Investigating the Impact of Audio Degradations on Users: Subjective vs. Objective Assessment Methods

Gillian M. Wilson and M. Angela Sasse
Department of Computer Science
University College London
Gower Street, London, WC1E 6BT
{g.wilson, a.sasse}@cs.ucl.ac.uk

Abstract

Low-cost multimedia conferencing (MMC) is increasing in popularity, but it is often questioned whether the quality of the audio and video provided is usable. Traditionally, subjective methods have been employed to assess this. However, recent findings suggest that subjective ratings, which are cognitively mediated, may not reliably detect the impact of quality on users. To address this problem, we are taking physiological indicators of stress as a measure of user cost. In a study with 24 participants, physiological and subjective responses were taken to six types of audio degradation. Results show that the most physiologically stressful condition (audio recorded using a bad microphone) was not subjectively rated as poor. This discrepancy between subjective and physiological responses illustrates the peril of using subjective assessment alone, and supports our proposal for a three-tier approach to media quality assessment of task performance, user satisfaction and user cost.

Keywords: evaluation methods, empirical evaluation, subjective assessment, user cost, audio, multimedia conferencing, physiological measurements.

1. Introduction

Multimedia conferencing (MMC) over the Internet is increasing in popularity. It facilitates communication between two or more users, through the tools of audio, video and a shared workspace. It is used in areas such as distance education and remote business meetings. High quality MMC solutions are available, but at a price that is out of reach to many users: lower quality is often sufficient for a range of purposes. To provide the benefits of MMC to a wider user community, determining the levels of audio and video quality required for users to effectively and comfortably complete their tasks is essential.

Currently, subjective methods are mainly used to assess media quality. However, results obtained with these methods may not always be a reliable indicator of usability. This paper details a new approach: physiological responses are being measured as an indicator of *user cost*. We propose that task performance, user satisfaction, and user cost should all be considered as part of a three-tier approach to evaluating multimedia quality.

We present the background to this approach in sections 2 and 3. Section 4 describes an experiment that examined the subjective and physiological effects of a number of audio degradations on users. Section 5 presents the results of this study, which are discussed in section 6. Finally, section 7 presents the conclusions and implications of this research.

2. Evaluating Multimedia Quality

The ITU (International Telecommunications Union) recommended subjective rating scales are widely used to assess audio and video quality. 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, recent research has highlighted their ineffectiveness in evaluating MMC audio and video [22, 23]:

- The scales were designed to rate toll-quality audio and high-quality video, whereas MMC audio and video are subject to unique impairments such as packet loss and delay.
- The scales are mainly concerned with determining if a particular degradation in quality can be detected, whereas with MMC it is more important to determine if the quality is *good enough* for the task.
- The short time duration of the test material used means that there is not the opportunity for the viewer/listener to experience all the degradations that impact upon MMC. Subsequently, a dynamic rating scale for video is now recommended by the

ITU (ITU- BT 500-8)[10] in order to account for changes in network conditions.

- The vocabulary on the scales (Excellent, Good, Fair, Poor, Bad) is unrepresentative of MMC quality and the scales are not interval in many languages, therefore scores obtained can be misleading.
- Finally, the scales treat quality as a uni-dimensional phenomenon. This is questionable as there are many factors that are recognised to contribute to users perception of audio [12] and video [8] quality.

In order to address these problems, an unlabelled rating scale was devised at UCL [22], and studies showed that users were consistent in their quality ratings using the scale. However, it is a post-hoc method, therefore is subject to primacy and recency effects. A dynamic software version of this scale was subsequently developed, QUASS[1] (Figure 1), which facilitates the continuous rating of the quality of a multimedia conference [2]. The drawback of this method is that continuous rating can result in task interference.

