Hindawi Publishing Corporation
International Journal of Digital Multimedia Broadcasting
Volume 2008, Article ID 712380, 20 pages
doi:10.1155/2008/712380

Research Article

Acceptance Threshold: A Bidimensional Research Method for User-Oriented Quality Evaluation Studies

S. Jumisko-Pyykkö,[1] V. K. Malamal Vadakital,[2] and M. M. Hannuksela[2]

[1] *Tampere University of Technology, Human-Centered Technology, P.O. Box 553, 33101 Tampere, Finland*
[2] *Nokia Research Center, P.O. Box 1000, 33721 Tampere, Finland*

Correspondence should be addressed to S. Jumisko-Pyykkö, satu.jumisko-pyykko@tut.fi

Received 5 March 2008; Accepted 17 July 2008

Recommended by Harald Kosch

Subjective quality evaluation is widely used to optimize system performance as a part of end-products. It is often desirable to know whether a certain system performance is acceptable, that is, whether the system reaches the minimum level to satisfy user expectations and needs. The goal of this paper is to examine research methods for assessing overall acceptance of quality in subjective quality evaluation methods. We conducted three experiments to develop our methodology and test its validity under heterogeneous stimuli in the context of mobile television. The first experiment examined the possibilities of using a simplified continuous assessment method for assessing overall acceptability. The second experiment explored the boundary between acceptable and unacceptable quality when the stimuli had clearly detectable differences. The third experiment compared the perceived quality impacts of small differences between the stimuli close to the threshold of acceptability. On the basis of our results, we recommend using a bidimensional retrospective measure combining acceptance and satisfaction in consumer-/user-oriented quality evaluation experiments.

1. INTRODUCTION

Consumer acceptance is a critical factor in the adoption of new mobile multimedia products and services. Acceptance is defined as the minimum level of user requirements that fulfills user expectations and needs as a part of user experience [1, 2]. User experience as a broad concept refers to a consequence of user's internal state, characteristics of designed system, and the context within interaction occurs [3]. Modern mobile services are collective results of several product elements and combine the effort of several players in a field from content owners, producers, and service providers to platform developers [4]. In the product development process, the quality of critical components is adjusted or optimized separately from the end-product or prior to the completion of the end-product. For example, in streamed mobile multimedia, the quality of network connection may represent one of these elements. To ensure that qualities of components developed in isolation are not barriers to the adoption of end-products, their acceptability should be studied in their optimization process.

In the development of signal or system quality as product components, subjective quality evaluation experiments are conducted. Subjective quality evaluation, also called perceptual, affective, or experienced quality evaluation, or even more broadly referred to as sensorial studies, is based on human judgments of various aspects of experienced material based on perceptual processes [5–7]. For the consumer-oriented critical product component assessment, an overall quality evaluation approach is appropriate. It is suitable for the evaluation of multimodal and heterogeneous stimuli [5, 7], and also assumes that human knowledge, expectations, emotions, and attitudes are integrated into quality perception [5, 7]. The overall evaluation approach has been applied in subjective quality evaluations of mobile television to study different codecs, audio-video compression parameters such as frame rates, bitrates, and screen sizes [8–10].

Subjective overall quality is mainly measured as an affective degree-of-liking, whereas only little attention has been paid to acceptance of quality. Subjective quality is usually measured as one-dimensional satisfaction based on

the methodological recommendations of the International Telecommunication Union [11]. Recently, the Quality of Perception (QoP) model has been proposed to combine two dimensions, namely, satisfaction and cognitive information assimilation, into one measure of subjective quality [12, 13]. However, these methods have not paid any attention to acceptance of quality. There are only few studies in which measures of acceptance have been reported [14]. However, no extensive theoretical background has been presented. Furthermore, these methods are applicable only when the quality is close to the acceptance threshold, and are not discriminative above or below the acceptance threshold, that is, the methods cannot be applied for the comparison of good qualities. These approaches necessitate changing the data-collection method for the duration of quality evolution. In sum, there is a clear need to develop an overall quality evaluation method of acceptance to ensure fulfillment of user minimum quality requirements in quality optimization and to provide comparability between studies independently of levels of quality under continuous technical development.

The aim of this paper is to develop research methods for assessing overall acceptance of quality. We present a literature review of acceptability and research methods as a basis for development in Sections 2 and 3. We conduct three experiments to develop and test validity under heterogeneous stimuli in the context of mobile television. The first experiment examines the possibilities of using a simplified continuous assessment method for assessing overall acceptability. The second experiment explores the perceived boundary between acceptable and unacceptable quality in four error rates having clearly detectable differences between stimuli. The third experiment compares the impacts of four different error control methods on perceived quality close to the threshold of acceptability with small differences between the stimuli. Finally, we present a discussion on all the experiments, provide recommendations for use of the methods, and conclude the study in Section 7.

2. MULTIMEDIA QUALITY

Multimedia quality is a combination of produced and perceived quality. Produced quality describes the technical factors of multimedia which can be categorized into three different abstraction levels, called network, media, and content [15, 16]. Perceived quality represents user's or consumer's side of multimedia quality, which is characterized by active perceptual processes, including low-level sensorial and high-level cognitive processes. A typical problem in multimedia quality studies is to optimize quality factors produced under strict technical constraints or resources with as little negative perceptual effects as possible.

2.1. Produced quality

Huge amounts of data, limited bandwidth, vulnerable transmission channel, and constraints of receiving devices set specific requirements for multimedia produced quality. Network-level quality factors describe data communication over a network and are often characterized by loss, delay, jitter, and bandwidth [15, 17, 18]. Network-level quality factors are discussed in greater detail in the subsequent paragraphs as they have a central role in this paper. Media-level issues include media coding for transport over the network and rendering on receiving terminals [15]. Recent studies on media-level quality factors have addressed the compression capability of codecs [19, 20], temporal factors in terms of video frame rates [13, 19], spatial resolution [9, 10], bitrates, spatial factors (e.g., monophonic or stereophonic sound), and temporal parameters of audio, such as sampling rate [20]. Increasing interest has been expressed in the topic of audio-video factors, like skew between audio and video streams [21] and shared resources between the streams, like bitrates [8, 9, 19, 20], and audiovisual transmission error control methods [22, 23]. The content level quality factors concern the communication of information from content production to viewers [15]. The topics studied include impacts of content manipulations [24], content comparisons (e.g., [8, 10, 13]), and text size [25]. High level of optimization, especially in the network and media levels, can cause noticeable degradation in perceived quality.

Network-level quality factors relate closely to imperfections of transmission channels. In fact, erroneous transmission of data may occasionally occur in any transmission channel. The causes of errors depend on the transmission channel and its characteristics. For example, in many wired-line networks, the main causes of errors are queue overflows at network nodes, while in a wireless network, the main cause of data corruption is due to the physical characteristics of the radio channel. Furthermore, the statistical characteristics of errors may also vary. They may be either isolated individual errors, burst errors, or a combination of both. Therefore, any methods to resolve errors in a transmission channel must take into consideration the cause of error as well as the nature of error that corrupts the data.

In wireless channels, the radio channel properties, such as interference from other cochannel signals, multipath propagation due to signal reflection from different natural, and man-made structures in the vicinity of the receivers, together with fading are the major causes of errors. If the receiver is a mobile terminal, errors may also occur due to the Doppler effect caused by the speed of the receiver. These errors typically occur as bursts rather than isolated individual errors [26, 27]. The nature, frequency, and duration of errors may vary regardless of the cause of errors.

Broadcast services typically fix transmission errors with forward error correction (FEC) coding, such as Reed-Solomon FEC codes [28]. FEC repair symbols are appended to the actual data such that when errors are encountered, the combination of the data and the FEC repair symbols can be used to obtain the correct data. The correction capability of FEC codes is limited, however, and once the number of transmission errors exceeds the correction capability of the FEC code, typically no lost data can be recovered. Consequently, the use of FEC codes causes an abrupt threshold between produced quality free of network-level errors and severely impaired quality due to transmission errors.

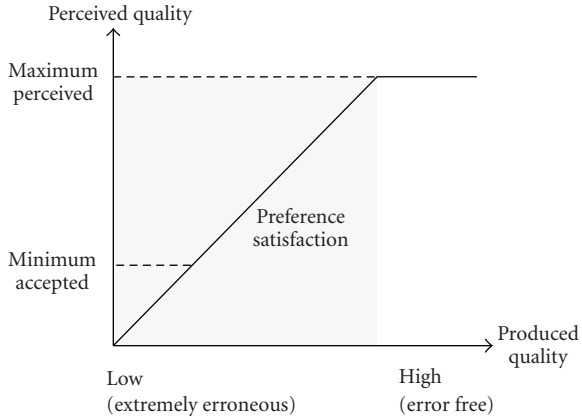

Figure 1: The levels of produced and perceived quality.

2.2. Perceived quality

Quality perception is constructed in an active process. Early sensory processing extracts relevant features from the incoming sensory information. In vision, brightness, form, color, stereoscopic, and motion information are distinguished in the early perceptual process while pitch, loudness, timbre, and location are attributes of auditory processing [29, 30]. However, the final quality judgment is always a combination of low-level sensorial and high-level cognitive processing. In cognitive processing, stimuli are interpreted through their personal meaning and relevance to human goal-oriented actions. This process involves individual emotions, knowledge, expectations, and schemas representing reality, which affect the importance of each sensory attribute and more broadly enable human contextual behavior and active quality interpretation [31–33]. For example, quality evaluations are not restricted to the characteristics of interpreted stimuli. The assessment of usefulness or fitness to purpose of use is included in human evaluations of quality [34].

2.3. Levels of produced and perceived quality

Multimedia quality can be presented as a relation between produced and perceived quality. We present this relation by applying basic conventions of psychophysics (originating from Fechner 1860 overview, e.g., [7, 35]), but widening the view to actual user quality requirements. The quality produced may have a wide range from low and extremely erroneous to extremely high fidelity and error-free presentation (Figure 1). However, the human perceptual processes cannot detect all levels of produced quality. In addition, the whole quality range is not appropriate for the consumer products.

When the produced quality is extremely high, the threshold of maximum perceived quality is reached. This means that an increase in produced quality does not improve the perceived quality since the differences in produced quality become undetectable and impossible to recognize. In psychophysics, this is called terminal threshold [7]. In consumer products, top-end multichannel audio or high-definition visual presentations may reach these thresholds under certain rendering constraints in the near future.

Below the maximum perceived quality, the levels of produced quality can be organized into orders of preference if the difference threshold between the stimuli is reached. Perceived quality at this stage represents satisfaction or pleasantness. Preferences can be compared until the stage, at which the decrease of produced quality no longer decreases, perceived quality. The lower edge of detection and recognition threshold is reached [7]. Produced quality that is close to lower thresholds is not appropriate for studying consumer products or services.

