The Good, the Bad, and the Muffled: the Impact of Different Degradations on Internet Speech

Anna Watson
Department of Computer Science
University College London
Gower Street
London, WC1E 6BT
+44 (0)20 7679 3643
A.Watson@cs.ucl.ac.uk

M. Angela Sasse
Department of Computer Science
University College London
Gower Street
London, WC1E 6BT
+44 (0)20 7679 7212
A.Sasse@cs.ucl.ac.uk

ABSTRACT

This paper presents an experiment comparing the relative impact of different types of degradation on subjective quality ratings of interactive speech transmitted over packet-switched networks. The experiment was inspired by observations made during a large-scale, long-term field trial of multicast conferencing. We observed that user reports of unsatisfactory speech quality were rarely due to network effects such as packet loss and jitter. A subsequent analysis of conference recordings found that in most cases, the impairment was caused by end-system hardware, equipment setup or user behavior. The results from the experiment confirm that the effects of volume differences, echo and bad microphones are rated worse than the level of packet loss most users are likely to experience on the Internet today, provided that a simple repair mechanism is used. Consequently, anyone designing or deploying network speech applications and services ought to consider the addition of diagnostics and tutorials to ensure acceptable speech quality.

Keywords

Internet audio, speech, media quality assessment, subjective assessment, multicast conferencing.

1. INTRODUCTION

Over the past 5 years, there has been increasing interest and growth in the use of multicast technology over the Internet in areas such as distance education and remote project meetings. It is well established that good-quality audio is a necessary condition for usable multimedia conferencing [e.g. 8, 16], and a great deal of research effort in the telecommunications arena has been directed at combating the effects of packet loss, jitter and delay [e.g. 2, 3, 9]. To date, there has been an implicit assumption in the networking community that many of these issues will be resolved through increased bandwidth [6, 24]. If this is true, given the level of provision in the US and western Europe today, the quality of speech users experience in Internet conferences should be good. Yet, in a recent large-scale field trial, users reported speech quality problems in one out of three multimedia conferencing activities, where sufficient bandwidth was available.

The PIPVIC-2 (Piloting IP-based VideoConferencing) project [12] involved 13 UK academic institutions and 150 participants in a range of educational activities running from December 1998 to October 1999. The project gathered both subjective (user opinion) and objective (network behavior) performance data, and developed methods of matching these two types of data more closely.

Subjective data was gathered through paper-based questionnaires at the end of each particular course; group workshops with the tutors and students; and through web-based opinion scales completed at set points during a conference. This latter method was used since it is notoriously difficult to gather reliable subjective opinions of the quality delivered in lengthy multimedia conferences. Waiting until the end of a conference leaves the user open to primacy and recency memory effects, whilst taking continuous readings throughout a conference leads to task interference [18]. Collecting subjective quality data at certain set points during the conference seemed a reasonable compromise. In an hour-long conference, these data were collected approximately every 20 minutes – after the sound and volume check at the start of the conference, midway through the conference, and at the end of the conference.

Objective data was collected through modification of the audio and video conferencing tools such that they logged the reception reports received from the other participants in the conference. These statistics could then be matched in time with the web-based opinion ratings.

The remainder of this paper describes an experiment informed by the main finding from the PIPVIC-2 project: that many reported speech quality problems are due not to network conditions, but rather to end-user behavior and equipment problems. We investigated both the subjective ratings and the physiological responses (blood volume pulse and heart rate) of listeners to samples of Internet speech degradations. Results confirm that the impact of volume discrepancies and voice feedback affect perceived quality more adversely than the levels of packet loss typically experienced in the project. Initial physiological results indicate that poor-quality microphones and too high volume levels are particularly stressful to users.

2. BACKGROUND TO EXPERIMENT

2.1 Rationale for conditions chosen

The user assessment of the conferencing sessions in the PIPVIC-2 trials showed that three factors were most often reported as problematic: missing words or incomplete sentences; variation in volume between participants; and variation in quality between participants. These problems, and their likely causes, are summarized in Table 1.