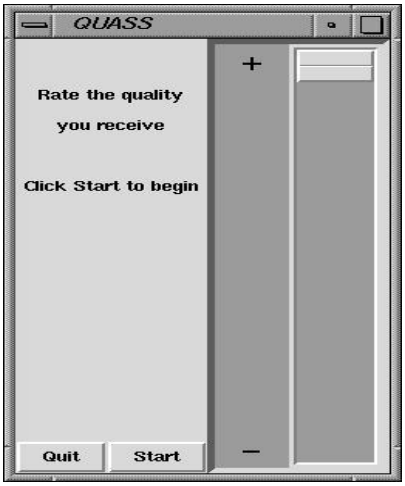

Figure 1. QUASS tool

2.1. Problems with subjective assessment

In addition to the specific problems with the rating scales, there is a fundamental problem with subjective assessment: it is cognitively mediated. This means that it is influenced by variables other than "perceptual adequacy".

This is illustrated in a study which found that users accepted significantly lower levels of media quality when financial cost was attached - the accepted quality levels were below the threshold previously established as necessary for the task [1]. In addition, Wilson & Descamps [27] showed that the level of task difficulty can influence the rating given to video quality: the same quality received a lower rating when the task being performed was difficult. Thus, it can be concluded that users may not always be able to accurately determine/judge the quality they need to complete a particular task when contextual variables are operating.

Moreover, Knoche et al. [13] conclude that subjective methods are fundamentally flawed, as it is not possible for people to register what they do not consciously perceive. Consequently, they recommend that task performance should be the measure by which media quality is judged.

Task performance is an essential element of usability, yet to rely on it solely would be unwise. Subjective methods capture the degree of user satisfaction with quality, which is important - but are not necessarily a reliable indicator of the impact that quality has on the user. Therefore, we argue that both task performance and user satisfaction should be used in conjunction with a measure of *user cost*, as part of a 3-tier approach. User cost is an explicit - if often neglected - element of the traditional Human Computer Interaction (HCI) evaluation framework.

2.2. User cost

User cost can be measured through subjective methods (rating scales assessing comfort, fatigue, etc.). Yet given the drawbacks of subjective assessment, we decided to look at objective methods of assessing the impact of media quality on users. One way of doing this is to measure the physiological levels of stress or discomfort experienced by users at different levels of quality.

When users are presented with insufficient audio and video quality in a task context, they must expend extra effort on decoding information at the perceptual level. If they struggle to decode the information, this should induce a response of discomfort or stress, even if they remain capable of performing their main task. Autonomous physiological responses are not subject to cognitive mediation, and collecting such measurements need not interfere with task completion.

2.3. Physiological measurements

The nervous system of humans is separated into the central nervous system (CNS) and the peripheral nervous system (PNS). The PNS comprises the somatic nervous system (SNS) and the autonomic nervous system (ANS). The ANS is divided into the sympathetic and the parasympathetic divisions.

The sympathetic division activates the body's energetic responses. When faced with a stressful situation, the ANS immediately mobilises itself without the need for conscious instruction. This is referred to as the 'fight or flight' response [4]. The sympathetic division prepares the body for action by e.g. speeding up the heart rate, dilating the walls of the blood vessels to

[1] Quality Assessment Slider

speed up blood flow to the limbs, and releasing glucose into the bloodstream for energy. Once the stressful situation has passed, the parasympathetic division takes over to restore the body to its equilibrium.

We decided to take measures of Heart rate (HR), Galvanic Skin Response (GSR) and Blood Volume Pulse (BVP), for the purposes of this research. These signals are unobtrusive, in that they do not require blood samples to be taken to measure stress hormones, and are easy to measure with specialised equipment.

2.4. Physiological responses to stress

Heart rate is a valuable indicator of overall activity level, with a high heart rate being associated with an anxious state and vice versa [7]. Seyle [19] has linked GSR to stress and ANS arousal. GSR is also known to be the fastest and most robust measure of stress [3], with an increase in GSR being associated with stress. BVP is an indicator of blood flow. The BVP waveform exhibits the characteristic periodicity of the heart beating - each beat of the heart forces blood through the vessels. The overall envelope of the waveform pinches when a person is startled, fearful or anxious, thus a decrease in BVP amplitude is indicative of a person under stress, and vice versa.

