Does Context Matter in Quality Evaluation of Mobile Television ?

Satu Jumisko-Pyykkö
Tampere University of Technology
P.O. Box 589
33101 Tampere, Finland
+358 50 361 0038
satu.jumisko-pyykko@tut.fi

Miska M. Hannuksela
Nokia Research Center
P.O. Box 1000
33721 Tampere, Finland
miska.hannuksela@nokia.com

ABSTRACT
Subjective quality evaluation is used to optimize the produced audiovisual quality from fundamental signal processing algorithms to consumer services. These studies typically follow the basic principles of controlled psychoperceptual experiments. However, when compromising compression and transmission parameters for consumer services, the ecological validity of conventional quality evaluation methods can be questioned. To tackle this, we firstly present a novel user-oriented quality evaluation method for mobile television in its usage contexts. Secondly, we present the results of an experiment conducted with 30 participants comparing acceptability and satisfaction of quality as well as goals of viewing in three mobile contexts and under four different residual transmission error rates, when the participants also performed simultaneous assessment tasks. Finally, we compare the results with a previous laboratory experiment. The studied error rates impacted negatively on all measured tasks with some contextual differences. Moreover, the evaluations were more favorable and less discriminate in the mobile contexts compared to the laboratory.

Categories and Subject Descriptors
H.5.1 Multimedia Information Systems: Evaluation/ methodology.

General Terms
Experimentation, Human Factors

Keywords
Mobile TV, mobile television, ecological validity, subjective quality, context, evaluation, transmission quality.

1. INTRODUCTION
Mobile television (TV) is expected to become a popular mass media service. Commercial services are already available in several countries, and numerous field trials are in progress all over the world [5].

Digital Video Broadcasting for Handhelds (DVB-H) is one of the most commonly used technologies for mobile TV. In DVB-H, an audiovisual service is transmitted in time-slicing bursts to achieve power savings in receiving terminals. Despite efficient mechanisms for error correction, the received data streams over DVB-H may contain residual transmission errors caused by well-known phenomena of radio communications, such as noise and interference.

In addition to obvious physical limitations of the rendering devices, such as screen size and resolution, the relatively high level of compression required for over-the-air transport and the impairments caused by transmission errors affect perceived video fidelity. Most of the recent studies, such as [11],[13],[14],[23],[24],[31],[32],[33],[35],[44] have examined the impact of compression in relation to experienced quality, whereas there are fewer studies analyzing the impact of transmission errors to perceived quality [14],[16],[26],[27].

Subjective research methods can be used to conclude perceptual multimodal acceptability, preferences, and critical quality factors [20],[21],[35]. There are two main approaches to measure the excellence of the stimuli quantitatively. The first is based on methodological recommendations of International Telecommunication Union (ITU) and is popular among the engineering society [20],[21]. This approach mostly follows common guidelines of psychophysical experiments, such as preferring short stimuli with several repetitions. The second approach attempts to gain a high ecological validity by constraining the tests to potential users and stimuli and evaluating acceptability or goals of viewing in parallel to quality assessment [10][11][13][14][26] [35]. However, usage context or situation is often not taken into account in subjective research methods.

There are only few comparisons published about usability in mobile and laboratory environments [28],[29]. These studies included a simple walking task and a more complicated traveling task. Interestingly, long loading times in web browsing were more rarely mentioned in real usage contexts than in a controlled environment. This may indicate that people are less sensitive for transmission delay or errors in real context than in laboratory.

MobileHCI 2008, September 2–5, 2008, Amsterdam, the Netherlands.

In this paper, we firstly present a novel user-oriented quality evaluation method for mobile TV to be used in contexts of use. Secondly, we present the results of an experiment which was conducted with 30 participants, compared four different residual transmission error rates with several television programs, and evaluated satisfaction, acceptability of quality, and goals of viewing in three contexts. Thirdly, we compare the results with a previous laboratory experiment.

The paper is organized as follows. Section 2 gives on overview to current knowledge of use of mobile TV as a base for the development of user-oriented contextual quality evaluation method. Section 3 describes the status of quality evaluation research and its methods in relation to mobile TV. Section 4 presents the developed method and section 5 summarizes the results. Finally, section 6 presents discussion of the study.

2. MOBILE TV IN ITS CONTEXT

Context of use, also known as usage context, comprises user characteristics, task, as well as technical, physical, and social environment [19]. In mobile usage contexts, physical and social environments are heterogeneous and may change during the usage session. For example, people may move between personal and group spaces, and they may perform actions as a response to unexpected events in addition to planned goals [42]. Furthermore, the intensity of actions varies and can be classified e.g. to acceleration, normal, and waiting [42]. Mobile TV has been studied within its context with two approaches: consumer studies and field trials. This section provides an overview of many studies performed using these approaches.

People are interested in viewing the same popular television programs from mobile TV as from their normal TV. News, music videos, sport, animations and series are the most popular types of content in order to become informed or entertained [1],[2],[30],[36],[39],[41]. A typical duration of a continuous mobile TV viewing session is relatively short, but it also varies according to the type of the program. Södergård et al. [41] reported that a typical viewing time is less than 10 minutes. Later, O'Hara et al. [36] confirmed with a field study that people viewed from short 30-second clips to 30-minute shows and movies. Mobile TV is viewed in different times of day from morning to late in the evening [1],[37].

Mobile TV enables both an individual and a shared viewing experience. Motivations for individual viewing are killing time, fighting loneliness, keeping up-to-date, browsing content and unavailability of other possibilities to watch television [36],[37],[41]. In public settings, mobile TV is also used to gain one's own space and privacy [36]. Even though screen size and use of headphones give practical limits to the use of mobile TV among multiple viewers sharing the same device, mobile TV is used in shared situations to entertain children, to carry the piece of information to new context as well as to store, collect and share it [36],[39]. As a difference to normal TV viewing, mobile TV is typically not used as a background noise [37].