Table 1 Key audio problems reported by users in the PIPVIC-2 field trials

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

Although 'missing words' were frequently cited as a problem, the outcome from the project's network monitoring activities showed that, in general, the level of packet loss on both the SuperJANET[1] multicast service, and participants' local area networks, was low during the trial. The project collected RTP (Real-time Transport Protocol) reception report statistics from the participants in various conferences. Reception reports are generated once every 2-5 seconds, and can be used to produce an overall picture of the level of loss experienced at different end sites in a conference. One such picture is shown in Figure 1, where reception reports reflecting the level of loss received by a participant in Glasgow from a participant in London are shown.

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

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

2.1.1 Codec, packet size and packet loss repair method

In the PIPVIC-2 project the Robust Audio Tool, RAT [13], was used. RAT is one of the two most commonly used software audio tools in the multicast community (the other being LBL's Visual-Audio Tool, VAT [5]). RAT offers greater functionality than VAT, in particular packet loss repair mechanisms.

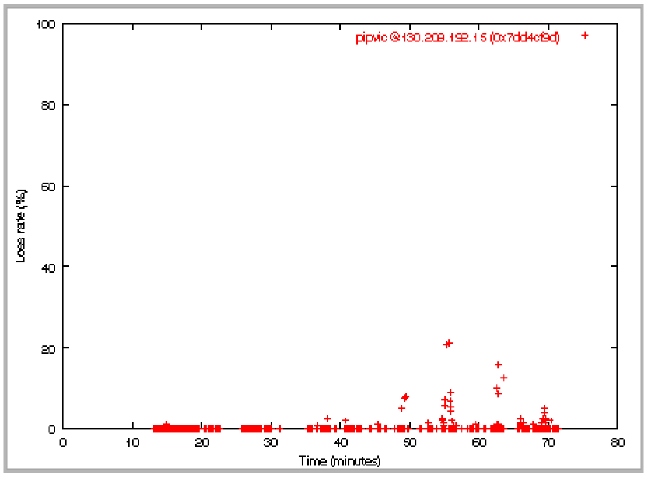

Figure 1 Loss reported from UCL by a Glasgow conference participant over a one-hour project meeting

In selecting the codec and packet size to investigate in the experiment, we decided to use the RAT version 3 default settings that had been available to users in the field observations, even though in recent months these defaults have been changed and are now likely to produce better perception[2]. Therefore the experimental speech material that was generated was coded in DVI [14], using 40 ms packets, and repaired with the receiver-based packet loss repair method packet repetition.

Packet repetition (also known as waveform substitution) fills in the space from a missing packet by repeating the last received packet. This technique works best when the packet size is small (20 ms): when the packet is large (80 ms), the speech signal is likely to have changed significantly within the missing packet, meaning that the repaired speech can sound faintly synthetic or metallic where the repaired waveform does not connect smoothly.

2.1.2 Packet loss rates

As a result of the PIPVIC-2 findings reported above, for the experiment we selected 5% as a lower level of packet loss, which is representative of the level of packet loss users are likely to experience on the SuperJANET multicast service today. The figure of 20% was chosen as a higher level for the experiment presented here because it is known from previous research that this is the level at which perceived quality of repaired speech starts to drop significantly, but where speech intelligibility is maintained [18, 19][3].. Since this level of packet loss is known to cause severe degradation, it would act as a reference point in the planned study.

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

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

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

2.1.3 *'Bad' microphone*

A poor-quality microphone was chosen as a condition because during the field trials users had reported and complained about 'tinny' or 'hummy' microphones. The selection of a 'bad' microphone is, of course, somewhat subjective. In addition, a microphone that produces 'bad' audio when used with one soundcard will not necessarily be a 'bad' microphone for another, making the matter more complicated. However, the effect of microphone distortion was still felt to be worthwhile investigating, since so many subjective comments refer to how the voice sounds, and whether it is pleasant to listen to [11]. The microphone chosen for the experiment was an Altai A087F.

2.1.4 *Volume differences*

Many users in the field trials complained of extreme volume differences between participants in multi-way conferences. Although it is possible to alter the incoming volume from a particular participant by adjusting the incoming volume slider in RAT, users tend not to adjust the slider when the next speaker is louder or softer, since it becomes tedious and interferes with the ongoing purpose of the conference. We decided to investigate the subjective effects of one speaker at 'normal' volume, and the other at 'too loud', and also the impact of 'normal' and 'too quiet'. Again, it is recognized that determining what is 'too loud' or 'too quiet' is a subjective decision to be taken by the experimenter, but by piloting the experiment with both network audio experts and novices we were able to determine levels that were commonly agreed to be 'too loud' or 'too quiet'.

