>
</tr>
<tr>
<td>Fair</td>
<td>3</td>
</tr>
<tr>
<td>Poor</td>
<td>2</td>
</tr>
<tr>
<td>Bad</td>
<td>1</td>
</tr>
</tbody>
</table>

(a) Listening -quality scale

<table>
<thead>
<tr>
<th>Effort required to understand the meaning of sentences</th>
<th>Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>Complete relaxation possible; no effort required</td>
<td>5</td>
</tr>
<tr>
<td>Attention necessary; no appreciable effort required</td>
<td>4</td>
</tr>
<tr>
<td>Moderate effort required</td>
<td>3</td>
</tr>
<tr>
<td>Considerable effort required</td>
<td>2</td>
</tr>
<tr>
<td>No meaning understood with any feasible effort</td>
<td>1</td>
</tr>
</tbody>
</table>

(b) Listening -effort scale

<table>
<thead>
<tr>
<th>Did you or your partner have any difficulty in talking or hearing over the connection?</th>
</tr>
</thead>
<tbody>
<tr>
<td>Yes</td>
</tr>
<tr>
<td>No</td>
</tr>
</tbody>
</table>

(c) Conversation difficulty scale

<table>
<thead>
<tr>
<th>Image quality</th>
<th>Score</th>
<th>Image impairment</th>
<th>Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>Excellent</td>
<td>5</td>
<td>Imperceptible</td>
<td>5</td>
</tr>
<tr>
<td>Good</td>
<td>4</td>
<td>Perceptible, but not annoying</td>
<td>4</td>
</tr>
<tr>
<td>Fair</td>
<td>3</td>
<td>Slightly annoying</td>
<td>3</td>
</tr>
<tr>
<td>Poor</td>
<td>2</td>
<td>Annoying</td>
<td>2</td>
</tr>
<tr>
<td>Bad</td>
<td>1</td>
<td>Very annoying</td>
<td>1</td>
</tr>
</tbody>
</table>

(d) Image quality scale

<table>
<thead>
<tr>
<th>Image impairment</th>
<th>Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>Imperceptible</td>
<td>5</td>
</tr>
<tr>
<td>Perceptible, but not annoying</td>
<td>4</td>
</tr>
<tr>
<td>Slightly annoying</td>
<td>3</td>
</tr>
<tr>
<td>Annoying</td>
<td>2</td>
</tr>
<tr>
<td>Very annoying</td>
<td>1</td>
</tr>
</tbody>
</table>

(e) Image impairment scale

Figure 2.1 ITU recommended speech and image quality measurement scales
2.1.2 Image Quality Scales
For the assessment of image quality, single
stimulus methods are rated using the quality
scale or impairment scale, and comparisons to
reference conditions are made using the dou-
ble-stimulus continuous quality scale (DSCQS) or the double stimulus impairment
scale (ITU-R BT.500-8). These scales are
shown in Figure 2.1 (d-f).

2.1.3 Audio-visual Quality Scales
Methods for the assessment of audio-visual
communications are presented in (ITU-T
P.920). The overall methodology is based on
conversation opinion tests. The 5-point qual-
ity scale is recommended for assessing the
video quality, the audio quality and the over-
all audio-visual quality. A 5-point 'effort
needed to interrupt' scale can also be used.

The utility of these scales with respect to
speech and video in real-time multimedia
communication over the Internet is considered
later in the tutorial.
3. ASSESSMENT OF TASK AND CONTEXT

In order to assess the impact of quality levels of real-time interaction via audio and video conferencing we have to put the interaction into the context within which it will be used.

To this end we split the factors that may have an influence on the interaction into three groups: task, user and situation. The tasks are then split into the further groups as shown below in Fig 3.1.

The characteristics of these tasks and how they impact on the quality requirements are discussed through this section.

Figure 3.1: Flow chart defining relevant task characteristics at a simplistic level
3.1 Task Characteristics
The term “task” could be defined at one of several levels of granularity from as high a level as “videoconference” to as low-level an action as performing a greeting or attempting to make eye contact (Rimmel, Hollier & Voelcker, 1998). The highest level is too broad to be useful as it combines several types of activity that might have different quality requirements while the lowest level is impractical. We therefore classify tasks at a level between these two extremes, according to the characteristics of the main activities of an individual session.

At the highest level, applications can be divided into Telepresence and Teledata and into Foreground and Background tasks, a division derived in part from Buxton (1995).

3.1.1 Telepresence vs. Teledata
This is the distinction between applications where the audio and video channels are used to support communication or at least awareness between users (Telepresence) and applications where they are used to carry other sorts of information (Teledata). Telepresence covers uses such as the Portholes system (Dourish & Bly, 1992) while teledata covers images of shared work objects (e.g. Nardi et al., 1993; Gaver et al., 1993) and includes Whittaker’s Video-as-

3.1.2 Foreground vs. Background
Buxton (1995) divides tasks into foreground tasks which take place with the user’s full attention and background tasks which do not. This may have implications for quality requirements as if a user pays less attention to a channel then quality may not need to be as high (Hollier & Voelcker, 1997). Foreground tasks would include tasks where the user has to react to or monitor or interact through the audio and video channels and background tasks would be activities where the user’s involvement is more passive and less purposeful.

As shown in Fig 3.2, these two main distinctions divide applications into four main types:

3.1.2.1 Foreground Teledata (Video-as-Data/Audio-as-Data)
This includes tasks where the user reacts to or monitors auditory or visual data that does not involve direct interpersonal communication between users. An example of Video-as-Data would include neurosurgery data (e.g. see Nardi et al., 1993). Requirements for video quality will generally be high especially in safety critical applications such as neurosurgery although other applications might require slightly lower quality such as film clips of houses provided by an estate agent. The precise requirements will vary widely depending on the exact nature of the information but it is expected that the requirements would be for high quality. Audio-as-Data might include music or even situations where the audio channel is being used to transmit verbal information but where the actual sound of the words is very important e.g. learning a foreign language.
Requirements for the audio-as-data function will probably be higher than for normal speech with users requiring lower levels of packet loss and a higher sample rate.

3.1.2.2 Background Teledata
This might include web cameras in the case of video data or an ambient soundtrack to a video clip in the case of audio data. Audio and video quality would not necessarily need to be so high for these sorts of tasks.

3.1.2.3 Foreground Telepresence (Personal Communication)
Foreground Telepresence is broken down further with applications divided into interactive or non-interactive tasks and into tasks where the main purpose of communication is problem solving or the exchange of information (called “cognitive” tasks for lack of a better name) or where the task is more social or negotiative, potentially requiring more access to non-verbal information. Therefore foreground telepresence applications are divided into Interactive Cognitive (for example the map task (Brown et al., 1984) which requires information exchange in an interactive conversational setting), Interactive Social (e.g., a business meeting involving negotiation, chatting with a relative on a video phone), Non-interactive Cognitive (for example remote lecturing situations where listeners may not have an opportunity to interact with the lecturer as they would in a normal conversation) and Non-interactive Social (this might include situations where only one person speaks but there is still a social element – the purpose of the message might, for example, be persuasion).

This includes all situations where the audio and video channels are used for interpersonal communication of some form e.g. videoconferencing, distributed lectures etc.

3.1.2.4 Background Telepresence (Personal Awareness)
This includes applications such as the Portholes system (Belotti & Dourish, 1997) which allow passive awareness of others’ activity without direct interaction. Audio and video requirements will be somewhat lower than for other telepresence applications. Users may tolerate smaller image sizes, lower frame rates and possibly poorer image quality. Audio need not be as important.

3.1.3 Media quality requirements by task type

3.1.3.1 Foreground teledata
Probably the main distinction is between whether it is the visual or auditory (or both) components that are important to the task. Aptekar et al. (1995) found that judgements of the acceptability of the quality of a video clip were affected by the relative importance of video and audio information for understanding the message. When video was of little importance compared to audio (e.g. advertisements or a clip from a talk show) subjects were willing to accept lower frame rates than when it was more important (e.g. clips of snooker). Video and audio quality requirements are therefore likely to depend on the actual importance of that medium to the task (see Fig 3.3). When either audio or video is relatively unimportant to the task the quality required will be reduced. In some cases of course only one channel (video or audio) will be present.

![Figure 3.3: Video in foreground teledata is further categorised into high and low temporal](image-url)
Temporality/Degree of motion

Apteker et al. (1995) found that the effect of frame rate on video acceptability was affected by what they called the temporality of the data – the amount of motion as measured by the amount of change in the image between frames and the resulting compression ratio. They found that low temporal video-data which changed little between frames was rated more poorly at low frame rates than high temporal data. However this may have been due to some of the materials they used e.g. clips of snooker. The total degree of change between successive frames may not always be the most important aspect. In some cases the change may be relatively small but crucial to understanding the information. For example in snooker clips much of the screen may remain static but the relatively small movements that do occur such as when a ball is in motion, may be relatively important. A better distinction for our purposes may be the extent to which motion is important for following events, although temporality as defined by Apteker et al. may be important.

It might be predicted that frame rates can be lower if changes occur less often or if motion detection is less important to the task. There will however be an absolute lower acceptable limit. In a review of several studies, Kies et al. (1997) suggest a lower limit of around 5Hz for most tasks.

Degree of visual detail

The degree of visual detail required to carry out the task may also be important. Neurosurgery might require greater visual detail than seeing a video clip of a potential house purchase and require better spatial and contrast resolution. Again the amount of visual detail necessary to successfully perform the task may be more important than the degree of detail in the object being viewed though this may also be important.

3.1.3.2 Background teledata

Requirements may be less stringent for background teledata. Several of the low audio and visual importance stimuli used by Apteker et al. might be classed as background teledata e.g. test patterns. They found that even quite poor video was acceptable in these cases. Similarly quality is less important for background audio data. In both cases however very low quality audio and video may prove a distraction from the user’s main task and thus be unacceptable.

3.1.3.3 Foreground telepresence

For foreground telepresence, the speech channel is the primary means of communication making audio particularly important. When video is used to display other participants, speech perception can be helped by the ability to lip-read, especially when it might otherwise be difficult such as when there is a lot of background noise (Summerfield, 1992) or when speech is in a foreign language (Reisberg et al., 1987). The influence of vision on speech seems to be maximised at frame rates above 15-16.7 Hz (Barber & Laws, 1994; Frowein et al., 1991; Nakazono, 1998). Barber & Laws also recommend a spatial resolution of at least 128x128 pixels and a contrast resolution of at least 64 grey levels for speech based tasks.

Low video quality can actually impaire speech perception with an increase in mishearing for certain syllables at low frame rates (less than 5 frames per second) caused by a mismatch between visual and auditory cues (Nakazono, 1998). Frame rate seems to be the most important determinant of video quality in verbal tasks with users willing to sacrifice other aspects of image quality to keep the frame rate above 5Hz (Pappas & Hinds, cited by Kies et al., 1997).

The extent to which the results of these studies will extend to real tasks is unclarified – in normal conversation participants spend relatively little time looking at each other and can also adjust the clarity of their speech to compensate for the lack of visual cues (Anderson et al., 1997). They also do this in response to poor video (Blokland & Anderson, 1998) though this will almost certainly require greater effort on the part of speakers.
Communication involves more than the simple reception of verbal information. Conversation also requires co-ordination of communication content and process (Clark & Brennan, 1991; Whittaker, 1995). This involves provision of feedback on speakers’ utterances through back-channels etc to indicate understanding and maintenance of smooth turn taking. We will discuss the methods for analysing this flow of conversation later on. Both of these processes are affected by audio delay (O’Connail et al., 1993) and possibly by other aspects of audio and video quality. The extent to which this affects task performance or user perceptions of quality will depend on the task.

Visual feedback often seems to have little effect on task performance (e.g. Anderson et al., 1996; Doherty-Sneddon et al., 1997) although it may become more important if the task or communication process is more difficult (Rudman et al., 1997; Veinott et al., 1997). Certain tasks e.g. those involving negotiation may require access to non-verbal information (Short et al., 1976). This may have implications for video and probably audio quality requirements. However, given the low impact of visual access to other participants on most co-operative tasks, video quality may be relatively unimportant if lip-reading or access to social cues is not required, although it is assumed that there is a minimum acceptable threshold.

### 3.1.3.4 Background telepresence

As these applications involve only passive awareness of others and not direct communication, audio and video requirements will be somewhat lower than for other telepresence applications. There will be no need, for example, to provide good synchronisation between audio and video for lip movements. Users may tolerate smaller image sizes, lower frame rates and possibly poorer image quality. Audio need not be as important.

### 3.1.4 Foreground Telepresence Task Categories

Foreground telepresence covers many types of task and can be broken down further (Fig 3.4). There are a number of other task distinctions which apply specifically to foreground telepresence.

#### 3.1.4.1 Interactive v. Non-interactive

Tasks will vary in their interactivity with some such as presentations or lectures mainly involving one speaker talking in monologue with few or no changes of speaker or instances of overlapping speech. Other tasks such as group discussions will be much more conversational. Certain situations will fall somewhere in between such as formal business meetings with a chair or group tutorials.