The most commonly mentioned physical environments for viewing are vehicles, such as public transportation and private cars, waiting halls or lounges, work, home and cafes [30],[36],[39],[41]. Specific contextual patterns have also been reported, some of which are briefly reviewed next. People do not watch mobile TV during short journeys [36], and in noisy environments they would prefer textual information over video [37]. In motion, audio is the preferred media, whereas during stationary reception, text, and video are the most pleasant media [30],[37]. A need for fluent transitions between video and audio are also reported when changing the physical environment, e.g. from a bus to a bus stop [36].

3. PERCEIVED QUALITY EVALUATION

The optimization problem for multimedia quality in a communication system is a complex function of parameters including produced or encoded quality and capability of correcting transmission errors. The main goal in this optimization is to reproduce and render a multimedia presentation with as little negative perceptual effects as possible under the technical constraints of the communication system. From a technical perspective, the optimization problem can be often simplified to three interrelated sub-problems: quality optimization under the constraints of feasible rendering capabilities of receiving devices, quality optimization of the encoded material (free of transmission errors), and quality optimization of the received multimedia material. The related research problems concern, e.g., display technology, screen size, and resolution, compression of multimedia material, and methods for robust data transmission, respectively. Perceived quality is a result of a complex psychoperceptual processing chain from incoming stimuli through active processes combining sensory information to final interpretations of stimuli including personal meaning and relevance to intentions and goals [12]. Individual emotions, knowledge, expectations and schemas representing the reality give different weightings for sensory attributes and they also enable human contextual behavior and active quality interpretation [12],[25]. The research methods to tackle the complex relation between reconstructed quality with a receiving device and perceived quality are called as subjective or perceived quality evaluation methods [11],[20],[21].

There are two main approaches in the measurements of perceived quality. Next, we summarize these approaches from the point of view of the evaluation procedure including scales and evaluation tasks, sample selection criteria and studied produced quality factors or variables for mobile TV.

3.1 Psychoperceptual approach

Evaluations are most commonly conducted with quantitative test methods based on Telecommunication Union (ITU) recommendations [20],[21]. The first recommendations were published in 1970's for television image quality research, but nowadays there are several recommendations for different quality evaluation fields and they are well-spread among the engineering society. For multimedia quality evaluation, ITU-T Recommendation P.911 provides the research methodology for three different retrospective research methods [21]. The most notable method, called Absolute Category Rating (ACR), uses short test stimuli (<10s) presented one by one and rated independently and retrospectively. The evaluations are given using a 5/9/11-point scale labeled from bad to excellent by naïve evaluators who are not working with picture or audio quality. ACR is applicable especially for a wide quality range and

therefore it has been used in several audiovisual quality studies for mobile video purposes [23],[24],[44].

3.2 User-oriented approach

The second experimental approach to measure multimedia quality can be considered as a user-oriented approach. Currently, there are neither methodological guidelines nor a full agreement about the used methods, but there is an aim to set the quality preferences or requirements according to context of use, taking into account users, performed tasks, and usage environments [19]. This approach is becoming popular among human-computer interaction professionals, and there are two main ways to establish quality evaluation – either in the terms of acceptability or quality of perception (e.g.[10][13][14][23][24][25][26][27][31][33]).

The main idea behind acceptability measures are in identifying the lowest useful quality level. McCarthy et al.'s [35] method describes acceptance percentage in relation to time based on Fechner's psychophysical method of limit. The threshold of acceptance is achieved by gradually decreasing or increasing the intensity of the stimulus in discrete steps in every 30 seconds. While watching, participants assess quality continuously and say aloud the point when quality becomes unacceptable or acceptable, respectively. Binary acceptance ratings are transformed to a ratio calculating the proportion of time during each 30-second period that quality was rated as acceptable and final results states acceptance percentage of time. The method has been used to examine compression parameters, screen sizes and shot types [31],[32],[33],[35]. Special attention has also been paid to sample selection in one of the reported studies, in which soccer fans corresponding to potential viewers were selected [[35]]. Later, Jumisko-Pyykkö et al. [26] and Hannuksela et al. [16] have also measured retrospectively acceptability nominally parallel to satisfaction, following some principles from ACR in transmission quality measures with 60-second-long stimuli material.

The model called Quality of Perception (QoP) pays attention on goals of viewing in quality evaluation tasks [3],[10],[11],[13],[14]. Multidimensional QoP is a combination of information assimilation and satisfaction formulated from dimensions of enjoyment and subjective, but content-independent objective quality (e.g. sharpness). Information assimilation is measured with questions about semantics of media content. Both of the satisfaction factors (enjoyment and objective quality) are assessed with a scale from 0 to 5. Stimuli ranging from 30 to 45 seconds in duration are presented one by one in a controlled environment. QoP is a sum of the scores for information assimilation and satisfaction, and it can be used to arrange the stimuli into preference order. The method have been used to assess both compression and transmission parameters. In few of these studies, evaluators have also been categorized according to their cognitive strategies [3],[10].

There have been attempts towards using a user-oriented quality evaluation scheme for mobile TV environment. For example, some preliminary work has been done in potential sample selection and also in finding the genre preferences of users [3],[10],[35]. Stimuli materials for the experiments have originated from television broadcasts suitable for mobile TV. Some studies, such as [3],[10],[11],[13],[14], have measured the goals of mobile TV viewing, to become informed or entertained, with QoP. In some studies, such as [26],[31],[32],[33],[35], measures of satisfaction and acceptability as an anchor of useful quality level have been used. However, there are still necessary steps to be taken to accomplish a truly user-oriented evaluation scheme, mainly because all evaluation methods and experiments have taken a place in a controlled laboratory environment. The laboratory environment, as an artificial setting, enables the accurate control of variables and replicable experiments, but laboratory studies suffer from limited realism and unknown level of generalizeability [45]. The fact that laboratory environments differ from heterogeneous physical and social mobile TV environments creates a need to verify the quality requirement derived from controlled experiments in real usage contexts.

In the area of usability studies, some experiments in mobile usage context have been reported. The main challenges of these kinds of usability studies are in the selection of proper usage contexts and realistic situations, applicability of evaluation methods, data-collection, and the number of unknown variables [28]. One aim of these experiments has been to compare the results between laboratory and field usability tests. For example, Kaikkonen et al. [28] compared usability in a laboratory test with usability experienced during a short-term travel task including crossing a street, taking the subway and escalators, as well as finding one's way. It was found that the field results remained equivalent compared to the laboratory results, but did not perfectly match. For example, long loading times in web browsing were mentioned less in the real usage contexts than in the controlled environment [28]. From the viewpoint of transmission delay or errors, this conclusion could be an indication about a lower level of sensitivity in a real usage context.