Discrimination testing is used to gather data on conventional thresholds. There are different types of discrimination tests and their further applications, such as method of limit, constant stimuli, and adjustment. Common to all of these methods is the binary data collection form. Either there is sensation or there is not "no sensation or yes, I perceive something" [7, 35] .

We assume that there are also other types of meaningful thresholds between those located at the extremes of perceived quality. When the produced quality approaches the level of very poor and erroneous presentation, there is the area of minimum acceptable quality within the perceptual preferences. The concept of minimum accepted quality can be expected to be relevant in system quality assessments for consumer electronics as an indicator of useful level of produced quality and as an anchor for user requirements. A more detailed conceptual presentation for acceptability is given in Section 3 from the perspectives of acceptance as technology adaptation and acceptance as sensorial experience.

3. ACCEPTANCE AND QUALITY EVALUATION METHODS

3.1. Technology acceptance—the wide audience approach

In the broadest sense, acceptability refers to the market decision whether to accept or reject products or services characterized by willingness to acquire the technology, use it, and pay for it [36, 37]. This approach is popular in the fields of consumer studies and human-computer interaction. In one of the most widespread theories, called the Technology Acceptance Model (TAM), factors predicting the intention to use information system and adoption behavior are formed [38, 39]. TAM was originally developed to measure the acceptance to use information systems for mandatory usage conditions, but later, it was adapted and modified for consumer products and mobile services (e.g., [40–42]).

In TAM, the main predictors of behavioral intention to use the tested technology are usefulness and ease of use. Usefulness refers to the degree to which a person believes that a certain system will help perform a certain task. Ease of use is defined as a belief that the use of the system will be relatively effortless. Low produced quality may be one of the obstacles in the acceptance of technology [38, 39]. In the context

of mobile multimedia, failures of produced quality factors, such as screen size and capacity, interface characteristics of mobile devices, wireless network coverage, as well as capabilities and efficiency of data transform [40, 42–44], may have indirect effects on usage intentions or behavior by affecting perceived usefulness and ease of use [38, 39]. From the broad viewpoint of acceptability, subjective quality evaluation experiments on certain techniques should ensure that perceptually minimum accepted quality level is reached for the developed information systems or services to be an enabler of wide audience technology adaptation.

3.2. Quality evaluation methods

Subjective quality evaluation experiments are conducted for signal or system development purposes. Information about these studies is used in the optimization of a system, like network or media parameters, or in the development of objective metrics. In the literature perceptual, hedonic, or experienced quality evaluation are typically used as synonyms for these measures depending on the different emphases [5–7]. These studies are conducted in a controlled environment to ensure a high-level of control over the tested variables and repeatability of measures. For consumer-oriented quality evaluation, overall quality judgments are used. Evaluations of excellence of stimuli are based on human perceptual processes. As the evaluations are based on human perception of the excellence of stimuli, knowledge, expectations, emotions, and attitudes are integrated into the final quality perception of stimuli [5, 7]. The overall quality evaluation can be used to evaluate heterogeneous stimuli material (e.g., multimedia) because it is not restricted to the assessment of a certain quality attribute, such as brightness, but rather based on a holistic view of quality [5].

There are three main approaches to evaluate subjective perceived overall quality which can be applied in the measures of relatively low produced multimedia quality. A summary of the essential properties of the methods is given in Table 1. The International Telecommunication Union Recommendation [11] provides a reliable research method called Absolute Category Rating (ACR), which is applicable for performance or system evaluations with a wide quality range [11]. In ACR, also known as the single-stimulus method, test sequences are presented one at a time and rated independently and retrospectively. The short stimuli materials and mean opinion score (MOS) using labeled scales to set the evaluations into order of preference in ACR. One of the ultimate aims of method development has been to create a very reliable subjective method providing comparable data for the construction of objective or instrumental multimedia quality evaluation metrics [11]. It is maybe not surprising that the method is especially widespread in engineering.

Quality of Perception (QoP) is a user-oriented concept and evaluation method combining different aspects of subjective quality introduced by Ghinea and Thomas [12, 13]. QoP is a sum 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 on audio, video, or text in different content and in the analysis right answers are transformed into the ratio of right answers per number of questions. Both satisfaction factors are assessed on a scale 0–5. Final QoP is the sum of information assimilation and satisfaction that sets the stimuli into order of preference. Both ARC and QoP result in subjective evaluations in the form of a preference order and can be applied in studies on low produced quality, but they are not restricted to it. However, these methods do not indicate any threshold of acceptance among these preferences.

McCarthy et al. [14] tackle the problem of quality acceptability on the basis of the classic Fechner psychophysical method of limit. The threshold of acceptance is achieved by gradually decreasing or increasing the intensity of the stimulus in discrete steps every 30 seconds. At the beginning of the test sequence, participants are asked if the quality is acceptable or unacceptable. While watching, participants evaluate quality continuously. They report the point of acceptable quality when quality of stimuli is increasing or the point of unacceptable quality when quality is decreasing. In the analysis, binary acceptance ratings are transformed into a ratio calculating the proportion of time during each 30-second period that quality was rated as acceptable. The results are expressed as acceptance percentage of time. This method is powerful when studying variables around the threshold but not those clearly below or above it [7].

The duration of stimuli differs between the three overall quality evaluation methods. The ACR recommends to use short stimuli (10 seconds). This approach pays attention to the constraints of the human working memory, which is about 20 seconds in duration and has limited capacity for units [45, 46], also, it assumes that it is possible to remember all impairments of a stimulus when assessing quality. In contrast, QoP and the method of limit use longer-lasting stimuli materials. They focus more on user and aim to maximize the ecological validity of the viewing task in the experiments and therefore stress less about an ability to remember each of the imperfections the stimulus had [12–14]. It is also worth mentioning that the use of short-stimuli material might be constrained by the measured phenomena, for example, they might fit for measuring compression, but not for transmission quality factors.

In contrast to the overall quality evaluation methods presented, there has been interest in studying instantaneous changes of real-time variation in quality. Originally, the method was developed to go beyond the limitations of the working memory and to enable the use of long material, even up to the duration of a full television program, for testing of time-varying image quality [49–51]. In continuous assessment, participants express their quality evaluation moving the slider on a graphical 5-point labeled MOS scale while watching the content. It has been used to assess the excellence of video and audiovisual quality [50–52]. Similarly to ACR and QoP, the acceptance threshold is hard to locate on this scale. Later, continuous monitoring has been reported to be too demanding evaluation task, especially for multimedia quality evaluation [52]. It may also impact on the natural strategy of human information processing [53].

Table 1: Overview to overall quality evaluation methods.

Method	ACR	QoP	Method of limit
Presentation	Single stimulus, Independently	Single stimulus, Independently	Continuous, gradually decreasing or increasing the intensity of the stimulus
Duration of stimuli	≤10 s	App. 30 s	210 s, quality changes every 30 s
Scales	5/9/11-point scales, MOS	Satisfaction (0–5) Enjoyment Objective quality Information assimilation: ratio of right answers	Binary acceptable/unacceptable
Applied	Audio-video bitrates, codecs, resolution, packet loss [8, 19, 20, 22, 23, 47]	Framerate, delay, jitter, devices, [12, 13, 17]	Framerate, quantization, audio-video bitrate, resolution, text quality, [9, 10, 14, 25, 48]

3.3. Acceptance evaluation

In most consumer-oriented quality evaluation or sensorial studies, acceptance represented refers to affective measurements and represents degree of liking. These measures are used to gather the subjective responses of potential customers or users to a product, product idea, or specific product characteristics [35]. Typically, acceptance is measured on an ordinal scale of overall preference of product or specific preference for a certain sensory attribute [35]. For example, in the context of video or audiovisual quality studies, Apteker et al. [54] and Wijesekera et al. [55] both used ordinal acceptance scales to study framerates whereas Steinmetz [56] studied acceptance of media synchronization on a nominal scale (acceptable, dislike, and annoying). When measuring acceptance as a degree of liking, it lacks of the same detail of threshold of acceptance as quality preferences derived from ACR methods. In contrast to the preference approach, there are only few studies by McCarthy et al. [14] and later Knoche et al. [9, 10, 25, 48] in which acceptance has been seen as a binary phenomenon representing the nature of conventional thresholds (Table 1). Apart from these few studies, acceptability has not typically been measured in the quality assessment of mobile multimedia.

Recent studies have assessed preferences of low produced qualities to optimize the quality of service parameters for mobile devices and networks. Most of the studies compare compression parameters, like low framerates, bitrates or audio-video bitrate share, modern codecs, small display size, and their interactions [8, 10, 19, 20, 57]. Impacts of transmission errors on perceived quality is less reported [58]or the studies focus on one media at a time [59, 60]. Independently of the source of impairments in produced quality, some of these studies compare extremely poor qualities [8, 9, 20] and, therefore, their feasibility can be questioned as follows. How relevant are comparisons of poorness of quality when evaluations are clearly targeted at consumer services? Where is the threshold of minimum accepted level in these preference evaluations?

This leads to the connections between acceptability and preference. As Jumisko-Pyykkö [8] has concluded earlier that "*to improve the connections between the quality preferences or pleasances to the real usage, the anchor of binary acceptability is necessary to…set parallel to quality preferences.*" This is important in quality evaluation studies comparing several parameters, media, and their interaction at the same time. Further, it becomes even more significant when studying the novel optimization problems derived from technology totally lacking previous knowledge about perceptual impacts of parameters. "*This (acceptability) would show the useful quality levels…and target the focus in this field to the meaningful and necessary parameter comparisons*" [8]. In the long term, the goal is to ensure that the produced quality is set in a way that constitutes no obstacle to the wide audience acceptance of a product or service.

For the sake of clarity, we call degree-of-liking or ordinal measured preference of quality satisfaction in this paper. Acceptance of quality refers to the binary measure to locate the threshold of minimum acceptable quality that fulfills user quality expectations and needs for a certain application or system.

4. EXPERIMENT 1

The first experiment had two goals. Firstly, the aim was to develop a new subjective quality evaluation method. Our main focus was on an assessment method for the overall evaluation of acceptance and satisfaction. We also wanted to develop a simplified continuous assessment method for instantaneous quality evaluations which would avoid the previously reported problems of conventional methods being too demanding [52, 53]. Secondly, we wanted to study the impact of simplified continuous assessment on retrospective evaluations between two samples.

4.1. Research method—test set-up

4.1.1. Participants

Two samples, each with 15 participants (equally stratified by age between 18–45 years and gender) conducted a study in a controlled laboratory environment. The samples contained mostly (80%) naïve or untrained participants. They had no previous experience of quality evaluation experiments, they were not experts in technical implementation, and they were not studying, working, or, otherwise, engaged in information technology or multimedia processing [11, 61]. In addition, they did not belong to any group of innovators and early adopters regarding their attitudes to technology [62].