Group size plays a part. One-to-one and small group situations where each user can interact with the others will often involve a conversational style of interaction and be highly interactive while situations where one site transmits to many (e.g. lectures, presentations) will be non-interactive. Many-to-many situations with larger groups will fall in between. This may be unacceptable in conversational settings and may affect performance on tasks that require moment-to-moment co-ordination or a high level of understanding (e.g. O’Malley et al., 1996). In more formal situations such as lectures, delay may be less noticeable and therefore more acceptable.
There may be implications for speech quality. In an interactive situation a listener has the opportunity to ask the speaker to repeat or clarify if the audio signal is unclear or unintelligible. In a non-interactive setting, listeners will not have the opportunity to interrupt and if key parts of the message are inaudible, may lose track of conversation flow. The same level of audio degradation may have greater impact in non-interactive situations, requiring higher audio quality.

Certain aspects of video quality, such as image size, appear to affect the perceived or actual interactivity of the conversation (e.g. Anderson et al., 1996; Monk & Watts, 1995). It may be that required levels of video quality differ between interactive and non-interactive tasks.

3.1.4.2 Social vs. Cognitive tasks

Generally tasks involving the transfer of information or co-operative problem solving (“Cognitive” tasks) seem to benefit little from the addition of a video channel (Short et al., 1976; Whittaker, 1995; Williams, 1977) with no differences found between audio-only and video conditions (e.g. Chapanis et al., 1972; Gale, 1990), video channels of different quality (e.g. 5Hz frame rate versus 30Hz, Masoodian et al., 1995), large or small video-windows, (Anderson et al., 1996) or when video is replaced by a still image (Vons, Vertegaal & Van der Veer, 1998). This suggests that for this sort of task there may be little benefit in improving video quality beyond the point at which it ceases to impair speech perception or prove distracting. However video quality may be more important when the task is more complex (e.g. Olson et al., 1995; Rudman et al., 1997) or when communication is more difficult (e.g. Veinott et al., 1997). In certain tasks an improvement was found face-to-face compared to an audio-only condition but not in video.

Possibly the results for video may have been affected by users lack of familiarity with video and it may be that experienced users benefit from higher quality video in ways that novice users do not (Rudman et al., 1997). Finally Monk et al. (1996) suggest that although users may be able to perform equally well without video it may be at the expense of greater effort. This may lead to greater stress during the interaction which is measurable – see Section 4.4 on Assessing User Cost. This may indicate that higher quality video may provide subjective benefits not apparent from these studies. The audio quality will need to be at least adequate though the precise quality requirements needed will probably vary according to other factors such as the user group. The extent to which users will tolerate delay also may be dependent on many factors.

Cognitive tasks cover at least two subtypes of tasks – information transfer tasks such as lectures and presentations which would correspond to the explaining, sharing and understanding information category in Olson’s (1994) taxonomy – and problem solving which would cover co-operative tasks such as the ones studied by Chapanis et al. (1972) and Gale (1990) and would correspond to McGrath’s (1984) Intellective, Creativity and Planning types. Problem solving tasks where the goal is to select a preferred consensus decision and where there is no objectively correct solution, e.g. McGrath’s (1984) Decision Making tasks, will not fall into this category as they may involve more social and negotiative processes where it is necessary to take others’ attitudes and affective state into account.

Social tasks appear to rely more on the presence of visual information and non-verbal communication (Short et al., 1976; Williams, 1977; Whittaker, 1995) and will probably be more affected by the quality of the video channel as a result. Tasks in this category might include conflict resolution or interpersonal perception, both of which appear to be influenced by communication modality (Short et al., 1976). Observational studies indicate that the video channel is used extensively in handling negotiation and affects judgements of others and feelings of trust (Rudman et al., 1997). These may all depend to some extent on information about others’ affect and attitudes con-
veyed by facial expression. Frame rate may need to be higher than for cognitive tasks as while viewers can see a speaker’s body movements and facial expressions at a relatively low frame rate of 6Hz, their ability to do so is improved at the higher rate of 30Hz (Kies et al., 1997). Low-resolution images with poor colour contrast may also affect users’ ability to use non-verbal information and these aspects of image quality may also need to be higher for social tasks.

As affective information is also carried by the voice it may be that high audio quality is also important. Social tasks would include the types of conflict noted by Olson as leading to use of the video channel – conflicts of motives and conflicts of viewpoints which correspond to McGrath’s Mixed Motive and Cognitive Conflict types respectively. The level of delay may also be important.

### 3.1.4.3 Difficult/Complex tasks versus Simple/Routine tasks

Rudman et al. (1997) consider the complexity of a task to be a factor in determining a group’s need for video and found that subjects engaged in a complex task found it to be useful in monitoring each other’s current level of understanding. While many studies have found no advantage in providing a video channel, this may be because they have tended to use relatively simple tasks (e.g. Chapanis et al., 1972). This may minimise the need to monitor each other’s understanding and reduce the need for non-verbal feedback, and it may be that use of the video channel is increased when the task becomes more difficult.

Boyle et al. (1994) found that subjects performing the map task tended to gaze at each other more at times when communication was made more difficult (points where map features were not shared by both subjects). Those studies that have found video to be advantageous have tended to be the most difficult either because communication is harder (Veinott et al., 1997) or because the task is complex (Olson, Olson & Meader, 1994). Users performing difficult tasks may therefore benefit more from higher quality video than easier ones, even if only subjectively.

Difficult tasks may also benefit from an interactive conversational style, which allows users to provide evidence about their level of understanding in the form of backchannels and acknowledgements and allows interruptions to ask for clarification or correct misunderstandings. Audio delay reduces the occurrence and effectiveness of such verbal feedback (O’Connail et al., 1993) and may impair performance to a greater degree on more difficult tasks. Difficulty or complexity may be defined in several ways. Rudman et al. (1997) consider it to refer to the number of variables that need to be considered in a task and also link it to a group’s levels of ability. Difficulty could also refer to situations where there is uncertainty in communicating such as the map task.

### 3.1.4.4 Urgent Tasks

Rudman et al. (1997) consider time pressure to be a factor in determining the need for video as non-verbal signals such as expression and nods allow users to check each other’s understanding and leave the audio channel free for other information. It is likely that when a task is particularly urgent or under particular time pressure then it may be particularly important that communication is clear and that participants in a conversation can tell if they are being understood and repair any misunderstandings or difficulties quickly. This may well increase the need for video. It seems likely that low delay will be more useful as this will allow users to provide more verbal feedback and interrupt if they have questions or relevant information. Good audio quality may be important to make communication maximally problem free.

### 3.1.4.5 Emotive Tasks

Olson (1994) proposes that tasks with a strong emotional content will require a greater degree of visual information in that they may require access to facial expression and other indicators of affect. This seems to supported by evidence that users prefer video communication over
conventional telephone systems for conversations with a high emotional content (Tang & Isaacs, 1994; Short et al., 1976). It may be that these tasks require a greater degree of "social presence" (Short et al., 1976) and will require a higher level of video service than less sensitive or emotive tasks. As emotion is also carried by tone of voice, higher audio quality may also be useful.

In the course of a particular session, users’ activities may fall into several of these categories. For example there may be visual material (tele-data) used in the course of a lecture. In such cases the requirements for the most important activity might be used. In most cases this will probably the one that takes up the largest amount of time. In others however an activity may take up relatively little time but be important to the overall success of the task. For example in medical consultation much of the time may be spent talking but some of it spent looking at medical charts or scan results. While this may take up a small amount of time it may be crucial to the consultation as a whole. This will mean that there will be high demand for video quality in this session. In such cases the most demanding set of requirements should probably be used.

3.2 User Characteristics
A number of user group characteristics may contribute to audio and video characteristics including sensory impairments, heterogeneity (or differences between user characteristics, such as knowledge or language background), age and experience with using the technology.

3.2.1 Sensory Impairments
Users with impaired hearing are likely to benefit from higher video quality. The ability to see a speaker and lip-read can counteract the effects of mild hearing loss (Summerfield, 1992). Frowein, Snooresburg, Pyter & Schinkel (1991) found in a study involving hearing impaired subjects, that the benefits of lip-reading, though present even at 5 frames per second, were maximised at frame rates above 15Hz, a similar finding to Barber & Laws' (1994) figure of 16.7 Hz. Image size and resolution (either CIF or QCIF format) were not found to have an impact within the range tested. Probably then hearing-impaired users will benefit from improved video for telepresence applications. For both teledata and telepresence audio quality may have to be good so as not to compound the problem.

Users with poor eyesight may have different video requirements, with potential benefits from increased image sizes in both telepresence and teledata applications. Resolution requirements may also be affected.

3.2.2 Heterogeneity of Participants
User groups can be more or less homogeneous (or similar to each other) in their background (Olson & Olson, 1997) and may share different amounts of common ground (Clark, 1996). This common ground potentially includes knowledge of a particular shared language, cultural and professional background and the personal common ground shared by individuals with shared experiences and a history of previous interaction. It might be predicted that audio and possibly video requirements will be higher for more heterogeneous groups and that there will be interactions if groups are heterogeneous in more than one of these ways.

Audio quality may need to be higher for heterogeneous groups. Homogeneous groups should be able to use their shared knowledge of the semantic context and topic vocabulary to guess what was said when audio is poor. Heterogeneous groups may be less able to compensate for degraded audio and the minimum acceptable audio quality may be higher.

3.2.2.1 Knowledge Background
Speakers’ use feedback from their audience to adapt their speech to listeners' levels of background knowledge (Kraut, Lewis & Swezey, 1982; Krauss, Fussell & Chen, 1995) and heterogeneous groups may be more reliant on verbal feedback to communicate effectively. They may therefore benefit from low audio delay and an interactive conversational style, which al-
allows participants plenty of opportunity to repair misunderstandings, while homogeneous groups may still be able to communicate well with higher levels of delay. Heterogeneous groups may also make greater use of non-verbal feedback. Rudman et al. (1997) reported that their subjects’ use of visual feedback appeared to increase when their levels of background knowledge differed.

Communication between individuals requires them to share an amount of mutual knowledge (Clark & Marshall, 1981; Krauss & Fussell, 1990; Clark, 1996). With only two participants, speakers can adapt quickly to their audience's expertise using verbal feedback alone (Isaacs & Clark, 1987). However with larger group sizes, it is not possible for all participants to provide verbal feedback (Fussell & Benimoff, 1995; Rudman et al., 1997) and speakers may benefit from the ability to visually check listener's understanding.

3.2.2.2 Language Background

Users may differ in their linguistic backgrounds, particularly if communication occurs across national boundaries. Language backgrounds may vary by differing amounts, from having different first languages to having different accents or dialects.

Non-native speakers are affected by degradation of the audio signal to a larger degree (Mayo, Florentine & Buss, 1997). Given that non-native speakers cannot compensate for poor audio to the same degree as native speakers it is likely that they will require higher audio quality than first language speakers. To a lesser degree, unfamiliar accents may have the same quality requirements.

Video quality may also be more important for non-native speakers. Veinott et al. (1997) found that pairs of non-native English speakers with different first languages, performed better at the Brown et al. (1984) map task in a video condition than in an audio-only condition. There was no difference for native speaker pairs. There are two ways in which video may have been more useful for non-native pairs – by assisting speech perception by displaying participants’ lip movements and by allowing participants to monitor each other’s level of understanding.

Reisberg, McLean & Goldfield (1987) found that the ability to lip-read increased comprehension by up to around 20% when subjects listened to speech in a foreign language. A smaller non-significant effect was found when speech was in subject’s native language but in an unfamiliar accent. This suggests that video of sufficient quality to allow users to lip-read (e.g. frame rate of over 16.7Hz) will be useful for linguistically heterogeneous user groups. The second way in which video may be useful is that it allows participants to monitor each other for signs of misunderstanding and to quickly initiate a repair if necessary. This may be more important in situations where a first language is not shared and there is greater chance of not being understood.

3.2.2.3 Personal Familiarity

As well as the common ground shared by members of particular communities and groups, individuals can also share personal common ground based on shared experiences and their previous history of interaction (Clark, 1996). Familiar individuals can make use of this additional common ground and can communicate more effectively on a range of tasks (e.g. Boyle et al., 1994; Krauss & Fussell, 1990). This improved ease and efficiency of communication between familiar participants may affect audio and video requirements.

While familiar users may be able to cope with reduced audio and video quality, they may benefit more from increased video quality. Rudman et al. (1997) suggest that groups learn to interpret person specific gestures and develop their own “visual language”. While there has been little observed advantage in providing familiar pairs with visual access for certain tasks e.g. the map task (Boyle et al., 1994; Doherty-Sneddon et al., 1997) there may be more of an effect in tasks such as negotiation and de-
cision making where it becomes more important to infer others’ attitudes and level of agreement. This may mean that in social tasks that familiar groups may gain additional benefits from higher video quality in some circumstances.

