Quality in Context - an ecological approach to assessing QoS for mobile TV

M. Angela Sasse & Hendrik Knoche

Department of Computer Science
University College London
a.sasse@cs.ucl.ac.uk, h.knoche@cs.ucl.ac.uk

Abstract

The paper presents an overview of Quality of Service (QoS) definitions, and reviews approaches to identifying quality requirements for audio and video quality in the context of novel multimedia services. We advocate an approach that maximizes ecological validity. It is important to recruit participants from the expected user population, create realistic tasks and using real material, and use an assessment method that creates minimal load on study participants. Finally, it is important to replicate all physical factors that affect quality in the eventual context of use. We present an example of how we apply this approach to identifying video quality requirements for mobile TV viewing.

1. Introduction

With the spread of Internet connectivity and a growing number of applications and services, the research literature on Quality of Service (QoS) in communications has seen rapid growth over the past 10 years. Yet, any reader approaching the burgeoning body of literature is bound to become confused

> *"Quality of Service (QoS) is one of the most elusive, confounding, and confusing topics in data networking today. […] The trade press, hardware and software vendors, consumers, researchers, and industry pundits all seem to have their own ideas and definitions of what QoS actually is …"* [1]

Networking researchers and practitioners see QoS as a technical feature of the network, and define it through a number of technical parameters [2]:

> *"Quality of Service (QoS) refers to the capability of a network to provide better service to selected network traffic over various technologies […] The primary goal of QoS is to provide priority including dedicated bandwidth, controlled jitter and latency (required by some real-time and interactive traffic), and improved loss characteristics."*

Application developers and service providers, on the other hand, are more concerned with the QoS experienced by the end-user; after all, if the intended customers are not satisfied with the quality they experience, they are unlikely to subscribe to a service, or to continue using it. At the same time, higher quality usually comes at a higher cost.

The International Telecommunications Union (ITU) provides recommendations for assessing the likely impact of technical parameters - such as bandwidth and delay – onto users' subjective experience of quality. The aim of carrying out such assessments is *"to determine the subjective effects of some new transmission equipment or modification to the transmission characteristics"* [3], and with its recommendations, the ITU aims to provide a standard method that yields results that can be compared across a number of studies. Whilst standardization has benefits, usability researchers have pointed out that the results of these assessments may not be a good predictor as to whether a particular level of QoS is acceptable or safe – as with all usability questions, "it depends" – on who the user is, the task she is trying to accomplish, and the context of use.

Thus, we have seen an increasing number of assessments of audio and video quality carried out, and there is a somewhat uneasy co-existence in this field between engineers who adhere to the standard ITU methods and calculate a Mean Opinion Score (MOS), and usability researchers (including the authors) who maintain that such a score is largely meaningless, and that more goal- and context specific assessment is needed to predict the acceptability or safety of a service.

The aim of the paper is to briefly summarise the debate, explain how the context-based assessment approach was developed over the past 10+ years, and to demonstrate how we apply this approach in the development of one particular service: Mobile TV. In section 2 of this paper, we present a review of previous quality assessment research, and explain how our approach has been developed. The argument is essentially that peoples' perception of multimedia quality is influence by a range of factors, which vary with what a particular application and service is used for, and the context of use. To provide valid predictions of acceptability, the key factors need to be identified and replicated as part of the assessment. Section 3.1 presents the application of that approach to a particular service – mobile TV. The study we cover in Section 3.2 used acceptability to study the effects of video size and audio quality on mobile TV content. Qualitative feedback gathered in this study sparked two more studies. The influence of text quality on the acceptability of video quality is presented in Section 3.3. The effect of small video sizes on the acceptability of the video quality of the shot types used in the content is explored in Section 3.4. The paper concludes with Section 4.

2. Background

When networking researchers started to transmit audio and video over the Internet in the early 90s, the response of many in the telecommunications community was that it would never be possible to deliver satisfactory audio and video quality over a best-effort, packet-switched network. Early experiments with Internet videoconferencing identified the need to provide audio without noticeable degradations, rather than expend bandwidth and processing power to improve video quality. In the first Internet audio tool, vat, lost packets were replaced with silence. Even when the resulting speech was intelligible, we noticed that gaps in the speech made participants irritable after relatively short periods of time [4]. This led us to consider ways of masking or repairing audio with lost packets, and we conducted an experiment to test the effect of several different methods on perceived speech quality [5]. The work eventually led to the design of a new multicast audio tool, rat [6], which in its current version is still used for multicast audio today (e.g. as part of the Access Grid toolkit [7]).

In the early work aimed at producing more usable packet audio quality [5], we used the ITU-recommended approach [3] and carried out listening tests and subjective assessment, resulting in Mean Opinion Scores (MOS), which could be compared to identify the most effective repair methods. The comparisons produced some key findings on the performance of repair methods: it showed the effectiveness of cheap, receiver-based repair techniques (packet repetition) at low loss rates and for small packet sizes. Whilst intelligibility was clearly improved at higher loss rates with Linear Predicitive Coding (LPC), participants sometimes subjectively preferred packet repetition since it sounded more like the speaker's own voice. The difference between intelligibility and perceived quality was a first indication that it cannot be appropriate to rely on subjective assessment alone.

When applying the method in these trials, some of its limitations emerged. Many participants were unsure how to map their perceptions onto the 5–point scale with labels Excellent, Good, Fair, Poor, Bad. There was a clear reluctance to describe any quality coming out of the computer as Excellent, and many asked *"What exactly is the difference between Poor and Bad?"*. There were clear indications that the way in which participants were using the scale meant the data were not interval-scaled, which, in turn, means it is not valid to calculate a mean score. In [8], we summarised our and other published concerns. In addition to labeling, our key concern was about using the ITU recommendations to assess new multimedia services centered on the lack of ecological validity of tests conducted:

the short duration of test material,

the absence of a task (other than assessing the quality), and

that assessing audio and video in isolation neglected interaction effects.