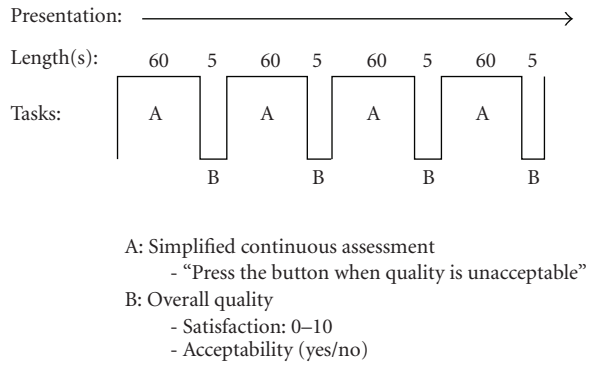

Figure 2: Experimental setup: simplified continuous assessment and retrospective ratings of quality and acceptance.

4.1.2. Test procedure

The test procedure was divided into pre-test, test, and post-test sessions. In the pre-test session, vision and hearing tests with demographic data collection took a place. All participants had normal or corrected-to-normal visual acuity (20/40) as well as normal color vision and hearing. In the combined training and anchoring, participants were shown the extremes of the sample qualities as examples of the quality scale and they became familiar with the contents and the evaluation task.

In the test, the test group evaluated quality with simplified continuous assessment parallel to retrospective ratings (Figure 2: Tasks A + B). The control group used only retrospective ratings (Figure 2: Task B). The sample material was shown using the Absolute Category Rating method where clips are viewed one by one and rated independently [11]. During the clip presentation, the test group used a simplified continuous assessment method in which instantaneous unacceptable quality was indicated by pressing a button on a game controller while viewing the content. After each clip, participants marked retrospectively the overall quality satisfaction score of a clip on an answer sheet using a discrete, unlabeled scale from 0 to 10 and the acceptance of quality (yes/no choice). 9 and 11-point scales are recommended over narrower scales because they compromise the end-avoidance-effect and problems of labeled scales [7]. The widely used labeled MOS scale was not used because it has been criticized for having unequal distances between the labels [49] and the meaning of these labels are not the same between cultures [63, 64]. Acceptance was measured on a binary scale imitating the measures of thresholds [7, 35].

The instructions for the quality evaluation tasks were as follows. For gathering the quality satisfaction score, the participants were asked to assess the overall quality of the presented clip. The measure of acceptance of quality was instructed by asking whether the participants would accept the overall quality presented if they were watching mobile television. No other evaluation criteria were given.

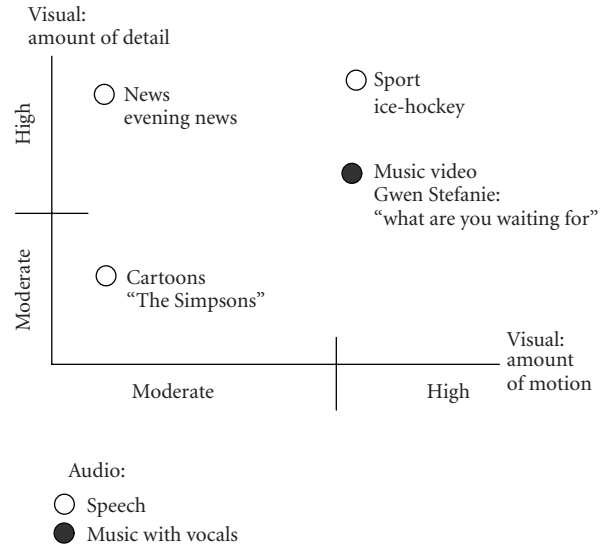

Figure 3: Genre of stimuli, contents, and their audiovisual characteristics.

The post-test session gathered qualitative data on experiences of erroneous streams. One test session lasted for about 1.5 hours.

4.1.3. Selection of test material

Four types of content, news, sport, music video, and animation were selected for test clips according to their potential for mobile television [48, 65, 66], popularity, and audiovisual characteristics (Figure 3). Each clip contained a meaningful segment of a TV program without cutting the start or end of a sentence, some textual information, several shots with different distances and angles to be representative of mobile television content.

The length of stimuli was approximately 60 seconds (61–63 seconds). The chosen duration enabled at least one impairment to appear with the lowest error rate. The use of shorter stimuli is recommended due to the limitations of human-working memory [45, 46], but with the chosen impairment rate, shorter stimuli would have been meaningless.

4.1.4. Network-level characteristics of mobile television

The target application for which the test was setup was mobile television. One of the most prominent standards for mobile television is the Digital Video Broadcasting-Handheld (DVB-H) standard [67], the characteristics of which are briefly reviewed in this section. DVB-H uses Internet Protocol (IP) packet encapsulation for datacasting. These IP packets are further encapsulated into User Datagram Protocol (UDP) packets, Real-Time Protocol (RTP) packets, and lastly Multi-Protocol Encapsulation (MPE) sections before being segmented into 188 byte (inclusive of 4 byte header) transport stream (TS) packets. DVB-H uses time-slicing for reducing power usage in receivers. The error

Table 2: Number of errors, mean durations, and standard deviation (in seconds) of burst errors for error patterns in different error rates.

Error rate		Error rate 1.7%		Error rate 6.9%	
Content		N	Mean(SD)	N	Mean(SD)
Cartoon	Audio	0–3	0.33(0.28)	3–6	0.37(0.20)
	Video	1	1.57(0.51)	3-4	1.06(0.54)
Music video	Audio	0–3	0.27(0.38)	3–7	0.70(0.17)
	Video	1	1.65(0.38)	2-3	1.21(0.43)
News	Audio	2	0.33(0.29)	2–6	0.38(0.20)
	Video	1	1.94(0.45)	2–4	1.08(0.35)
Sport	Audio	0–3	0.34(0.28)	4–6	0.34(0.21)
	Video	1-2	1.10(0.34)	2–4	1.06(0.44)
Error rate		Error rate 13.8%		Error rate 20.7%	
Cartoon	Audio	11–14	0.32(0.19)	9–22	0.30(0.15)
	Video	7-8	1.61(0.97)	13–15	1.31(0.75)
Music video	Audio	11–14	0.31(0.19)	9–22	0.31(0.19)
	Video	7–9	1.27(0.74)	12–15	1.27(0.75)
News	Audio	11–14	0.32(0.19)	9–22	0.30(0.15)
	Video	7–9	1.41(1.00)	11–13	1.40(0.99)
Sport	Audio	12–15	0.30(0.18)	13–22	0.30(0.14)
	Video	7-8	1.61(0.81)	11–14	1.50(0.90)

correction system of DVB-H, known as MPE-FEC, is based on Reed-Solomon FEC codes computed over the IP packets of a time-sliced burst of data [68].

4.1.5. *Production of test materials—transmission error simulations*

The test setup simulated DVB-H reception. The goal of the error simulations was to produce four detectable different transmission error rates with varying number, length, and location of errors. To achieve this goal, 6 pilot experiments were conducted to make a final decision about the final error rates. 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 for the simulations. 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.

The selected test materials were encoded using recommended codecs for IP datacasting over DVB-H [67]. Visual content was encoded using a baseline H.264/AVC encoder with the quarter common interchange format (QCIF), a bitrate of 128 kbps, and a frame rate of 12.5 frame per second [8, 19, 67, 69]. For audio encoding, Advanced Audio Coding (AAC) was used with a bitrate of 32 kbps and sampling rate of 16 kHz as monoaural. An Instantaneous Decoder Refresh (IDR) frame was inserted per time-sliced transmission burst to minimize tune-in delay to new receivers tuning in to the channel and to provide better error resilience under DVB-H channel error conditions. The protocol stack of DVB-H was applied conventionally. The length of transmission burst interval was set at approximately 1.5 seconds, and a code rate of 3/4 was used for MPE-FEC [70].

At the receiver, simple error concealment procedures were used. When a picture of video was lost, all subsequent pictures were replaced by the last correctly received picture in presentation order until the arrival of the next IDR picture. Thus, errors in video produced discontinuous motion. Similarly, the lost audio frames were replaced by silence, resulting in gaps during playback. The error characteristics are presented in Table 2.

4.1.6. *Presentation of test materials*

The experiments were conducted in a controlled laboratory environment [71]. The stimuli materials were viewed on a Nokia 6630 handset with a Nokia player. During the viewing, the device was enclosed in a stand and adjusted to eye level with a viewing distance of 44 cm [8]. For audio playback, headphones were used and the level of audio loudness was adjusted to 75 dBA.

A game controller (Logitech Dual Action gamepad) was used to instantaneously mark unacceptability in the simplified continuous evaluation. A logging program was run on a laptop (Fujitsu Simens Lifebook Pentium 3, Windows 2000) to collect the user input. The logging program run on Python 2.3.5 and used PyGame 1.6 module for accessing the game controller button events. When the button of the game controller was pressed, the program saved the number of seconds elapsing from the reference time at the beginning of the presentation. All clips were played three times in random order and the positions of the transmission errors varied in each repetition.

4.1.7. Method of analysis

Acceptance

To compare the acceptance ratings between the samples, we used Chi-square test, which is applicable to measure the differences of categorical data in independent measures [72].

Satisfaction

To compare the differences in satisfaction ratings between the samples, we used the Mann-Whitney U test as a nonparametric method (Kolmogorov-Smirnow: $P < .05$). The Mann-Whitney U test to measure differences between ordinal measured two independent samples [72]. A significance level of $P < .05$ was adopted in this study.

4.2. Results

We examined the effect of simplified continuous assessment on retrospective overall quality evaluation of acceptance and satisfaction. We compared the retrospective evaluations between the test group and the control group.

Acceptance

When the effects in all combined evaluations of acceptance were compared, the effect was not significant ($\chi^2 = .803$, df = 1, $P > .05$, nor was there any significant effect in the comparison samples in different error ratios ($P > .05$). Moreover, in the comparisons between the samples in each content and error ratio, there was no significant effect of continuous assessment on evaluation of acceptance in 15/16 cases ($P > .05$). The only exception appeared in the sport clip with error ratio 20.7 ($\chi^2 = 4.05$, df = 1, $P < .05$).

Satisfaction

There was no significant difference in the retrospective overall quality assessment of satisfaction. There was no significant effect in the comparison of all given evaluations ($U = 246999$ $P > .05$, ($P = .12$), nanoseconds), nor was the effect significant in the comparison of all error ratios ($P > .05$) or in the comparisons of each content in each error ratio between the two research methods ($P > .05$).