3.2.3 Age of Users
Studies of video-mediated communication have tended to look at adults but children may receive a greater benefit from video. Compared with adults, children often use gestures to communicate information they understand but cannot verbalise and are less able to adjust their style of communication when they cannot see their partners (Doherty-Sneddon & Kent, 1996). As a result they are less effective communicators in a range of situations, when they do not have visual access to other participants.

Adults may adapt to lower audio quality by speaking more clearly and by taking greater care to check their mutual understanding. Children are less likely to do this and may therefore be more affected by poor audio, and by half-duplex and delayed audio. The difficulties in communication faced by children probably only apply in interactive situations.

Older users may also have different requirements through the probability of their being less familiar with the technology and the increased probability of hearing loss and other sensory impairments.

3.2.4 Users’ experience with technology
The normal context for communication is face-to-face conversation and participants may need to learn to adapt in situations which differ from this in terms of their basic characteristics (Clark, 1996). Newlands, Anderson & Mullin (1996) found, for example, that users of a text based system initially performed more poorly at the map task than subjects communicating face-to-face but were able after practice to perform as well by adapting the way they communicated to the medium. Users of multimedia communication systems may have to develop skills to overcome the constraints of the technology to take full advantage of the medium.

Studies that have found video to be advantageous have generally studied groups over a long period of time where users have gained considerable experience of video (Tang & Isaacs, 1993; Rudman et al., 1997) while those that have not have tended to be short laboratory studies of novice users (e.g. Doherty-Sneddon et al., 1997). This is despite the fact that the video and audio used in the laboratory studies was often of higher quality than that used in field studies of experienced users.

These differences may be because experienced users have learned to adapt to the medium and can communicate more effectively over video. For example, Isaacs & Tang (1993) noted that users exaggerate their gestures, to overcome the fact that gestures are less noticeable over video. Rudman & Dykstra-Erickson (1994) observed that users became more sensitive to visual information over a ten-week period and became more adept at using it to co-ordinate conversation. Studies of experienced and novice users have tended to differ in their measures and methodology, making comparisons difficult.

Experienced users may also be able to cope with lower quality levels by adapting to them over time. Watson & Sasse (1996) observed that user’s perceptions of audio quality changed over a period of time, with an audio channel subject to moderate packet loss receiving better ratings at later points in the study although it is not clear whether experienced users behaviour would change as well as their perceptions or whether their minimum requirements will be different.

These different categories of user are not mutually exclusive and users may fall into more than one. When calculating users’ requirements, those for sensory-impairments will probably take priority over others.
3.3 Situation Characteristics
A number of factors inherent in the situation or environment may affect the requirements for quality. Typical examples of this include the geographical remoteness of participants from each other, the amount of background noise in the environment (both auditory and visually) and the number and distribution of users. Other characteristics of the environment specific to individual situations may also interfere.

3.3.1 Geographical Remoteness of Participants
Users may be widely dispersed geographically, sometimes being separated by thousand of miles. Anderson et al. (1996) found that when subjects communicated with partners over a long distance (between Scotland and the Netherlands) they benefited from a low level of video service (at 5 frames per second) with subjects in a video condition worrying less about losing contact with their partners and rating themselves as more aware of their partners. This effect will probably be diminished at very low frame rates (less than 5 frames per second) as the image is updated less frequently and becomes more static. Conversely the benefits may be increased with higher frame rates though studies have found 5 frames per second to be adequate for communication over large distances (Tang & Isaacs, 1993).

Other aspects of video quality may be important. Low image quality (e.g. low pixel resolution, blockiness due to packet loss) may decrease the subjective benefits of video.

Audio delay is likely to be another important factor. Users consider their partners to be less responsive when the audio signal is lagged (Zhang, Wolf, Darjand & Touna, 1998) and this may affect perceived feelings of contact with other participants.

Alternatively users may be more tolerant of low quality in all media because they have lower expectations of a long-range communication link or because of its increased utility due to savings in travel time etc.

3.3.2 Amount of Background Noise
The environment in which an application may be used will vary widely in their amount of background noise. Speech comprehension may be impaired in noisy situations. The signal-to-noise ratio of speech can be improved if listeners can lip-read. This may require frame rates over 16.7Hz (Barber & Laws, 1994; Vitkovich & Barber, 1994) and adequate pixel and contrast resolution to be effective (Barber & Laws, 1994). Therefore users in noisy situations may benefit from high quality video for tasks involving verbal communication.

Environments can also suffer from visual “noise”. If there is a lot of motion in the background this may affect requirements for frame rate (Apteker et al., 1995). Other environmental factors e.g. lighting conditions may affect other aspects of image quality.

The following apply specifically to Foreground Telepresence tasks:

3.3.3 Number of Users
It is harder to co-ordinate conversation with more than two speakers and it has been suggested this may increase reliance on video (Fussell & Benimoff, 1995; Rudman et al., 1997). In multiparty conversation it is essential to know who is currently speaking and without video it can be harder to keep track of the conversation (O’Connail et al., 1993; Sellen, 1995). O’Connail et al. reported that it was often hard to identify speakers with a low quality video system so there will probably be a minimum video quality level at which this is possible.

Video may also be used to handle the increased complexities of turn taking (Rudman et al., 1997). In two-party conversations video appears to have little effect on turn-taking (e.g. Masoodian et al., 1995, Anderson et al., 1996) but in multiparty interactions it may play a greater role as the situation is complicated by the possibility of several participants competing for the floor, potentially leading to an increase in disruptive simultaneous starts. 

Ex-
Experimental studies have shown mixed results with some finding no objective differences between video and audio-only conditions in multiparty interactions (Sellen, 1995) and others finding the conversation to be more fluent in video (Daly-Jones et al., 1998).

In a two-party conversation, participants can give feedback through the verbal channel and do not necessarily need a video channel (Krauss et al., 1977; Fussell & Benimoff, 1995). With larger groups, verbal feedback becomes less useful as participants may be unable to provide simultaneous feedback without individual responses becoming inaudible. In situations where verbal feedback is unavailable, participants can use the visual channel more, which may also increase use of video (Krauss et al., 1977; Rudman et al., 1997).

While there may be a greater reliance on video, the actual quality required to coordinate groups may not be that high (Isaacs & Tang, 1993; Tang & Isaacs, 1993). Generally, larger groups are likely to be affected by video quality and possibly audio to a greater extent.

### 3.3.4 Distribution of Users

When there is more than one user per terminal this may affect required video quality. Daly-Jones et al. (1998) found that high quality video increased the fluency of conversation between two groups of two, communicating over a video link. When there are two or more participants at each terminal remote users need to be able to identify which person is speaking, require feedback from silent participants etc. This will increase the need for video and possibly require higher audio quality as before. The pattern of interaction between participants is likely to be different when several users share a terminal with differences between interactions with co-present and remote participants. This pattern of interaction may be affected by the quality of the audio and video link between sites.

Each of the task types outlined is likely to have different basic audio/video quality requirements. These may in turn be affected by other factors involving the type of user - the taxonomy lists the heterogeneity of users, the age of users, experience with the technology

<table>
<thead>
<tr>
<th>Factors applying</th>
<th>Task</th>
<th>User</th>
<th>Situation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Background Teledata</td>
<td>N.A.</td>
<td>N.A.</td>
<td>N.A.</td>
</tr>
<tr>
<td>Foreground Teledata</td>
<td>-Degree of motion</td>
<td>-Sensory impairments</td>
<td>Background noise</td>
</tr>
<tr>
<td></td>
<td>-Degree of visual detail</td>
<td>-Age of users</td>
<td></td>
</tr>
<tr>
<td></td>
<td>-Experience of users</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Background Telepresence</td>
<td>N.A.</td>
<td>N.A.</td>
<td>N.A.</td>
</tr>
<tr>
<td>Foreground Telepresence</td>
<td>Interactive</td>
<td>Social</td>
<td>N.A.</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Cognitive</td>
<td>-Difficulty</td>
<td>-Heterogeneity of users</td>
</tr>
<tr>
<td></td>
<td></td>
<td>-Urgency</td>
<td>-Age of users</td>
</tr>
<tr>
<td></td>
<td></td>
<td>-Emotive</td>
<td>-User Experience</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>-Sensory Impairment</td>
</tr>
<tr>
<td>Non-Interactive</td>
<td>Social</td>
<td>N.A.</td>
<td>-Age of users</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>-User Experience</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>-Sensory Impairment</td>
</tr>
<tr>
<td></td>
<td>Cognitive</td>
<td>-Emotive</td>
<td>-Age of users</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>-User Experience</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>-Sensory Impairment</td>
</tr>
</tbody>
</table>

*Figure 3.5: An indication of the task, user and situation factors that may apply when considering audio and video systems for different situations*
and the presence of any sensory impairments (e.g. being hard of hearing, sight problems etc) - and factors involving the general situation the application is to be used in – background noise, the number of users, the distribution of users (one user per PC and hence one video window each, or several users at one PC etc) and the geographical remoteness of users from each other (see Fig 3.5).

3.4 Applying the Taxonomy

The taxonomy can be envisaged as a cube (Fig 3.6) with the three axes representing the different categories of task, user and situation and each cell representing a particular combination of these factors with its own minimum and optimal quality requirements.

To calculate the requirements for each cell first take the quality requirements to perform the basic task with normal, adult, homogenous users with no situational factors. These can be considered the default requirements for this type of task. Then take the requirements for any relevant user characteristics and situational factors and if these are higher than the default task requirements then raise the quality levels accordingly. Otherwise it can be assumed that the default task requirements are high enough to cope with the additional demands (if any) of the user group and situation.

Some tasks may have very high default requirements and these may not need to be altered to meet user demands at all. Similarly some user groups (e.g. linguistically heterogeneous) may have the same high demands across the relevant task types.

For some types of task, many situational and user factors will not be relevant. In certain other cases different factors may interact. For example the requirements of heterogeneous groups may differ above a certain group size, requiring better video quality but they may not require video to the same degree if there are only two participants.

Apart from the task categories, the user and situation categories are not mutually exclusive and several cells may apply. In this case a probable solution will be to take only the most important factors into account or take the cell with the highest requirements (probably this will be the same).

In some cases requirements of different factors may conflict. For example certain tasks/users may be helped by a larger, good quality image but with a large group, screen space requirements or bandwidth restrictions may make this less practicable. There will probably be some aspects of the image that users are more willing to trade off than others (e.g. users are willing to sacrifice contrast resolution to prevent the frame rate from becoming too low, Pappas & Hinds, cited by Kies et al.).

Fig 3.6 The Taxonomic Application Cube
4. ASSESSMENT METHODS

4.1 Choice of methods
There are a number of factors to consider when selecting which methods to use in an evaluation study:

4.1.1 Context
A general tool for assessing evaluation issues relevant to a context of use of a multimedia system is available in Appendix A. The factors involved include the following:

- **Status of system development:** Some methods are more suited to certain aspects of the development cycle than others.

- **Scope of the system:** The scope of a system can influence the choice of evaluation options that are viable. If the scope of the system is small then the evaluation can be constructed in a fairly simple manner. However, if the scope is larger with an increase in the number of users, groups and organisations, the administration of evaluation process may become more complex and restrictive.

- **Characteristics of specific tools**
  The characteristics of the data-collection techniques that are being used should also be considered for the following reasons:
  - **Validity:** the extent to which the tool measures what it is really intended to measure.
  - **Reliability:** the extent to which the tool provides stable and repeatable results across repeated administrations.
  - **Sensitivity:** the capability of a technique to measure even small variations in what it is intended to measure.
  - **Intrusion:** the degree of interference of an instrument in the task being performed: the degree to which it disturbs users or tasks.
  - **Acceptance:** the extent to which people are willing to work with the measuring tool.
  - **Ease of Use:** required expertise to apply the method.
  - **Costs:** i.e. labour intensity and the accessibility to test sites and users.
  - **Availability:** Whether the measuring tool is free or commercially available.

4.1.2 Subjective and Objective – the differences
Any information that originates from users, experts or observers can be considered to be ‘subjective’ data. Whereas information which can be recorded without potential bias or which is stored in archives is considered ‘objective’.

Subjective methods that are specifically unstructured in nature include open interviewing or participative observation. Subjective methods have the advantage of providing a wealth of information. The disadvantage of course is that all this information has then to be analysed extensively. Structured subjective methods such as questionnaires and checklists can provide concise data and valuable results very efficiently.

Objective methods of data collection often result in highly reliable information but are usually quite limited in the scope for interpretation. Certain system-based measures can be recorded objectively and relatively automatically via the system in question but complex processes, such as user interaction are difficult to record in this way. The analysis of objectively recorded data can often be time consuming.

4.1.3 Structured v unstructured methods
Structured methods are where expected user responses are already defined. The data is gathered in frequency counts: e.g. number of positive answers to questionnaire questions, or frequency of automatic recorded system usage.

Unstructured methods produce free response answers as results. These can include messages, behaviour sequences, descriptive narratives etc. This information has to either be interpreted intuitively or expertly coded into relevant categories. The results can then be analysed in the same way as structured data.