Under stress, HR rises in order to increase blood flow to the working muscles, thus preparing the body for the 'fight or flight' response [4]. GSR increases under stress: the precise reason this happens is not known. One theory is that it toughens the skin, thus protecting it against mechanical injury [26] as it has been observed that skin is difficult to cut under profuse sweating [5]. A second theory is that GSR increases to cool the body in preparation for the projected activity of 'fight or flight'. BVP decreases under stress. The function of this is to divert blood to the working muscles in order to prepare them for action. This means that blood flow is reduced to the extremities, like a finger.

2.5. How are these responses measured?

A ProComp unit, manufactured by Thought Technology Ltd. [20], is used in this research to measure physiological signals. In measuring GSR, two silver-chloride electrodes are placed on adjacent fingers and an imperceptible small voltage is applied. The skin's capacity to conduct the current is measured.

Photoplethysmography is used to measure HR and BVP. This involves a sensor being attached to a finger and a light source is applied: the light reflected by the skin is measured. At each contraction of the heart, blood is forced through the peripheral vessels, which produces an engorgement of the vessel under the light source. Thus, the volume and rate at which blood is pumped through the body are detected.

2.6. Research problems

Measuring physiological signals in response to media quality can be problematic. One of the main issues is how to separate stress and other emotions, such as excitement about the situation or task, in an experiment. This is a problem as the physiological patterns accompanying each emotion are not clearly understood [3], however recent research at the Massachusetts Institute of Technology Media Laboratory has shown that eight emotions can be distinguished between with eighty-percent accuracy [21], which is an encouraging result. We are using the following methods to address this problem in our experiments, by attempting to ensure that there is no stress placed on participants by factors other than the quality:

- In our lab-based trials, we hold the environment as constant and minimally stressful as possible. An example of this is that we make sure that environmental events, like the phone ringing, do not occur: we need to determine the effects the quality has in isolation before we can account for environmental events in the field.
- We measure the baseline responses of participants for fifteen minutes, prior to any experimentation occurring. This allows participants and the sensors time to settle down and gives us a set of control physiological responses.
- We administer subjective assessments of user cost, i.e. scales of discomfort, to allow people to comment on how they feel during experiments. Physiological measurements identify problems, but do not aid problem resolution when used in isolation.
- Finally, we carefully design the tasks used in our experiments to ensure that they are engaging, yet minimally stressful. The tasks used in our experiments are taken from the taxonomy of tasks performed in networked multimedia environments developed by the ETNA project [6] (section 7.2).

2.7 Video frame rate study

A study conducted as part of this research [28] investigated the subjective and physiological responses to 5 frames per second (fps) and 25fps of twenty-four participants, when they had to perform an engaging task. Results showed that participants had an increase in stress responses at 5fps as opposed to 25fps, yet they did not subjectively notice that the frame rate had changed. Thus, a discrepancy between subjective and physiological responsess was highlighted. Having looked at a parameter of video, it was then decided to investigate the impact of audio degradation on users.

3. Internet Audio

It is well established that good audio quality is important in MMC [11, 18], and much effort has been expended to protect audio from network degradations [e.g. 9]. The network research community has assumed that increasing the amount of bandwidth - and thus reducing the amount of packet loss – would ensure sufficient audio quality. Yet, in a large-scale field trial where sufficient bandwidth was available[2], users still reported audio problems in 1 out of 3 sessions [25]. Subjective assessment of user opinion and objective details about the network behavior were gathered throughout the project.

The most commonly reported problems in this field trial were attributed to packet loss, differences in volume between participants, echo and poor headset quality. Interestingly, the network statistics from the trials showed that audio packet loss was rare, and was mainly in the region of 5%, with occasional short bursts of 20%. Therefore, we decided to conduct an experiment to determine the subjective and physiological responses to a number of audio degradations caused by the network, end-user behavior and equipment problems.

4. Experiment

This experiment investigated audio in isolation, as we wanted to investigate the effects of its degradations, without the video channel causing a distraction.

