Straight from the Heart:
Using Physiological Measurements in the Evaluation of Media Quality

Gillian M. Wilson & M. Angela Sasse
Department of Computer Science, University College London
Gower Street, London, WC1E 6BT
G.Wilson/A.Sasse@cs.ucl.ac.uk

Abstract

Subjective methods are widely used to evaluate Multimedia Conferencing quality, yet when used in isolation the results can be misleading. Therefore, this research is measuring physiological indicators of stress (GSR, HR & BVP) as an indicator of the user cost of the reception quality of a multimedia conference. These measurements are taken as part of a 3-Dimensional approach, which incorporates task performance, user satisfaction and user cost. The main results have shown that subjective and physiological responses do not always correlate and that specific physiological signals respond differently depending on the degradation and task. The ability to measure stress subconsciously and unobtrusively has many potential applications.

Keywords: Subjective assessment, Multimedia conferencing, physiological measurements, 3-D approach.

1 Introduction

Multimedia conferencing (MMC) allows real-time communication between two or more users through the tools of audio, video and a shared workspace. In recent years MMC has become widely used – the number of Internet users is expected to triple between 1998 and 2002, (Cullinane, 1998) largely because of new applications like MMC. It is valuable in areas like distance teaching and remote business meetings. However, most research in this area has been at the level of the network provider, as opposed to the end user. This is unwise considering it is the end user who will determine the uptake, and pay for, such applications.

Despite MMC being subject to unique degradations, such as packet loss, computer workstations and high-bandwidth networks can deliver good audio and video quality at a higher financial cost. Since most users do not want to pay more than necessary for their communications and the fact that there will always be a market for lower, less expensive quality, the optimum and minimum levels of quality that support users completing specific tasks need to be determined. The point at which increased quality is of no further potential benefit to the user should also be investigated as this allows efficient use of bandwidth.

Establishing such quality thresholds is essential for network providers and application designers. To date, there has been little research investigating how best to assess MMC quality, and there exist no guidelines on the quality levels that should be delivered for a particular task, or on the assessment methods that should be used to determine these thresholds.

Currently subjective methods are mostly used in the assessment of audio and video quality. However, there are problems associated with the use of those methods, and when used as the single means of assessing whether quality is adequate, the results obtained may give a misleading impression about the impact of the quality on the user.

This paper presents a novel method for assessing multimedia quality in the context of networked applications: physiological responses to media quality degradations are being taken as an objective measure of user cost. Such methods should be part of a traditional HCI evaluation approach that considers task performance, user satisfaction and user cost, in order to obtain a reliable indication of how media quality affects users.

Section 2 of this paper details the status quo in MMC assessment, and the drawbacks of this. Section 3 presents the novel assessment method this research is utilizing and section 4 describes the research approach. Section 5 presents the empirical studies that have been performed to date. Section 6 summarises and discusses the main findings and finally, section 7 details the specific and general contributions this research is making.

2 Current methods of assessment

2.1 Subjective assessment

Subjective methods are widely used to assess MMC quality. In particular, these often take the format of the rating scales recommended by the International Telecommunications Union (ITU). Typically, a short section of material is played, after which a 5-point quality/impairment rating scale is administered and a Mean Opinion Score (MOS) calculated. However, research has highlighted their ineffectiveness in evaluating multimedia audio (Watson & Sasse, 1998) and video. For example, the test material is not long enough to allow the user to decide if the quality is *good enough* in the context of longer, more interactive use. In addition, the scale labels can be viewed as unrepresentative of MMC quality. For example, it is unlikely that it would ever be classed as 'excellent'.

In addition to problems with the rating scales, subjective assessment has a more fundamental problem - it is subject to cognitive mediation. This means that contextual variables such as budget (Bouch & Sasse, 1999), or task difficulty (Wilson & Descamps, 1996) can influence users assessment of quality. Moreover, Knoche 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. Therefore, utilising subjective assessment in isolation can be misleading.

2.2 Task performance

Knoche et al. (1999) argue that measures of task performance should be used as an alternative to subjective assessment. Task performance is an essential element of usability, yet should not be used as its only measure. For example, a user may be able to complete his/her task under a particular quality level, but how does it impact upon them *physically* – is it uncomfortable/do they have to strain more? Additionally, what are their *opinions* on the quality?