The lack of task and context when asking participants to assess quality in short test sequences led other researchers to suggest that people's ability to complete typical tasks must be ascertained as well [9].

Whilst the ITU recommendations [3] promote the use of the labeled 5-point scale as the standard methods, they actually allow that for a specific assessment, another scale may be more appropriate:

"Other opinion scales that may be suitable are variants of the methods of "magnitude estimation" and "cross-modality matching". The responses on these scales may be one of the following:

a) one of a numerical series of categories labelled 1, 2, 3, 4, 5 (and denoted as such to the subject), but with descriptions attached only to the first and the last, to identify the subjective dimension;

b) a numerical mark on a scale from one to a number much greater than five – say 10 or 100; or

c) a length proportional to some property (e.g. quality), marked manually along a given straight line."

We explored alternatives b and c, and after a series of trials with various alternatives, Anna Watson [10] arrived at 100-point scale whose ends could be labeled "+" and "-", or with words relevant to the specific assessment. This scale was validated in a number of experiments, most notably [11], in which we demonstrated that the effects of volume differences, bad microphones and echo affected perceived audio quality significantly more than network degradations.

The same study also reported our first attempt to monitor actual impact of quality on users, as opposed to just recording their perceived assessment of the quality. Monitoring participants' heart rate and skin conductance, we found strong reactions (which are indicative of stress) to audio sample affected by volume differences and echo. In subsequent studies of physiological responses to audio and video quality we found that physiological responses can be detected in passive viewing, but it is difficult to obtain clear results in interactive tasks.

We subsequently carried out experiments in which users could continuously rate perceived quality using a software-based slider tool version of the continuous rating scale e.g. [12]. Again, the method worked when users were passively viewing or listening, but did not work in interactive tasks (such as videoconferencing) because the need to continuously respond to changes in quality distracted users from their main activity. However, we noticed that users would use the slider to indicate when they did find the quality annoying or unacceptable. These observations led us to consider a method which did not rely on a scale, but simple binary judgement of whether a quality is acceptable or not. This is a direct representation of when quality is not good enough, and one that is much simpler for the user to make. A detailed rationale of the method, and how it can be applied to derive utility curves that allow a service provider to see how many of their customers will likely "switch off" at what quality level, can be found in [11]. This is the method we have successfully used in a number of studies since 2003, and are continuing to use today. We still have to be aware that in the real world, users may react differently to quality levels than in the lab. In their seminal book on QoS, Zeithamel et al. [13] point out that expected quality of a service determines its perceived quality – QoS will be perceived as

positive when expectations are exceeded
satisfactory when expectations are met, and
negative when expectations are not met.

Expectations about a multimedia service may be formed on the basis of pricing of a service or the way it is marketed. Thus, service providers should be careful to create unrealistic expectations in their marketing. In many of our studies, we present quality levels with a notional cost attached. When pricing a scheme, it is important to identify what elements or characteristics of a service are most valuable to customers, and price the service accordingly. This information can be obtained through qualitative data (from interviews and focus groups) for novel services, or through quantitative data collected in surveys.

In addition to framing quality assessment in an ecologically valid manner, it is important to replicate the context of use as far as possible. Firstly, it is important that participants making the assessment are interested in what they are viewing. Asking people to assess the video quality of many repeats of a video on bricklaying will induce boredom in anyone not particularly interested in bricklaying. We always aim to recruit participants for our studies who are interested in the material we use for testing, and who are potential customers of the service we are evaluating.

Secondly, it is important to replicate all relevant factors of the viewing experience. Our approach is to create a use context that is as close as possible to the real use context. As Mellers and Cook [14] point out, *"Preferences do not occur in a vacuum, they are always formed relative to a context."* Participants' preferences and their judgments occur in a context which may be clearly defined or is implied by an experimental setup or assessment approach. Engeldrum stressed that integrative attributes like image quality are more context- or application-dependent than perceptual attributes such as sharpness, graininess etc [15].

The next section presents an example of how we have applied this assessment approach in a series of recent studies to identify QoS requirements for mobile TV.

3. Mobile TV studies

3.1. Pilot study

In order to get a first idea about how multimedia content would come across on mobile devices, we designed a pilot study on the perceived video quality of different content types on mobile devices. It is important that tests are carried out on a mobile device as it cannot be assumed that the experience of watching a small TV window on a 17" monitor at a fixed distance is the same as watching the same size window on a mobile device. With a hand-held device, users can easily move the screen closer to them. When watching on a large screen, they must move their whole body closer to the display, which requires more effort. TVs are usually watched in a posture where the head is upright. Handheld devices are operated with the head tilting down.

3.1.1. Material

We used five short clips (15 sec.), one of each of the content type weather, news, music, movie and football. We encoded the clips at up to five different encoding bitrates for a PDA (audio: WMA V9, video: WM V8) and a 3G mobile phone (audio: GSM AMR video: PV MP4). The video encoding bitrates ranged from 32 to 448 kbps with a nominal frame rate of 12.5 fps.

3.1.2. Procedure

We told the participants that a technology consortium was investigating ways to deliver TV content to mobile devices, and that they wanted to find out about quality requirements for video and audio for watching different types of content.

Twenty participants watched the clips in decreasing quality and rated them one by one. The video clips were presented on the full screen of a PDA (iPAQ 2210) at QVGA resolution of 320x240 and on a 3G phone at QCIF resolution of 176x44. In the study by McCarthy et al., participants called out the acceptability ratings while watching the clips [16;16]. To remove the requirement of having an experimenter note down the ratings, we had the participants first watch the clip and then provide quality ratings for audio and video on a score sheet. For both dimensions, they provided a rating on a labeled scale from 0 to 100 and a binary rating whether they found the quality acceptable or not. The participants watched the five clips repeatedly up to five times. We presented the clips always starting with the highest and then successively lower encoding bitrates.