4.3. Discussion

The results showed that the simplified continuous assessment method did not affect the evaluations of retrospective acceptance and satisfaction between the studied samples. Earlier continuous assessment methods have been criticized for requiring a high level of involvement on the part of the evaluator and for possibly changing the way of information processing while evaluating quality [52, 53]. It is known that the difficulty, similarity, and practicing of tasks are the basic factors affecting performance of dual tasks [73]. Our study indicates that the simplified continuous assessment task developed is easy enough to be used parallel to retrospective evaluations without negative impact. Our results are also supported by Reiter and Jumisko-Pyykkö [74]. They concluded that while viewing the content, simple parallel tasks like pressing the button or catching the object, did not impact on the requirements of quality in audiovisual applications. Based on these results, we will use simplified continuous assessment in parallel with other methods to evaluate overall quality in different transmission simulations.

5. EXPERIMENT 2

To apply the developed overall quality evaluation methods, we used them to measure the impact of transmission errors. As in experiment 1, we assumed a mobile television usage scenario using the DVB-H standard. The goal of the experiment was to study the effect of four clearly detectably different residual transmission error rates on perceived quality. We aimed to locate the threshold between acceptable and unacceptable quality, examine the quality satisfaction, and also express acceptance percentage of time. In addition, we examined the relations between the results of these three different methods to evaluate their reliability.

5.1. Research method—test setup

5.1.1. Participants

30 participants, recruited according to the same criteria and meeting the same sensory requirements as in experiment 1, participated in the experiment.

5.1.2. Test procedure

The test procedure was identical to the test sample procedure in experiment 1 (Figure 2: Tasks A + B). The simplified continuous assessment was used parallel to retrospective ratings of acceptance and satisfaction.

Test materials, Test material production—transmission error simulations, and material presentation were identical to those in the experiment 1.

5.1.3. Method of analysis

Acceptance

McNemar's test was applied for the nominal retrospective acceptance evaluations to test the differences between two categories in the related data [72].

Satisfaction

Satisfaction data were analyzed using Friedman's test and Wilcoxon matched-pair signed-ranks test because the presumption of parametric methods (normality) was not met (Kolmogorov-Smirnow $P < .05$) [72]. Friedman's test is applicable to measure differences between several and Wilcoxon's test between two related and ordinal datasets [72].

Acceptance percentage of time

To formulate the data of simplified continuous assessment in the form of overall Acceptance percentage of time, nominal data was converted to a scale variable using the conversion introduced by McCarthy et al. [14].

$$(1 - (\text{unacceptable pressings/length of the clip})) * 100 \quad (1)$$

After the conversion, each of the stimuli was given a score showing the percentage of acceptable quality of stimuli presentation. Friedman and Wilcoxon's tests were then used in the actual analysis.

Relations between different measures

To analyze the connections between the different overall quality evaluation measures, Spearman's correlation as a nonparametric method for ordinal data was used and the Chi-square test of independence evaluated independence between distributions of two variables measured on a categorical scale [72].

5.2. Results

5.2.1. Acceptance

The results of acceptance measurements showed that error rates 1.7% and 6.9% of uncorrectable time-slices were experienced as giving acceptable subjective quality, while error rates of 13.8% and 20.7% were perceived as unacceptable. The differences between the error ratios were significant (All comparisons $P < .001$; Animation: 13.8% versus 20.7% $P < .05$) except the difference between the error rates 13.8% and 20.7% in the news, music video, and sport clips evaluations (Figure 4 $P > .05$, nanoseconds).

5.2.2. Satisfaction

In terms of satisfaction, the order of preference in all combined evaluations of error ratios was 1.7%, 6.9%, 13.8%, 20.7%. Error rates had a significant effect on quality scores ($F_R = 437.6$, df = 3, $P < .001$) and the differences between the error rates were significant ($P < .001$).

The preferred order of satisfaction was the same in the content-by-content examination but there were some variations in the pairwise comparisons of the highest error rates (Figure 5). Error rates had significant effect on all satisfaction evaluations in all contents (Animation: $F_R = 183.3$, df = 3, $P < .001$, Music video: $F_R = 145.2$, df = 3, $P < .001$, News: $F_R = 183.4$, df = 3, $P < .001$, Sport: $F_R = 203.6$, df = 3, $P < .001$). The evaluations differed significantly between all error rates in animation ($P < .001$), sport ($P < .001$), and music video content presentations (between 13.8% and 20.7% $P < .01$; all others $P < .001$). In the presentation of news content, the differences were significant ($P < .001$) excluding the ratios 13.8% and 20.7% ($P > .05$, nanoseconds).

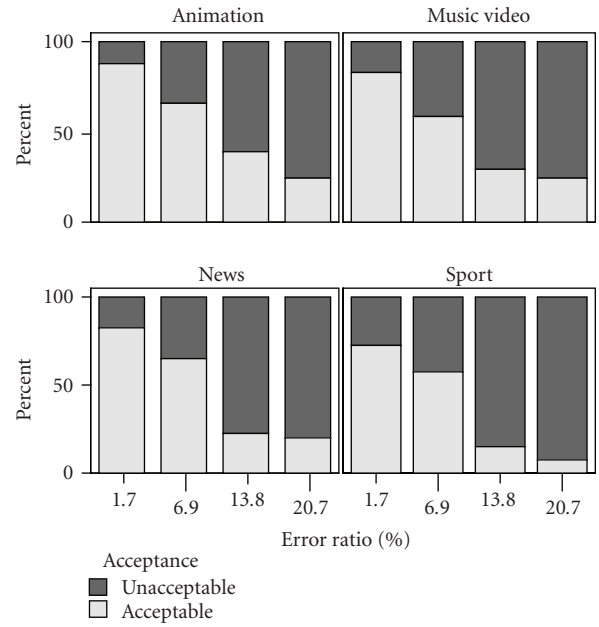

Figure 4: Acceptance of different error rates for all contents.

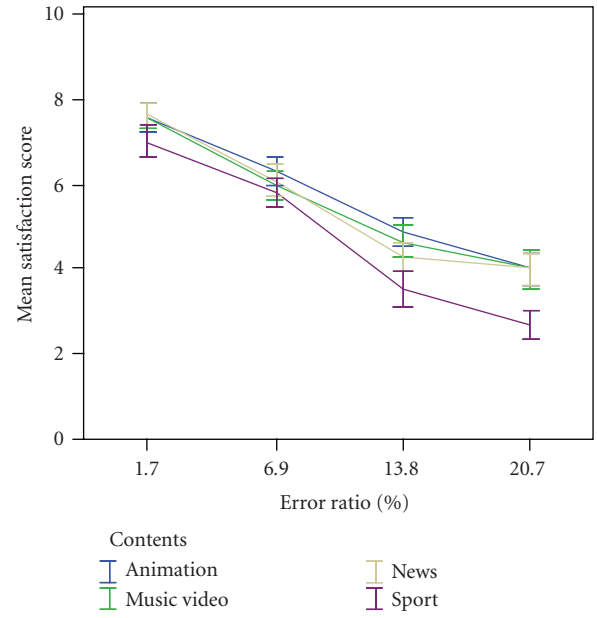

Figure 5: Mean satisfaction scores for all contents. Error bars show 95% CI of mean.

5.2.3. Acceptance percentage of time

Three outliers were removed from the data because they either expressed unacceptable quality very rarely during the presentation or they expressed it infinitely. Similar personal variation has also been expressed in the use of conventional continuous assessment [51].

The acceptance results based on a combination of continuous assessment were similar to the results of retrospective ratings.

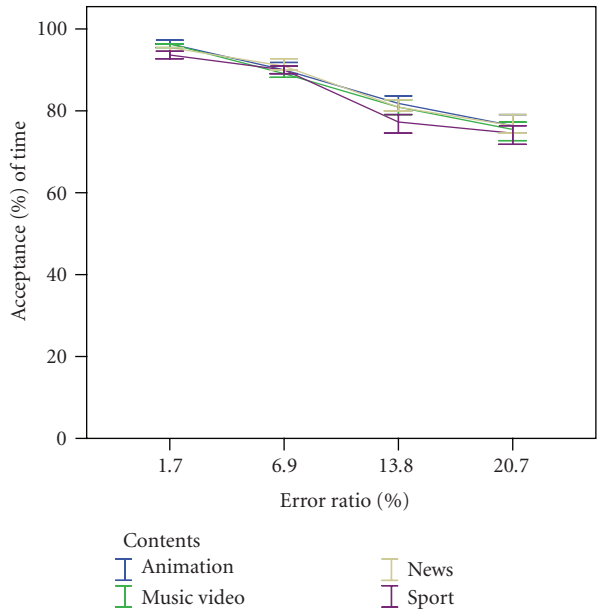

Figure 6: Acceptance percentage of time for all contents. Error bars show 95% CI of mean.

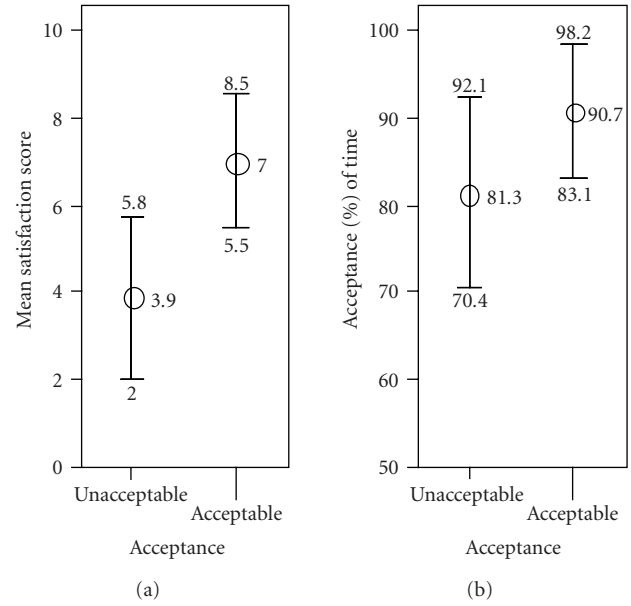

Figure 7: Relations on the scale between retrospective acceptance and satisfaction; and retrospective acceptance and acceptance based on continuous assessment. Bars show mean and standard deviation.

The lowest error rate 1.7% gave acceptable viewing experience for approximately 95% of the time whereas the highest error rate gave the acceptable experience only approximately 75% of the time (Figure 6). The acceptance evaluations were significantly affected by the error rates (F_R = 774.4, df = 3, $P < .001$) and the evaluations differed significantly between all tested error rates ($P < .001$). The effects of different error rates were similar to the combined evaluations in content-by-content examination. In the animation (F_R = 210.9, df = 3, $P < .001$), music video (F_R = 190.5, df = 3, $P < .001$), and news (F_R = 176.5, df = 3, $P < .001$) content evaluations differed significantly between all error rates ($P < .001$). In the sport content evaluation (F_R = 208.1, df = 3, $P < .001$), the differences between the evaluations varied significantly between error rates ($P < .001$; and 13.8% and 20.9% $P < .01$).