2.3 The 3-Dimensional (3-D) approach

In tackling this problem, a traditional Human Computer Interaction (HCI) evaluation framework of task performance, user satisfaction and user cost (Figure 1) has been revisited. User cost is an explicit - if often disregarded - element of this framework.

The weighting given to each of these dimensions depends on the context in which MMC is being used. For example, if the context is entertainment, then the most important dimension to consider is user satisfaction – the opinion the user has and the enjoyment they felt from interacting with the application. If a user interviews daily using MMC, the most important dimension to consider is the *cost* to the user, as interacting frequently with degraded quality whilst performing an important task may be harmful to health. If MMC is being used in a distance education context, the most important dimension is task performance, as the students must be able to effectively learn using the application.

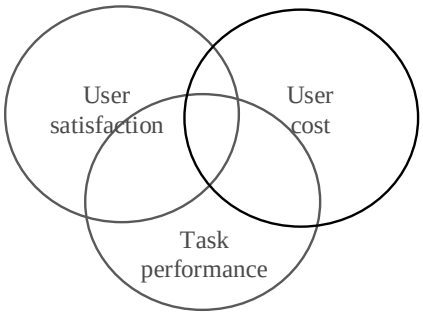

Figure 1: The 3-D approach

3 User Cost

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 stress and discomfort. 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 a response of stress, even if the user remains capable of performing his/her main task. Thus for the purposes of this research, user cost is defined as *physiological stress*.

3.1 Physiological Stress

To measure stress in this research, the following signals have been adopted (Wilson & Sasse, 2000a): Galvanic Skin Response (GSR), Heart Rate (HR) and Blood Volume Pulse (BVP). These signals were chosen as they are physically non-invasive, are good indicators of stress and are easy to measure with specialised equipment - the ProComp, manufactured by Thought Technology Ltd[1] is being used. Under stress, GSR and HR increase, whereas BVP decreases.

3.2 Positive and negative stress

There are two types of stress that exist: *distress*, which is negative and *eustress*, which is a positive type of stress. Could it be that degraded quality, as being investigated here, is arousing and positive for users as it forces them to pay more attention? This research believes this to be

[1] www.thoughttechnology.com

false. It is know that everyone has an optimal level of arousal for performance - this is known as the Yerkes-Dodson law (1908) - and that going over this level will result in an impairment of task performance and possibly stress. The working hypothesis of this research is that there is an inherent amount of arousal in every task and to increase this with degraded quality will result in distress. It is also posited that interacting frequently with such poor quality levels could be damaging to health in the long term as the user is constantly straining more.

To address this issue, three physiological signals are measured, as opposed to relying on solely one - this was recommended by researchers (Healey & Picard, 2000) at the MIT[2] Media Laboratory working in the area of detecting driver stress physiologically. In addition, the results are interpreted along with subjective measures of user cost (section 4.2).

4 Research Approach

4.1 Research questions

The following questions are being tackled in this research:

- Can physiological responses to media quality degradations be detected? This is the most fundamental question as taking such measurements in this area is novel.
- What aspects of objective quality are stressful for users? Quality is not uni-dimensional and encompasses more than parameters affected by the network, thus effects of other contributing factors, which have largely been overlooked, (e.g. degradations such as echo and volume differences between speakers, which are due to the hardware set-up and end-user behavior) must be investigated.
- If physiological responses can be detected, what are they indicative of and how do we know they are detrimental for users?

4.2 Research strategy

From pilot trials, it was found that participants' physiological signals were indicating stress when an experiment began. This was attributed to the participant being apprehensive about the experiment and their performance in it. To combat this, physiological responses from the first five minutes of experiments are disregarded. The baseline gathering session, which occurs fifteen minutes prior to any experimentation, also allows participants time to settle down and to acclimatize to the environment.