4.1. Material

The material used was a dialogue between two male speakers, which had been taken from previous project meetings conducted via MMC. The material was recorded, then played back to the participants so that all participants heard exactly the same degradations. Additionally, listening passively is less stressful than being actively involved in a real-time task.

The material was recorded using a 16 bit linear codec and silence suppression. Degradations were then induced onto the stream and the recordings were split into two-minute files.

The conditions were:

1. **5% audio packet loss** on both speakers.
2. **20% audio packet loss** on both speakers.
3. Audio recorded by one speaker with a **bad microphone.**
4. Audio recorded by one speaker that was **quiet.**
5. Audio recorded by one speaker that was **loud**.
6. One speaker used an open microphone and speakers, as opposed to a headset, which meant that the other speaker generated **echo.**

[2] The PIPVIC-2 (Piloting IP-based VideoConferencing) project involved 13 UK institutions in educational activities [16]

We accept that the judgement of the non-network factors – such as whether a microphone is "bad" or not - is subjective. However, since it had been reported as a problem in the PIPVIC-2 trial, it was important to investigate further to determine the physiological and subjective influence it had on users. In addition, the samples were checked independently by three Internet audio experts, who found them representative of the distortions we wanted to mimic, whilst remaining intelligible. A pilot trial with six participants also showed that the subjective responses to all the samples were as expected.

4.2. Procedure

Twenty-four novice Internet audio users participated in the study. They wore a Canford DMH12OU headset and were played a one-minute volume test file first. They then listened to the experimental conditions, which were six two-minute files. Each file was played twice in order to determine the consistency of participants' subjective ratings. The order of the files was randomised, with a reference condition always being played first and eighth: the six conditions were heard once all the way through before being repeated. All the files were played through a Sun Ultra workstation.

After each condition, participants had to rate the quality they heard on a 100-point scale. They also had to explain why they gave the rating. Physiological measurements were taken to all conditions, with fifteen minutes of baseline measurements being taken prior to the experiment commencing. The following hypotheses were posited:

1. There will be different physiological responses to the conditions.
2. These will not always correlate with subjective responses.

5. Results

5.1. Physiological results

The mean physiological results of each participant to each condition were combined and are shown in Figures 3, 4 and 5.

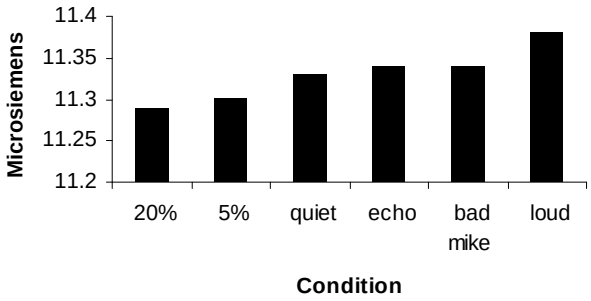

Figure 3. Mean GSR of all participants

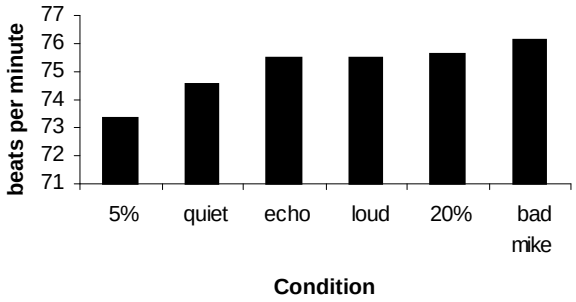

Figure 4. Mean HR of all participants

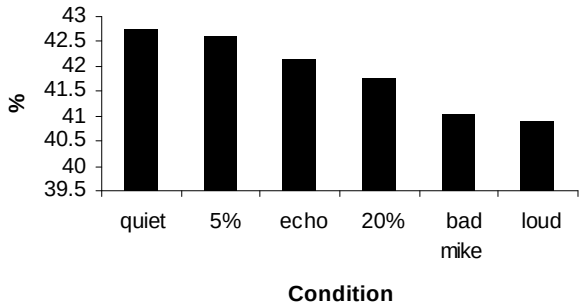

Figure 5. Mean BVP of all participants

A Multivariate Analysis of Variance (MANOVA) was performed on the data with the independent variable audio degradation. There was a significant effect of condition on HR and BVP signals, but not on GSR: HR ($F_{(5,115)}$=4.106, p=.002), BVP ($F_{(5,115)}$=3.316, p=.008). Pairwise comparisons revealed where the differences were:

- **Bad microphone** was significantly more stressful than **quiet** and **5% loss** in both HR and BVP at the .05 level.
- **Loud** was significantly more stressful than **quiet** and **5% loss** in both HR and BVP at the .05 level.
- **20% loss** was significantly more stressful than **5% loss** and **quiet** in both HR and BVP at the .05 level.
- **Echo** was significantly more stressful than **quiet** in the HR signal only at the .05 level.
- There were no significant differences in the GSR signal (see section 6).

5.2. Subjective Results

A two-factor with replication ANOVA at the 1% probablility level showed that there was a significant effect of condition (F 6, 322 = 62.25, p<0.01), and that there was no significant difference between the 1st and 2nd presentation ratings (F 1, 322 = 0.799).

Analysis of the mean subjective results (Figure 6) showed that there was no significant difference between the **5% loss** and **quiet** conditions (Qobt = 2.39), but that the **5% loss** condition was rated significantly higher than **echo** (Qobt = 9), **loud** (Qobt = 12.41) and **20% loss** (Qobt = 13.43) at the 1% probability level and at the 5% level for **bad microphone** (Qobt = 4.17, Qobt = 4.33). In addition, there was no significant difference between the **20% loss** condition and the **echo** and **loud** conditions at the 1% level (Qobt = 4.43 and 1.02 respectively), despite **20% loss** being subjectively rated the lowest. Table 1 shows the differences between the conditions subjectively and physiologically.

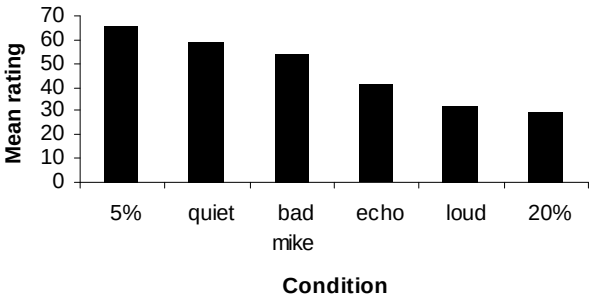

Figure 6. Mean subjective rating for each condition

Cond. 1	Cond. 2	Significant subjective difference?	Significant physiological difference?	Concur?
5% loss	Quiet	No	No	Yes
5% loss	Bad mike	Yes	Yes	Yes
5% loss	Echo	Yes	No	No
5% loss	Loud	Yes	Yes	Yes
5% loss	20% loss	Yes	Yes	Yes
Quiet	Bad mike	No	Yes	No
Quiet	Echo	No	Yes	No
Quiet	Loud	No	Yes	No
Quiet	20	No	Yes	No
Bad mike	Echo	No	No	Yes
Bad mike	Loud	No	No	Yes
Bad mike	20	No	No	Yes
Echo	Loud	No	No	Yes
Echo	20% loss	No	No	Yes
Loud	20% loss	No	No	Yes

Table 1. Showing the subjective and physiological differences between two conditions, and whether the two assessment methods concurred.

6. Discussion of Results

The first and most important point to make is that a **bad microphone** is the first and second most stressful condition physiologically, yet subjectively it is not rated as being poor (3rd best out of 6 conditions). Secondly, although subjectively the **20% loss** condition is rated as the worst, physiologically this is not the case.

We had expected that the **bad microphone** condition would be subjectively rated poorer than it was, from the results of the PIPVIC-2 trial. The simple reason for the discrepancy between the **bad microphone** condition subjectively and physiologically could be due to the task.