5.2.4. *Relations between the overall quality evaluation methods*

All quality evaluations based on three different evaluation methods were related to each other. Retrospective acceptance was discriminative on a scale of satisfaction, but not on the acceptance based on simplified continuous assessment. Related or correlated measures indicate that results measured on one scale can be used to interpret the results in another scale. Discrimination between the scales, such as the independence of the acceptable and unacceptable ratings from the satisfaction scales, can be examined in a further analysis for locating the threshold of acceptability. The idea resembles the classical Thurstonian scaling, aiming to construct nonoverlapping concepts with equal intervals on the attitude scale (e.g., [7]) .

Acceptable quality was expressed between scores of 5.5 and 8.5 (Mean = 7.0, SD = 1.5; Figure 7) and unacceptable quality was located between scores of 2.0–5.8 (Mean = 3.9, SD = 1.9). The distribution between acceptable and unacceptable ratings on the satisfaction scale differed significantly ($\chi^2(10) = 683.2$, $P < .001$). In relation to evaluations based on continuous assessment, acceptable quality was located between 83% and 98% (M = 90.7% of time SD = 7.6; Figure 7) of total acceptances of time, overlapping with unacceptable quality evaluations (M = 81.3% of time SD = 10.8). The distributions between acceptable and unacceptable ratings on a scale of acceptance % of time likewise differed significantly ($\chi^2(36) = 319.1$, $P < .001$). The retrospectively rated satisfaction and acceptance based on continuous assessment were positively and linearly related (Spearman: $r = .725$, $P < .001$). In practice, the acceptance threshold is located in the range of 5.5–5.8 on the satisfaction scale in this experiment. It is not justifiable to draw a similar conclusion for the measures of acceptance percentage of time because the threshold is located between 83.1 and 92.1 and the confidence intervals of unacceptable and acceptable percentage of time overlap to a great extent.

5.3. *Discussion*

The perceived preference order in all measured scales for error rates was 1.7%, 6.9%, 13.8%, and 20.7%, respectively, indicating clearly detectable differences between stimuli. Acceptance ratings give a quality anchor for this preference order showing that the threshold between acceptable and unacceptable quality lies between error rates of 6.9% and 13.8% and this result is not dependent on content. In practice, acceptable quality can be reached when approximately 4/60 seconds are corrupted, resulting altogether in a maximum 10 detectable errors [59, 60] in different media.

In the literature, an error rate of 5% is the conventionally used limit value of operative quality of restitution (QoR) for mobile reception [68] but our result showed a slightly higher tolerance of errors.

The order of preference for different error rates collected using different methods was similar in all contents with few exceptions. Exceptions were found especially in the comparisons of acceptance ratings of the highest error rates. In these error rates, the produced quality is relatively modest. The evaluation criterion of acceptance may be much tighter compared to the task of evaluating quality satisfaction or it may be hard to accept any such erroneous presentations as the goal of viewing can no longer be achieved [34]. In addition, a binary acceptance scale may be useful only in the identification of the threshold, not in detailed comparisons of preferences regarding low qualities. In summary, the assessment results were closely related between all three measures indicating good reliability, and they had good discriminative capability when differences between stimuli were distinguishable and the stimuli not extremely erroneous.

6. EXPERIMENT 3

For further estimation of the reliability and discriminative ability of the overall quality evaluation methods presented, we continued the work with heterogeneous error characteristics, realistic in multimedia broadcasts. The third experiment aims to compare two different error rates on both sides of acceptability by pre- or postprocessing them with four different error control methods. This combination was assumed to produce detectable, but relatively small differences between stimuli.

Few studies have reported comparisons of error control methods related to DVB-H to improve experienced quality. Hannuksela et al. [23] have compared unequal and equal error protection methods with two different error rates. Unequal error protection (UEP) method uses priority-based segmentation of media streams in which audio and the most important coded video pictures have the best protection under harsh channel transmission conditions. By contrast, all media data are of equal importance in the conventional equal error protection method (EEP). The experiment compared these methods with error rates of 6.7% and 13.8% and concluded that in the highest error rate UEP improved the subjective quality. Further, Hannuksela et al. [22] also compared audio redundancy coding and conventional error protection methods with two different error rates (6.7% and 13.8%). Audio redundancy coding (ARC) aims to ensure audio continuity in very erroneous channel conditions and their results showed it to improve perceived quality, especially with the harshest error rate. Earlier studies have shown that error control methods can provide some quality improvements depending on error rate, but no extensive study of different error control methods and error rates has been published.

The aim of the experiment is to compare the interactions of four different error control methods and error rates close to the threshold of acceptability with small differences between the stimuli. In addition to measuring overall satisfaction of quality and acceptance percentage of time, we are interested to ascertain if the boundary of acceptability can be affected by error control methods. To evaluate reliability, we also examine the relations between results of three different methods.

6.1. Research method—test setup

6.1.1. Participants

Our participants were 45 participants, recruited according to same criteria as in experiments 1 and 2.

6.1.2. Test procedure

The test procedure was identical to that of experiment 2. The total duration of the experiment was approximately 2 hours.

6.1.3. Selection of test material

Test materials were identical to experiment 2.

6.1.4. Material production process—transmission error simulations

The aim of the error simulations was to produce stimuli material with relatively small, but detectable differences between stimuli in various forms. As a base for error simulations, two different error rates known to be perceived around a boundary between acceptable and unacceptable (experiment 2) qualities were selected and further four different error concealment methods were applied to these. The simulated error rates produced a varying number, length, and location of errors, and error concealment methods caused different audiovisual appearance form for these errors (Table 3).

Four different error resiliency methods were tested. While one of the error resiliency methods gave more importance to audio, another gave video error resiliency more importance. The remaining one used channel-assisted error resiliency based on unequal error protection. These methods are described in greater detail below.

The first method, called conventional transport with picture freeze (CT-PF), did not use any kind of additional error resiliency measures apart from the protection provided by DVB-H MPE-FEC. The method was used as a base for comparing other error resiliency methods tested. It assumed a compliant audiovisual decoder, albeit with no intelligence. In this method, when the decoder encountered errors in a video stream, it stopped decoding any subsequent pictures until an Intra Decoder Refresh (IDR) picture arrives. IDR pictures use no other pictures as a prediction reference and therefore provide a resynchronization point in an erroneous bit stream. During the period when the decoder stopped decoding, it presented the last uncorrupted decoded picture. Subjectively, when this method was used, an error was perceived as jerky motion in visual streams. The duration of these jerks in visual streams depended on the IDR interval

Table 3: Number of errors, mean durations and standard deviation (in seconds) of burst errors for error patterns in different error rates and error control methods.

Concealment content		N	Mean(SD)	N	Mean(SD)
CT-PF			Error rate 6.9%		Error rate 13.8%
Cartoon	Audio	7–9	0.15(0.08)	18	0.18(0.09)
	Video	3	1.2(0.58)	5-6	1.5(0.71)
Music video	Audio	7–9	0.15(0.08)	18	0.17(0.09)
	Video	3–5	0.92(0.55)	5	1.72(1.01)
News	Audio	8	0.15(0.07)	18	0.17(0.09)
	Video	3-4	1.53(1.13)	5–7	1.63(1.04)
Sport	Audio	8	0.15(0.07)	18	0.17(0.07)
	Video	2–4	1.22(0.72)	6	1.29(0.07)
SAR-PF			Error rate 6.9%		Error rate 13.8%
Cartoon	Audio	2	0.11(0.06)	7–11	0.11(0.04)
	Video	2-3	2.41(1.26)	4-5	1.90(1.01)
Music video	Audio	2	0.11(0.06)	7–11	0.12(0.04)
	Video	2–4	2.11(1.52)	5	1.82(0.80)
News	Audio	2	0.11(0.06)	7–11	0.11(0.03)
	Video	1–3	2.28(1.53)	1–3	6.03(3.60)
Sport	Audio	2	0.11(0.06)	7–11	0.12(0.03)
	Video	1–4	2.30(2.00)	3	3.04(1.47)
CT-EC			Error rate 6.9%		Error rate 13.8%
Cartoon	Audio	7–9	0.15(0.08)	18	0.18(0.09)
	Video	7–9	0.18(0.06)	18	0.18(0.09)
Music video	Audio	7–9	0.15(0.08)	18	0.17(0.09)
	Video	7–9	0.17(0.07)	15–18	0.19(0.09)
News	Audio	7–9	0.15(0.08)	18	0.18(0.09)
	Video	7–9	0.18(0.07)	17–19	0.18(0.10)
Sport	Audio	8	0.15(0.07)	18	0.17(0.07)
	Video	7-8	0.20(0.08)	17–19	0.19(0.11)
UEP-PF			Error rate 6.9%		Error rate 13.8%
Cartoon	Audio	4-5	0.29(0.14)	11	0.34(0.19)
	Video	7–12	0.43(0.65)	14-15	0.54(0.69)
Music video	Audio	3–5	0.27(0.16)	10	0.38(0.24)
	Video	8–12	0.32(0.42)	11	0.72(1.21)
News	Audio	3-4	0.32(0.18)	9–11	0.36(0.22)
	Video	8–10	0.34(0.44)	13–17	0.44(0.49)
Sport	Audio	3	0.35(0.17)	9-10	0.39(0.22)
	Video	6–12	0.34(0.47)	9–12	0.60(0.92)

and the position of the error between two IDR intervals. The audio compression scheme used in the tests encoded 1024 samples of every audio channel as frames. These frames were all independent of each other and a loss of any one frame of the bit stream did not affect any other subsequent frames of an audio channel. When an audio frame was lost, it was replaced with a null frame perceived as silence by the listener. Subjectively, audio frame losses were perceived as discontinuous audio.