2.1.5 *Echo*

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

2.2 Measurement methods

The most common listening quality rating scale in use is the ITU-recommended 5-point listening quality scale (resulting in a Mean Opinion Score, or MOS) [4]. This scale has come under criticism from an increasing number of researchers in recent years [7, 10, 18] for a number of reasons, not least of which is the fact that the labels on the scale (*Excellent, Good, Fair, Poor, Bad*) are not appropriate for the level and type of degradation experienced in speech over the Internet, since the quality encountered will rarely be described as *excellent*. The other key reasons are:

1. Although treated as such, the scale is not an interval scale as represented by its 5 qualitative labels (*Excellent, Good, Fair, Poor, Bad*) [7, 17, 23]. *Fair*, for example, is not indicative of a midpoint to most people.
2. Use of the 5-point scale leaves the experimenter ignorant of the subject's perspective and rationale for positioning on the scale [10].
3. Quality is a *multi*dimensional phenomenon [11, 23], and means are required by which the dimensions that have the largest effects can be identified.

On-going research at UCL has been investigating a number of novel methods [18, 19, 1, 21] for measuring received quality in conferences over the Internet, and in particular has developed what we believe to be a more suitable rating scale for the subjective assessment of Internet media. This method will be discussed in the following section.

2.2.1 *Subjective measurement*

The unlabelled scale that was described in [18, 19] has evolved into a scale where the end-points are bounded by 1 and 100 (see Figure 2). (This development was necessary when gathering data in the PIPVIC-2 project via web-based evaluation forms.)

We fully subscribe to the point made by [11], that speech quality should not be treated as a unidimensional phenomenon, since one or many different dimensions may affect the listener's opinion. This is why there are no descriptive labels other than at the end points on the 100-point scale. Instead we ask subjects to describe how the sample sounded, and why a certain rating was awarded. This allows us to gain a deeper insight into factors that affect perceived quality, with a long-term view to producing a series of diagnostic scales along different quality dimensions. The background to this research lies in the observation that the vocabulary that is used by Internet audio experts is rarely matched by novice users when describing how Internet communication sounds to them [18].

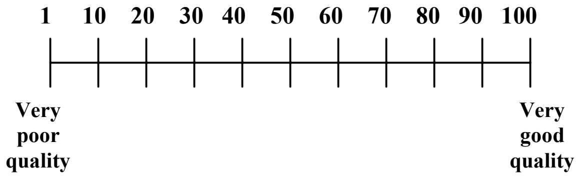

Figure 2 The 100-point quality rating scale

2.2.2 *Objective (physiological) measurement*

A traditional Human-Computer Interaction (HCI) approach should take into account *user cost* as well as user satisfaction [21]. User cost addresses the level of fatigue, discomfort, physical strain etc. that people experience in performing a certain task. We have recently begun to investigate the effect of different media quality on user stress, as measured by Blood Volume Pulse (BVP) and Heart Rate (HR). Using a ProComp[4] unit, we place sensors on the fingers of subjects, making it possible to monitor their physiological responses to different types of impairment. This enables us to assess the relationship between expressed opinion and user cost.

BVP is an indicator of blood flow which decreases when a person is under stress. Under stress, HR increases in order to increase blood flow to the muscles (the 'fight or flight' reflex). Therefore the physiological indicators of increased user cost would be a decrease in BVP and an increase in HR, compared to the levels recorded in a resting state.

[4] Manufactured by Thought Technology,

http://www.thoughttechnology.com/

3. THE EXPERIMENT

3.1 Experimental material

A two-person conversation was scripted from recordings of multicast project meetings, with names and locations changed from the original recordings. This script was acted out by two male actors without regional accents. The actors sat at Sun Ultra workstations at different locations on the same local network for the duration of the recording. (Only audio was recorded: video did not play a part in this study.) The recording was made at 16 bit linear quality and recorded via the record facility in RAT. Silence suppression was left on and both microphones were kept open during the recording. The actors wore identical Canford DMH12OU headsets. Different parts of the conversation were subject to manipulation by the experimenter such as the volume and feedback of one of the speakers, and the headset in use. The resulting recordings were then split into 2-minute files and coded into DVI, at 8kHz sampling rate, and 40 ms packets. Packet loss and repair (packet repetition) were generated on the files where required, using the software program **test_repair**[5].