Observed stress in the experiments could be due to environmental factors, such as a phone ringing or people moving around in the room. Therefore, the environment in the laboratory-based trials is kept constant and minimally stressful: the effects of quality need to be determined in isolation before external variables can be accounted for in the field. In addition, the tasks used throughout this research have been carefully designed to ensure that they are engaging whilst remaining minimally stressful - a stressful task could drown out responses to the quality. Again, the effects of the quality alone need to be determined before real-world tasks can be used.

Finally, subjective assessments of user cost are administered to allow the participants to comment on how excited or stressed they felt throughout the experiment - this aids interpretation of the physiological results. Subjective assessments of quality are also gathered: physiological measurements identify problems, yet when used in isolation they do not aid problem resolution - we still need people to tell us *why* they found a degradation stressful.

5 Empirical Studies

5.1 Video frame rate

The first parameter of MMC that was investigated was video frame rate (Wilson & Sasse, 2000b). Television quality video is transmitted at 25 frames per second (fps) in the UK, however frame rates lower than this are necessary in MMC, as high frame rates require a large amount of bandwidth. Previous research by Anderson et al. (2000), found that participants do not subjectively notice the difference between 12 and 25 fps when they are engaged in a task. Additionally there was no effect of frame rate on task performance. If such a difference is still not noticed when it is made more extreme, does this imply that low frame rates have no effect on people? Such a finding would have positive implications for the conservation of bandwidth.

In this experiment, the difference used by Anderson at al was made more extreme: 5 & 25 fps were investigated. Twenty-four participants were involved and their task was to watch recorded interviews, which were acted between a university admissions tutor and candidates for the computing degree course at University College London. The participants' task was to decide if the candidates should be offered a place on the course. This task was carefully designed, with the interaction between the interviewer and interviewees being scripted, in order to keep the content consistent between interviews and minimally stressful.

[2] Massachusetts Institute of Technology

The interviews began at 16 fps for 5 minutes. The data under this frame rate was discarded in order to account for any change in physiological responses due to the experiment beginning. Participants then saw two interviews at either 5-25-5 fps or 25-5-25 fps. Each frame rate was held for a period of 5 minutes & the frame rate changed twice to counteract any expectancy effect.

Physiological measurements were taken throughout the experiment. At the end of each interview, a questionnaire was administered where participants expressed their opinions on the quality, the task, and how they felt (stressed or otherwise) throughout the experiment.
The following hypotheses were posited:
1. There will be different physiological responses to the two frame rates: 5 fps will cause physiological signals to indicate stress.
2. Participants will not subjectively notice the frame rate change, as the task is engaging.

5.1.1 Results and discussion

The results showed that video frame rate had a physiological impact upon the participants in the direction that 5 fps caused a statistically significant increase in stress responses from 25 fps in all three signals, with 75% of results going in the direction predicted. Subjectively, only 16% of participants noticed that the frame rate had changed. Thus, both the hypotheses were supported.

These results show that physiological responses to video quality can be detected. Yet, when users are engaged in a task, they do not subjectively notice the difference between two extreme frame rates. The direct implication of these results is that at very low frame rates, users have to work harder at the same task. Application designers and network providers should consider this information.

The findings from this experiment also imply that acceptable quality levels required for a task and those that result in unacceptable user cost should not be determined by subjective assessment alone, as it may not pick up important effects occurring in a participant's physiology. If solely subjective assessment had been used in this experiment, it would have been assumed that the low frame rates were adequate in this context. This supports the argument for the 3-D approach.

5.2 Audio degradations in a passive listening task

Having investigated a parameter of the video channel, for the second experiment it was decided to investigate the impact of a number of audio degradations. Previous research has determined that good audio quality is a necessary condition for usable MMC (Sasse et al., 1994). To date, it has been assumed that many audio problems in MMC were due to network packet loss and that increasing the amount of bandwidth available would ultimately improve the application for the end-user. However, a large-scale field trial involving thirteen UK academic institutions (Watson & Sasse, 2000a) found that in addition to the many problems reported due the network, there were also reported problems due to end-user behavior and the hardware set-up. Therefore, it was decided to perform an experiment to determine the physiological and subjective responses to such degradations and their relative importance (Wilson & Sasse, 2000c).