Figs 4.1 & 4.2 illustrate examples of structured and unstructured assessment methods which can be utilised according to the objectivity of the data source.
Subjective Methods

<table>
<thead>
<tr>
<th>DATA SOURCE:</th>
<th>STRUCTURED</th>
<th>UNSTRUCTURED</th>
</tr>
</thead>
</table>
| Users       | • questionnaires/rating scales  
• structured interviews  
• systematic contact, data logs | • open interviews  
• think aloud protocols  
• post hoc comments  
• diaries |
| Expert judgement | • formal, theory based analysis  
• rating against checklists | • brainstorming |
| Observation | • systematic observation and coding of user interaction | • impressions  
• participative observation |

Figure 4.1 Examples of structured and unstructured assessment methods which can be used to assess subjective data

Objective Methods

<table>
<thead>
<tr>
<th>DATA SOURCE:</th>
<th>STRUCTURED</th>
<th>UNSTRUCTURED</th>
</tr>
</thead>
</table>
| Recording equipment | • data from experimental equipment  
• physiological registration | • general audio or video recording  
• computer logging |
| Archives | • recorded data | • personal documents |

Figure 4.2 Examples of structured and unstructured assessment methods which can be used to assess objective data

4.1.4 Real vs. Contrived tasks

This distinction is similar to the distinction between field and laboratory studies. Real tasks are normally assessed in a field trial while contrived tasks are assessed in the laboratory.

Although the majority of our studies are lab based, we have also begun to conduct field studies in the workplace to give us a truly multifaceted picture of the impacts of multimedia technologies on users.

To see the difference between these types of task as black and white is however misleading. It is possible to create at least a good impression of a real task within a laboratory setting, and it is also possible for a real task in a field study to be affected by the assessment that is being carried out. With thorough research a task can be made to look and feel real to the user, and the control that is possible in a laboratory setting is invaluable.

Of course for assessing the value of multimedia for a specific task within a specific environment, it is not possible to conduct that assessment in a laboratory setting.

It is only when multimedia system is used in the environment for which it is intended that you find out about all the eccentricities of that situation, and of the users that will be taking part. See section 3 on assessment of task and context for more information.
4.2 Task performance
Task performance can be measured in a number of ways, each of which is appropriate only in certain circumstances. For instance with some tasks it may be important that they are completed very quickly, whereas for others accuracy is far more important. Alternatively you may not want to take any measures of task performance at all, especially in computer mediated communication where you are often more interested in the process of the communication than the performance at a task. When deciding what measures of task performance to take in a field study, it is important to think about how that task would be judged in the real world.

4.2.1 Benchmarking
Because of the problems involved with conducting laboratory experiments it is useful to have some standard by which you can compare your results. Using benchmarking involves using a specific task with fixed goals which can be time, output, errors etc. This enables you to test the performance across the communications media and compare your goals.

This can also be done if you have a specific task used in the organisation, for which the normal outcomes are very specific and documented. In our laboratory testing we use tasks that we have developed over a number of years and for which we know the normal outcomes. This allows us to compare studies against previous experiences.

Examples of real life tasks which may have a benchmark:
- Call centres – Number of calls dealt with in a shift
- Schools – Number of top grades in the exams
- Computer Help Desk – Time taken to deal with a problem
- Typing – Number of errors allowed

4.2.2 Output
Measuring the output from a task over a certain amount of time can be useful either if there are lots of similar small tasks, or if the task does not have a specific beginning or end.

Examples of real life tasks for which output can be measured include:
- Typing – Number of words per minute
- Assembly line – Number of components added per minute
- Doctor’s Surgery – Number of patients seen in a day
- Mortgage Applications – Number of applications processed in a day

4.2.3 Time
Time taken on the task can be a useful measure to take. It is worth taking this as a measure if the time taken is an important factor in the task (see examples below). On other occasions, time may be directly or indirectly opposed to other performance measures. For instance, when interviewing for a job, it is probably more important to get the right candidate for the job, rather than to conduct the interview in the shortest time possible.

Examples of real life tasks for which time might be measured include:
- Insurance – Time taken to make a single quote for insurance.
- Typing – Time taken to complete a words-per-minute test.
- Textiles – Time taken to manufacture one garment

4.2.4 Errors
The number of errors made in a task can provide you with indirect information about the ease of the communication. At a very basic level the more errors made during the task, the more the communication medium is likely to have had an effect. At a more complex level it is possible to look at the type of errors made and look at how critical they are.
An example of an occasion where we have used errors to judge the task performance was in a mortgage application study, where a user and an expert filled in a mortgage application form using computer mediated communication. Although our study was lab based in this case, errors were important, because of how they would impact on the same scenario in real life. For instance, if the wrong information was filled in, then the mortgage might not come through, or at the very least the information would have to be checked, thus slowing the application down.

Another study where we have used errors to compare task performance is in the Map Task (Brown et al., 1984), a collaborative problem solving task that we have used over a number of years. For this study, one person (the information giver) has a map with a route drawn on it and the other person has a map without a route. The information giver then describes the route to the information follower who has to draw the route on their map as accurately as possible. The task is made more difficult by slight differences between the two maps. The errors are judged by using the area of difference between the drawn route and the one on the information giver’s map.

4.2.5 Costs of task performance methods
The costs involved in Task Performance are generally lower than other methods, unless judgement is required to get the performance score. For measures that are specific to a particular task or domain, it can be the case that these measures are taken as a matter of course. In these cases the only extra cost involved is the time required to conduct the analysis on the data.

4.3 User satisfaction
Problems with subjective rating include high variability between subjects, possibly due to different expectations of the technology and different levels of user experience. When taking subjective measurements, particularly where subjects are rating different levels of quality, a within subjects design is preferred, allowing you to compare each individual’s rating over different levels of service. The following literature shows some of the pitfalls that need to be avoided when considering measuring user perception of quality.

4.3.1 Audio Quality
With respect to real-time communication over the Internet, audio quality can be roughly divided into the areas of speech intelligibility and perceived quality. The major factor that impacts on speech intelligibility is packet loss. The speech stream is digitised and divided into small units or packets, usually containing around 40 to 80 ms of speech each. Over a best-effort network such as the Internet, a degree of packet loss is inevitable.

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.

Subjectively, the effect of packet loss can be that the speech sounds broken up or ‘choppy’, as packets can contain phonemes, the smallest unit of speech intelligibility. However, when packet loss repair techniques are used at the receiver or at the sender, phonemic restoration (Warren, 1970) can be achieved and the listener will be less aware of network effects.

Although speech intelligibility can be improved via these techniques, it is not safe to assume that there will be a resulting increase in perceived speech quality. It is important to make the distinction between speech intelligibility and speech quality. It is relatively easy to make speech sound intelligible, but whether
the speech is pleasant to listen to is another issue, since different repair methods can make the speech sound somewhat artificial or robotic (see Watson & Sasse, 1997).

In addition to the effects of packet repair, subjective quality can be affected by issues such as echo and feedback (often caused by ‘leaky’ headsets), poor quality microphones (resulting in ‘tinny’ sounding voices) and volume differences between speakers. The effects of these aspects can sometimes be more detrimental than packet loss, as Watson & Sasse (2000) have demonstrated.

It is therefore important to view perceived quality as a multidimensional phenomenon, but as will be discussed below, this has rarely been considered.

### 4.3.2 Video Quality

The quality of interactive video in real-time transmission can vary tremendously. Elements such as pixel resolution, image size, display and frame rate have the most influence on the user’s perception of quality (Watson & Sasse, 1997).

There is little evidence to show that video aids the user in task performance unless there are specific communication problems to overcome, such as when the users do not share a common first language (Veinott et al., 1997). Although generally user performance in specific tasks has been found to be no different whether video is used or not (Finn et al., 1997).

Users do however consistently report subjective benefits of having video present e.g. (Tang & Isaacs, 1993; Daly-Jones et al., 1998). One suggestion to account for this is that the users may feel that maintaining task performance without video incurs greater effort (Monk et al., 1996).

### 4.3.3 Complementary quality of audio and video

Audio and video quality issues are usually assessed as separate entities. However there is substantial evidence that the quality of one medium can have an impact on the users perceived quality of the other (Watson & Sasse, 1996; Rimmel et al., 1998).

Other similar work has been conducted by Hollier & Voelcker (1997) where users were presented with several video clips with an accompanying audio commentary; the videos used varied at 6 levels of image quality and the audio had 4 levels of quality. Results showed that the identical audio segment would receive different quality ratings depending on the quality of the video that accompanied it. Also the audio quality was found to have an influence on the perceived quality of the video.

Users perception of quality is also likely to vary with the task (Hollier & Voelcker, 1997). If the users are involved in learning a foreign language then the audio quality may need to be substantially higher for success than if the task is to present a report at a routine meeting (Watson & Sasse, 1996). Video quality is likely to be more important in an intense interview situation than it might be in other relaxed scenarios. User perception of audio and video quality may be directly linked to the level of quality they assume is necessary for the situation.

### 4.3.4 Post-hoc rating – quality and adequacy

In our subjective measurements of video and audio we have distinguished between quality and adequacy of the media. Providing this distinction allows users to judge the audio and video against any criteria that they have already created from their real-world experience of these media, whilst also having the opportunity to judge the adequacy of the quality for the particular task they have been involved in.

Our subjective rating is usually applied at the end of individual tasks. Where possible we usually try to get people to give an initial rating of the quality after a short exposure to the media. This usually takes the form of asking subjects to briefly introduce themselves.
1. Please indicate anywhere on the scale below how you would rate the quality of the audio.

   Low | 10 | 20 | 30 | 40 | 50 | 60 | 70 | 80 | 90 | 100 | High

2. Please indicate anywhere on the scale below how you would rate the quality of the video image.

   Low | 10 | 20 | 30 | 40 | 50 | 60 | 70 | 80 | 90 | 100 | High

3. Please indicate anywhere on the scale below how adequate the audio was for the task.

   Low | 10 | 20 | 30 | 40 | 50 | 60 | 70 | 80 | 90 | 100 | High

4. Please indicate anywhere on the scale below how adequate the video image was for the task.

   Low | 10 | 20 | 30 | 40 | 50 | 60 | 70 | 80 | 90 | 100 | High

[Figure 4.3: An example of subjective rating scales.]

One problem with post-hoc subjective rating scales is that it can be relatively insensitive to fluctuations during the course of the test. This can of course be an advantage in that small blips in the audio or video do not become noise in the data, but if the properties of the network are such that fluctuations occur then it can be important to capture the users’ perceptions of the entire session, rather than being subject to primacy or recency effects. In order to measure fluctuations it can be preferable to use a dynamic assessment scale (see next section for more information).

The scales shown in Fig 4.3 have been used and tested over numerous studies and enable the users to rate the video and audio freely. The continuous nature of the scales make them particularly useful where you are asking subjects to rate the quality of the audio and video at various stages in the experiment, because it allows them to indicate even small differences in what they perceive. The omission of labels at each point of the scale avoids the potential problem of users being prejudiced towards or against a particular item through the use of wording.

4.3.5 Criticisms of the ITU scales

Earlier (Section 2) we introduced the ITU recommended speech and image quality measurement scales. These are again summarised in Figure 4.4. We wish here to examine the usefulness of these scales with respect to Internet delivered real-time interactive multimedia.

4.3.5.1 Real-time Internet Speech

Criticism of the recommended scales with respect to assessing the perceived quality of Internet speech falls into 3 main areas:

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

Internet speech is (in the main) narrowband and subject to a range of network and environmental degradations. Given these facts, the labels on the listening quality scale (i.e. Excel-
lent, Good, Fair, Poor and Bad) seem inappro-
priate. Even with training, it is likely that re-
sponses will be concentrated at the lower end of
the scale, which has been borne out in both ex-
perimental and field studies (Watson & Sasse, 1996).
With respect to the category la-

tables on the listening effort scale, it is even
easier to see how bias towards the lower end of
the scale might occur.

The variable network conditions that affect
some real-time services mean that speech qual-


ty can change rapidly and unpredictably. In
listening-quality tests the recommended test
material is short in duration – 10 seconds at
most. 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.

Finally, the binary difficulty scale is patently
unsuited for the assessment multimedia con-
ferencing (MMC) conversations, since even a
small amount of packet loss is likely to cause
difficulty in hearing or talking, even if short-
lived.

4.3.5.2 Real-time Internet Video
As with Internet speech, criticism of the rec-
commended scales with respect to Internet video
assessment falls into 3 main areas:
  • vocabulary of the scale labels
  • duration of the test material
  • artificiality of assessing video without au-
dio

The ITU-R recommendations are concerned
with establishing the subjective performance of
television systems. This means that in terms of
colour, brightness, contrast, frame rate etc., the
quality component under investigation is as-
sumed to be already of a high standard, which
is simply not the case for Internet video. Like
Internet speech, real-time Internet video is
characterised by a large variety and range of
impairments, which can change rapidly. This
trait means that the single- and double-stimulus
impairment tests are not suitable, since, as is
reflected in the terminology of the scale (im-
perceptible/perceptible), they have been de-
dsigned to determine whether individual small
impairments are detectable.