3.1.3. Results

We limit the findings of the pilot to the perceived video quality on the iPAQ. Further results can be found in [17]. We averaged the numerical quality ratings and the acceptability ratings of all 20 participants. For comparison we have plotted the acceptability and the mean quality ratings in Figure 2 and Figure 1 respectively. From the *Acceptability* scores we can see that the acceptability ratings of news and weather roughly plateau at an encoding bitrate of 128kbps. At higher *Encoding Bitrates*, acceptability increased only marginally if at all. The music video and movie clip reached a plateau at 224kbps. Football was the most demanding *Content Type* and required 320kbps to be acceptable to 85% of the participants and did not increase further at the higher encoding bitrate of 448kbps.

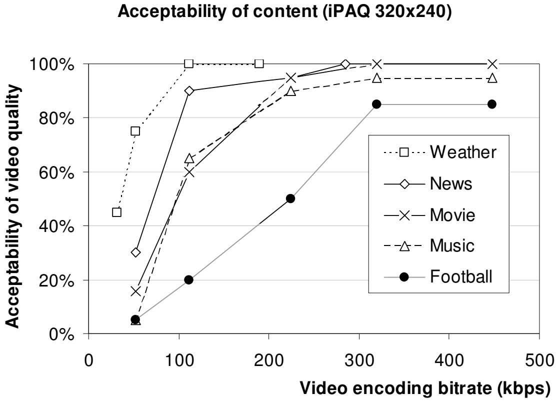

Figure 1: Acceptability ratings of video quality at different encoding bitrates by content types

The average video ratings were not as conclusive. All their graphs followed the logarithmic shape that one would expect from just noticeable differences in signal detection theory [18]. However, it is not clear from the average video rating results what encoding bitrate should be considered sufficient for a paying customer of a mobile TV service. The results tell us more about the participants' ability to reliably discriminate quality levels than providing guidance as to which encoding bitrates is satisfactory in a context of watching TV on a mobile device. We can see that - despite content types receiving the same average video rating - their perceived acceptability differed. For example, news and movie content received the same mean video rating (ca. 55) at an encoding bitrate of 112kbps. But news content was found to be of acceptable quality to 90% of the participants, whereas the movie content was only acceptable to 60% of the participants. This difference could partly be explained by the fact that the audio part of news carries much of the information that participants may deem most important when watching the news. They might therefore find lower video quality acceptable.

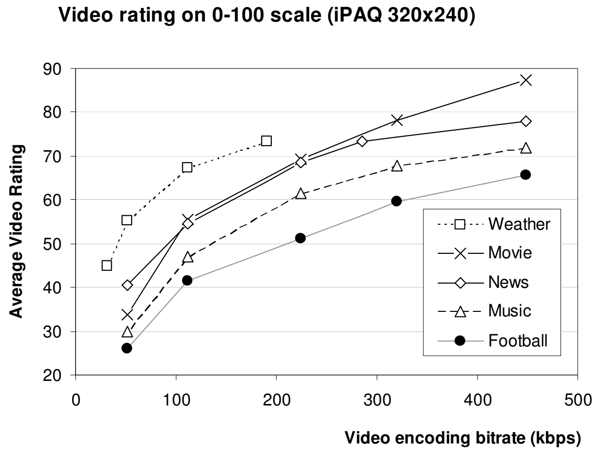

Figure 2: Average video quality ratings at different encoding bitrates by content type

Despite the rise in the video rating between 320kbps and 448 kbps the acceptability of the football content did not increase. Apparently, there were other factors than video quality that influenced the acceptability of mobile TV content. The size of the image is one potential factor that influences people's opinion of mobile TV. We will take a closer look on the influence of the size of the area on which the video is displayed on the acceptability of video quality of content types in study 1.

3.2. Study 1 (size)

Concerns about the size of video displays were noted in focus groups assessing the potential uptake of mobile TV services [19]. Users wanted a screen as large as possible for viewing, but they do not want their phones to be too big. We designed study 1 to explore how the context of mobile TV use might be sensitive to size since people could consume mobile TV on a range of devices with different sizes and resolutions. We did this by measuring how size affected the acceptability of video quality of mobile TV content.

3.2.1. Material

The clips used in this study lasted for two minutes and twenty seconds; according to previous research, a realistic duration for mobile TV interaction [20]. This study used 16 clips, four of each of the content types football, news, music and animation recorded from TV and DVDs. We encoded all clips at four different sizes (240x180, 208x156, 168x128, 120x90) that resulted in video window of different sizes as specified in Table 1.

Table 1: Image sizes used on PDA

Screen area (mm^2)	Pixels (P)	P/mm^2
53mm x 40mm (2,120)	(240 x 180) 43,200	20
46mm x 34.5mm (1,587)	(208 x 156) 32,448	20
37mm x 28mm (1,036)	(168 x 126) 21,268	20
26.5mm x 20mm (530)	(120 x 90) 10,800	20

The video encoding bitrate was manipulated in two different ways. Within a particular TV clip the bitrate allocated to video was gracefully degraded every 20 seconds by 32 kbps from a maximum of 224kbps down to 32kbps (Windows Media Video V8). The boundaries of the intervals were not pointed out to the participants. They were simply presented with a continuous clip that gradually decreased in quality. In addition to changing the video bitrate within a clip, two duplicate sets of clips were produced with different bitrates allocated to the audio channel.

The *Low Audio* clips coded the audio channels at 16kbps (Windows Media Audio V9) whereas the *High Audio* clips were coded at 32 kbps. Theses values were selected based on results of the pilot study in which participants' acceptability of 32kbps audio compared to 16kbps audio had declined from 95% to 80%.

3.2.2. Procedure

In this study, participants were able to give acceptability ratings with a stylus at any point in time on the same device that was presenting the clips.

As in the pilot study, we told the participants that a technology consortium was investigating ways to deliver TV content to mobile devices, and that they wanted to find out the minimum acceptable video quality for watching different types of content.