Twenty-four participants listened to the same six two-minute recorded audio files twice, of a dialogue between two speakers that had been taken from previous project meetings conducted by MMC. The quality did not change within each condition. At this point it was not desirable to have the video channel causing a distraction to the audio degradations, thus audio was examined in isolation. The task was one of passive listening, so as to place minimal stress upon the user.
The conditions were:

- 20% audio packet loss.
- 5% audio packet loss[3].
- Loud volume differences between speakers.
- Quiet volume differences between speakers.
- Echo.
- Audio recorded using a 'bad' microphone (mike).

It is acknowledged that the bad mike condition is difficult to quantify, yet it was viewed worthwhile to investigate as it was found to be problematic in the field trial mentioned previously. Moreover, three Internet audio experts agreed the condition to be identifiable of the degradation that we aimed to test. 20% packet loss generally has a bursty pattern of occurrence, as opposed to the stable nature of the condition in this experiment and it is known that at 20% loss there is a sharp drop in perceived quality of repaired speech (Watson & Sasse, 1997, 1998). Thus, it was of interest to determine if the effects of consistently high packet loss were as severe as would be expected, and to compare this to the other degradations.

[3] 5% is representative of the level of packet loss most users are likely to experience on the SuperJANET multicast service today (Watson & Sasse, 2000)

Physiological measurements were taken throughout, and at the end of each clip the participants rated their opinion of the quality on an unlabelled 100-point scale.
There were 2 hypotheses:
1. There will be different physiological responses to the conditions.
2. These will correlate with subjective responses, as the task is not engaging.

5.2.1 Results and discussion

The results from this experiment showed some interesting discrepancies between subjective and physiological responses in the BVP and HR signals. However, GSR failed to produce any significant results, the possible reasons for which will be discussed later.

Loud was both stressful and subjectively rated as being poor. The **bad mike** condition was stressful, but was not subjectively poor. **20% packet loss** was rated as being the poorest subjectively and was in the top three most stressful in both of the physiological signals, but was never *the* most stressful. **Echo** was subjectively poorer than it was stressful. Finally, **quiet** and **5% loss** were the least stressful and were rated as being of the best quality.

These results show that physiological responses to audio degradations can be detected. **20% packet loss** was never the most stressful in any of the signals, despite being rated as the poorest. This implies that more attention should be given to the relatively easily rectifiable problems caused by hardware set-up and end-user behavior, as these affect users as much as problems due to the network (section 6.1).

The subjective and physiological results did not correlate: this was not expected, as the task was not engaging. Thus, a more fundamental problem with subjective assessment may have been uncovered (section 6.3). These results provide support for the 3-D approach to media quality assessment: if solely subjective assessment had been used in this experiment, 20% loss would be treated at the expense of a bad mike. This experiment is a starting point in the investigation of audio degradations, so before any recommendations can be made, the video channel must be added and the task made more engaging.

A final issue to highlight is the fact that the GSR signal failed to produce any significant results, despite the means being generally in the same direction as the HR and BVP signals. There are 3 possible explanations for this:
1. It is known within the psychophysiology community that autonomic signals do not always correlate with each other all the time (Lacey & Lacey, 1958), as individuals have specific response patterns.
2. There may be different types of discomfort to media quality degradations, with audio degradations not affecting GSR.
3. It could be suggested that audio degradations do not affect GSR in a passive task, but they may in an engaging task with the video channel present.

5.3 Audio degradations in a recorded interview task

For the third experiment, it was necessary to determine if the results to the previous experiment held when the video channel was added, the task was made more engaging and the samples were increased in length.

Twenty-four participants watched four recorded interviews, which were again scripted and acted between a university admissions tutor and candidates for the computing degree course at UCL. The interviews were ten minutes in length each. Twelve participants experienced five minutes of normal quality followed by five minutes of degraded quality, whereas the other twelve experienced five minutes of degraded quality followed by five minutes of normal quality. The conditions were:

- 20% audio packet loss on both speakers.
- 5% audio packet loss on both speakers.
- Audio that was loud.
- Audio recorded using a 'bad' microphone.