The listening task was short in duration (2 minutes per condition) so it may be that this does not allow the full impact of a **bad microphone** to manifest itself upon the user. Additionally, the task was passive, thus the effects of a **bad microphone** may not affect people as much when they do not have to interact with others. However, there may be more interesting attribution effects occurring, which we will only hypothesize about until a full trial investigating the effects of a **bad microphone** is performed (section 7.1). Examination of the remarks made by participants may help with this (Table 2).

Bad microphone	Loud	20% loss
• distant • far away • muffled • on telephone • walkie talkie • in a box	• annoying • hear breathing	• robotic • cuts out • digital • electronic • metallic • broken up

Table 2. Common descriptions of three conditions

It may be that **20% loss** is less stressful than a **bad microphone,** because users do not have to strain themselves in order to determine what is being said. The effect of a **bad microphone** (being 'muffled' and 'in a box') may be more irritating for the user. On the other hand, the **20% loss** condition may be consistently bad, whereas the **bad microphone** condition may be more ' bursty' and thus more stressful.

In addition to the **bad microphone** and **20% loss** conditions, **loud** joins them as the 'top three' most physiologically stressful degradations. **Loud** audio is physically uncomfortable to listen to, much more so than audio that is **quiet**. **5% loss** is not viewed as a problem either physiologically or subjectively. **Echo** was rated poorly subjectively, yet physiologically it was only significantly more stressful than **quiet** in the HR signal. Subjectively participants found it annoying, yet physiologically this difference did not emerge to such an extent.

Interestingly, out of the three worst degradations, subjectively and physiologically, only one is caused by the network: **20% loss**. Network providers and application designers should take note that, even in a well-provisioned network with no packet loss, sub-optimal hardware, setup and end-user behavior could still adversely affect users' experience with the technology.

To minimise the occurrence of these problems, Watson & Sasse [24] recommend that firstly, audio tools incorporate a fault diagnosis option. This is where users would search though a list of terms that describes their problem in terms most commonly generated by users (e.g. fuzzy), and a list of potential actions to remedy this would be offered. Secondly, designers could offer an expert system style diagnosis on a speech stream to identify likely problems.

This results from this experiment provide support for the three tier approach to multimedia quality assessment, as presented in section 2.1. If solely subjective assessment had been used in this experiment, the importance of a **bad microphone** would have been missed at the expense of treating **echo**. Conscious rating and autonomic responses work in different ways, especially when the task being performed is engaging [28]. However, this passive listening task was not engaging, thus the differences between subjective and physiological responses are highlighted even more.

The finding that GSR did not produce any statistically significant results needs to be noted. The direction of the means corresponds to those of HR and BVP, with the exception that **20% loss** is the least stressful. However, the difference is tiny: 00.10 microsiemens. It is known in the psychophysiology community, that autonomic signals do not correlate with each other all the time [14]: this could be a plausible explanation. Alternatively, we could suggest that audio degradations do not affect GSR and that there are different types of discomfort to media quality degradations. Only further research will determine if there are different physiological responses to multimedia degradations.

7. Conclusions

Three main conclusions can be made from this study. Firstly, physiological responses to audio degradations can be detected. Secondly, subjective assessment does not always correlate with physiological responses. Therefore we recommend that the three-tier approach be adopted in order to give a rounded indication of how the user is affected by the quality. Finally, we propose that the neglected element of user cost be given more consideration in usability evaluation of any technology.

7.1. Future Studies

Two experiments are being conducted at present. The first is looking at four audio degradations in a full multimedia conference. It uses the main findings from this experiment, as we want to determine if similar results to this experiment are found when the task is a) longer in duration (samples are ten minutes each), b) engaging and c) incorporates the video channel. Twenty-four participants will watch four recorded interviews of school pupils applying for a degree place at UCL. Their task is to rate the quality of the conference and to determine the suitability of the candidates to the course. The conditions are:

- **Loud** audio, as it was both physiologically and subjectively poor.
- Audio recorded using a **bad microphone,** as it was physiologically, but not subjectively poor.
- **20% audio packet loss**, as it was more subjectively than physiologically poor.
- **5% audio packet loss**, as it was both subjectively and physiologically good.