The instructions stated: *"If you are watching the coverage and you find that the [video] quality becomes unacceptable at any time, please click the button labelled 'Unacc'. When you continue watching the clips and you find that the quality has become acceptable again then please click the button labelled 'Acc'.*

The participants' ratings, i.e. the taps on the 'Unacc.' and 'Acc.' buttons, were recorded on the device. The interface of the experiment is presented in Figure 3.

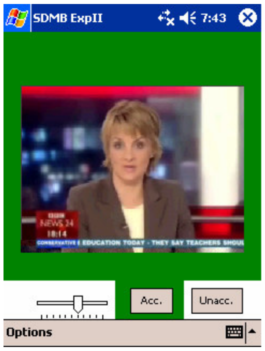

Figure 3: Experimental interface with buttons "Acc.", "Unacc." and volume control on the bottom left

We ran four different groups, each comprising 32 participants. Each group was presented 16 clips in total in groups of four clips at each of the four image sizes. The groups differed in whether they experienced *Increasing* or *Decreasing* image sizes and whether the audio quality was *High* or *Low*. Within each group, we also ran four variations to control for content using a Latin squares design such that the different content clips were tested at each of the different image sizes across participants. The dependent variable was *Video Acceptability*. Independent variables were *Image Size*, *Content Types*, *Video Bitrate*, *Audio Bitrate.* Control variables were *Size Order*, *Sex*, and *Corrected Vision*. The variable *Corrected Vision* coded whether participants considered themselves to have normal vision or whether they wore contact lenses or glasses.

3.2.3. Results

For a complete account of the results see [21]. Before analyzing the results, we conservatively coded each 20 second interval of a clip as *unacceptable* if they had given a rating of unacceptable at any point during that period. The resulting data was analysed using a binary logistic regression to test for main effects and interactions between the independent variables – *Image Size*, *Video Encoding Bitrate*, *Content Type* and *Audio Bitrate*. Control variables *Sex*, *Corrected Vision* and *Size Order* were also included in this analysis. The logistic regression showed that *Image Size* and *Content* were significant predictors of acceptability, [$\chi^2(3)=446$, $P<0.001$; $\chi^2(3)=1056$, $P<0.001$], and an interaction between *Image Size* and *Content type* [χ^2 (9) = 136, $P <0.001$]. A summary of this interaction is shown in averaged across all encoding bitrates.

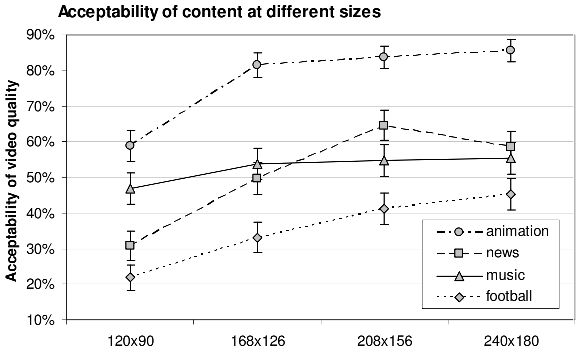

Figure 4: Image Size effects depended on the content.

Not surprisingly, the low motion animation clips received the best ratings – for this type of content there was no significant difference in acceptability as image resolution was reduced from 240x180 to 168x126 [χ^2 (2) = 0.468, n.s.], but at the smallest image resolution acceptability dropped off sharply [Z=-6.49, $P < .001$]. For News content the *acceptability significantly increases as the image resolution was reduced* from 240x180 to 208x156 [Z=-2.11, $P < 0.05$], after which point there was a steady decline in acceptability with decreasing image resolution. The curve for Music videos was relatively flat, and there was no significant difference in acceptability across the four image resolutions [χ^2 (3)=6.1, n.s.]. Finally, Sports coverage showed the lowest levels of acceptability. There was no significant difference in acceptability between the two largest image resolutions, but at image resolutions smaller than 208x156 acceptability significantly declined [χ^2 (2) = 25. 9, $p < 0.001$].

There was a significant effect of *Audio Bitrate* in the logistic regression [χ^2 (1) = 62.8, $p < 0.001$] but not in the direction expected. The participants were less likely to rate quality as unacceptable when the audio quality was low (16kbps) (see Figure 5). This was an unexpected result given the findings of previous studies on audio-visual interactions which show that increasing audio quality increases video quality ratings. The explanation may lie in the way the task is framed. Whereas many previous studies required participants to rate video quality on a scale we asked people to indicate when they find it unacceptable. In this context, low audio quality seems to set participants' expectations such that they are less likely to rate the video as unacceptable. By contrast, those given high audio quality have higher expectations and are more easily disappointed with the visual counterpart.

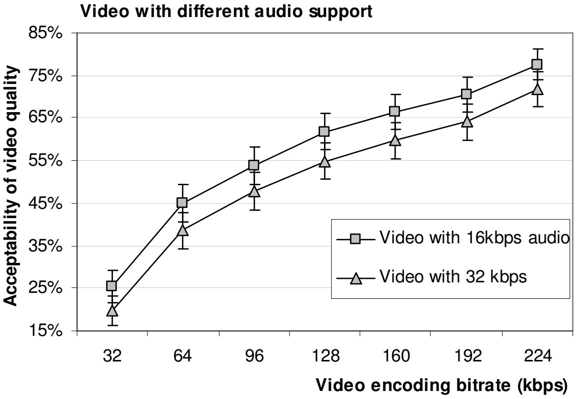

Figure 5: The effect of audio quality on video quality acceptability

From the qualitative feedback we gathered after the experiment, we learned that the biggest concern for unacceptability of video quality was illegible text. This was most prominent in the news content where text was used in logos, text tickers, headlines, inserts and diagrams.

When we looked at the acceptability ratings of the news content in more detail and compared them to the results obtained in the pilot study (both shown in Figure 6).