The video was held at a level established in experiment 1 that does not adversely affect people: 25 fps. Physiological measurements were taken throughout, and subjective assessment of the quality and how the participants felt, were administered at the end of each interview.
There were two hypotheses:

1. There will be different physiological responses to the conditions: bad mike will be most stressful, 5% will be the least stressful.
2. Subjective and physiological responses will not concur, as the task is engaging.

5.3.1 Results and discussion

The results showed that there was a distinction between what caused stress *most frequently* - **20% packet loss**, and what caused the *highest levels* of stress - **loud**. The **bad mike** condition was the third most stressful condition, and **5% packet loss** was the least stressful. Subjective results are still in the analyses stage.

The **5% loss** condition being the least stressful was expected - it does not subjectively or physiologically affect people much. From the results of experiment 2, it was predicted that the **bad mike** condition would cause more stress than it did here. This result illustrates the need to investigate degradations using a variety of tasks, as different tasks have different quality requirements. In the passive listening task the **bad mike** caused problems, but when the task is more engaging and the video channel is present, it does not affect people as much.

Loud causing the highest levels of stress is understandable: a volume level which people find uncomfortable will cause them high levels of stress, whereas with **20% loss**, this seems to be consistently irritating.

The expectancy effect which has been found subjectively (Bouch & Sasse, 2000) appeared to be occurring in the physiological data: this is where in the poor to normal quality group, the normal quality was more stressful than in the other group who experienced normal to poor quality. A possible reason for this could be that users strain more after degraded quality, as they are on-guard in case it occurs again. Alternatively, the effects of degraded quality may be long lasting, thus affecting the user under the normal quality. This is being investigated further as it has implications for quality delivery, in that good quality should not be delivered directly after very poor quality, as users will not perceive it as being as good as it is.

5.4 Audio and video degradations in an interactive task

In the final experiment that has been performed to date (performed as part of the ETNA[4] project), this research moved forward. Firstly, the participants were active, as opposed to completing the passive tasks employed previously. Secondly, audio and video were manipulated in the same condition. The questions of interest were; will meaningful results emerge to the quality when performing an active task, and what interactive influence do audio and video degradations have on each other?

Eleven admissions tutors (the participants) at UCL interviewed four candidates at Glasgow University applying for the computing degree course at UCL. The interviews were conducted over the network and in real-time. The participants were told that the candidates were real applicants, and that the interviews were to determine the feasibility of utilising MMC as an interviewing tool. In fact the 'candidates' were actors, in order to provide consistency of responses. There were no limits put on the interview length – they were usually in the region of ten minutes. There were four conditions:

- High (~0% packet loss) audio quality.
- Low (~15% packet loss) audio quality.
- High video quality (~20fps).
- Low video quality (~5fps).

There were 4 hypotheses:

1. The low-video quality with low-audio quality condition will be the most stressful.
2. The high-video quality with the low-audio quality will be the next most stressful, as it is known that audio is the more important channel in MMC (Sasse et al., 1994).
3. The low-video quality with the high-audio quality will be the third most stressful condition.
4. The high-video quality with high-audio quality will be the least stressful condition.

5.4.1 Results and discussion

Basic statistical analyses performed to date on the physiological responses have provided inconclusive results: few signals were in the direction predicted and there was little correlation between the signals, thus to this point no conclusions can be made from them. However, more in-depth analyses are being performed utilising more advanced statistics. This is necessary as there are many more effects to be teased apart than in the passive tasks.

There did appear to be some order effects, where the GSR signal increased with time, whereas the HR signal decreased with time. The reason for the this could be that the interviews were too long - some were twenty minutes in length each for the four conditions. This means that some participants were in the room for over one hour and thirty minutes in total (including the baseline measuring session). As a result, their fingers may have become very sweaty from having the sensors on, thus GSR would increase. This indicates stress, yet the fact that the HR decreased implies that the participants were more relaxed as time went on, which is understandable. The length of time participants are involved in an experiment needs to be considered in future experiments.

Subjectively, changes in the video channel were found to be less important than changes in the audio channel, as was expected. However, none of the 'candidates' were borderline in their academic records: would the video channel play a bigger role when a real character judgment is required? Finally, the admissions tutors thought that remote interviews of this type were feasible, which is positive in situations where it is impossible for the interviewer and interviewee to be in the same place at

[4] http://www-mice.cs.ucl.ac.uk/multimedia/projects/etna

the same time. However, face-to-face interviews were still preferred.