The second method used audio redundancy coding to achieve better audio reception in heavy DVB-H channel error conditions and is therefore called Synchronised Audio Redundancy coding with picture freeze (SAR-PF). When MPE-FEC frames were constructed with audiovisual data as input, audio packets that constitute the next MPE-FEC frame in transmission were replicated and sent in the current MPE-FEC frame. The audio decoder expected two copies of every coded audio frame. However, when errors destroyed an audio frame, the decoder looked for the second copy of the same audio frame and if received correctly, used this copy instead. This redundancy of audio packet coupled with their transmission in different time-sliced bursts greatly reduced the probability of any audio frame being completely lost. Video error concealment was identical to what was done in the CT-PF method described above. However, to account for the additional bit rate overhead incurred due to redundant audio packets, the video bit rate was dropped such that the overall bit rate was the same as the other error resiliency

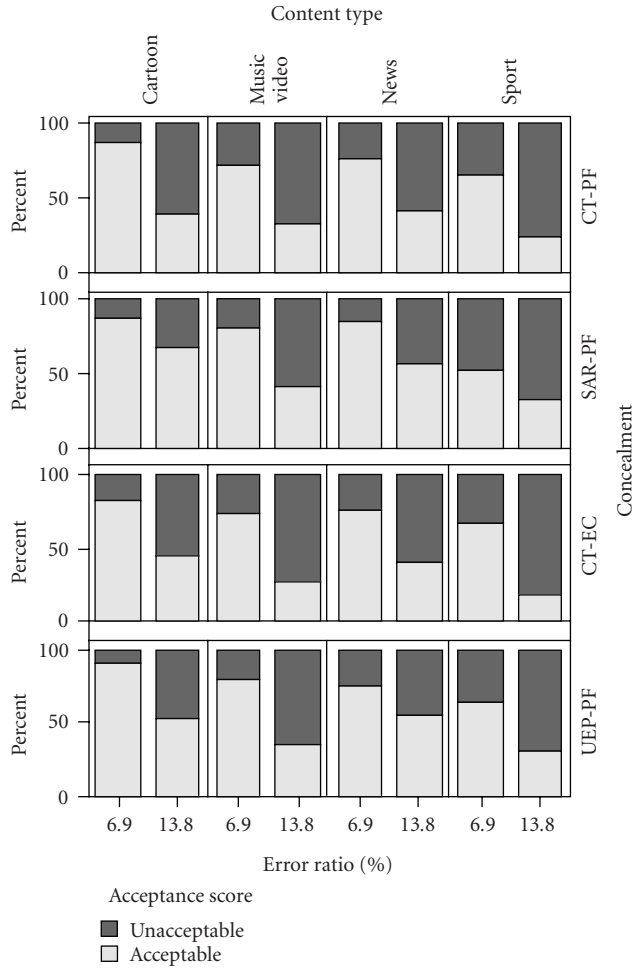

Figure 8: Retrospective acceptance of different error rates and concealment methods for all contents.

methods. In other words, the media-level-produced quality of the coded video was poorer than in the CT-PF method. More details of the SAR-PF method are available in [23].

The third error concealment method, called Conventional Transport with Error Concealment (CT-EC) used a very simple decoder-based visual error concealment method for concealing lost parts of the video sequence. When a picture of the sequence was lost, the decoded picture buffer (DPB) replicated the last correctly received picture (in presentation order) and used it instead of the lost picture. The reason for this replacement was the assumption that spatial video redundancy can be fairly high (depending on the video sequence) and the replaced picture is a good enough estimate of the lost picture. However, since the replaced picture was not the exact representation of the lost picture, motion compensation errors occured in pictures using the replaced picture as reference, and these errors propagated until an Intra picture and/or IDR picture arrived. For audio, the error concealment was similar to what was used in the CT-PF method, where the last audio frames were replaced with a silent frame.

The fourth method of error resiliency is called Unequal Error Protection with Picture Freeze (UEP-PF). First, the media datagrams covering a certain period of playback time were assigned priorities. In the tests, two priorities were used. Audio packets, video reference pictures (both IDR and reference predicted pictures) were assigned priority 1 (the highest), and nonreference pictures were assigned priority 2 (the lowest). The priority-assigned datagrams were grouped together such that all datagrams in a group had the same priority. The protection of the priorities was chosen such that priority 1 datagrams were protected with a 3/4 MPE-FEC-code-rate while the priority 2 datagrams were completely unprotected. These grouped and protected MPE-FEC matrices (called peer MPE-FEC matrices) were then sent back to back without any delay between these MPE-FEC frames. More details on the UEP-PF method are available in [23, 75]. The first and last five seconds of presentation were left error-free to avoid memory effect (primacy and recency) in evaluation of long test materials [49, 53].

6.1.5. Presentation of test materials

The presentation of the test materials was similar to that in the previous experiments. All clips were played twice in random order and the positions of the transmission errors varied in both repetitions.

6.1.6. Data-analysis methods

Selection of data-analysis methods followed the methods described for experiment 2.

6.2. Results

6.2.1. Acceptance

Between error rates

Lower error rate (6.7%) provided mostly acceptable and higher error rate (13.8%) unacceptable quality with significant difference between them in all studied concealment methods and contents ($P < .01$; Figure 8).

Between error concealments

All concealment methods were evaluated equally acceptable in error rate 6.9% ($P > .05$). In contrast, in error rate 13.8%, SAR-PF and UEP-PF ($P > .05$) were evaluated equally and more acceptable CT-PF and CT-EC ($P < .001$) which were in same level as well ($P > .05$).

In error rate 6.7%, mostly all error concealment methods were evaluated into same level, but there were some content-dependant variations. There were not differences between the concealment methods in animation and music video presentation ($P > .05$). News content, concealed with SAR-PF, was evaluated more acceptable than other methods (SAR-PF versus others $P < .05$; all other comparisons $P > .05$). In contrast, SAR-PF provided the most modest quality for sport presentation ($P < .05$; all other comparisons

$P > .05$), approaching the boundary between acceptable and unacceptable quality.

In error rate 13.8%, SAR-PF and UEP-PF provided more acceptable quality than CT-PF and CT-EC. For animation and news presentation, most of the participants considered SAR-PF and UEP-PF as equally acceptable ($P > .05$) and CT-PF and CT-EC as equally unacceptable ($P > .05$) with significant differences between them (SAR-PF versus CT-PC, CT-EC $P < .05$; UEP-PF, and CT-PF $P < .01$). In music, SAR-PF was significantly better than CT-EC ($P < .05$), while all other methods were in same level ($P > .05$). For sport presentation, SAR-PF and UEP-PF were rated as the most acceptable ($P > .05$) with significant difference to other methods ($P < .05$). Error rate 13.8% is in general evaluated as unacceptable, but in the case of cartoon and news with concealment method, SAR-PF quality can become acceptable or, with method UEP-PF, reach the boundary of acceptable and unacceptable ratings.

6.2.2. Satisfaction

Between error rates

Similar to the results for acceptance, error ratio 6.7% was reported more satisfying than error ratio 13.8% in all contents and error control methods ($P < .001$; Figure 9).

Between error concealments

Error ratios and error concealment methods affected satisfaction evaluations ($F_R = 982.1$, df = 7, $P < .001$) and error concealment strategies had a significant effect on evaluations within both error rates (6.9%: $F_R = 17.252$, df = 3, $P < .01$, 13.8%: $F_R = 94.381$, df = 3, $P < .001$).

In terms of satisfaction, CT-EC provided the lowest quality in comparison to other concealment methods ($P < .05$), which were equally evaluated ($P > .05$) for error rate 6.9%. In error rate 13.8%, the most satisfying quality was given by SAR-PF, followed by UEP-PF and the lowest quality by equally rated CT-PF and CT-EC ($P > .05$) with significant differences between all ($P < .01$).

There were also content-dependent preferences between the concealment methods in different error rates. For the lower error rate of 6.9% for animation content, all concealments were evaluated at the same level ($P > .05$). UEP-PF and SAR-PF were evaluated equally, giving the most satisfying quality in music video ($P > .05$), but only differences between UEP-PF and others were significant ($P < .001$). SAR-PF was evaluated as the most satisfying for news content compared to other methods ($P < .01$). In sports, CT-PF and UEP-PF were found equally good ($P > .05$) and significantly better than SAR-PF ($P < .05$).

In error rate 13.8%, error concealments SAR-PF and UEP-PEF were among the most satisfying methods in all contents. For animation presentation, SAR-PF and UEP-PF were evaluated equally being more satisfying ($P > .05$) than other methods ($P < .001$). In music video, SAR-PF, UEP-PF, and CT-PF ($P > .05$) were more satisfying than the concealment method called CT-EC ($P < .05$). The SAR-PF and UEP-PF were equally evaluated ($P > .05$) in news presentation in which SAR-PF was significantly better than CT-PF and CT-EC ($P < .01$) and UEP-PF significantly better than CT-PF ($P < .05$). For sport content, SAR-PF, and UEP-PF ($P > .05$) were more satisfying than the others with SAR-PF significantly outperforming both CT-PF and CT-EC ($P < .001$).

6.2.3. Acceptance percentage of time

Between error rates

Lower error rate (6.7%) was reported to give a higher acceptance rate percentage of time compared to higher error rate (13.8%) ($P > .001$; Figure 10). An exception was found in news presentation with error rate 6.7%, methods CT-EC and UEP-PF were evaluated at the same level with error rate 13.8% concealed with SAR-PF ($P > 0.05$, ns).

Between error concealments

Error ratios and error concealment methods affected acceptance evaluations based on simplified continuous assessment ($F_R = 1335.0$, df = 7, $P < .001$). The error concealment strategies also had a significant effect on within error examination (6.9%: $F_R = 48.5$, df = 3, $P < .001$, 13.8%: $F_R = 223.0$ df = 3, $P < .001$). In error rate 6.9%, SAR-PF yielded the highest acceptance percentage of time with significant difference ($P < .01$) to others being on the same level ($P > .05$). Similarly, SAR-PF yielded the highest acceptance % of time in error rate 13.8% ($P < .001$), followed by UEP-PF and CT-PF ($P > .05$) and UEP-PF and CT-EC ($P > .05$).

There were also some content-dependent variations between the concealment methods with the lower error rate of 6.9%. For presenting cartoons, the longest acceptable presentation for cartoon content was given by SAR-PF outperforming the others ($P < .05$), followed by UEP-PF (difference from others $P > .05$). In music video, SAR-PF and UEP-PF were evaluated at the same level ($P > .05$, difference from others $P < .05$). The concealment SAR-PF also provided the highest quality ($P < .001$) for news content with significant difference from other methods which were evaluated equally ($P > .05$). In sport content, there were no differences between the methods ($P > .05$) except the UEP-PF, which yielded the lowest quality ($P < .001$).

In the higher error rate (13.8%), CT-PF, SAR-PF, and CT-EC ($P > .05$) were more satisfying than the most modestly assessed UEP-PF ($P < .001$) for cartoon content. In music video, SAR-PF is the highest quality with a significant difference from the others ($P < .001$), UEP-PF is the second highest ($P < .05$), and the other methods were evaluated at the same level ($P > .05$). For the news, the concealment called SAR-PF yielded the highest quality ($P < .001$) and all other methods were on the same level ($P > .05$). As in news content, SAR-PF yielded the highest quality for sport content with significant difference from the others ($P < .001$), CT-PF and CT-PC the second highest ($P > .05$), and UEP-PF the most modest ($P < .05$).