The conditions that were generated were:

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

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

3.2 Subjects

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

3.3 Procedure

The subjects each listened to the seven 2-minute test files twice (to determine the consistency of subjects' scores on the 100-point scale). The files were played out through the program Audio Tool[6] on a Sun Ultra workstation. Each subject listened to the files wearing a Canford DMH12OU headset. There was no accompanying video image. The test files were preceded by a 1-minute file which had no degradations. The function of this file was for the subjects to assess whether the volume playout level was acceptable to them, but they were also instructed that the volume test file should be taken as indicative of the best quality they would encounter in the following test files. This ensured that the subjects knew what the upper limit of quality would be. The order of the test files was randomized, with one exception: the **reference** (no degradation) condition was always heard first and eighth. The 7 conditions were therefore all heard once before they were repeated in a different order.

Baseline physiological readings were taken for each subject for 15 minutes before the listening part of the study, using the Procomp measurement device[7]. Sensors were placed on the left hand of each subject, taking measurements of blood volume pulse (BVP) and heart rate (HR).

After each test file the subject was asked to provide a quality rating, for the file as a whole, from the 1-100 scale where 1 represents *Very Poor Quality* and 100 represents *Very Good Quality* (see Figure 2). The subject was then asked to explain why that rating was awarded i.e. how the speakers had sounded to him/her. These answers were tape-recorded.

4. RESULTS

4.1.1 Quantitative results

The mean results and standard error for the perceived quality ratings are shown in Figure 3. The graph suggests that a 'normal' level of packet loss (5%) when repaired with packet repetition has little impact on perceived quality when compared to the **reference** (no degradation) condition. As expected, **20% loss** repaired with packet repetition has a profound effect on perceived quality, but it appears a loud-normal volume discrepancy, and an echo effect also affect perceived quality adversely. Are these apparent differences statistically significant?

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

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

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

[7] The Procomp, manufactured by Thought Technology, encompasses physiological measurement sensors and software.

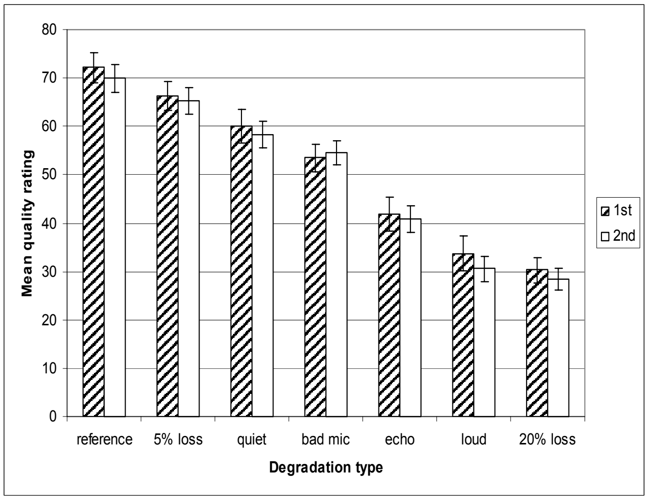

Figure 3 Mean quality rating awarded for different degradation types, on 1st and 2nd occasion of hearing

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

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

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

4.1.2 Physiological results

The mean HR and BVP readings for each condition were calculated, and are shown in Figures 4 and 5 respectively. The graphs show a different order of severity for the different conditions compared to that seen in the subjective results (Figure 3). Are these differences significant?