4. RESEARCH METHOD

4.1 Participants

30 participants, equally stratified by age (18-45 years) and gender, took part into the experiment between spring 2006 and summer 2007 in the city of Tampere. Less than 20% of the participants were categorized as innovators, early adopters, or professional evaluators according to their attitude towards technology or profession [20],[40]. All participants reported to have normal vision and hearing.

4.2 Test procedure

Each test session started with two preparatory parts. First, demographic data collection took place. Second, in the combined anchoring and training phase, the extremes of produced quality range were presented and participants got to know evaluation tasks and themes of contents.

The actual tests were conducted with a similar procedure in three different usage contexts. Absolute Category Rating, [20],[21] was applied partly. In this procedure, the stimuli are viewed one by one and they are rated retrospectively and independently. After presentation of each approximately 60-second-long stimulus, the participants were given 25 seconds to finish a quality evaluation containing four parts [Figure 1]. First, the participants answered whether or not the experienced quality was acceptable.

Measuring acceptability provides an anchor of lowest useful quality level for novel techniques that have not been examined perceptually [16],[26],[27]. , the participants gave an overall satisfaction score of quality on a continuous and unlabelled 11-point scale [43]. The last two quality evaluation tasks measured two goals of television viewing, to become entertained and informed [2]. Third, they assessed how entertaining they experienced the stimulus on a continuous 11-point scale. Fourth, in an information recognition task, the participants answered to three multiple choice questions, tackling the main points of the stimulus. The questions were presented from audio and visual part of stimuli. The information recognition is more efficient bringing back strategy than information recall in television watching situation [34].

After the actual test part in each context, the experiences and impressions of presented quality and context were collected with a semi-structured interview. The duration of the whole experiment was approximately two hours.

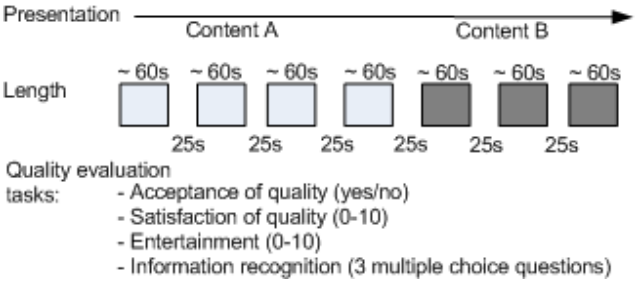

Figure 1 Presentation contents and evaluation tasks

4.3 Selection of Test Material

The test material was chosen among news, sport, music videos and cartoons, which are expected to be among the most popular genres for mobile television [1],[30],[36],[39],[41]. In each genre, the selected material belonged to the most popular broadcasted programs in Finnish television between 2005 to spring 2006 [9]. In addition, popular content was selected because it is evaluated more critically than unfamiliar content [22].

Audiovisual characteristics, such as the amount of spatial details and movement and the type of the audio track, varied [Figure 2] in the selected test clips. A group of four test stimuli, presented one by one, formed a continuous story in order to give as realistic a viewing experience as possible [Figure 1]. Three stories were cut from each selected program, and each story was cut into four test clips. The duration of the clips was approximately 60 seconds in order to have at least one simulated impairment during each stimulus. The clips were semantically meaningful segments of the programs. Due to the short duration of music videos, three different music videos were selected to represent three different stories. One story from each content category was shown in each usage context. The use of stories as stimuli differs from controlled psychophysical experiments conducted in a laboratory environment in which the same stimuli are repeated several times [20],[21].

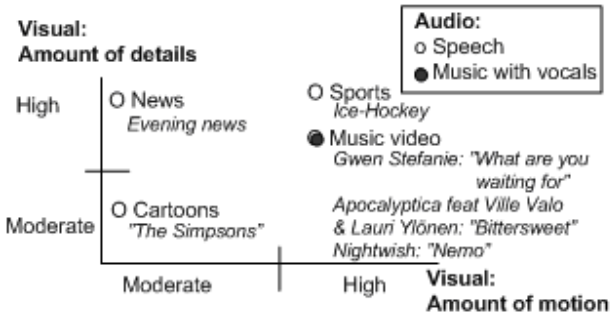

Figure 2 Genre of stimuli, contents and their audiovisual characteristics

4.4 Test environments – contexts and tasks

Test material was viewed in three predefined mobile TV usage contexts. The participants were given different tasks in order to simulate real usage situations in parallel to the quality evaluation task. Waiting for a friend in a railway station, traveling by local bus, and spending time in a café are among the most mentioned usage situations for mobile TV [30],[36],[39],[41]. The usage contexts and tasks are described in detail in Table 1.

Table 1 Context descriptions: physical environment and tasks

Physical environment	Task
Train station	Participant is waiting for a friend arriving by train . While waiting, the participant watches mobile TV and keeps an eye on the active timetable in order to meet the friend on time.
Bus	Participant takes a bus to the local library. The participant spends traveling time by watching mobile TV and keeps an eye on the environment in order to get off at the right stop.
Cafe	Participant spends time by watching mobile TV and needs to concentrate on the content despite the noise and fuss around.

4.5 Production of Test Material - Transmission Error Simulations

Four different transmission error rates, resulting to a varying number, length, and location of transmission errors were simulated. The preparation of the test stimuli was similar to previously reported studies [26],[27] and therefore only the key factors are summarized here.

The clips were encoded using Advanced Video Coding (H.264/AVC) and Advanced Audio Coding (AAC) as recommended for the IP data casting service over DVB-H [7],[8],[17],[18]. The video was coded at a bitrate of 128 kbps and a frame rate of 12.5 frames per second and the audio bitrate was 32 kbps with a sampling rate of 16 kHz as monaural [24]. At least one Instantaneous Decoding Refresh (IDR) frame was inserted per time-sliced transmission burst to decrease the tuning-in delay at the receiver and improve error resiliency. The

protocol stack of DVB-H was applied conventionally. The transmission burst interval was set to approximately 1.5 seconds, and a code rate of ¾ was used for the Multiprotocol Encapsulation – Forward Error Correction method (MPE-FEC). [6].