<table>
<thead>
<tr>
<th>Quality of the speech/connection</th>
<th>Score</th>
<th>Did you or your partner have any difficulty in talking or hearing over the connection?</th>
</tr>
</thead>
<tbody>
<tr>
<td>Excellent</td>
<td>5</td>
<td>Yes 1</td>
</tr>
<tr>
<td>Good</td>
<td>4</td>
<td>No 0</td>
</tr>
<tr>
<td>Fair</td>
<td>3</td>
<td></td>
</tr>
<tr>
<td>Poor</td>
<td>2</td>
<td></td>
</tr>
<tr>
<td>Bad</td>
<td>1</td>
<td></td>
</tr>
</tbody>
</table>

(a) Listening-quality scale

<table>
<thead>
<tr>
<th>Effort required to understand the meaning of sentences</th>
<th>Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>Complete relaxation possible; no effort required</td>
<td>5</td>
</tr>
<tr>
<td>Attention necessary; no appreciable effort required</td>
<td>4</td>
</tr>
<tr>
<td>Moderate effort required</td>
<td>3</td>
</tr>
<tr>
<td>Considerable effort required</td>
<td>2</td>
</tr>
<tr>
<td>No meaning understood with any feasible effort</td>
<td>1</td>
</tr>
</tbody>
</table>

(b) Listening-effort scale

<table>
<thead>
<tr>
<th>Image quality</th>
<th>Score</th>
<th>Image impairment</th>
<th>Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>Excellent</td>
<td>5</td>
<td>Imperceptible</td>
<td>5</td>
</tr>
<tr>
<td>Good</td>
<td>4</td>
<td>Perceptible, but not annoying</td>
<td>4</td>
</tr>
<tr>
<td>Fair</td>
<td>3</td>
<td>Slightly annoying</td>
<td>3</td>
</tr>
<tr>
<td>Poor</td>
<td>2</td>
<td>Annoying</td>
<td>2</td>
</tr>
<tr>
<td>Bad</td>
<td>1</td>
<td>Very annoying</td>
<td>1</td>
</tr>
</tbody>
</table>

(c) Conversation difficulty scale

<table>
<thead>
<tr>
<th>Did you or your partner have any difficulty in talking or hearing over the connection?</th>
</tr>
</thead>
<tbody>
<tr>
<td>Yes 1</td>
</tr>
<tr>
<td>No 0</td>
</tr>
</tbody>
</table>

(f) Double stimulus continuous quality scale

Figure 4.4 ITU recommended speech and image quality measurement scales
With respect to use of the quality scale, the same criticism can be levelled as to its use with Internet speech: the vocabulary is unsuitable, and therefore we can expect responses to 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 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).

The quality tests typically require the viewer to watch short sequences of approximately 10 seconds duration, and then rate the material. It is not clear that a 10-second video sequence is long enough to experience the types of degradations common to best-effort Internet video. In addition, the quality judgements are intended to be made entirely on the basis of the picture quality. It should be queried whether it makes sense to assess Internet video on its own (i.e. without audio) since it is rarely the case that the video image in real-time communication over the Internet is the focus of attention in the same way that the picture is when we watch television. We believe that the utility of the low frame rate video commonly found in real-time Internet communication 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 networks at present.

For this reason, the audio-visual quality recommendations should be better suited to assessing Internet video. However, since it is the 5-point scales that are recommended again, the criticisms raised above remain valid. ‘One-off’ quality ratings gathered at the end of an audio-visual session are also unable to capture the changing perceptions users may have during communication across a packet network with varying conditions.

We therefore believe that the assessment methodologies recommended by the ITU are not suitable for subjective quality assessment of real-time communication over packet networks such as the Internet. In particular we have argued that the 5-point quality scales are not viable due to their vocabulary. However, there is a further concern over the 5-point quality scale – how legitimate is it as an interval scale?

4.3.5.3 The nature of the 5-point quality scale
The 5-point quality 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. There are a growing number of researchers however, who question whether such trust in this scale is warranted.

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 & McManus, 1986). Investigations have also been carried out to validate the ITU assumption that the scale labels have been adequately translated into different languages, such that the scale is ‘equal’ in different countries, so that quality results can be generalised across the world.

4.3.5.4 Internationally interval, or internationally ordinal?
Investigations of the interval nature of the rating scales have generally been carried out using the graphic scaling method. Subjects are presented with 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. By measuring
the distance of the marks from the bottom of the scale, the means and standard 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 good news for Japanese speakers, it is a different story for English, Dutch, Swedish and Italian speakers.

Jones & McManus (1986) used the same method to investigate whether the intervals represented by the quality scale 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 i.e. 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 to be a “3-point, 2-interval scale”.

Jones & McManus also carried out their study in Italy. 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 note that a supposed ‘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 equate ‘OK’ with ‘Good’.

Other researchers have found similar results. Virtanen et al. (1995) found that there was a flattened lower end (i.e. the Swedish terms

---

**Figure 4.5 Mean positions for quality terms placed on a 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.**
equivalent to ‘Bad’ and ‘Poor’ were perceived as very similar), and there was a 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.

The results of a study conducted at UCL involving 24 British English speakers asked to position the ITU terms and other descriptive adjectives are shown in Fig 4.5. Again it can be seen that the ITU terms do not represent equal perceptual intervals.

4.3.5.5 Summary
The ITU-recommended quality scale is not the international interval scale it is purported to be. But the quality scale is also not internationally ordinal, since the positional rankings of the qualitative terms in different languages are not equal. However, there is another, more complex, issue at hand, and that is the overall concept of quality: the 5-point scale treats quality as a single measurable dimension, despite evidence to the contrary.

4.3.6 What is quality?
When we talk about investigating perceived quality, what exactly do we mean by 'quality'? With respect to speech, 'quality' can be conceived of as an umbrella term. There are many variables that contribute to forming a perception of speech quality, such as listening effort, loudness, pleasantness of tone and intelligibility. With respect to video and image quality, we can identify a set of factors that affect overall perceived quality, such as frame rate, lighting, image size, 'blockiness', packet loss and degree of synchronisation with the audio.

As Preminger & Van Tasell (1995) observe with respect to speech quality, "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." Knoche et al. (1999) agree, arguing that use of the traditional 5-point scale leaves the experimenter ignorant of the subject’s perspective and rationale for positioning on the scale.

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, "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" (Jayant, 1990). However, while increasing bandwidth would undoubtedly solve many quality issues, it should not be treated as a panacea. It is undoubtedly the case that many quality issues, such as echo, poor quality microphones, and disruptive volume differences can be settled without resorting to increasing bandwidth. Since bandwidth is a valuable resource, attending to these issues is important, for both the HCI and networking communities (Watson & Sasse, 1996; Watson & Sasse, 1997; Podolsky et al., 1998).

4.3.7 Unlabelled scale
As a result of the issues described above regarding use of the ITU rating scales, we have explored and developed the use of an unlabelled (100 point) rating scale (see Fig 4.3). This scale does not have quality terms since we recognise that the quality is a multidimensional phenomenon. Instead we often ask users of the scale to describe why a certain rating was awarded, in order to try to establish which criteria are most important to users. 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.

The unlabelled scale has been proven to be used reliably and consistently (e.g. Watson &
Sasse, 1997; Watson & Sasse 2000), but it does not allow us to capture users’ perception of quality as and when it changes. 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. For this reason we developed a software continuous quality tool, QUASS (QUality ASsessment Slider) (Bouch et al., 1998). This tool is somewhat similar to the Single Stimulus Continuous Quality Evaluation method recently recommended in ITU-R BT.500-8, but QUASS is not a hardware tool, and does not use the ITU quality labels.

4.3.8 Continuous Rating
Continuous rating scales allow perceptions of the quality of the video or audio in the current session to be collected throughout the session. The International Telecommunications Union (ITU) has adopted such a method for measuring time-varying image quality under the name SSCQE (Single Stimulus Continuous Quality Evaluation). During a continuous assessment trial, the user will view information on the screen and move a slider up and down according to their perception of a pre-specified attribute. The position of the slider is recorded at regular intervals throughout the trial and plotted as a function of time.

Since the unlabelled scale proved to be reliable in early studies (e.g. Watson & Sasse, 1997), a software tool was developed at UCL 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 Fig 4.6.

Although QUASS is similar to the SSCQE method discussed above, it differs in two important respects. Firstly, the tool 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.

QUASS has been successfully used in laboratory settings to investigate users’ perceptions of different levels of audio quality (Bouch & Sasse, 1999). However, although QUASS allows an examination of the subjective effects of quality fluctuations to be conducted, 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 Internet com-
communication 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. 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 real-time communication 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 can also function in a second mode, whereby movement of the slider controls the objective quality delivered. The rationale for this approach is that full attention can be paid to the task at hand, and only when the quality becomes so poor as to be intrusive to the accomplishment of the task, would the user have to direct attention to the slider. Using QUASS in this mode allows us to determine minimum levels of quality required by users to complete certain tasks under laboratory conditions.

4.3.9 Real-world interactive communication

Although QUASS can be used effectively in passive listening/viewing environments, or in experimental set-ups to permit increases in quality when required, in real-world interactive situations task interference is likely to occur.

We have therefore been using two unlabelled scales in our interactive task studies, one which asks people to rate the quality of the audio or video stream, and the other which asks them to assess the adequacy of this quality for the purposes of the task at hand. Examples of these scales are shown in Figure 4.3). We have shown that these scales are used in a reliable fashion, and can also be used across a variety of audio and video impairments e.g. echo, volume differences, poor quality microphones and differences in frame rate (see Watson & Sasse, 2000).

Using these scales to investigate the subjective effects of different types of impairments is helping us to establish which types of impairments are more intrusive than others. For example, in a recent study we found that levels of packet loss commonly found on the Internet today, when repaired to produce phonemic restoration, do 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 (Watson & Sasse, 2000).

4.3.10 Costs of Subjective Assessment Methods

The capital outlay for the methods described above is minimal, the main cost being in time required to analyse the data gathered.

4.4 Assessing User Cost

4.4.1 Physiological Measurements

Subjective methods are widely used to assess MMC quality. However, there are drawbacks with using subjective assessment in isolation. The main problem centres on the fact that it is cognitively mediated. This means that contextual variables such as task difficulty (Wilson & Descamps 1996) or budget (Bouch & Sasse 1999) can influence users’ assessments of quality. Moreover, Knoche, DeMeer et al. (1999) argue that subjective assessment is fundamentally flawed, as it is not possible for users to register what they do not consciously perceive. So how can a reliable indicator of the impact of MMC quality on users be gained?

In tackling this problem, a traditional Human Computer Interaction (HCI) evaluation framework of task performance, user satisfaction and user cost (Figure 4.7) has been revisited. User cost is an explicit - if often disregarded - element of this framework.
Subjective approaches to measuring user cost exist (e.g. mood scales), yet due to the problems with cognitive mediation, it was decided to focus on finding an objective method. One way of doing this is to measure physiological signals that are indicative of arousal. When a user is presented with insufficient audio and video quality in a task, he/she must expend extra effort on decoding information at the perceptual level. If the user struggles to decode the information, this should induce an arousal response, even if the user remains capable of performing his/her main task.

Arousal is viewed here as a negative event. We predict that there is an inherent amount of arousal in each task and that adding to this with degraded quality will result in too much arousal. This can interfere with task completion and result in people becoming tense, anxious and unproductive. If someone interacted with degraded quality frequently when performing an important task over a period of time, this could have adverse effects on health and result in psychological stress.

This research using physiological measurements began in 1998 at UCL, with the commencement of the PhD of Gillian Wilson.

4.4.2 The Physiology Set-up at UCL

The system we use is called the ProComp, which is manufactured by Thought Technology Ltd (http://www.thoughttechnology.com/). It is capable of digitising data from up to eight sensors simultaneously. The sensor information is digitally sampled and the resulting information is sent to the computer via a fiber-optic cable. It is battery operated. The sensors require little or no skin preparation for use. The selection of physiological sensors includes devices specialised for electromyography (EMG), electroencephalography (EEG), electrocardiography (EKG), blood volume pulse (BVP), heart rate (HR), skin conductance (SC), respiration and temperature. The basic software required for use with the ProComp runs under DOS, which we use. This allows readings to be displayed as polygraph type traces or bargraphs. Real-time displays can be adjusted and tonal feedback is available either through a headset or on headphones. A more sophisticated, yet expensive, multimedia biofeedback software is available - Biograph.

For our research, it was decided to measure SC, BVP and HR because they are physically non-invasive and are good indicators of arousal. To measure SC, two sensors are placed on the fingertips on the same hand (see Fig 4.8). An imperceptibly small voltage is passed between the sensors, and the skin's capacity to conduct the current is measured. An increase in the conductance of the skin is associated with an increase in arousal.