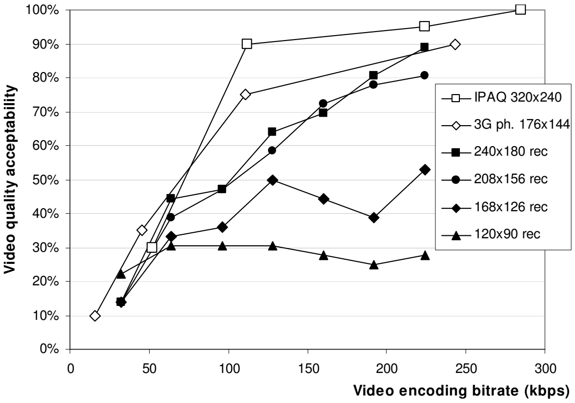

Figure 6: Acceptability of recorded TV news content (rec.) at different encoding bitrates by size in comparison to the pilot data (displayed with white markers)

We also saw that the acceptability of the clips in study 1 was much lower than the ones in the pilot study in general and most notably for the clips that were roughly equal in size (176x144 in the pilot and 168x126 in study 1). This comparison might be confounding different devices and codecs, methodologies and was based on different content not only in terms of length. Nevertheless, prompted by the participants' feedback we reviewed the news clip of the pilot study in terms of text legibility. It only contained illegible text in Arabic language which the participants could neither read nor understand. Its headlines and other text items in Western font were legible at both resolutions. We therefore decided to research the effect of text legibility on video quality in more detail and designed the study presented in Sec. 3.3.

Another complaint voiced in the feedback rounds was that certain shot types did not render well on the small screens and that important detail got lost. This pertained particularly to the football content in which the pitch was often shown from a great distance. We present the influence of video size on shot types in Sec 3.4.

3.3. Study 2 (text)

Since text seemed to be an influential factor for mobile TV quality we designed this follow-up study to assess the effect of text legibility and quality on overall perceived video quality. It was not clear from the results of study 1 to what extent text quality and legibility influenced the overall video quality perception because:

1. it employed illegible text (at 120x90 and 168x126) which was smaller than five pixels in height - the minimum for rendering fonts and
2. participants were not tested for their visual acuity.

3.3.1. Material

We included the news material from study 1 for comparison purposes and added another four clips that were recorded in the same way. Before encoding the clips we modified the material such that the text presented in the ticker on the bottom of the screen, the logo and text inserts that appeared temporarily in the area right of the logo would be legible at all target sizes. For a complete description of the modifications and how we carried them out see [22]. Figure 7 shows the material used in study 1 and the material after the modifications for the study at hand.

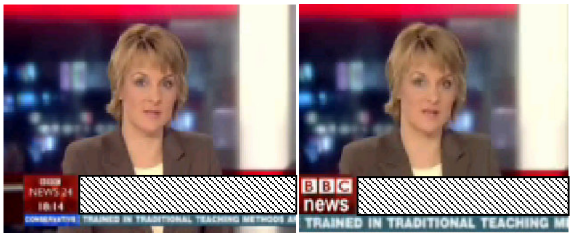
BBC NEWS 24 18:14
CONSERVATIVE TRAINED IN TRADITIONAL TEACHING METHODS A
BBC news
TRAINED IN TRADITIONAL TEACHING M

Figure 7: Content before (l.) and after (r.) modifications. Text inserts appeared in the hatched area.

After the modifications were made, we encoded these base clips using the same sizes and video encoding bitrates employed in study 1 but only one audio encoding bitrate of 32kbps (WMA V9). After encoding the clips in the same way as in study 1 we produced a second set with high text quality in which the ticker line, the BBC logo, and text inserts above the ticker were replaced with the footage before the encoding. As a result we had one set of eight clips in which the text quality remained high throughout the clip despite the remaining screen degrading in quality every twenty seconds and a second set in which the text degraded along with the rest of the image (see Figure 8 for an example).

Two of the eight base clips contained text in the main window that was rendered illegible by smaller sizes. For better comparison with study 1 we chose to include these clips in the tested set and included a control variable for them in the analysis.

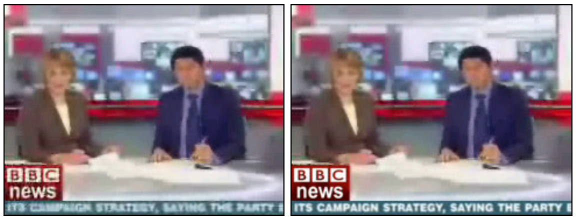
BBC news
ITS CAMPAIGN STRATEGY, SAYING THE PARTY
BBC news
ITS CAMPAIGN STRATEGY, SAYING THE PARTY

Figure 8: News with degrading (l.) and high text quality (r.)

3.3.2. Procedure

We used the same equipment, interface, instructions and methodology as in study 1 but in this study the participants had to complete a two-eyed Snellen test for 20/20 vision. A total of 64 participants (31 women and 33 men) provided acceptability ratings of the video quality. We ran four groups. Each group of 16 participants viewed eight clips in groups of two clips at each of the four sizes. The groups differed in whether they experienced *increasing* or *decreasing* image sizes and whether the text quality of the ticker, the headline inserts, and the news logo was *degrading* along with the video quality or of constant *high quality*.

The dependent variable was *Video Quality Acceptability*. The independent variables were *Image Size*, *Video Encoding Bitrate*, *Text quality.* Control variables were *Size Order*, *Sex*, *Native English Speaker, Text in Content*, and *Normal Vision*. We used the control variable *Text in Content* to identify the two aforementioned clips that contained small text in the main window. The variable *Normal Vision* coded whether participants had 100% visual acuity according to the administered Snellen test [23].

3.3.3. Results

As in study 1, we conservatively coded for each participant each 20 second interval of a clip as *unacceptable* if the video quality had been unacceptable at any point during that period.