The simulation of the DVB-H channel was done with a Gilbert-Elliot model that was trained according to a field trial carried out in an urban setting with an operable DVB-H system. Four rates (1.7%, 6.9%, 13.8%, 20.7%) for erroneous time-sliced bursts after FEC decoding (known as MPE-FEC frame error ratio, MFER) were chosen into the simulations. The duration of erroneous audio-visual content was therefore approximately 1, 4, 8, and 12 seconds for the MFER values 1.7%, 6.9%, 13.8%, and 20.7%, respectively. It is noted that these residual error rates do not represent typical DVB-H reception, but rather are examples of extremely harsh radio conditions. Such severe radio conditions were selected for the test to discover the threshold between acceptable and unacceptable quality.

Simple video error concealment was used: when a picture was lost, all subsequent pictures were replaced by the last correctly received picture in the presentation order until the arrival of the next IDR picture. Thus, errors in video were perceived as discontinuous motion. The lost audio frames were replaced by silence resulting to perceived gaps during the playback.

4.6 Presentation of Test Materials

The clips were viewed on a Nokia 6630 handset. Headphones were used for audio playback not to disturb the people in the close proximity [39] and participants were free to adjust the level of loudness from the starting level of 75dBA. Four stimuli of each content formed a story and they were presented after each other in the chronological order. The order of simulated error rations were randomized between the stimuli within the story. In addition, the starting context was randomized.

4.7 Data-analysis Methods

Different methods of analysis were used. McNemar's test was used in the analysis of the nominal acceptance evaluations to test the differences between two categories in the related data [4]. Data of satisfaction and entertainment evaluations were normally distributed (Kolmogorov-Smirnov $p>.05$) allowing the use of parametric methods. To examine the main effects and interactions for satisfaction and entertainment we built a 3x4x4 (number of contexts x error rates x content types) design to be analyzed with repeated measures of ANOVA. Repeated measures of ANOVA can be used to compare differences between three or more conditions for interval data in related design [4]. Data of information recognition was not normally distributed (Kolmogorov-Smirnov $p<.05$) inferring a non-parametric method. As a non-parametric equivalent for repeated measures ANOVA, Friedman test was used to test to differences of several conditions of related ordinal data and Wilcoxon matched pair signed rank test was used to measure the differences between two related data sets as suggested in [4]. To analyze the connections between different measures, Pearson's correlation as a parametric method for interval data and Spearman correlation as a non-parametric method for ordinal data were used [4]. To compare acceptance ratings between studies we used Chi-square test, which is applicable when testing the frequencies between categories in independent measures. In comparisons of satisfaction ratings between studies we used independent samples t-test, exploitable when two different samples are tested with a same task but in different conditions [4].

In statistical tests, the significance level of $p<0.05$ was adopted.

5. Results

This section presents the results from impact of different usage contexts, error ratios and contents on measured dimensions of acceptance, satisfaction, entertainment and information recognition. In addition correlation between these methods is presented. Results of this study are also compared to those of previous studies.

5.1 Context

Acceptance: There were no differences in acceptance evaluations between contexts ($p>.05$).

Satisfaction: Context did not impact on satisfaction ratings ($F_{2,56}=0,53$ $p>.05$, ($p=.59$, ns), $\eta 2=.02$).

Entertainment: Contexts impacted on entertainment evaluations ($F_{2,56}=5,61$ $p<.01$, $\eta 2=.17$: Figure 3). Quality was assessed as less entertaining in the bus environment ($p<.01$) compared to other environments which were equally evaluated ($p>.05$).

Information recognition: Contexts impacted on information recognition scores ($F_r=19,67$, df=2, $p<.001$; Figure 4). The level of information recognition was higher in the café than other contexts ($p<.01$). The bus and station were equally evaluated ($p>.05$).

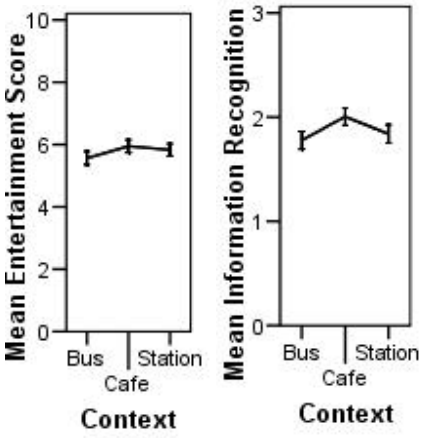

Figure 3, Figure 4 Mean information recognition and entertainment scores for different contexts. Bars show 95% CI of mean.

5.2 Error ratio

Acceptance: There was a significant difference ($p<.001$) between error rates 1.7%, 6.9%, 13.8% and 20.6% in descending order of acceptance except that 13.8% and 20.6% were equally evaluated ($p>.05$; Figure 5).

There were some content dependant variations in acceptance ratings (Figure 6). In cartoons and news, two lowest error rates were the most acceptable (p>.05) differing significantly from two highest error rates (p<.001) which were also rated into the same level (p>.05). For music video, the differences were significant between all error rates (p<.05). For sport content, two highest error rates were the most unacceptable (p>.05) differing significantly from others (p<.01). It is worth of noticing that the acceptability ratings of sport content (13.8%, 20.6%) and music video (20.6%) are the only stimuli which were rated clearly below 50% boundary between unacceptable and acceptable ratings.

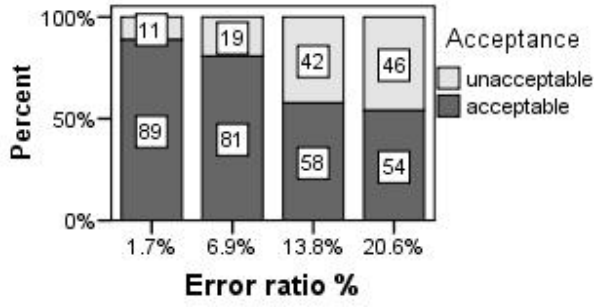

Figure 5 Acceptability ratings for different error ratios.