BVP and HR can be measured simultaneously using the same sensor, called a photoplethysmograph (see Fig 4.9). It is placed on the index finger. This applies a light source to the skin and measures the amount (BVP) and rate (HR) at which blood is pumped round the body. The unit of measurement of BVP is percent, and heart rate is measured in beats per minute (bpm). An increase in heart rate is asso-
associated with an increase in arousal, as blood is being pumped round the body quicker to reach the working muscles in order to prepare them for action. A decrease in BVP is associated with an increase in arousal, as less blood is being pumped to the extremities – it is being used by the working muscles.

4.4.3 Some problems we have encountered
When we started to work with the ProComp, setting it up was a relatively easy task. What has been more difficult is learning to understand the signals. For example, knowing that a sensor is giving false readings because it is misplaced. To overcome this, many pilot trials were run. Now, taking measurements in an experiment is simple and quick to set-up, as the equipment is portable.

Another problem we experienced was with movement. If a participant moves their hand whilst performing a task, it will interfere with the readings. For this reason, we have to have people sitting down with their hands still on a table or on their lap. Some people find this uncomfortable. In addition, some find it difficult to interact with someone e.g. interviewing them, without making use of their hands to add expression to the conversation. In these situations the equipment can be constraining. In addition, typing is not possible as the participants only have one hand free.

Finally, for the last experiment that has been completed to date, the task was to interview candidates in real-time. We have found interpreting the data from this difficult, as the task in itself is stressful. Thus, it seems that responses to the task are obscuring any responses to the quality. We are still working on trying to tease this data apart, thus no conclusions about task stress can be made at present.

4.4.4 Some of our findings so far
To date 4 main experiments have been performed:

- Experiment 1: Investigating the physiological and subjective effects of low and high video frame rates in a recorded interview task (Wilson and Sasse 2000a)
- Experiment 2: Investigating the physiological and subjective effects of audio degradations in a passive listening task (Wilson and Sasse 2000b)
- Experiment 3: Investigating the physiological and subjective effects of audio degradations in a recorded interview task (see Fig 4.10)
- Experiment 4: Investigating the physiological and subjective effects of audio and video degradations in an interactive task

The main findings from these have been:

4.4.4.1 Audio degradations due to hardware set-up and end-user behaviour
Interestingly, it was discovered that audio problems due to the hardware set up and end-user behaviour affected users just as much as network problems. In particular, problems such
as loud volume, affected users much more than the normal level of packet loss in a multimedia conference, 5%, and as much as 20% packet loss, which does not occur frequently and when it does it is usually of a bursty nature as opposed to being stable over time, as it was in this experiment. Even if perfect quality is delivered in terms of the network, the user's experience with the technology could still be marred by easily rectifiable hardware problems.

4.4.4.2 Differences in signal responses
Another interesting finding is that the three physiological signals respond differently to audio and video degradations. For example, in the video frame rate experiment, all of the signals responded strongly, but it was HR that responded the strongest. However, in experiment 2, SC did not respond at all: there were no significant differences between conditions. Yet, when the video channel was added in experiment 3, SC did produce significant results. Thus, it seems that there may be different patterns of arousal for the different degradations and that these partially depend on the task being performed.

4.4.4.3 Lack of correlation between subjective and physiological measures
We have discovered that subjective and physiological results do not always correlate with each other. When the research began, it was posited that this would only be the case when the task the user was performing was engaging, as the participant would pay more attention to the task, rather than rating the quality. This was supported, for example in the video frame rate experiment. However, some interesting discrepancies were also found in the non-engaging passive listening task.

The latter finding could indicate that when a participant becomes bored, they do not pay enough attention to rating the quality: their mind may wander. Yet to counteract this argument in the experiment, the subjective ratings were consistent for the first and second presentation of the degradations.

Additionally, the most recent versions of subjective assessment scales were used, so the argument that the problem lies with rating scales being insensitive does not hold. Thus, a fundamental flaw of using subjective assessment in isolation may have been uncovered: users cannot consciously evaluate the impact quality has on them in short lab-based trials. If this lack of awareness persists in long-term studies, it would be worrying, as prolonged exposure to degraded quality could be harmful. To address this result, it is advocated that the 3-tier approach be utilised in multimedia quality evaluation, and also in assessment in other areas of HCI.

4.4.5 How does physiology fit in with our other methods?
Physiological measurements can be used in isolation. However, we would not recommend it. In our research it is used with subjective measures of user satisfaction. This allows us to determine if there is a factor other than the quality that could be influencing results. For example, the person is anxious generally, or very excited about participating in the experiment.

Until now, simple questionnaires have been used to probe these issues. However, for the 5th experiment that has just been completed, we used the POMS scales (Profile Of Mood States, see www.edits.net). We used these, as we want to gently investigate if people’s mood changes after interacting with degraded quality. The results from this are still being analysed at present.

In addition, taking the task into consideration is important. For example, if a participant seems highly anxious at specific points, we can look at the material at that point and determine what they were responding to.

4.4.6 Methodology Guidelines
From our experience in testing subjects participating in MMC sessions, we have derived a set of methodology guidelines. These are presented in Fig 4.11.
Sensors should be placed on the participant’s non-dominant hand to allow one hand free to fill out questionnaires etc.

The environment in the testing location should be minimally stressful, e.g. with no phones ringing and no people moving around, as this can interfere with results.

Make sure the temperature in the testing location is comfortable – a room too hot or cold can influence results.

Measure baseline physiological responses for 15 minutes prior to any experimentation. This allows a set of control data with which to compare responses in an experiment, and gives the participants and sensors the opportunity to settle down.

During the baseline session, give the participant a newspaper to read and avoid any prolonged interaction with them – the purpose is for them to relax.

People’s SC will naturally rise as they adjust to wearing the sensors – thus the baseline session allows ‘true’ readings during the experiment to occur.

Ensure that the participant moves little as this can produce artefacts in the results.

Ensure that the sensors are snug fitting – this can be difficult in people with smaller hands, thus adjustments to the sensors need to be made.

Export and back-up data files immediately after each participant.

It is not good practice to keep the software running for long periods of time. Re-start it regularly.

Check the encoder battery level frequently – if they run flat during a session, any data saved will be meaningless.

Zero the SC sensor before each participant: this cancels any offsets.

| Figure 4.11: Methodology guidelines for physiological measurements |

**4.4.7 The future of physiology at UCL**

We have found using physiological measurements gives us a valuable added set of data that remains ‘untapped’ by subjective assessment.

Thus, we have purchased more of the ProComp units, which will enable us to take readings from multiple sites simultaneously.

We are interested in investigating the effects on the user when interacting with degraded quality in the long-term, as opposed to short lab based experiments. To do this we are investigating the use of other signals, which are good indicators of psychological stress, and we are also considering measuring stress hormones in the blood.

**4.4.8 Cost for physiology: capital outlay and time**

*Capital Outlay:* To purchase a unit such as ours costs in the region of three thousand pounds. After this there are no on-going costs, except if additional sensors are required.

*Time:* It is necessary to invest some time in learning how the signals respond in different situations, and which readings are false. Analysing the results can take a lot of time, as there is a lot of data (we take 20 samples per second on 3 signals). However, not all experiments will require such a high degree of granularity – that depends on what is being investigated and over what time period.
4.5 Impact on User Behaviour

4.5.1 Eye tracking
Eye tracking has been used in psychological research for many years and has been effective in measuring eye movements in a range of cognitive processes such as reading and perception tasks (see Mullin et al. 2001 for more examples). Most of the published research has used a fixed eye-tracking device where the subject’s head is kept completely still by employing head rests and bite bars. Some less restrictive eye-tracking technology has allowed subjects to move their heads more freely by using head mounted devices, enabling eye-tracking research to explore other areas such as driver behaviour.

It was not until the current non-invasive type of eye-tracker was developed that we considered using it for our research, although it would be possible to experience full communication wearing a helmet eye-tracking device. In our research, where participants have to do collaborative tasks using multimedia, a helmet device would have proved unsuitable because we feel that it is important to recreate the real-world context as accurately as possible. We felt that the technology itself would impact too heavily on the communication, by making the participant wearing the helmet look ‘strange’.

4.5.1.1 The Eye-tracking setup in Glasgow
The system we use is a Remote Eye-tracking Device (RED) manufactured by SMI GmbH (http://www.smi.de/). This system is supposed to be best positioned directly below the monitor and as close to it as possible. The image below shows a subject sitting in the reclining chair being eye-tracked. She is wearing an audio headset. The eye-tracker is the black box below the screen. A video camera sits on top of the screen transmitting the head and shoulders image to the other participant.

Participants sit at approximately 70cm from a computer screen displaying a window containing task relevant information such as a video picture of the other person (See Fig 4.12).

Infrared light is directed at the subject’s eye from a panel on the side of the eye-tracker and the reflected radiation is received by program-
field. A third computer controls the mirrors to compensate for mild head movements to physically track the head/eye movements and to receive the resulting eye-position data. The point of regard is calculated as a function of the distance between the centre of the pupil and the corneal reflection (Fig 4.13).

Software has been written for the operation of the co-operative computer working environments and to enable the calibration of physical points of regard on the eye-tracked participant’s screen with readings received from the RED. This calibration enables the computer to take a reading for the difference between where it calculates the user is looking and where the point of gaze is measured to be. We use a 6-point calibration, whereby each of the points is checked by the computer to see that it concurs with previously calibrated points.

4.5.1.2 Some problems we encountered
It took much longer than we expected to properly set up the eye-tracker and the testing environment, to check our calibration techniques, and to acquire sufficient dexterity in manipulating the tracker whilst live. Tracking is lost at various times, and, when the RED cannot re-discover the eye position, a human operator who is continually monitoring the operation, must manually put it on track again. It takes acquired skill and experience to do this quickly.

During an initial testing phase of the eye-tracking device we spent a large number of hours testing the device on different subjects and running mock user sessions. After some time, we eventually became confident that we were able to accurately track gaze across the screen. It was very reassuring to conduct pilot sessions and to consistently see the user looking at the current point of interest.

Although these eye-trackers are remote, this does not mean that they do not place any restrictions on the user. The camera that focuses on the eye to measure the eye movements cannot compensate for movement of the head in the radial plane towards and away from the computer screen. During our initial testing phase we found that it was very difficult for a person to keep their head stationary while still engaging in free conversation. To counter this we arranged the testing room in such a way that the user did not have to concentrate on keeping their head still.

We did this by providing a reclining chair, which supports the head in the reclined position. In this way the subject’s own weight naturally restricts their movement towards the eye-tracker. Although they are lying back, the subject’s head is propped forwards on a cushion, so that it is approximately perpendicular to the floor, giving a clear view of the eye in the Remote Eye-tracking Device. The computer monitor and the eye-tracker are then cantile-
vered at a position above their hips to allow room for the subject’s feet and legs underneath. We have found this to be very useful, as it removes any need for the subject to concentrate on what their head is doing and with this setup we have managed to get reasonable amounts of data from most subjects.

4.5.1.3 Some of our findings so far

In a replication of our map task experiment (see section 4.2.4 for a description of the map task) where eye gaze was measured (Mullin et al., 2001), the data suggest that instruction givers spend less time monitoring the visual link to their collaborators than they do in face-to-face interactions (see Fig 4.14).

This may be because the quality of the visual information is inferior and so less useful. Alternatively, it may be because the increased cognitive demands or unfamiliarity of communicating via a remote communication system mean that they feel less able to devote time to what seems a less critical component of the task than studying the map in order to formulate their next instruction. In a visually more complex and richer task we found similar distribution of gaze, such that the onscreen resources receive the great majority of the subject’s attention. In this study the amount of eye gaze directed towards the other participant was even more modest, perhaps due to the complexity of the on-screen materials.

We have also found that markedly different levels of video quality (25 frames per second vs. 5 frames per second), which users could reliably discriminate in isolation, had no significant impacts on their subjective quality ratings in a questionnaire study, (Anderson et al., 2000). The patterns of eye gaze that we have recorded in these experiments, showing fairly infrequent glances at videoconference windows, suggest that different quality levels may have little impact on users of multi-component interfaces.

These results, showing very little gaze directed towards the video window of the other person, can go some way towards explaining why changes in the quality have little impact on the participant. We are confident that our findings so far and in the future will have a great deal of useful information for multimedia service providers.

Figure 4.14: Screen-shot of map task with overlaid data captured from the eye-tracker. The screen shot has been compressed to fit into the eye-tracking software window.
4.5.1.4 How does eye-tracking fit in with our other methods?

Eye tracking alone cannot tell you about the impact of the technology on communication between two people. It can give you very specific data on what an individual is looking at during the course of the communication, but you need to use other methods to tell you why. For instance, we can tell when and how long the subject is looking at the video window containing the face of their partner, but we do not know if this is because they have nothing else to do, or because they are listening intently to the other party. We have found that we can use eye tracking for one of two reasons, either 1) we have created a hypothesis about the user behaviour that we can test with the eye-tracking equipment, or 2) we use the eye-tracker to look at the behaviour and create various hypotheses that we can test through other methods.