Across all participants text quality was not a significant predictor of the acceptability of video quality [$\chi^2(1)$ =2.4, n.s.]. This is due to the fact that the opposing ratings of the non-native and native speakers cancelled each other out. Post-hoc tests revealed an interaction between *Text Quality* and *Native Speaker* [$\chi^2(1) = 40.1$, $P < 0.001$]. This effect came as a surprise. Native speakers who watched clips supported by high text quality rated them higher in terms of acceptability than the non-native speakers. The non-native speakers rated video quality higher when video was accompanied by text that degraded with the video (illustrated in Figure 9).

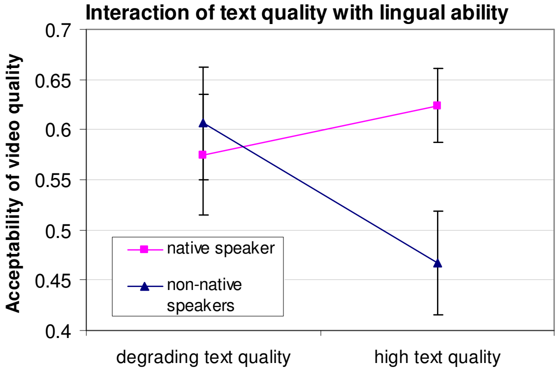

Figure 9: The effect of text quality on the overall perceived video quality depends on the lingual ability of the beholder

We partitioned the data set and looked separately at the two groups. Two non-parametric Mann-Whitney tests showed significant differences for *Text Quality* for both the native speakers [Z=-2.1, P<0.05] and the non-native speakers [Z=-5.3, P<0.001]. We ran the original binary logistic regression without the variable *Native English Speaker* on the partitioned data sets. Along with all the previously described variables *Text Quality* turned out to be a significant predictor of acceptability in the analysis of the native speakers [$\chi^2(1)$=8.2, P<0.01] and the non-native speakers [$\chi^2(1)$=21.7, P<0.001] - but in opposing directions as described above. Similarly, the control variable *Text in main window* was a significant predictor of acceptability [$\chi^2(1)$=17.4, P<0.001] for the native speakers but not for the non-native speakers [$\chi^2(1)$=0.01, n.s.]. Considering the impact of the non-native speakers we will limit the presentation of results to the 36 native speakers. Averaged across all encoding bitrates and sizes the acceptability of news content increased from 50% with degrading text to 57% when presented with high text quality instead. In Figure 10 we have plotted the acceptability scores of the news clips in this study supported by high text quality in comparison to the news clip used in the pilot study.

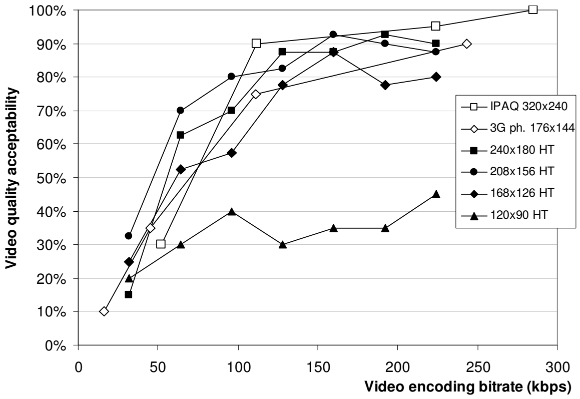

Figure 10: Acceptability of news content with high text quality (HT) in comparison to the pilot study (white markers)

Despite the legibility of the text at the smallest size (120x90) in terms of size in this study compared to study 1 the acceptability of video quality of news content still dropped dramatically when image size was reduced to 120x90 pixels. High text quality increased the acceptability of news content especially for all sizes larger than 120x90. This was especially true for the clips at 168x126. Their acceptability now reached levels similar to that of the QCIF (176x144) clip of the pilot study. Based on these results a conservative advice to service providers would be to not deliver news content at resolutions lower than 168x126 pixels. The gains in video quality by displaying high quality text are substantial and service providers should consider separate delivery of text e.g. through formats like SMIL, QuickTime etc. which allow for rendering of text at the receiver.

3.4. Shot type analysis

In study 1, participants complained about the lack of detail in certain shot types (e.g. extreme long shots in football) or that they could not identify people or objects when presented on small displays. Shot types differ in the degree of detail and the amount of context in which the subject is situated. The way in which objects are shot, edited, presented and decoded by the audience follows established conventions [24]. The different shot types used in film-making help the audience to "read" the message the director wants to convey.

To the best of our knowledge no previous research has addressed the question of how small display sizes affect the different shot types used in video material. However, previous research on picture quality showed that sharpness is judged differently for a portrait of a person and the depiction of a landscape [25].

TV and cinema content use a mix of shot types with varying lengths. Creating a fully counterbalanced set of stimuli with real content clips of considerable length is therefore hard to achieve. We decided to drop this requirement for an initial study on the effect of resolution on the acceptability of the video quality of shot types and classified each shot of the video clips of study 1 according to Thompson's classification [24] (see Figure 11 for example shot types of football content). We then looked at the acceptability ratings from study 1 aggregated by shot types.

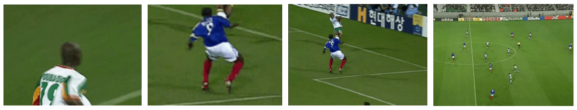

Figure 11: Shot types used in football content from left to right: medium shot (MS), long shot (LS), very long shot (VLS) and extreme long shot (XLS)

There were two possible caveats with this approach. First, due to the use of the method of limits the experimental design did not present all parts of the video clips at all encoding bitrates. Consequently, the average encoding bitrate at which shot types were encoded were not identical. Second, many video encoders compress e.g. low motion video clips better than clips that include a lot of motion. Some shot types might contain more motion on average than others and therefore look better after encoding in terms of visual quality, e.g. sharpness. Thus even if the shot types had been encoded at identical average encoding bitrates that would have not guaranteed equal visual quality of the shot types after encoding.