6 Discussion of Results

6.1 Audio degradations due to the hardware set up and end user behavior

Interestingly, it was discovered that audio problems due to the hardware set up and end-user behavior affected users as much as network problems. In particular, a 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 here. Thus, even if perfect quality is delivered in terms of the network, the users' experience with the technology could still be marred by easily rectifiable hardware problems.

Watson & Sasse (2000b) suggest that to address this issue, a fault diagnosis option could be built. For example, a help menu on an audio tool could include a list of problems described in the terms that users most commonly generate e.g. fuzzy, which refers to packet loss. The user could then search down the list for the description of the problems, and then follow the solution suggested.

Another idea put forward is to playback the user's audio as heard by other participants. At present, it is not possible for participants in a multimedia conference to hear what they sound like. This could occur at the start of each conference to allow time to rectify any problems that may exist. Following on from this, designers could develop a tool to perform an expert system style diagnosis of a users' speech stream and point them to the likely cause of the problem. Users could be required to record sample sentences & only be allowed onto the network once the quality of the sample files is matched or recognized as providing satisfactory quality for the task the user is performing.

6.2 Differences in signal response

Another interesting finding is that the individual physiological signals respond differently to audio and video degradations. It is known that physiological signals do not always correlate with each other (Lacey & Lacey, 1958). For example, in the video frame rate experiment, all of the signals responded strongly, but it was HR that responded the strongest. However, in the passive listening task, GSR did not respond at all: there were no significant differences. When the video channel was added in experiment 3, GSR did respond well. Thus, it seems that there may be different patterns of stress for the different degradations and that these partially depend on the task being performed. This is understandable: poor audio in a passive, non–engaging task may not affect GSR, yet when the task is made more engaging, and the participants are paying more attention, a degradation which causes them to strain more to complete their task will put them under more stress. This finding offers support for the use of three signals, as opposed to relying on solely one (Healey & Picard, 2000).

6.3 Relationship between subjective and physiological measures

Another main finding of this research is that subjective and physiological results do not always correlate with each other. When this research began, it was posited that this would only be the case when the task the user was performing was engaging. 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 (experiment 2).

This 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 rating scales were used, so the argument that the problem lies with the scales being insensitive does not hold. Thus, a fundamental flaw with subjective assessment may have been uncovered: users cannot consciously evaluate the impact that media quality has upon 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 strongly recommended that the 3-D approach be utilized in multimedia quality evaluation and also in assessment in others areas of HCI.

6.4 Future work

Future experiments investigating users' responses to video packet loss in an audio and video streaming entertainment application will be investigated. These experiments will incorporate the investigation of habituation to degraded quality and the impact of stable and variable quality.

The fact that quality degradations can have a physiological impact on users, without them being aware of it, raises another important issue. If the physiological response is due to users working harder on the perceptual task, it is likely that they will get tired or irritated during longer stretches of the task. This means they may

experience an affective response, and – given that they do not realise it is due to the audio or video quality – they may misattribute it to another source. This type of misattribution could have serious implications in a real-world context. For example, if a user was interviewing a candidate over a multimedia conference link, he/she might attribute the effect of the degraded quality to the candidate, and subsequently judge them as being unsuitable for the position. We did not observe this effect in experiment 4 (section 5.4), but none of the "candidates" in this experiment were borderline in their application.

Another scenario is if a user was interacting with an e-commerce web site under poor quality conditions, the stress that the user is likely to experience could be attributed to a lack of trustworthiness of the company, thus the user's opinion of the company could be damaged. This effect has been found subjectively (Bouch et al., 2000). In this paper, an experiment was performed, where there was latency between the request for a web page and receiving the page. It was discovered that cumulative slowness on web pages suggested to the participants that the product being sold was of an inferior quality, and that the security of the purchase was compromised. Once the participants perceived that the security had been compromised, no purchase was made, thus the main purpose of the commercial web site was critically damaged.