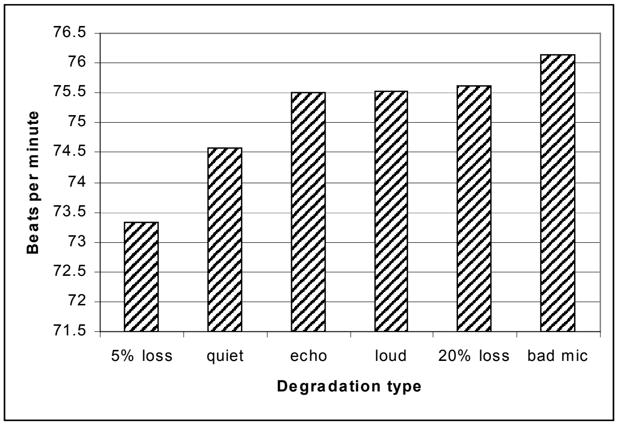

Figure 4 Mean HR of all participants

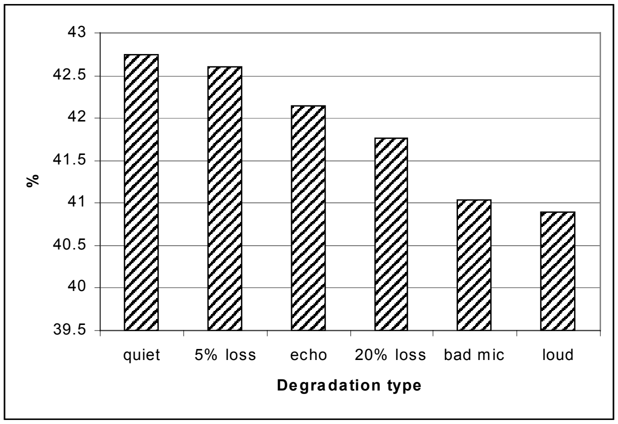

Figure 5 Mean BVP of all participants

A multivariate analysis of variance (MANOVA) revealed significant effects of condition for both HR (F 5,115 = 4.106, p < 5) and BVP (F 5,115 = 3.316, p < 0.05) signals. Pairwise comparisons revealed that, for both HR and BVP, **bad mic**, **loud** and **20% loss** were all significantly more stressful than **quiet** and

5% loss at the 5% level of probability. **Echo** was found to be significantly more stressful than quiet in the HR signal only, at the 5% level of probability.

In contrast to the subjective ratings, therefore, the **20% loss** condition did *not* produce the worst physiological ratings: the **loud** and **bad mic** conditions produced significant increases in HR and significant decreases in BVP, indicating that these two conditions were the most stressful for the subjects to listen to. The least stressful conditions appear to be the **quiet** and **5% loss** conditions.

The significance of these findings will be discussed in section 5. (For a detailed presentation and discussion of the physiological results, the reader is referred to [22].)

4.1.3 Qualitative results

In addition to providing a rating on the 100 point scale, subjects were asked to describe why they had awarded each rating. The primary aim of this part of the study was to search for common descriptive terms used by non-expert users to describe different types of degradations to aid in the building of diagnostic scales, as discussed in section 2.2.1. The descriptions also functioned as a check on the experimental conditions by enabling us to check that users had perceived and reacted to the effect intended.

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

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

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

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

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

5. DISCUSSION

The results of the experiment have shown that the typical PIPVIC-2 level of packet loss (which was generally below 5%), when repaired with a method such as packet repetition, does not affect users' subjective ratings adversely when compared to a no-loss condition, whereas non-network factors such as volume discrepancies between speakers, poor quality microphones, and echo or feedback do. It is not the case that the users do not *notice* the degradation in the **5% loss** condition (since their descriptions of the files are different from those of the **reference** condition), but rather that it has less impact on perceived quality than other types of degradation. We have demonstrated that users will rate the different conditions consistently on a 100 point rating scale. However, we have observed that, although their ratings and descriptions may be consistent, users often attribute impairments inaccurately, suggesting there is a need for a diagnostic tool to aid users in correctly identifying the source of different impairments, and then enable them to take appropriate steps to correct them.

The physiological results are intriguing in that they indicate that users are more adversely affected by the **bad mic** condition and less affected by the **20% loss** condition than the subjective rating results suggest. In a previous study looking at the impact of video frame rate on both subjective ratings and user cost, it was found that viewers did not notice (subjectively) the change in frame rate from 5 to 25 frames per second (fps), but their physiological measurements changed significantly in the direction of stress [21]. These types of findings emphasize the importance of carrying out research of this nature, combining subjective ratings with measurements of user cost - we believe that subjective results alone do not provide a wholly accurate picture.