We have used the eye-tracker in a number of studies now to good effect. Our next step for our gaze analysis involves correlating the eye-gaze data in time with the video stimulus on the computer screen and the audio recording of the participants’ conversation. This will enable us to tell what was being looked at in correspondence with what was being said.

All in all, we have found eye-tracking to be a useful methodology in teasing out the intricacies of human behaviour across communications media, as long as it is combined with other methodologies. We are still learning how best to use eye-tracking within the context of our tried and tested methodologies. Some of our future work will be directed at the amount of use of the video window when participants complete different tasks.

4.5.1.5 Costs for eye-tracking

Capital Outlay: The capital outlay required for a Remote Eye-tracking Device is quite high. However, there are no ongoing costs for materials. Considerable outlay is required in terms of time to learn how to use the system and to get it set up exactly as required. Once the eye-tracker is set up and one person has become an ‘expert’ with the system, transfer of the knowledge required to run experimental sessions is relatively straightforward. As mentioned above, we ran extensive testing sessions before the system was run on experimental subjects.

Time: Eye-trackers produce a lot of data. The difficult part of running an eye-tracking experiment is sometimes deciding exactly what you are interested in. Therefore, you should spend some time prior to the experiment deciding on the analysis you want. The most common form of analysis that we use is to set up objects on the screen (overlaid on areas of interest such as the video window, or specific elements of the task). These are then used in any analysis you do whether you want to get a percentage value for how long they were looking at that object throughout the whole test or view a time line showing when they were looking at that object. The other major decision is to decide what length of fixation you are interested in.

Costs for correlating eye-tracker data with media records: Video analysis of any video data is likely to be time consuming. An estimate of the time involved is similar to that for audio transcription (i.e. between 4 and 8 hours per hour of tape). The variability in the time taken is dependent on the amount and depth of analysis that you wish to conduct. As yet, we have no real notion of the time involved in correlating eye-tracker data with the video record of what was going on on the computer screen from moment to moment and simultaneously correlating the eye-tracker data with audio recordings of the participants’ dialogue.

4.5.1.6 Methodology Guidelines

Based on our experience, we list in Fig 4.15 some indicative guidelines we have derived and which we recommend should be taken into consideration when contemplating eye-tracking in CSCW.
<p>| | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Where participants can see one another, eye-tracking devices that cause participants to look ‘strange’ may influence the communication task and hence such devices should be avoided. Using a RED device avoids this problem.</td>
<td>10</td>
</tr>
<tr>
<td>2</td>
<td>A RED device allows a participant to move relatively naturally whilst still maintaining tracking. However:</td>
<td>11</td>
</tr>
<tr>
<td>3</td>
<td>Although an eye-tracked individual can move about to some extent and the device will compensate and follow the head and eye movements, gross movements are difficult to follow and data will be lost.</td>
<td>12</td>
</tr>
<tr>
<td>4</td>
<td>If the individual moves significantly to a new position with respect to the viewed scene during the eye-tracked session, then the calibration, and hence the data, may be invalid for some or most of the session. Thus calibration is required at least at the beginning and the end of a session.</td>
<td>13</td>
</tr>
<tr>
<td>5</td>
<td>Notwithstanding the foregoing, if the individual moves to a new position for a while, then moves back to the original position, the fact that they moved, and that their data whilst in their new position may be invalid, may not have been noticed. The operator must watch out for this.</td>
<td>14</td>
</tr>
<tr>
<td>6</td>
<td>Movement in the radial plane (towards or away from the RED) cannot be automatically dealt with. The depth of focus is only a few centimetres at best. An operator must intervene to re-focus on the fly. This is relatively difficult and, in general, operators must be skilled, as learned pattern recognition and manual dexterity are necessary to restore tracking when the device loses track.</td>
<td>15</td>
</tr>
<tr>
<td>7</td>
<td>Learning the keyboard shortcuts required to operate the eye-tracker is important skill. Fast manipulation of the camera during an experiment can make a big difference to the results.</td>
<td>16</td>
</tr>
<tr>
<td>8</td>
<td>In order to maintain a natural communication situation or working environment, the eye-tracked participant must be free to move, but contrarily must be restrained in order to get reliable data. Methods must be devised (such as the reclining chair we have described) to “naturally” restrict movement of the participant. This does not always work to the same extent for all participants.</td>
<td>17</td>
</tr>
<tr>
<td>9</td>
<td>We have found that up to 50% of gaze is still lost for many participants in these situations. We had to discard between about 20% and 40% of participants’ data, as they did not meet the criterion of being successfully tracked for 50% of their gaze.</td>
<td>18</td>
</tr>
</tbody>
</table>

*Figure 4.15: Some practical guidelines for eye-tracking*
4.5.2 Assessing impact on communication

4.5.2.1 Conversation analysis
Communications technology and the quality at which it is used can have a marked impact on users, without necessarily affecting their subjective view of it (i.e. how they would rate it). One way to assess this is to look at the actual make up of the conversation. For instance, we have found that tasks where communication is delivered over an audio-only link can tend to be a lot more formal in that there are longer turns, and fewer interruptions than in face-to-face conversations. See table 4.16 for definitions of conversation analysis measures and examples of how they can affect the findings.

A typical result is that of Olson et al. (1995), who, when studying three-person groups, reported that subjects spent less time explaining and clarifying issues and rated the overall quality of the discussion as higher when video was present.

One of the problems with much of the research designed to assess the quality requirements of audio and video is that studies have tended to focus on only one or two objective measures in isolation - for example, task outcome (Chapanis; 1975; Short, Williams & Christie, 1976; Williams, 1977), or structural aspects of the communicative process such as turn taking (Sellen, 1992). Field studies of users (such as O’Connaill, Whittaker & Wilbur; 1994), have found benefits of at least some forms of video mediated communication either in terms of objective measures of the communicative process or in terms of subjective impressions of users (Tang and Isaacs, 1993). However, as writers such as Monk et al. (1995) argue, a more complete picture of the impacts of technologies requires a more multidimensional approach using a broader range of evaluation data to assess the benefits and costs to users. Such an approach has been adopted by researchers such as Olson et al. (1992), (1993) and Strauss & McGrath (1994). We need to understand the relationship between these variables in order to get a clearer picture of how technology mediates communication and collaboration.

For any collaborative task or interaction, the content of the dialogue is analysed in terms of the pragmatic functions which the speakers are attempting to convey as the dialogue progresses. This involves coding of all the communicative behaviours or ‘Conversational Games’ (Kowtco, Isard & Doherty, 1991) which are attempted and how these are distributed across dialogues when speakers communicate face-to-face, in video mediated communication or in audio-only conditions. We can also explore aspects of the non-verbal communication (gaze) on task. For other tasks we can examine the decision-making process and how frequently ‘clients’ change their plans and decisions. We can examine the lengths of the dialogues in the different conditions and the turn-taking behaviour of speakers. Detailed post-task questionnaire on aspects of user satisfaction with the task, communication and technology can be conducted.

The communication processes have been examined in a number of ways including quantitative measures of the amount of talk needed to complete tasks using different communication media, qualitative assessments of the structure and contact of interactions, detailed assessments of the way turn taking is managed, and even investigations of the articulatory quality of the speech produced in different communicative contexts.

In earlier research on collaborative problem-solving we had found that in face-to-face interaction, participants needed to say significantly less to achieve the same level of performance than in audio only conditions (Boyle, Anderson, & Newlands, 1994). This study is unusual in showing subtle but significant benefits of the availability of visual signals for problem-solving. Most earlier studies which focused only on task outcome showed no advantage for face-to-face problem solving (Davis, 1971;
Chapanis et al., 1972; Williams, 1977). Only in tasks involving conflict or negotiation was there some evidence of a benefit for communication with visual contact (Morley & Stephenson, 1969; Short, 1974). The task we used in the study by Boyle et al and in subsequent explorations of video-mediated communication, is a form of collaborative problem solving known as the Map Task (Brown et al., 1984).

4.5.2.2 Some conversation analysis findings
In the Boyle et al. (1994) study, 32 pairs of subjects tackled the Map Task sitting at opposite sides of a table, communicating face-to-face or with a screen between them. In face-to-face dialogues speakers used 28% fewer turns and 20% fewer words than in the audio only condition. Yet face-to-face participants achieved equally good levels of performance with this reduced verbal input. The interaction was also managed more smoothly in face-to-face collaboration, with 8.7% of turns containing interruptions compared to 12% of turns in the audio only condition. These advantages suggest that speakers can use visual signals to supplement the information presented verbally to assist managing the process of turn-taking.

In another study using the Map Task, we have found that performances where an audio delay was introduced (and the video was synchronised with the audio) were significantly poorer than those without an audio delay. The delay also caused a significant rise in the total number of turns and the rate of interruptions with nearly three times as many interruptions occurring compared with the no-delay conditions. This is slightly at odds with findings reported by O’Connail et al. (1993) who found that audio delay introduces a more formal style of interaction with longer turn-lengths, fewer speaker changes or overlaps and reduced verbal feedback.

In a different collaborative task involving a travel agent (actor) and a client, whilst we found no differences in task performance, in the face-to-face condition the travel agent used 22% fewer words than when they were using only an audio link between rooms. So the length advantage found previously in face-to-face interaction had been replicated in another problem-solving task. However, video mediated communication again failed to deliver the efficiency gains of face-to-face interaction.

<table>
<thead>
<tr>
<th>Dimension</th>
<th>Definition</th>
<th>What it tells you</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of turns</td>
<td>The total number of individual turns (or continuous speech by one individual) in the conversation.</td>
<td>If the number of turns is particularly high this is normally complemented either by more words indicating a longer conversation, more interruptions or by shorter turns. Fewer turns works in the opposite way.</td>
</tr>
<tr>
<td>Number of words</td>
<td>The total number of words spoken by both individuals during the conversation.</td>
<td>The more words used to complete a specific task can tell you how efficient the communication was.</td>
</tr>
<tr>
<td>Number of interruptions</td>
<td>The number of times that speech was interrupted during the conversation.</td>
<td>More interruptions in a dialogue can indicate that conversational cues are being misinterpreted or missed. Different transcribing methods can mean that interruptions are difficult to compare across studies.</td>
</tr>
<tr>
<td>Number of words per turn</td>
<td>Calculated by dividing the total number of words by the total number of turns, this identifies the average length of turn during the conversation.</td>
<td>Longer turns can indicate a more formal, almost presentation style, conversation. Shorter turns usually indicate an informal conversation.</td>
</tr>
<tr>
<td>Backchannels</td>
<td>Backchannels occur when one party acknowledges the other without making an intelligible word. Backchannels may or may not be counted as interruptions depending on how you wish to analyse the dialogues.</td>
<td>Where backchannels are used less this may be as a result of poorer audio rendering them less effective.</td>
</tr>
</tbody>
</table>

Table 4.16: Terms used in conversation analysis
4.5.2.3 How to transcribe dialogues
An example of a real dialogue is shown in Fig 4.17. All text is written in lower case with no punctuation. “Ums” are included, as are all repeated words. The forward slash (i.e. /) indicates an interruption; angle brackets (i.e. < … >) enclose passages of interrupted speech.

If it is feasible it is normally preferable to use an audio typist to make a first pass at transcribing dialogues (with strict written instructions about how dialogue should be written. In a second pass the dialogues can then be checked and coded for interruptions etc. One pass at the dialogues with an attempt to transcribe and code in one will often miss quite a large part of the conversation.

2 is it nice down there
1 <its a pretty nice day /
2 oh it's raining up here /
1 its been nice thats a pity um so in place of that um kind of introduction have you had a chance to look at the website and get get some ideas about it>

2 yeah yeah
1 okay okay um so let me begin by asking you um um what you're doing now and um why you wanna you're taking a year out right you say
2 um ive taken a year out yeah

Figure 4.17: A snippet of dialogue taken from one of our experiments

4.5.2.4 Conversational games analysis
Conversational Games Analysis charts the way speakers achieve their communicative goals. The analysis is derived from the work of Power (1979) and Houghton and Isard (1987), which proposed that a conversation proceeds through the accomplishment of speakers’ goals and subgoals - these dialogue units being called Conversational Games. (For example, an instruction is accomplished via an INSTRUCT Game.) Conversational Games Analysis was developed to detail patterns of pragmatic functions in Map Task dialogues. Utterances are categorised according to the perceived conversational function which the speaker intends to accomplish. This involves taking several sources of information into account: the semantic content of the utterance, the prosody and intonational contour accompanying the utterance, and the utterance location within the dialogue. So for example "Go right" could function to instruct, to elicit feedback or to provide feedback depending upon its dialogue context and intonation.

We have used the following games in our analysis of dialogues:

INSTRUCT: Communicates a direct or indirect request for action or instruction.
CHECK: Listener checks their own understanding of a previous message or instruction from their conversational partner, by requesting confirmation that the interpretation is correct.
QUERY-YN: Yes-No question. A request for affirmation or negation regarding new or unmentioned information about some part of the task (not checking interpretation of a previous message).
QUERY-W: An open-answer Wh-question. Requests more than affirmation or negation regarding new or unmentioned information about some part of the task (not checking interpretation of a previous message).
EXPLAIN: Freely offered information regarding the task, not elicited by co-participant.
ALIGN: Speaker confirms the listener's understanding of a message or accomplishment of some task, also checks attention, agreement, or readiness.
We have also analysed dialogues using referential analysis. This is another type of content analysis, which identifies references to objects from the task and analyses the ‘quality’ of the references in further detail. See Jackson et al. (2000) for more detail.