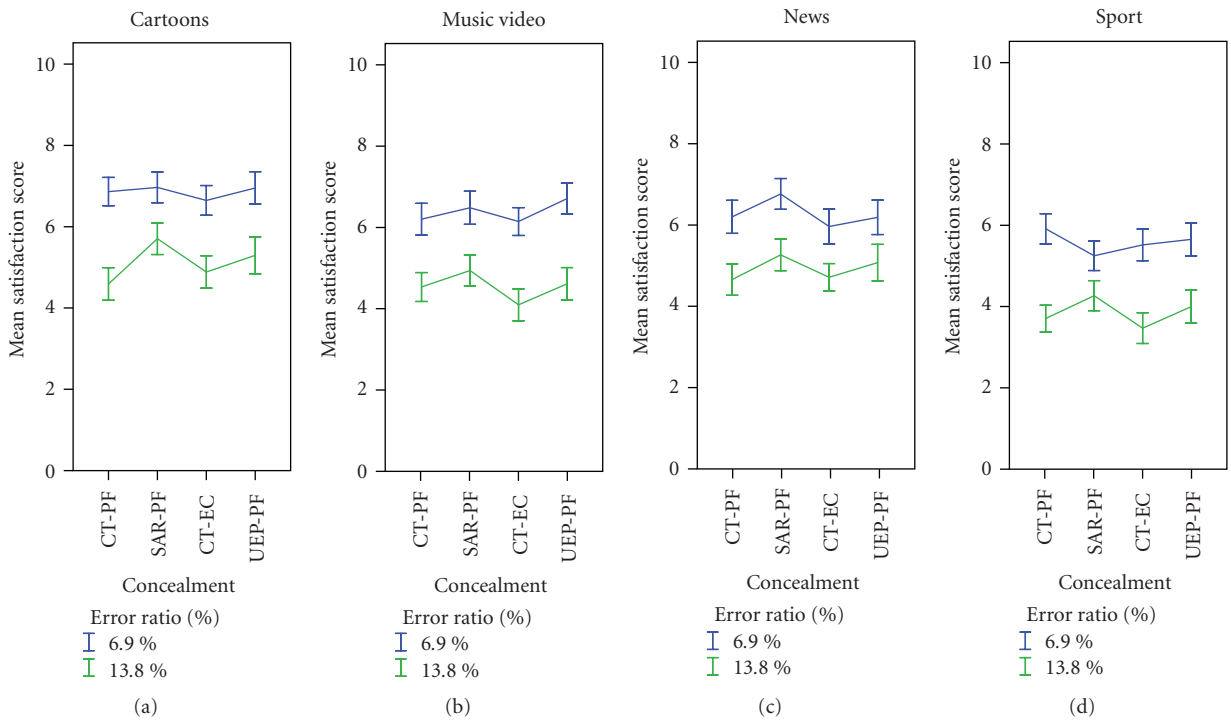

Figure 9: Retrospective satisfaction of different error rates and concealment methods for all contents. Error bars show 95% CI of mean.

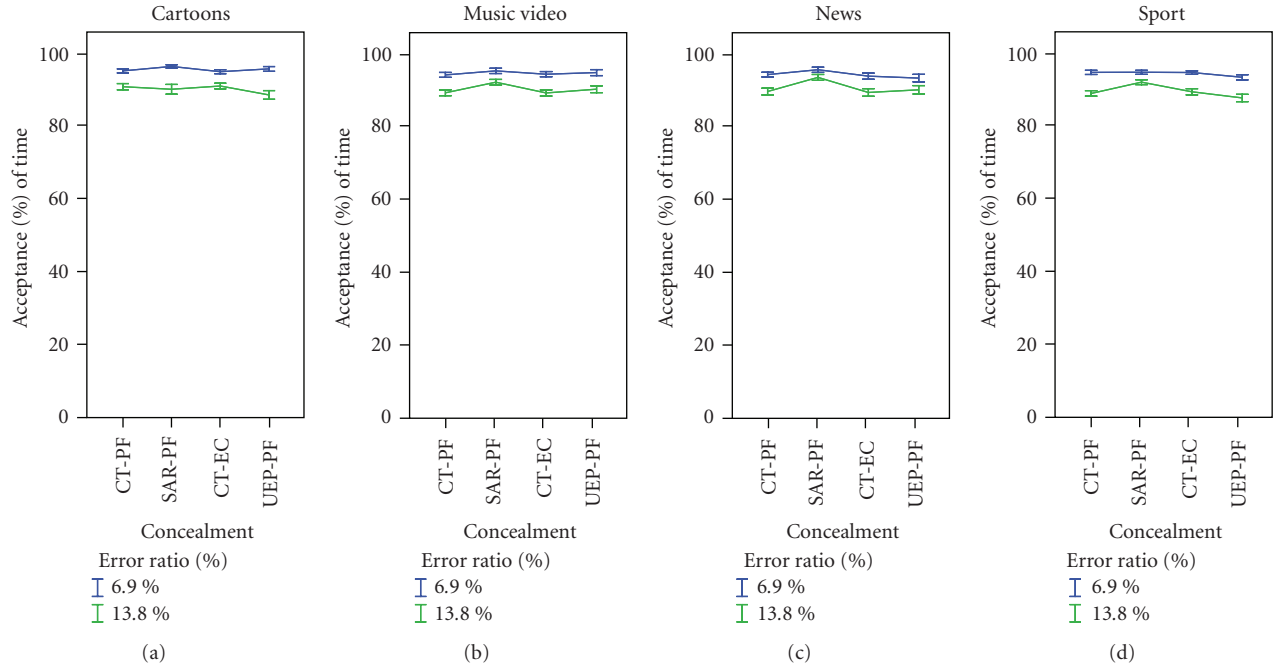

Figure 10: Acceptance percentage of time of different error rates and concealment methods for all contents. Error bars show 95% CI of mean.

6.2.4. *Relations between the overall quality evaluation methods*

As in experiment 2, the three different evaluation methods were related to each other. Acceptable and unacceptable quality was clearly detectable on a scale of satisfaction, but not on a scale of acceptance percentage of time. Acceptable quality was connected to scores between 5.2 and 8.1 (Mean = 6.6, SD = 1.45; Figure 11) on a satisfaction scale and unacceptable quality to scores between scores of 2.1–5.4

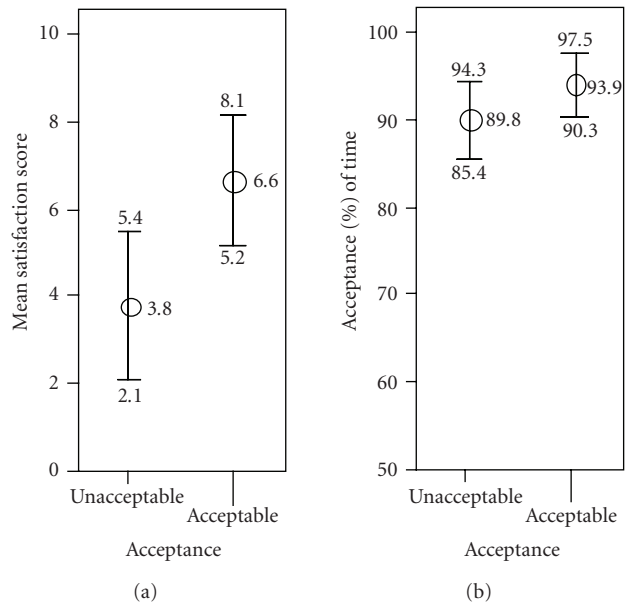

Figure 11: Relations on the scale between retrospective acceptance and satisfaction; and retrospective acceptance and acceptance based on continuous assessment. Bars show mean and standard deviation.

(Mean = 3.8, SD = 1.67). In the examination of the relation between acceptance and acceptance percentage of time, acceptable quality was located between 90 and 97% (M = 93.9% of time, SD = 3.6) on a scale of acceptance percentage of time with widely overlapping unacceptable quality range (M = 89.8% of time, SD = 4.4). As in the previous experiment, both the distributions of retrospectively rated satisfaction and acceptance ($\chi^2(10) = 1370.3$, $P < .001$) and the distributions between the retrospectively rated acceptance and acceptance based on continuous assessment ($\chi^2(49) = 632.0$, $P < .001$) differed. The retrospectively-rated satisfaction and acceptance based on continuous assessment were also positively and linearly related (Spearman: r = .542, $P < .001$). In practice, according to this experiment, the threshold between acceptable and unacceptable ratings is between the scores 5.2 and 5.4 on the satisfaction scale. The threshold on a scale of acceptance percentage of time is between 90.3 and 94.3 in which the overlapping of the confidence intervals constrains the interpretation of results.

6.3. Discussion

All the evaluation methods were able to detect the differences in the level of error rates confirming the results of experiment 2. Higher error rate was experienced giving poorer quality compared to lower error rate in all methods measured. In the measures of acceptance percentage of time, only one exception appeared in which the poorest quality of lowest error rate was evaluated equal with the highest quality of most erroneous error rate.

When error control methods were compared, variations were found in the results gathered using retrospective and continuous methods. In error rate 6.9%, the requirements for different error control methods varied content dependently. For example, in news content, SAR-PF outperformed the other methods in all measures, whereas all methods were equally retrospectively evaluated for cartoons. CT-PF and UEP-PF were among the methods that provided highest quality for sport content in the retrospective measures, whereas UEP-PF was the poorest method according to acceptance percentage of time measures. In high error rate, retrospective methods had excellent agreement in acceptance and satisfaction revealing that SAR-PF and UEP-PEF were among the most satisfying methods in all contents. These error control methods even enabled cartoons and news to reach the 50% acceptance threshold. In contrast, according to simplified continuous assessment, SAR-PF provided the highest acceptance percentage of time while UEP-PF did not produce the highest quality in any of the cases. In all of the cases measured with continuous assessment, SAR-PF was among the methods producing the highest acceptance percentage of time.

From the viewpoint of research methods, there are two main conclusions. Firstly, good agreement between the retrospective methods indicates that detailed analysis is not needed for both of the measures. Both of the methods are needed in data collection, but different emphasis is given in the analysis. As quality satisfaction is measured using an ordinal scale and therefore providing a chance to use sophisticated and efficient methods of analysis [72], it should be used as a primary data source for analysis. Data on acceptance of quality may only be analyzed to locate a certain threshold of acceptance and these thresholds can be used as references in the interpretation of the results of quality satisfaction. Secondly, simplified continuous assessment may not be a reliable method for overall quality evaluation to discriminate stimuli having small noticeable differences. The results of simplified continuous assessment differed from the results of retrospective measures when the differences between the stimuli were small.