Thus, the effects of affective responses that are likely to occur unconsciously in a user under degraded quality levels need to be investigated further. This can be done by analysing task performance (e.g. in the interviewing task), and qualitative data analysis of the way in which users describe the interaction (e.g. words they use to describe the candidate).

7 Contributions

7.1 Specific Contributions

This research aims to produce two substantive contributions. Firstly, the minimum levels of multimedia quality for certain tasks at which users can successfully perform, without significant user cost, will be determined. These findings will be incorporated into the ETNA Project, which aims to produce a taxonomy of real-time multimedia tasks and applications, and to determine the maximum and minimum audio/video quality thresholds for a number of these tasks. This will greatly assist network providers and application designers, as they will have guidelines on the quality they need to deliver for specific tasks, thus improving applications for the end-user.

Secondly, a masters student at UCL has been working alongside this research on providing feedback to the user when they are engaged in an application. For example, a user involved in a multimedia conference will be able to see an animated face in the corner of the screen that is happy when the user is relaxed and sad when the user becomes stressed. Such basic feedback will give control & awareness back to the user of effects they are not normally conscious of. At present, the interface is a prototype and does not respond in real-time, however it is an interesting development and promotes greater awareness of the stress degraded media quality may cause.

A methodological contribution will also be made: guidelines for further research in this area will be provided e.g. it may appear that certain signals are better indicators of some impairments than others, as mentioned previously. This will assist further research in this area, which at present is sparse, yet this may be about to change with other institutions in the UK adopting this technique.

7.2 General Contributions

This research is also providing a general contribution to HCI methodology, by promoting the measurement of user cost. It has largely been neglected in this area yet is vital to the uptake and prolonged use of applications. Designers need to ensure that products users interact with in everyday life and use to perform important tasks do not put them under any adverse pressure. Stress levels in the workplace are already very high so any attempt to reduce them should be considered

Thus, this technique is not solely for use in multimedia quality evaluation. It can be used in other areas, such as distance teaching. For example, in a tutorial performed over a multimedia conference link, the physiological signals of the students could be measured - if a student became stressed, then a warning light could be displayed above their image on the tutors machine. This would allow the tutor to slow the pace of the lecture or to clarify a point.

The measurement of physiological signals could also be used in the area of product assessment. An example of this is that researchers at British Telecom have developed the Motivational User Interface (MUI) (Millard et al., 1999). It is for use by call center operators and attempts to make their jobs less stressful. They are keen for us to perform a physiological evaluation of the interface and a determination of environmental stressors.

The detection of stress could also be beneficial for illnesses like RSI (Repetitive Strain Injury). For example people who do a lot of a specific activity, like typing, could wear muscle sensors to alert them to take a break and to rest before any lasting damage is done. This would be extremely useful when it is considered that users may not notice the stress they are under, as has been illustrated by this research.

Stress in the workplace is an ever-increasing problem - over 6 million working days per year in the UK are lost due to stress. At an individual level, stressed workers could be given feedback on their physiological signals as they work, which would allow them to determine the areas of their job and the times of the day at which they become most stressed. They could then adopt measures to combat this e.g., taking a break or reducing their stress levels through the technique of biofeedback - this is a technique for monitoring physiological activity and converting it into an auditory or visual message. It has been shown that e.g. tension headaches and heart rate can be controlled by giving feedback without any external reward as the feedback and perceived control are intrinsically rewarding.

Alternatively, employers could monitor stress levels. This is usually done subjectively, but due to the problems with subjective assessment, the physiological measurement of stress may contribute some new and possibly more truthful data (e.g. people may not feel they can openly criticise areas of their job they dislike). Action could then be taken by employers to improve the areas of employees' jobs, which includes the working environment, in which stress levels are highest.

These examples illustrate that this novel research approach, which unconsciously and unobtrusively measures stress levels, has many areas in which it can offer a valuable contribution.

Acknowledgements

Gillian Wilson is funded through an EPSRC CASE studentship with British Telecom Labs - award GR/98002907.

References

Allanson, J., Rodden, T. & Mariani, J. (1999) A toolkit for exploring electro-physiological human-computer interaction. Proceedings of INTERACT '99, IOS Press, pp231-238.