The second experiment is examining audio and video degradations in an interactive task. This study is being carried out as part of the ETNA project (section 7.2). It will involve eleven admissions tutors at UCL interviewing four candidates in Glasgow over the network and in real-time. Video frame rate and audio packet loss are being varied in the same condition, so that each interview will have either high or low video frame rate along with high or low levels of audio packet loss. This study is a step forward for this research as the task being performed is active, as opposed to passive. Thus, the results will determine the efficacy of utilising physiological measurements in field trials. In addition, both audio and video are being manipulated in the same condition: this will allow us to determine the interactive effects of one upon the other.

The final experiment we hope to conduct will examine the effects of a **bad microphone**. In this experiment the **bad microphone** condition was subjective, however due to the fact that it clearly does impact upon people physiologically, we want to examine it in more detail e.g. considering the signal to noise ratio.

7.2 Contributions

Our continuing work in this area aims to produce two substantive contributions. Firstly, the minimum levels of multimedia quality at which users can successfully perform their tasks, without significant user cost, will be determined. The impact of problems caused by the network will be investigated, such as delay and jitter. However, quality is not uni-dimensional and encompasses more than variables affected by the network. Thus, the effects of other contributing factors must be examined, e.g. image size, and problems due to the hardware set-up. This will allow network providers to allocate resources with the end users' requirements clearly specified, which will ultimately improve applications for the end user.

These findings will be incorporated into the ETNA Project [6], 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.

Secondly, we are working on providing feedback to the user in an application. For example, a user could be involved in a multimedia conference and would have their physiological responses displayed in the format of an animated face in the corner of the screen. If the user were under stress, the face would become sad and if the user were calm, the face would become happy. Such basic feedback would give an increased awareness and control back to the user of effects they are not usually conscious of.

A methodological contribution will also be made - guidelines for further research in this area will be produced. For example, it may become apparent that some signals respond better to specific degradations than others - in our studies GSR responded strongly to video frame rate [28], yet did not respond significantly to audio degradations. This will aid further research in this area which at present is sparse, yet this may be about to change with other institutions adopting this technique.

This research is also providing a general contribution to HCI methodology, by promoting the measurement of user cost. This has largely been neglected in the area of HCI, 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 assessment: it can also be used in areas such as product assessment.

8. Acknowledgments

Our grateful thanks go to our co-experimenter, Anna Watson, of the Computer Science Department at UCL. Gillian Wilson is funded by a joint EPSRC CASE studentship with BT Labs.

9. References

1. Bouch, A. & Sasse, M. A. (1999), "Network Quality of Service: What do Users Need?", Proceedings of the *4th International Distributed Conference*, 22nd – 23rd September 1999, Madrid.
2. Bouch, A., Watson, A. & Sasse, M. A. (1998), "QUASS – A Tool for Measuring the Subjective Quality of Real-time Multimedia Audio and Video", in J. May, J. Siddiqi & J. Wilkinson (eds.), *HCI '98 Conference Companion*, pp.94-95, 1st – 4th September 1998, Sheffield, UK.
3. Cacioppo, J. T & Louis, G. T. (1990), "Inferring Psychological Significance from Physiological Signals", *American Psychologist* **45**(1), 16-28.
4. Cannon, W.B. (1932), "*The Wisdom of the Body*", (Reprinted 1963.) New York: WW Norton.
5. Edelberg, R. & Wright, D. J. (1962), "Two GSR Effector Organs and their Stimulus Specificity", Paper Read at the *Society for Psychophysiological Research*, Denver, 1962.
6. ETNA Project web site