To control for both the differences in encoding bitrate as well as possible correlations between shot types and encoder performance we used the objective quality measure peak signal-to-noise ratio (PSNR) to obtain a rough estimate of the content's visual quality. We rescaled all degraded clips up to the resolution of the original clips and employed Avisynth's built-in PSNR compare function to compute the degradation of these encoded clips in comparison to their originals [26]. Since we compared up-scaled versions of the low resolution clips with the reference clip we can expect that the lower resolution clips will in general yield lower PSNR scores. For example a clip with a resolution of 120x90 would be up-scaled by a factor of about four which will result in higher peak signal-to-noise ratio than a clip up-scaled from 240x180 by a factor of two. We only used the PSNR scores as indicators of visual quality between the shot types in clips of the same resolution.

3.4.1. Results

The data were generated from the acceptability replies of the participants from study 1 but this time on a per second basis. For example, if a participant had been in the unacceptable state during a second it was marked 'unacceptable' for this participant. We decided to exclude all ratings in the three seconds following a scene change to allow for participants' adjustment to the new picture. In doing so we excluded shots that lasted less than three seconds. In addition to the variables analysed in the original study we included *Shot Type* as an independent and *Native Speaker* as a control variable. The latter variable denoted native English speakers.

We analysed the data using a binary logistic regression to test for main effects and interactions between the independent variables – *Image Resolution*, *Video Encoding Bitrate*, *Content Type, Shot Type* and *Audio Bitrate.* Control variables *Gender, Corrected Vision, Resolution Order* and *Native Speaker* were also included in this analysis. The variable *Corrected Vision* indicated whether participants had uncorrected vision or wore contact lenses or glasses.

The regression revealed significant effects of all of the control and independent variables as in study 1. Non-native English speakers were less likely to rate the quality of a clip unacceptable than the native English speakers. We excluded the data from the non-native speakers and repeated the regression. All results we present from here on are based on the 72 native speakers that took part in the study.

We report only the acceptability scores of shot types that each participant had watched for a total of at least 40 seconds. *Shot type* was a significant predictor of acceptability [$\chi^2(1)=148.4$, $P<0.001$]. All shot types became more acceptable with increased sizes. Furthermore, the regression revealed an interaction of *Shot Type* and *Content Type* [$\chi^2(1)=1337.1$, $P<0.001$]. We will limit our account of shot types to football. More detail and the results of the other content types can be found in [27].

Almost all of the scenes in the football footage depicted players in motion or camera pans of the pitch. Shot types closer than a medium shot are not common in football coverage. It is hard to zoom in on and follow players because they often move in unpredictable ways. The extreme long shot provides the viewer with an overview of what is going on in the playing-field. It is very popular and even in the highlights material used in the study this shot was used approximately 50% of the time.

Non-parametric tests showed that there was no significant difference in acceptability of the XLS at the highest resolution when compared to the other shot types [$\chi^2(3)=2.34$, *n.s.*]. However, at all sizes lower than 240x180, the results confirmed the qualitative feedback in study 1 about the XLS - the XLS was the least acceptable shot type.

Surprisingly, the acceptability of the medium shot depicting the greatest amount of detail in the football material declined much more than the long and the very long shot at sizes smaller than 208x156 (see Figure 12).

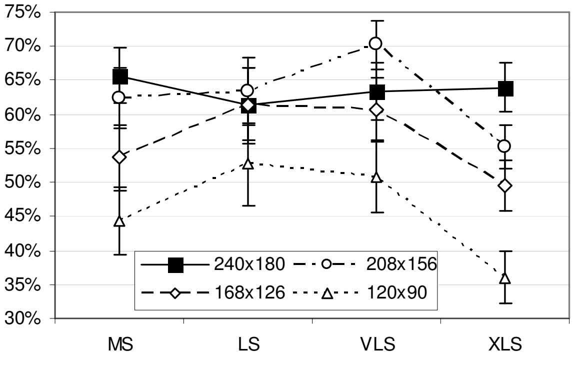

Figure 12: Acceptability of shot types of football content

In the computed PSNR values depicted in Figure 13 we find no evidence that the lower acceptability of MS and XLS might be induced by lower visual quality. Both the MS and the XLS yielded considerably higher PSNR values in comparison to the LS and VLS.

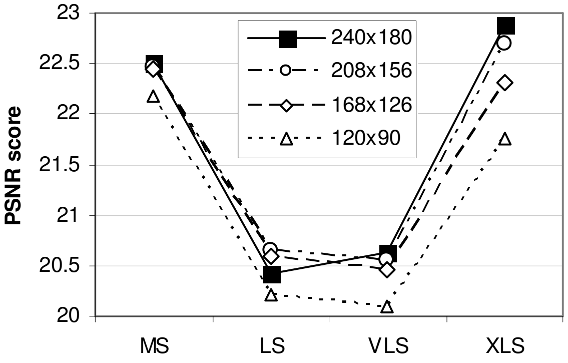

Figure 13: PSNR scores of football content at different sizes by shot types

When shown on fewer than 240x180 pixels, the XLS might benefit from cropping off the safe area around the middle of the screen in TV footage [24] or intelligent cropping schemes e.g. [29] or [30], all of which would show a part of the screen in more detail. Clearly, the results at hand warrant more research that could control for movement and other possible covariates of shot types. More insight will aide mobile content producers in making informed choices in this novel area of multimedia consumption.

4. Conclusions

The studies on mobile TV presented in this paper exemplify the assessment approach that we advocate. Our studies identified number of factors that have an impact on the perceived video quality in mobile TV services. Apart from the video encoding bitrate, we found that the level of quality required varies depending on

the content of video depicted,
the size at which the content is displayed,
the shot types that are used to depict the content,
the audio quality which accompanies the video
and the legibility of text if present in the video.