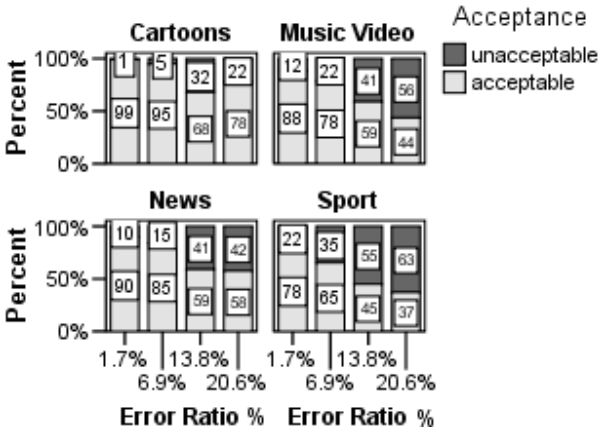

Figure 6 Acceptability ratings for different error ratios and contents.

Satisfaction: Error rates had a significant impact on satisfaction evaluations ($F_{2,57}$=112,0 p<.001, ŋ2=.80; Figure 7) and they interacted significantly with evaluations of content ($F_{5,149}$=2,63 p<.05, ŋ2=.087; Figure 9). There was a significant difference (p<.001) between error rates 1.7%, 6.9%, 13.8% and 20.6% in descending order of satisfaction except for the pair of 13.8% and 20.6% (p>.05). There were no content dependant variations in satisfaction.

Entertainment: Error rates impacted on entertainment evaluations ($F_{2,46}$=15,2 p<.001, ŋ2=.35; Figure 8) and interacted as well with content ($F_{5,153}$=6,36 p<.001, ŋ2=.19; Figure 10). Two lowest error rates were the most entertaining (p>.05) and two highest error rates the least entertaining (p>.05). The differences were significant between these groups (p<.01).

A more detailed analysis revealed the following dependencies in entertainment evaluations. Cartoons and music video were the most entertaining pieces of content, and in addition cartoons were more entertaining than any other piece of content independently on error rates. In cartoons, there were no difference between error rates 6.9% and 13.8% (p>.05) with a difference to others (p>.001). In music video, 6.9% error rate was the most entertaining with a significant difference to others (p<.05). Two least entertaining contents were news and sport. In news, the highest error rates were the least entertaining and equally rated (p>.05, vs. others p<.05). In sport, the two lowest error rates (p>.05) were the most entertaining with a significant difference to two highest error rates (p<.05).

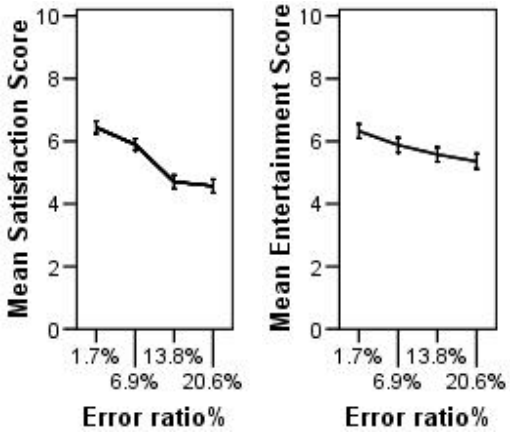

Figure 7 & Figure 8: Mean satisfaction and entertainment scores for different error ratios. Bars show 95% CI of mean.

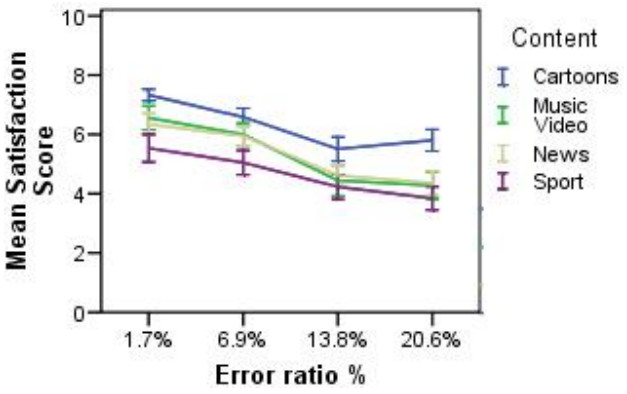

Figure 9 Mean satisfaction scores for different error ratios and contents. Bars show 95% CI of mean.

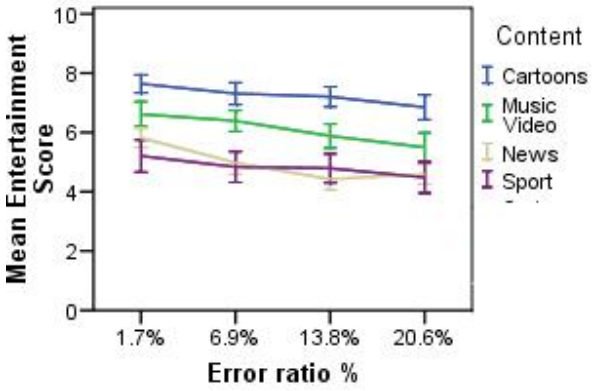

Figure 10 Mean entertainment scores for different error ratios and contents. Bars show 95% CI of mean.

Information recognition: Error rates impacted on information recognition scores (F_R=23,0, df=3, p<.001;Figure 11). The three least erroneous ratios resulted equally high level of information recognition (p>.05). In contrast, the highest error ratio collected lowest information recognition scores (p<.01). There were also some variations between contents for information recognition (Figure 12). For cartoons, the error rate of 6.9% was the hardest for information recognition with significant difference to others (p<.01) while all others were equally scored (p>.05). For music video presentation, two lowest error rates were easiest (p>.05) for information recognition compared to two highest error rates (p>.05; difference p<.05), and, for news, error rates 6.9% and 13.8% are in the same level (p>.05) with a difference to others (p<.05). Finally, sport content with error rate 20.6% is clearly worse than others (p<.05).