4.5.2.5 Some conversational games analysis findings

In a study based on the map task, the analyses of face-to-face dialogues showed that speakers less often check that their listener understands them (ALIGN) or that they have understood their partner (CHECK) than in audio-only interactions. There were significant increases in the frequency with which speakers used ALIGN and CHECK games in audio-only conditions, these games occurring 50% and 28% more often respectively. Where visual signals are not available speakers do more verbal checking, whilst in face-to-face conversations non-verbal signals may be substituted, (for a full account see Doherty-Sneddon, et al., 1997).

In a similar study using video mediated communication and audio only conditions, conversational game analysis showed only one significant difference between video mediated communication and audio only dialogues: there were significantly more ALIGN games in audio only than in video mediated conditions. So, as we would have predicted from our analysis of face-to-face and audio dialogues, speakers check that their listener has understood what they are saying (ALIGN games) more frequently when they only have an audio link than when visual signals are available.

In this respect video mediated communication seems to deliver the same type of dialogue benefit as face-to-face communication. Video mediated communication failed to deliver the other advantage of face-to-face interaction: the significant reduction in the number of CHECK games, where listeners check on their understanding of what the speaker has just said.

This comparative study then indicates that with further analysis of the dialogues we can uncover differences that are not apparent from simple observation of the conversation or from regular conversation analysis.

4.5.2.6 Costs

**Capital Outlay:** A transcribing machine with a foot pedal, speed control and auto backspace makes it a great deal easier to transcribe the tape. If you are not using this setup, the transcription will take longer. Audio recording equipment (including an audio mixer) will need to be purchased if this is not already available.

**Materials:** Audio cassettes are required for every experimental session. Make sure they are well labelled and that the tab is pressed in to write protect them. Where the audio is being transcribed it is sensible to make copies of the tapes before the transcription is started as the excessive forwarding, rewinding and pausing of the tape can wear it out. It is also preferable to use good quality tapes to reduce any chance of losing data.

**Time:** Collecting the data requires minimal time just to prepare the audio cassettes. There is an initial setup time for getting the audio mixed onto one tape. If possible two audio feeds should be fed into left and right so that when transcribing, it is easier to pick out who is talking and hear what they are saying when they talk over each other.

By far the most expensive time for conversation/dialogue analysis is the transcription. A good audio typist will take between 4 and 8 hours to transcribe an hour of tape depending on the quality of the audio. Usually this then needs a second pass to check the accuracy of the transcription and to add any further coding required. For basic checking and coding of interruptions and backchannels allow 4-6 hours per hour of tape. This can be reduced if you can find a good audio typist, who is accurate.
Additional Costs: For content analysis of dialogues, you can take all the costs shown above for conversation analysis, and then add on extra time for coding the content of the dialogues. This is a particularly lengthy method of data analysis, and before you start on this you should aim to have a clear idea of what you want to achieve. The level of the content analysis will impact greatly on the time required to complete the coding, there is not really any standard.

5. CONCLUSIONS
Methods for measuring perceived media quality such as the ITU scales are inadequate insofar as they do not take into account task parameters and contexts in which real world users are operating when participating in real-time interactive multimedia communication.

We suggest that the task, the users and the context of use must be taken into account explicitly in order to evaluate the efficacy of a communications system. We further suggest that using objective methods of measuring task performance and better subjective techniques for assessing user satisfaction, together with deeper analyses of user behaviour and user cost can throw light on the underlying processes and the possible consequences of adopting different media quality levels.

The adoption of more rigorous or deeper evaluation methods does often imply an increase in the cost of the assessment operation. Practitioners must balance these assessment costs against the alternative cost of getting an inadequate or misleading evaluation.

It is important to appreciate that our advocacy of a more rigorous and valid evaluation methodology does not imply that the resulting “required quality levels” of the media incorporated into any envisaged real-time interactive communication system would necessarily be higher. In contrast, we suggest that these methods may well assure those developing and implementing such systems that though the incorporated media would not pass traditional quality tests, they may indeed be adequate for their purposes in the contexts envisaged for their use.
6. REFERENCES


Veinott, E. S., Olson, J. S., Olson, G. M. & Fu, X. (1997). Video matters! When communication ability is stressed, video helps. *Proceedings of CHI’97*


Watson, A. and Sasse, M.A. (1998): Measuring Perceived Quality of Speech and Video in Multi-
Bristol, England, 55-60.

Watson, A. and Sasse, M.A. (2000a): Distance Education via IP Videoconferencing: Results from a

Del Rey, CA; 269-276.


Whittaker, S. & O'Connail, B. (1997). The role of vision in face-to-face and video-mediated com-
munication. In K. E. Finn, A. J. Sellen & S. B. Wilbur (Eds.), *Video-Mediated Communication*,

Williams, E. (1977). Experimental comparisons of face-to-face and video-mediated communication:

Communication *International Broadcasting Convention*, Amsterdam

Wilson, G.M. and Sasse, M.A. (2000a). Do users always know what’s good for them? Utilising

Wilson, G.M. and Sasse, M.A. (2000b). Investigating the impact of audio degradation on users:
Subjective vs. objective assessment measures. *OZCHI 2000. Interfacing reality in the new millen-
nium*. Sydney, Australia.

7. APPENDIX A

Assessing context of use in video and audio conferencing: The MARC Evaluation Checklist: Reference and Assessment

The MARC (Multimedia Assessment Reference Checklist) is based on some general work we have done on evaluation of systems and assessing the context in which systems are used. The tool (a derivative of a more general assessment reference checklist produced by the EC TAP funded project MEGATAQ) provides a method for illuminating the overall evaluation implications of new applications or services. As new multimedia products are usually embedded in a context of interactions, a variety of consequences are possible. The MARC can be utilised to highlight potentially important evaluation issues.

The introduction of a new product or system can affect three main areas:

Area 1. *Context of Use*: user characteristics including tasks and responsibilities, new technology and the general working environment.

Area 2. *Interactive Processes*: human computer interaction, user interaction mediated by technology in the framework of task performance and social/organisational processes. We use the methodologies in section 4.3 (Assessing User Cost), 4.4 (Impact on User Behaviour), 4.5 (Impact on Communication) to analyse the interactive processes.

Area 3. *Outcomes*: task effectiveness, product quality, user satisfaction, social and organisational impact. The methods outlined in sections 4.1 (Analysing Task Performance) and 4.2 (Assessing User Satisfaction) are invaluable for measuring and assessing the outcomes of new multimedia systems.

The following checklist attempts to alert the evaluator to the wide spectrum of evaluation issues that may be involved with audio and video conferencing systems, and to assess the extent to which the new system will affect that context. Please note that not all questions will be relevant in all cases.

As an example of issues that may be covered by context of use analysis, consider how multimedia conferencing may be used across timezones. This may be particularly relevant where most of the previous communication has been asynchronous (such as email) or by travel. The new system may change the balance of the power in the office, with the ‘technology expert’ gaining more power. It may make the arrangements for communication more formal, or it may increase the amount of interworking. It may change the way in which the company makes up working groups and teams. It is therefore very important to take the context of use into account when considering new media.
# MARC Area 1. Context of Use

## 1.1 User Characteristics
The new system and the end user

1.11 Will the users of the multimedia system require any of the following:
- New skills or qualifications?
- Additional abilities?
- Change in attitude to task/organisation?
- Increase in team-orientation/co-operation?
- Increase in awareness of new technology?

1.12 Does the multimedia system have any impact on
- The age range of users
- Gender distribution
- First language of users

## 1.2 Group characteristics
The affects of the new system on the existing group structure

1.2.1 Will the multimedia system require changes in any of the following:
- Number of people involved per group
- Communality of the task
- Intensity of interaction (frequency of contact )
- Formality and continuity of team membership
- Duration of the group-life

1.2.2. Group structure
Does the multimedia system result in any changes in the following conditions
- Group composition
- Role/task distribution within the groups
- Power distribution

1.2.3 Task co-operation
Is the multimedia system likely to result in changes in
- Task co-operation within the groups

1.2.4. Group culture:
Is the multimedia system likely to result in changes in
- Attitudes of the groups

1.2.5. Co-ordination
Is the multimedia system likely to result in changes in
- The way the group is co-ordinated
- The ease/difficulties of co-ordinating the group
### 1.3 Tasks and Activities
Is the multimedia system compatible with existing task characteristics:

#### 1.3.1. Job characteristics
Is the multimedia system likely to result in changes in
- Job content
- Hours of work
- Other job characteristics

#### 1.3.2. Task Characteristics
Is the multimedia system likely to result in changes in
- Task goal
- Task duration
- Task complexity

### 1.4 Group Task Characteristics
Is the multimedia system likely to result in any changes in
- Task interdependence within the group

### 1.5 Organisational Environment
Compatibility with existing organisational characteristics

#### 1.5.1 Organisational Structure and Procedures
Is the multimedia system likely to result in changes in
- Departmental/team distribution
- Hierarchical distribution
- Work procedures
- Organisational culture

#### 1.5.2 Technical environment - The introduction of the multimedia system
Did the introduction of the multimedia system result in changes in the following
- Existing tools and applications
- Existing hardware.

### MARC Area 2. Interaction

#### 2.1 Task Interaction

##### 2.1.1 Task Performance
Does the multimedia system have any effect on:
- Speed of task completion
- Errors involved in task performance

##### 2.1.2 Usability
Direct affects of the multimedia system on the user base
Do the users:
- Require more or less mental effort to complete tasks
- Require to invest time and effort learning the system
- Feel satisfied with the system
- Feel they have control when dealing with the system
- Understand the system functionality
- Find the system attractive and exciting to use
- Find the system suitable and useful for performing tasks
- Find the system compatible with their work environment
- Use the system frequently
### 2.1.3 Network performance
The affects of networking quality on system use
If the multimedia system is dependant on remote network connectivity, was the system affected by
- Network delays
- Apparent loss of network connectivity

### 2.1.4. Information access
Does the multimedia system make relevant information
- More accessible
- Easier to obtain

### 2.1.5. Quality of information processes
Is the information
- Accurate
- Adequately detailed
- Current

### 2.1.6. Co-operative Task Interaction
Does the multimedia system appear to affect
- The process of user consensus
- The speed of decision making
- The effectiveness of co-operative problem solving
- The flexibility of task interaction
- The structure of meeting procedures

### 2.2 Communication Issues
#### 2.2.1 Information Exchange and Communication
Are the users who communicate through the system
- Increasing the frequency of information exchanges
- Altering the nature and content of the information exchanges

#### 2.2.2. Quality of interpersonal communication
Has the multimedia system resulted in
- Changes in the use of other communication media
- Changes in the quality of the communication processes
- An increase in the use of non verbal communication

### 2.3. Social interaction
#### 2.3.1 Social Interaction
Has the multimedia system affected
- Participation equality
- Dominance by certain participants
- Social seclusion of certain individuals
- Formality or informality of group meetings
- Pressure for conformity
- Conflict arousal and handling
- Increase in team spirit
- Mutual perception of group members
### 2.4 System Introduction

**2.4.1 System Introduction Phase**

During the design and introduction of the system was there adequate

- Participation of users in the design process
- Information about the multimedia system available
- Training in the use of the multimedia system
- Management involvement and backing

### MARC Area 3. Consequences and Outcomes

**3.1. The goals, services and products of the system:**

- What are the most important products or services of the system
- What are the end goals for the users of the system
- What are the criteria for measuring product or goal success
- Has there been a change in the quality of the products
- Has there been a change in the quantity of the products

**3.2. The costs and efficiency of the system**

3.2.1 Have there been alterations in

- User travel time
- Time involved accessing information
- Time involved in communicating with colleagues

3.2.2 Using high, medium and low, what are the expected costs of

- User training
- Ongoing user support
- Initial disturbances in user performance

**3.3. User outcomes and consequences:**

Has the multimedia system resulted in any changes in

- The quality of user tasks
- Increase in user skill base and learning opportunities
- User requirement for support and help
- General user satisfaction indicators, (motivation, task involvement, absenteeism, stress)
- User behaviour monitoring processes (performance logging)
- Job security issues (staff reduction or availability of promotion opportunities)

**3.4. Group consequences.**

Has the multimedia system resulted in any changes in

- The scope and intensity of user contacts with other individuals
- Group cohesion and knowledge base

**3.5. Organisational consequences**

Has the multimedia system resulted in any changes in

- Organisational culture
- Organisational communication
- Flexibility
- Organisational innovation
- New market opportunities
- Inter-organisational collaboration
- Organisational competitiveness
- Organisational technology policy
- Organisational aims
- Industrial relations