Anderson, A.H., Smallwood, L., MacDonald, R., Mullin, J., Fleming, A. & O'Malley, O. (2000) Video data and video links in mediated communication: what do users value? International Journal of Human Computer Studies, 52 (1), 165-187.

Bouch, A. & Sasse, M.A. (1999) Network quality of service: what do users need? Proceedings of the 4th International Distributed Conference, 22-23 September 1999, Madrid, Section 9, pp78-90.

Bouch, A. & Sasse, M.A. (2000) The case for predictable network service. In K. Nahrstedt & W. Feng (eds.) Proceedings of MMCN 2000, January 24-27 2000, San Jose, California.

Bouch, A., Kuchinsky, A. & Bhatti, N. (2000) Quality is in the eye of the beholder: meeting users' requirements for Internet quality of service. Proceedings of CHI 2000, April 2000, The Hague, The Netherlands, ACM Press, pp297-304

Cullinane, P. (1998) Ready, steady, crash. Telephony, 3, pp3-13.

Healey, J. & Picard, R. (2000) Smart car: detecting driver stress. Proceedings of ICPR'00, Barcelona, Spain.

Knoche, H., De Meer, H.G. & Kirsch, D. (1999) Utility curves: mean opinion scores considered biased. Proceedings of 7th International Workshop on Quality of Service, 1-4 June 1999, University College London, London, UK.

Lacey, J.I. & Lacey, B.C. (1958) Verification and extension of the principle of autonomic responses stereotypy. American Journal of Psychology, 71, 50-73.

Millard, N., Coe, T., Gardner, M., Gower, A., Hole, L. & Crowle, S. (1999) The future of customer contact. British Telecom Technology Journal, http://www.bt.co.uk/bttj/vol18no1/today.htm

Sasse, M.A., Bilting, U., Schulz, C-D. & Turletti, T. (1994) Remote seminars through multimedia conferencings: experiences from the MICE project. Proceedings of International Networking Conference, 13-17 June 1994, Prague, Czech Republic.

Watson, A and Sasse, M.A. (1997) Multimedia conferencing via multicast: determining the quality of service required by the end user. Proceedings of AVSPN '97 – International Workshop on Audio-visual Services over Packet Networks, pp 189-194.

Watson, A. & Sasse, M.A. (2000a) Distance education via IP videoconferencing: results from a national pilot project. Proceedings of CHI 2000 Extended Abstracts, April 2000, The Hague, The Netherlands, ACM Press, pp113-114.

Watson, A. & Sasse, M.A. (2000b) The good, the bad, and the muffled: the impact of different degradations on internet speech. Proceedings of the 8th ACM International Conference on Multimedia, October 30th - November 3rd, Marina Del Rey, CA; pp. 269-302.

Wilson, F. & Descamps, P.T. (1996) Should we accept anything less than TV quality: visual communication. Paper presented at the International Broadcasting Convention, Amsterdam, 12-16 September 1996.

Wilson, G.M. & Sasse, M.A. (2000a) Listen to your heart rate: counting the cost of media quality. In A. Paiva (ed.) Affective Interactions – Towards a New Generation of Computer Interfaces. Lecture Notes in Artificial Intelligence. Springer, ISBN 3-540-41520-3, pp9-20.

Wilson, G.M. & Sasse, MA. (2000b) Do Users Always Know What's Good For Them? Utilising Physiological Responses to Assess Media Quality. In S. McDonald, Y. Waern & G. Cockton (eds.) Proceedings of HCI 2000: People and Computers XIV - Usability or Else! Springer, pp. 327-339, September 5th - 8th, Sunderland, UK.

Wilson, G.M. & Sasse, M.A. (2000c) Investigating the impact of audio degradations on users: subjective vs. objective assessment methods. In C. Paris, N. Ozkan, S. Howard & S. Lu (eds.) Proceedings of OZCHI 2000: Interfacing Reality in the New Millennium, pp135-142, December 4th - 8th, Sydney, Australia.

Yerkes, R.M. & Dodson, J.D. (1908) The relation of strength of stimulus to rapidity of habit formation. Journal of Comparative and Neurological Psychology, 18, 459-482.