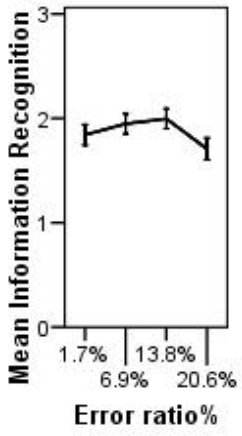

Figure 11 Mean information recognition scores for different error ratios. Error bars show 95 CI of mean.

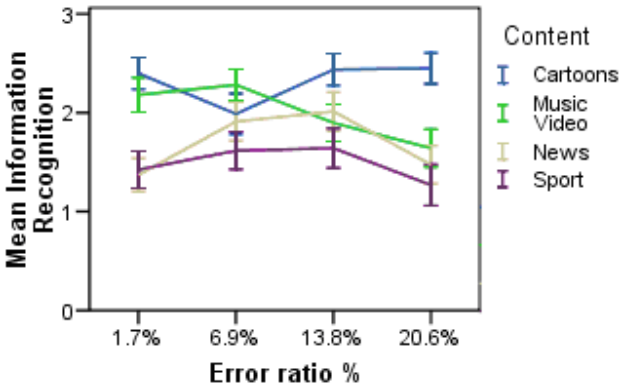

Figure 12 Mean information recognition scores for different error ratios and contents. Error bars show 95% CI of mean.

5.3 Correlations between satisfaction, entertainment and information recognition

To evaluate the connections between the measures, correlations were computed. Evaluations of satisfaction and entertainment were positively linearly correlated (r=.48 p<.001). In contrast, satisfaction (r_S=.07 p<.05) and entertainment (r_S=.18 p<.001) scores were not correlated to information recognition scores.

5.4 Context vs. Laboratory

Data from all contexts and a previous laboratory experiment were combined for analysis [27]. The previous controlled laboratory quality evaluation experiment was conducted with 30 participants. Sample selection criteria were the same in both studies, as far as the range of age, gender, technology attitude, and numbers of professional evaluators were concerned. In the laboratory experiment, a one-minute-long stimuli material originated from the same genre or program, contained similar audiovisual characteristics, and were processed with the same error rate simulations as in the contextual study. The stimuli were played on the same mobile device and headphones in both experiments. In the laboratory experiment, the stimuli were also viewed one by one and assessed retrospectively and independently [21]. The participants rated acceptance of quality on a nominal yes/no scale and quality satisfaction on an unlabelled 11-point scale after each stimulus. Any other questions were not presented. In the context study, in addition to acceptance and satisfaction ratings, entertainment and information recognition was judged. To estimate comparability between studies due to different assessment tasks, we relied on Hands' [15] study of multimodal quality perception when dividing attention between content and quality evaluation task. The results showed that simultaneous content recall task does not impact on quality assessment task in quality evaluation experiments. Hands conducted an experiment with several audiovisual impairments under the multimodal television quality assessment. Based on the conclusion that quality ratings are independent of content recall tasks, we can compare acceptance and satisfaction ratings between context and laboratory experiments. In the analysis, data from different contexts were integrated to represent the contextual data.

Acceptance - In the context study, 71% of all stimuli were experienced as acceptable, whereas only 48% of all stimuli were rated as acceptable in the laboratory experiment (χ^2 = 139.4, df=1, p<.001:Figure 13). This tendency appeared in all four error rates (p<.01).

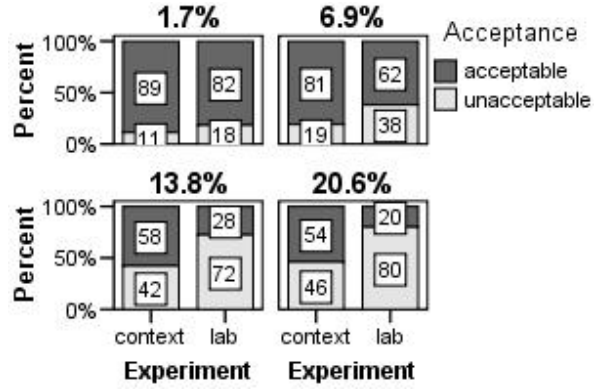

Figure 13 Comparison of acceptance ratings of four error rates between context and laboratory experiments.

Satisfaction - There were no difference between overall satisfaction ratings between the studies when averaging over other variables (t(2874)=-.327 p>.05, ns, (p=.74); Figure 14). However, a more detailed analysis revealed that the differences between error rates were smaller in the context test than in the laboratory study and there were also differences between the satisfaction evaluations in three error rates [Figure 14]. In 1.7% error rate, ratings given in the context test were lower than in the laboratory test (t(717)=7.90, p<.001), but error rate 6.7% did not reveal any difference between experiments (t(717)=6.48, p>.05 (p=.19, ns). Two highest error rates (13.8% and 20.6%) showed higher ratings in the context test than in the laboratory experiment (13.8%: t(717)=-2.70, p<.01; 20.6%: t(717)=-6.02 p<.001).

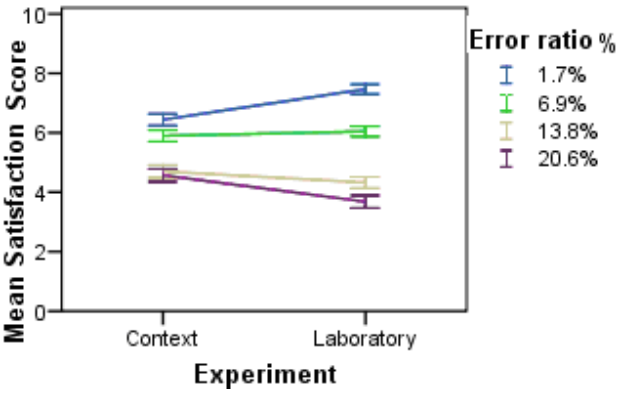

Figure 14 Comparison of satisfaction ratings of four error rates between context and laboratory experiments. Error bars show 95% CI of mean.