All of these affected participants opinion on whether they found a given video quality acceptable or not.
Our methodology of using simple acceptability ratings, and collecting qualitative data (we asked participants a why they labeled video quality unacceptable) provided us with pointers to the factors that influence people's experience of watching mobile TV, which could then be investigated in further studies. The advantage of this approach is an increase in ecological validity, since we identify all factors that are relevant to this particular service, rather than assuming that we know them. We can be confident that the quality thresholds are a good predictor of how users will respond in the field, as long as service providers bear in mind that marketing and pricing of a service influence customer expectations.

5. References

[1] Ferguson, P. and Huston, G., *Quality of Service: Delivering QoS on the Internet and in Corporate Networks* Wiley, 1998.

[2] Cisco, "Quality of Service Networking," *Internetworking Technologies Handbook* 2006.

[3] International Telecommunications Union (ITU). P.800.Methods for subjective determination of transmission quality. ITU-T P.800. 2004.

[4] Sasse, M. A., Biltung, U., Schulz, C. D., and Turletti, T. Remote seminars through multimedia conferencing: experiences from the MICE project. INET '94/JENC5. Proceedings of INET '94/JENC5 , 251-258. 1994.

[5] Hardman, V., Sasse, M. A., Handley, M., and Watson, A. Reliable Audio for Use over the Internet. International Networking Conference. Proceedings of INET'95 , 171-178. 1995. Reston, VA, ISOC.

[6] Hardman, V., Sasse, M. A., and Kouvelas, I., "Successful Multi-party Audio Communication over the Internet," *Communications of the ACM*, vol. 41, no. 5, pp. 74-80, 1998.

[7] http://www.accessgrid.org/software . 2006.

[8] Watson, A. and Sasse, M. A. Measuring perceived quality of speech and video in multimedia conferencing applications. ACM Multimedia. Proceedings of ACM Multimedia'98 , 55-60. 1998. ACM.

[9] Knoche, H., de Meer, H., and Kirsh, D. Utility curves: Mean opinion scores considered biased. Proceedings of IWQoS'99 , 12-14. 1999.

[10] Watson, Anna, "Assessing the Quality of Audio and Video Components in Desktop Multimedia Conferencing." PhD Thesis University of London, 2001.

[11] Watson, A. and Sasse, M. A. The Good, the Bad, and the Muffled: The Impact of Different Degradations on Internet Speech. Proceedings of the 8th ACM International Conference on Multimedia , 269-276. 2000.

[12] Bouch, A. and Sasse, M. A. Why Value is Everything: A user-centred approach to Internet Quality of Service and Pricing. Wolf, L., Hutchison, D., and Steinmetz, R. Quality of Service - Proceedings of IWQoS 2001 , 49-72. 2001. Springer. (Lecture Notes in Computer Science 2092).

[13] Zeithaml, V. A., Parasuraman, A., and Barry, L. L., *Delivering Quality Service: Balancing Customer Perceptions and Expectations* The Free Press, 1990.

[14] Mellers, B. J. and Cook, A. D. J., "The role of task and context in preference measurement," *Psychological Science*, vol. 7, no. 76, 1996.

[15] Engeldrum, P. G. Psychometric Scaling: Avoiding the Pitfalls and Hazards. IS&T's 2001 PICS Conference Proceedings , 101-107. 2001.

[16] McCarthy, J., Sasse, M. A., and Miras, D. Sharp or smooth? Comparing the effects of quantization vs. frame rate for streamed video. Conference on Human Factors in Computing Systems CHI '04. Proceedings of CHI , 535-542. 2004.

[17] McCarthy, J., Miras, D., and Knoche, H. TN01-1.1.03_UCL_MAESTRO_bandwidth_study_V02. 2004.

[18] Fechner, G. T., *Elemente der Psychophysik* Leipzig: Breitkopf und Härtel, 1860.

[19] Knoche, H. and McCarthy, J. Mobile Users' Needs and Expectations of Future Multimedia Services. WWRF12. Proceedings of the WWRF12 . 2004.

[20] Södergård, C. Mobile television - technology and user experiences Report on the Mobile-TV project. P506. 2003. VTT Information Technology. ESPOO 2003.

[21] Knoche, H., McCarthy, J., and Sasse, M. A. Can Small Be Beautiful? Assessing Image Resolution Requirements for Mobile TV. ACM Multimedia. Proc.of ACM Multimedia 2005 , 829-838. 2005. ACM.

[22] Knoche, H., McCarthy, J., and Sasse, M. A. Reading the Fine Print: The Effect of Text Legibility on Perceived Video Quality in Mobile TV. ACM Multimedia. Proceedings of ACM Multimedia'98 . 2006.

[23] Bennett, A. G., "Ophthalmic test types," *Br.J.Physiol.Opt.*, vol. 22 pp. 238-271, 1965.

[24] Thompson, R., *Grammar of the shot* Elsevier Focal Press, 1998.

[25] Freiser, H. and Biederman, "Experiments on image quality in relation to the modulation transfer function and graininess of photographs," *Photographic Science and Engineering*, vol. 7, no. 28, 1963.

[26] Avisynth. http://www.avisynth.org/ . 2005.

[27] Knoche, H., McCarthy, J., and Sasse, M. A. A close-up on Mobile TV: The effect of low resolutions on shot types. EuroITV 2006 - Beyond Usability, Broadcast, and TV. Proc.of EuroITV '06 . 2006.

[28] Knoche, H., McCarthy, J., and Sasse, M. A., "How Low Can You Go? The Effect of Low Resolutions on Shot Types in Mobile TV," *MTA*, 2006.

[29] Dal Lago, G. Microdisplay Emotions. http://www.srlabs.it/articoli_uk/ics.htm . 2006.

[30] Holmstrom, D., "Content based pre-encoding video filter for mobile TV." Umea University, 2003.