We believe that the method outlined in this paper, using field trials to inform the design of controlled experiments, is a meaningful and practical way forward in terms of understanding the complexity of factors affecting perceived quality in multimedia conferencing across the Internet. As discussed in section 1, gathering subjective opinions of the quality delivered in multimedia conferences is not straightforward, since memory effects or task interference can occur, depending on the method used. The approach described here allows us first to identify the main effects affecting real users performing real tasks, then to perform controlled experiments to confirm and assess the relative impact of these effects. The experimental step is of great importance since users are often unable to correctly identify what is responsible for the problem in a conferencing environment, due to the complexity of many interacting factors and media. For example, it is known that audio quality can affect the rating of video quality [15]. The experiment reported here has shown that speech quality problems can be attributed to the wrong source, highlighting the importance of ascertaining as much subjective explanatory data as possible.

However, the work reported here is merely the first step in a logical research progression – there is a danger in making assumptions about quality required without careful consideration of the task being undertaken. It is very likely, for example, that both speech and video quality variables would be rated differently in an interactive experiment, and depending on the task being performed in that experiment, different factors will be important.

6. CONCLUSIONS AND RECOMMENDATIONS

The experiment has clearly demonstrated that the perceived quality of network audio is *not* primarily affected by the level packet loss we observed in the large-scale field trial (provided that a packet loss repair method such as packet repetition is in use). Volume discrepancies, poor quality microphones and echo have a greater impact on the user, meaning that it is possible to have perfect transmission from a network viewpoint, but still have poor quality audio from a user's viewpoint. The solutions envisioned mainly involve raising and improving user awareness, both of what the problem is, and how to solve it. These can be low-cost solutions – a huge amount of people-support should not be required once audio tools are better set up to support non-expert users.

By further analyzing how people describe different types of degradation, it should be possible to provide improved fault diagnosis to novice users. For example, a help menu on an audio tool should provide a list of problems described in terms that users most commonly generate, such as *'fuzzy'* and *'buzzy'* which, as we have seen, are related to a specific type of packet loss repair (packet repetition). The user could search down this list for the terms that describe his or her problem, then follow the solution suggested (e.g. change the receiver-based packet loss repair method to another, such as pattern matching).

There is perhaps less that a user can do about someone else's bad microphone, other than tell them that they sound *'muffled'*, *'distant'*, or like they're *'on the phone'*. One solution would be a pre-session diagnostic that would reflect the user's audio as heard by other participants, since at present the user cannot hear what he/she sounds like. We propose developers design a tool to perform an expert-system style diagnostic of a user's speech stream and point to likely causes of problems. After a system is initially set up, users could be required to record sample sentences – as in a voice recognition package for word processing, for example – and only be allowed onto the network once the quality of the sample files is matched or recognized as providing satisfactory quality.

The key problems highlighted in the study also provide a strong case for the inclusion of aspects such as automatic gain control and reliable echo suppression in Internet audio tools. These are already present in RAT version 4, but they are optional settings – users need guidance on when to apply them.

7. FUTURE WORK

There is a clear need to quantify the exact levels of degradation that were imposed in this study, in order to identify the levels that represent enough, too much, or too little of a certain quality variable. By establishing these levels, suitable input to designers of future tool diagnostics can be provided.

As discussed in section 1, the research community has focused on investigating the effects of objective degradations such as delay and jitter. Future work should therefore consider the relative weights of these factors against user and hardware variables, as has been done with packet loss in the present study.

An obvious further step will be to recreate the experimental conditions presented in this paper in an interactive task environment, and have people engage in active conversations as opposed to passive listening. It can be hypothesized that the effects of the factors investigated here will be altered in this setting. We can predict that the effect of echo, for example, will have an even more negative effect when a subject is trying to engage in a conversation with another person, but keeps hearing their own voice fed back to them. It will also be important to introduce a video channel into the set-up, and observe the impact of audio-visual interactions.

Another important aspect to investigate will be the effects of the interaction *between* different impairment types, for example one person with a bad microphone conversing with someone speaking too loudly. Again, presenting this scenario as an interactive experiment is likely to lead to different results.

Future work will continue to gather physiological data, to gain a better understanding of the user cost of different types of degradations, and the relationship between user cost, subjective opinion, and task performance.

8. ACKNOWLEDGMENTS

Our thanks go to Gillian Wilson at UCL, who carried out the physiological measurements. We gratefully acknowledge the technical advice and assistance of Orion Hodson and Colin Perkins, also at UCL. Partners on the UKERNA funded PIPVIC-2 project also merit recognition and thanks. The research reported here was funded under the EPSRC project ETNA (GR/M25599).