6. Discussion and conclusions

Subjective or perceptual quality evaluation tests are often conducted to optimize the reconstructed audiovisual quality. As these experiments take a place in a controlled laboratory environment, their sufficiency for optimizing compression or transmission parameters for mobile consumer services, to be used in heterogeneous environments, was questioned in this paper. The goal of our study was three-fold. First, we presented a novel user-oriented quality evaluation method for mobile TV to be used in real usage contexts. Second, we conducted an experiment with 30 participants including four different residual transmission error rates with television materials in three different contexts. Third, we compared the results of the conducted experiment with a previous laboratory experiment.

The quality evaluation was conducted under three different tasks and environments following selected typical mobile TV usage situations: waiting or killing time task at the railway station, relaxing in a café, and taking a local bus to the predefined location including transitions by foot. Even though the complexity of tasks and the nature of environments varied, they did not cause a difference in evaluations of acceptance or satisfaction of quality indicating that distinctions in reconstructed quality were clear enough to be separated in all contexts. When it comes to the goals of watching mobile TV, the entertainment evaluations were the lowest in the bus context and information assimilation the highest in the café context. The café context may provide the calmest and most pleasant environment for focusing on viewing, explaining the improved entertainment and information recognition evaluations. In the bus context, people may focus their attention on evaluation, but the complicated task on the move results to a generally unpleasant experience of entertainment. Previous work, conducted in a laboratory environment, emphasizes that simple dual tasks like navigation parallel to quality evaluation does not impact on quality evaluations itself, but even people themselves report tasks being more complicated or unpleasant [38]. It may also be possible that the bus context was not appropriate enough for mobile TV viewing, as it contained two short trips lasting 10 minutes each and walking and waiting on the bus stop in between. Recent studies, published during our experiments, have pointed out that people would not watch mobile TV during very short journeys, but rather they would change media in transitions between environments, and not view moving images on the move [36],[37]. Further work needs to be done to analyze and report the collected interview data of this study to confirm these suggestions as a base for improvements for context selection.

The studied error rates impacted on evaluation in a way that the least erroneous streams provided the highest quality in all measured dimensions. Highly erroneous ratios (13.8% and 20.6%) gave equally low acceptability, satisfaction and entertainment levels. In contrast, only the most corrupted presentation lowered the ability to assimilate information. These results confirm that very erroneous presentation with several noticeable cuts in both modalities not only impacts on quality but also prohibits reaching the goals of viewing.

The comparison between the context study and the laboratory study revealed that people accepted higher transmission error rates in real context. Overall acceptability ratings between the

studies differed and only few presentations were rated as unacceptable in the context study in contrast to the laboratory study in which all presentations at error rates 13.8% and 20.6% were unacceptable. In general, satisfaction was equally rated between the studies, but the evaluated differences between the error rates were smaller in the laboratory experiment. These results indicate that the evaluations were more favorable and less discriminate in the mobile contexts compared to the laboratory. In addition, the results imply that the subjective quality requirements drawn from laboratory environment might be higher than needed in actual contexts, which is also supported by a previous usability study [28].

To conclude, this study described a novel user-oriented quality evaluation method for mobile TV in its expected usage contexts. The study also highlighted that there is a difference between the quality requirements derived from laboratory and context studies. In the long term, our goal is to introduce an exhaustive user-oriented quality evaluation scheme that can also be applied in quality optimization of future services, such as in mobile three-dimensional television.

7. ACKNOWLEDGMENTS

We would like to thank RTT Oy (http://www.rtt.tv) for financial support and supervision of the work. The work of the first author is supported by the Graduate School in User-Centered Information Technology (UCIT) and this publication preparation work by HPY's Research Foundation.

8. REFERENCES

[1] Carlsson, C., Walden, P. Mobile TV - To Live or Die by Content, Proc 40th HICSS (2007) 51b.

[2] Casey, B., Casey, N., Calvert, B., French, L., Justin, L. Television Studies – The Key Concepts. Routledge. London, 2002.

[3] Chen, S. Y., Ghinea, G., Macredie, R. D. A Cognitive Approach to User Perception of Multimedia Quality: An Empirical Investigation. International Journal of Human-Computer Studies, 64(12), (2006)1200-1213.

[4] Coolican, H. Research methods and statistics in psychology ,4th ed, London: J. W. Arrowsmith Ltd, 2004.

[5] DVB-H Global Mobile TV: Services, Trials & Pilots. http://www.dvb-h.com/services.htm

[6] ETSI, Digital Video Broadcasting (DVB): DVB specification for data broadcasting, ETSI standard, EN 301 192 V1.4.1, 2004.

[7] ETSI, Digital Video Broadcasting (DVB): Transmission systems for handheld terminals, ETSI standard, EN 302 304 V1.1.1, 2004.

[8] ETSI, Digital Video Broadcasting (DVB); Specification for the use of video and audio coding in DVB services delivered directly over IP, ETSI standard, ETSI TS 102 005 V1.2.0, 2005.

[9] Finnpanel. www.finnpanel.fi, visited 20.5.2006.

[10] Ghinea, G & Chen, S. Y. The impact of cognitive styles on perceptual distributed multimedia quality. British Journal of Educational Technology. Vol 34,4 (2003). 393-406.

[11] Ghinea, G. & Thomas, J. P. QoS impact user perception and understanding of multimedia video clips, Proc. of ACM Multimedia '98". (1998) 49-54

[12] Goldstein, E. B. Sensation and Perception. United States of America: Wadsworth, 2002.

[13] Gulliver, G., Ghinea, G. Stars in Their Eyes, What Eye-Tracking Reveals About Multimedia Perceptual Quality. IEEE Transactions on Systems, Man, and Cybernetics, Part A: Systems and Humans. Vol. 34,4(2004) 472- 482

[14] Gulliver, S.R., Ghinea, G. Defining User Perception of Distributed Multimedia Quality. ACM Transactions on Multimedia Computing, Communications and Applications, 2(4), (2006) 241-257

[15] Hands, D. "Multimodal Quality Perception: The Effects of Attending to Content on Subjective Quality Ratings". Proceedings of IEEE 3rd Workshop on Multimedia Signal Processing, 1999 pp. 503-508, Copenhagen, Denmark 1999.