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

7. Frijda, N. H. (1986), "The Emotions, chapter Physiology of Emotion", *Studies in Emotion and Social Interaction*, Cambridge University Press, Cambridge, pp124-175.
8. Gilli Manzanaro, J., Janez Escalada, L., Hernandez Lioreda, M. & Szymanski, M. (1991), "Subjective Image Quality Assessment and Prediction in Digital Videocommunications", *COST 212 HUFIS Report*.
9. Hardman, V., Sasse, M. A., Handley, M. & Watson, A. (1995) "Reliable Audio for Use over the Internet", *Proceedings of INET' 95*, International Networking Conference, 27-30 June 1995, Honolulu, Hawaii, pp. 171-178, Reston, VA:ISOC.
10. ITU-R BT.500-8 "Methodology for the Subjective Assessment of the Quality of Television Pictures": http://www.itu.int/publications/itu-t/iturec.htm
11. Kawalek, J. (1995), "A User Perspective for QoS Management", *Proceedings of 3rd International Conference on Intelligence in Broadband Services and Network*, IS & N 1995, Crete, Greece.
12. Kitawaki, N. & Nagabuchi, H. (1998), "Quality Assessment of Speech Coding and Speech Synthesis Systems", *IEEE Communications Magazine*, October 1998, pp.36-44.
13. Knoche, H., De Meer, H. G. & Kirsh, D. (1999), "Utility Curves: Mean Opinion Scores Considered Biased", *Proceedings of 7th International Workshop on Quality of Service*, 1st - 4th June 1999, University College London, London, UK.
14. Lacey, J. I. & Lacey, B. C. (1958), "Verification and Extension of the Principle of Autonomic Responses Stereotypy", *American Journal of Psychology*, **71**, 50-73.
15. Picard, R. W. & Healey, J. (1997), "Affective Wearables", *Personal Technologies*, **1**(4), 231-240.
16. PIPVIC 2 web site:

http://wwwmice.cs.ucl.ac.uk/multimedia/projectS/pipvic2/

17. RAT (Robust Audio Tool). Available from http://www-mice.cs.ucl.ac.uk/multimedia/software
18. Sasse, M. A., Biltung, U., Schulz, C-D & Turletti, T. (1994), "Remote Seminars through Multimedia Conferencings: Experiences from the MICE Project", *Proceedings of International Networking Conference*, pp251/1-8, 13th-17th June 1994, Prague, Czech Republic.
19. Seyle, H. (1956), *"The Stress of Life"*, McGraw-Hill.
20. Thought Technology http://www.thoughttechnology.com/
21. Vyzas, E. & Picard, R. W. (1999), "Offline and Online Recognition of Emotion Expression from Physiological Data", *Workshop on Emotion-Based Agent Architectures, Third International Conference on Autonomous Agents*, 1st May 1999, Seattle, WA.
22. Watson, A. & 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, 15th - 16th September 1997, Aberdeen, Scotland, UK.
23. Watson, A. & Sasse, M. A. (1998), "Measuring Perceived Quality of Speech and Video in Multimedia Conferencing Applications", *Proceedings of ACM Multimedia '98*, ACM New York, pp.55-60, Bristol, UK.
24. Watson, A. & Sasse, M. A. (2000), "The Good, the Bad and the Muffled: the Impact of Different Degradations on Internet Speech", *To be presented at ACM Multimedia 2000*, 30th October - 4th November 2000, Los Angeles, California..
25. Watson, A. & Sasse, M. A. (2000), "Distance Education via IP Videoconferencing: Results from a National Pilot Project", *CHI 2000 Extended Abstracts*, pp.113-114, ACM Press, 1st - 6th April 2000, The Hague, The Netherlands.
26. Wilcott, R. C. (1967), "Arousal Sweating and Electrodermal Phenomena", *Psychological Bulletin* **67**, 58-72.
27. Wilson, F. & Descamps, P. T. (1996), "Should We Accept Anything Less than TV Quality: Visual Communication", Paper presented at *International Broadcasting Convention*, 12th – 16th September 1996, Amsterdam.
28. Wilson, G. M. & Sasse, M. A. (2000), "Do Users Always Know What's Good for Them? Utilising Physiological Responses to Assess Media Quality", in S. McDonald, Y. Waern & G. Cockton (eds.), *Proceedings of HCI 2000: People and Computer XIV - Usability or Else!*, pp. 327-339, 5th-8th September 2000, Sunderland, UK.