9. REFERENCES

[1] Bouch, A., Watson, A. and Sasse, M.A. QUASS: A tool for measuring the subjective quality of real-time multimedia audio and video. Poster presented at HCI '98 (Sheffield, England, September 1998).

[2] Gruber, J.G. and Strawczynski, L. Subjective effects of variable delay and speech clipping in dynamically managed voice systems. IEEE Transactions on Cummunications, 1985, 33(8), 801-808.

[3] Hardman, V., Sasse, M.A., Handley, M.J. and Watson, A. Reliable audio for use over the Internet. Proceedings of INET '95 (Honolulu, Hawaii, June 1995), 171-178.

[4] ITU-T P.800 Methods for subjective determination of transmission quality. Available from http://www.itu.int/publications/itu-t/iturec.htm

[5] Jacobson, V. vat manual pages, Lawrence Berkeley Laboratory, USA. Software available from http://www-nrg.ee.lbl.gov/vat/

[6] Jayant, N.S. High-quality coding of telephone speech and wideband audio. IEEE Communications Magazine, Jan. 1990, 10-20.

[7] Jones, B.L. and McManus, P.R. Graphic scaling of qualitative terms. SMPTE Journal, November 1986, 1166-1171.

[8] Kawalek, J. A user perspective for QoS management. Proceedings of 3rd International Conference on Intelligence in Broadband Services and Network (IS & N '95, Crete, Greece).

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

[10] Knoche, H., De Meer, H.G. and Kirsh, D. Utility curves: Mean opinion scores considered biased. Proceedings of IWQoS '99 (London, England, May 1999), 12-14.

[11] Preminger, J.E. and Van Tasell, D.J. Quantifying the relationship between speech quality and speech intelligibility. Journal of Speech and Hearing Research, 1995, 38, 714-725.

[12] PIPVIC-2 Project web site at http:// www-mice.cs.ucl.ac.uk/multimedia/projects/pipvic2/

[13] RAT (Robust Audio Tool). Available for download from http://www-mice.cs.ucl.ac.uk/multimedia/software

[14] Recommended practices for enhancing digital audio compatibility in multimedia systems (version 3.00). Technical Report, Interactive Multimedia Association, Annapolis, MD, 1992.

[15] Reeves, B. and Nass, C. The Media Equation. Cambridge University Press/CSLI Publications, 1996.

[16] Sasse, M.A., Bilting, U., Schulz, C-D. and Turletti, T. Remote seminars through multimedia conferencings: Experiences from the MICE project. Proceedings of INET'94/JENC5.

[17] Teunissen, K. The validity of CCIR quality indicators along a graphical scale. SMPTE Journal, March 1996, 144-149.

[18] Watson, A. and Sasse, M.A. Measuring perceived quality of speech and video in multimedia conferencing applications. Proceedings of ACM Multimedia '98 (Bristol, England, September 1998), ACM Press, 55-60.

[19] Watson, A. and Sasse, M.A. Multimedia conferencing via multicast: Determining the quality of service required by the end user. Proceedings of AVSPN '97 (Aberdeen, Scotland, September 1997), 189-194.

[20] Watson, A. and Sasse, M.A. Distance education via IP videoconferencing: Results from a national pilot project. Poster to be presented at CHI 2000 (The Hague, The Netherlands, April 2000).

[21] Wilson, G. & Sasse, M.A. Do users always know what's good for them? Utilising physiological responses to assess media quality. To be presented at HCI 2000, September 5th - 8th, Sunderland, UK.

[22] Wilson, G. & Sasse, M.A. Investigating the impact of audio degradations on users: Subjective vs. objective measurement methods. Submitted to OZCHI 2000. Available as UCL Computer Science research note RN/00/36.

[23] Virtanen, M.T, Gleiss, N. and Goldstein, M. One the use of evaluative category scales in telecommunications. Proceedings of Human Factors in Telecommunications, 1995, 253-260.

[24] Zhang, L., Deering, S., Estrin, D., Shenker, S. and Zappala, D. RSVP: A new resource ReSerVation Protocol, IEEE Network Magazine, 1995, 7(5), 8-18.