[16] Hannuksela, M.M., Malamal Vadakital, V.K., Jumisko-Pyykkö, S. Comparison of Error Protection Methods for Audio-Video Broadcast over DVB-H, EURASIP Journal on Advances in Signal Processing, vol. 2007, 2007. doi:10.1155/2007/71801

[17] ISO/IEC 14496–10:2003, Coding of Audiovisual Objects—Part 10: Advanced Video Coding," 2003, also ITU-T Recommendation H.264. Advanced video coding for generic audiovisual services. 2003.

[18] ISO/IEC 14496-3: Information technology -- Coding of audio-visual objects - Part 3: Audio, including amendment 1: "Bandwidth Extension" and amendment 2: "Parametric Coding for High Quality Audio". 2003.

[19] ISO 13407 Human-centred design processes for interactive systems. International Standard, the International Organization for Standardization. 1999.

[20] ITU-R BT.500-11 Methodology for the subjective assessment of the quality of television pictures, International Telecommunications Union – Radiocommunication sector, 2002.

[21] ITU-T P.911 Recommendation P.911, Subjective audiovisual quality assessment methods for multimedia application, International Telecommunication Union – Telecommunication sector, 1998.

[22] Jumisko, S., Ilvonen, V., Väänänen-Vainio-Mattila, K. The Effect of TV Content in Subjective Assessment of Video Quality on Mobile Devices. Proc. Multimedia on Mobile Devices, IS&T/SPIE Symposium on Electronic Imaging, (2005), 243-254.

[23] Jumisko-Pyykkö, S."I would like to see the subtitles and the face or at least hear the voice": Effects of Picture ratio and Audio-video Bitrate Ratio on Perception of Quality in Mobile Television. Personalized and Mobile Digital TV Applications in Springer Multimedia Tools and Applications Series. 2007.

[24] Jumisko-Pyykkö, S., Häkkinen, J. Evaluation of Subjective Video Quality on Mobile Devices. Proc. ACM Multimedia (2005), 535-538

[25] Jumisko-Pyykkö, S. Häkkinen, J., Nyman, G. Experienced Quality Factors - Qualitative Evaluation Approach to Audiovisual Quality. Proc IST/SPIE conference Electronic Imaging, Multimedia on Mobile Devices 2007

[26] Jumisko-Pyykkö, S., Vadakital, V., Liinasuo, M. , Hannuksela M. M., 2006. Acceptance of Audiovisual Quality in Erroneous Television Sequences over a DVB-H Channel. Proc. VPQM, USA, January 2006.

[27] Jumisko-Pyykkö, S., Vinod Kumar, M. V., Korhonen, J. Unacceptability of Instantaneous Errors in Mobile Television: From Annoying Audio to Video. Proc. Mobile HCI 2006, ACM Press (2006), 1-8.

[28] Kaikkonen, A., Kekäläinen, A., Cankar, M., Kallio, T., Kankainen, A. Usability Testing of Mobile Applications: A Comparison between Laboratory and Field Testing. Journal of Usability Studies, Vol. 1 (1), (2005)4-17

[29] Kjeldskov, J., Stage, J., New techniques for usability evaluation of mobile systems, International Journal of Human-Computer Studies, Vol 60,5- 6 (2004) 599-620.

[30] Knoche, H.O., McCarthy, J.D.: Good news for mobile TV, Wireless world research forum, Proc. WWRF14, 2005.

[31] Knoche, H., McCarthy, J. D., Sasse, M. A. Can Small Be Beautiful? Assessing Image Size Requirements for Mobile TV. Proc. ACM Multimedia 2005, (2005),561-

[32] Knoche, H., Sasse, M. A. Breaking the news on mobile TV: user requirements of a popular mobile content. Proc IS&T/SPIE Symposium on Electronic Imaging, 2006

[33] Knoche, H., McCarthy, J., Sasse, M. A. How low can you go? The effect of low resolutions on shot types. Personalized and Mobile Digital TV Applications in Springer Multimedia Tools and Applications Series (2007)

[34] Lang, A. The limited capacity model of mediated message processing. Journal of Communication, (2000) 50, 46-70.

[35] McCarthy, J. D., Sasse M. A. and Miras D. Sharp or Smooth?: Comparing the Effect of Quantization vs. Framerate for Streamed Video. Proc. CHI 2004, 535-542.

[36] O'Hara, K., Mitchell, A. S., and Vorbau, A. Consuming video on mobile devices. Proc. CHI '07. ACM Press (2007), 857-866.

[37] Oksman, V., Noppari,., Tammela, A., Mäkinen, M., Ollikainen, V., News in mobiles. Comparing text, audio and video. VTT 2007. http://www.vtt.fi/inf/pdf/tiedotteet/2007/T2375.pdf

[38] Reiter, U. & Jumisko-Pyykkö S. Watch, Press and Catch - Impact of Divided Attention on Requirements of Audiovisual Quality. Proc. 12th HCI Int 2007.

[39] Repo, P. Hyvonen, K. Pantzar, M. Timonen, P. Inventing Use for a Novel Mobile Service. International Journal of Technology and Human Interaction, vol. 2, 2 (2006) 49-62.

[40] Rogers E. M. Diffusion of Innovations, 5th ed., New York: Free Press, 2003.

[41] Södergård C. (ed.). Mobile television – technology and user experiences, Report on the Mobile –TV Project. Espoo: VTT Publications 506, 2003.

[42] Tamminen, S. , Oulasvirta, A. Toiskallio, K., Kankainen, A. Understanding mobile contexts, Personal and Ubiquitous Computing, vol 8,2 (2004) 135-143.

[43] Watson, A., Sasse, M. A. Measuring Perceived Quality of Speech and Video in Multimedia Conferencing Applications. Proc. ACM multimedia 1998. 55-60.

[44] Winkler, S., Faller, C. Perceived audiovisual quality of low-bitrate multimedia content. IEEE Transactions on Multimedia, vol. 8, no. 5, (2006) 973-980.

[45] Wynekoop, J. L., Russo, N., L., Studying system development methodologies: an examination of research methods. Information Systems Journal 7,1(1997)47–65.