There are two main conclusions about the error rates and error control methods we studied. Error rate seems to be a more important factor in perceived quality than an error control method. Further research may focus on error rates and more detail examination of different impacting error characteristics, such as duration, location, and modality within these error rates. In addition, the results of the comparisons of error rates and error control methods also reflected the relation between content dependency and level of quality. In the low error rates, some dependant preferences appeared. For example, the error control methods improving audio quality was emphasized in news presentation while improvements in visual quality were highlighted in sport content. By contrast, extremely erroneous quality seems to hide the content-dependent preferences highlighting the importance of audio quality in all contents. These results are supported by an earlier study comparing several audio-video bitrates. These authors concluded that relations between optimal audio-video bitrates are content dependent, but in low qualities audio qualities is emphasized [8].

7. DISCUSSION AND CONCLUSIONS

In this paper, we examined research methods for assessing overall acceptance of quality in three experiments. At first, we explored the possibilities of using simplified continuous assessment in the evaluation of overall acceptance parallel to retrospective measures. Secondly, we studied the boundary between acceptable and unacceptable quality using clearly detectable differences between stimuli. Finally, we studied the acceptance threshold with small differences between stimuli under heterogeneous conditions. We conducted these studies in the context of mobile television with varying error rates and error control methods with several television programs in a controlled environment. Our results showed that instantaneous and retrospective evaluation methods can be used in parallel in quality evaluation without causing changes to human information processing. All measures were discriminative and correlated when clearly detectable differences between stimuli were studied. By contrast, when small differences between stimuli were examined, the results of retrospective measures correlated but differed from the results based on the evaluation of instantaneous changes. In this section, we discuss the main results and make recommendations for the use of these methods.

7.1. Bidimensional method for retrospective evaluations of overall acceptance and satisfaction

As the main result of this study, we recommend parallel use of retrospective measures of acceptance and satisfaction in quality evaluation experiments. Acceptance, representing the first dimension, is needed to ensure that test variables reach the predefined thresholds depending on the goal of the study (e.g., 50%, 80%). However, the nature of measuring a threshold has some constraints. Firstly, the measure is discriminative when studying variables close to the threshold, but is not clearly below or above it. Secondly, as acceptability is measured on a binary scale, it imposes limitations on the use of efficient methods of analysis which are needed in careful pairwise comparisons [7]. To go beyond these constrains and broaden the use of the method to the other quality ranges, we recommend studying satisfaction of quality parallel to acceptability. Satisfaction, the second dimension, as a degree-of-liking is most commonly measured on a 9- or 11-point ordinal scale which enables the use of efficient methods of analysis [7]. In addition, it allows using same data-collection method for the duration of continuous quality evolution.

Data-collection and analysis using a combination of acceptance and satisfaction methods are summarized in Figure 12. We recommend a separate analysis for both of the measured dimensions, but as a starting point, the relation between the measures needs to be considered to ensure the reliability. There are two options for extracting the desired threshold. Firstly, the tested parameters can be dissected from the frequencies of acceptance data. In the second option, the value range of the threshold between acceptable and unacceptable scores can be identified on satisfaction scale in the case the measures are not strongly overlapping. Further, the located threshold can be used in the interpretation of results of a detailed analysis of preferences derived from the satisfaction data.

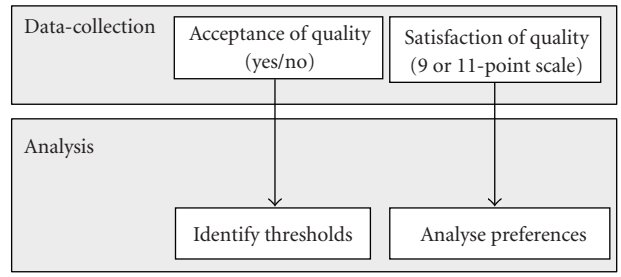

Figure 12: Data-collection and analysis for bidimensional measure combining retrospective overall acceptance and satisfaction.

This work focused on presenting a bidimesional research method, but it did not aim to model bidimensionality in the level of analysis. Our studies showed evidence that the location of acceptance threshold on satisfaction scale is relative to the measured phenomena. To name the constant values for the threshold on a satisfaction scale might be impossible and it might restrict the use of method for measuring different quality ranges. However, our study was limited to the evaluations of clearly detectable and small differences around the threshold. Further work needs to explore the behavior of these measures on the high or low levels of produced qualities for modeling the actual usage of the different scales. In addition, to validate the bidimensional method, further studies need to apply it for studying all multimedia abstraction layers and their interaction. This study targeted only the network and media levels while less attention was paid to the content layer. Finally, to broaden the presented method, there is a need to explore acceptability evaluations in relation to other user-oriented assessment tasks, like examination of goals of viewing.

7.2. Overall acceptance based on evaluation of instantaneous changes

As a minor result, a simplified continuous assessment task to evaluate instantaneous quality changes can be used in parallel with retrospective evaluation methods in quality assessment. This data-collection method can offer insights for the annoying factors in time-varying quality [76] without changing human information processing which has been the shortcoming of the previous methods [53]. When talking about constructing overall evaluations based on instantaneous assessments, there are some challenges. Our results showed that the overall acceptance scores of continuous assessment were relatively high all the time and were not very well distinguishable in the terms of retrospective acceptance. Moreover, their ability to differentiate small differences between stimuli was limited. All of these aspects might have been impacted by an additive approach for constructing the overall evaluations we used. In this trail, further work needs to examine other possibilities for the use of instantaneously

recorded data to predict overall quality by weighting the certain segments of evaluations, like peaks and ends [77]. This perspective can also reveal something new from the fundamental problem of relation between parts and whole in the human information processing. On the other hand, this approach does not necessarily erase the need of measuring a retrospective acceptance anchor. In the current phase, we recommend using a simplified continuous assessment method for tracking the acceptability of instantaneous changes (e.g., [76]) in parallel to retrospective methods, but not as the only method for evaluating overall acceptance.

7.3. Conclusions

This study presented an evaluation method of acceptance representing the minimum level of user requirements in which user expectations and needs are fulfilled. The proposed bidimensional evaluation method combining acceptance and satisfaction can be extended or integrated into any consumer- or user-oriented sensory studies to ensure the level of minimum quality of a relevant component. For example, in the context of multimedia quality, it can be added to an existing QoP model targeting the measurement of quality preference and goals of viewing [12, 13]. The method can also help system developers to test meaningful parameter combinations when testing a novel set of parameters, parameter combinations or several modalities (e.g., audio-video parameter combinations for mobile 3D television).

However, acceptability measurement is just one of the first steps on the way to understanding consumer- or user-oriented experienced multimedia quality. Our long-term aim is not only to focus on acceptance evaluation as method to ensure the quality of a critical system component, but also to understand the effect of user characteristics, system design, and the actual context of use on experienced quality.

ACKNOWLEDGMENTS

This study was funded by Radio- ja televisotekniikan tutkimus Oy (RTT). RTT is a nonprofit research company specialized in digital television datacasting and rich media development. Satu Jumisko-Pyykkö's work was funded by the UCIT graduate school and this article was supported by the HPY research foundation and Finnish Cultural Foundation. The authors wish to thank Hannu Alamäki and Kati Nevalainen about their work in the project.

REFERENCES

[1] R. J. Abbott, *An Integrated Approach to Software Development*, John Wiley & Sons, New York, NY, USA, 1986.

[2] V. Roto, *Web browsing on mobile phones—characteristics of user experience*, Ph.D. dissertation, Helsinki University of Technology, Helsinki, Finland, 2006.

[3] M. Hassenzahl and N. Tractinsky, "User experience—a research agenda," *Behaviour & Information Technology*, vol. 25, no. 2, pp. 91–97, 2006.

[4] S. J. Barnes, "The mobile commerce value chain: analysis and future developments," *International Journal of Information Management*, vol. 22, no. 2, pp. 91–108, 2002.

[5] S. Bech and N. Zacharov, *Perceptual Audio Evaluation: Theory, Method and Application*, John Wiley & Sons, New York, NY, USA, 2006.

[6] P. G. Engeldrum, *Psychometric Scaling: A Toolkit for Imaging Systems Development*, Imcotek Press, Winchester, Mass, USA, 2000.

[7] H. T. Lawless and H. Heyman, *Sensory Evaluation of Food: Principles and Practices*, Chapman & Hall/CRC, New York, NY, USA, 1998.

[8] S. Jumisko-Pyykkö, ""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," *Multimedia Tools and Applications*, vol. 36, no. 1-2, pp. 167–184, 2008.

[9] H. Knoche, J. D. McCarthy, and M. A. Sasse, "Can small be beautiful? Assessing image resolution requirements for mobile TV," in *Proceedings of the 13th Annual ACM International Conference on Multimedia (MULTIMEDIA '05)*, pp. 829–838, Singapore, November 2005.

[10] H. Knoche, J. McCarthy, and M. A. Sasse, "How low can you go? The effect of low resolutions on shot types in mobile TV," *Multimedia Tools and Applications*, vol. 36, no. 1-2, pp. 145–166, 2008.

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

[12] G. Ghinea and J. P. Thomas, "QoS impact on user perception and understanding of multimedia video clips," in *Proceedings of the 6th ACM International Conference on Multimedia (MULTIMEDIA '98)*, pp. 49–54, Bristol, UK, September 1998.

[13] S. R. Gulliver, T. Serif, and G. Ghinea, "Pervasive and standalone computing: the perceptual effects of variable multimedia quality," *International Journal of Human Computer-Studies*, vol. 60, no. 5-6, pp. 640–665, 2004.

[14] J. D. McCarthy, M. A. Sasse, and D. Miras, "Sharp or smooth?: comparing the effects of quantization vs. frame rate for streamed video," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '04)*, pp. 535–542, Vienna, Austria, April 2004.

[15] K. Nahrstedt and R. Steinmetz, "Resource management in networked multimedia systems," *Computer*, vol. 28, no. 5, pp. 52–63, 1995.

[16] G. Wikstrand, *Improving user comprehension and entertainment in wireless streaming media: introducing cognitive quality of service*, Ph.D. thesis, Department of Computer Science, Umeå University, Umeå, Sweden, 2003.

[17] S. R. Gulliver and G. Ghinea, "Defining user perception of distributed multimedia quality," *ACM Transactions on Multimedia Computing, Communications and Applications*, vol. 2, no. 4, pp. 241–257, 2006.

[18] S. Jumisko-Pyykkö, M. V. Vinod Kumar, M. Liinasuo, and M. M. Hannuksela, "Acceptance of audiovisual quality in erroneous television sequences over a DVB-H channel," in *Proceedings of the 2nd International Workshop in Video Processing and Quality Metrics for Consumer Electronics*, Scottsdale, Ariz, USA, January 2006.

[19] S. Jumisko-Pyykkö and J. H"akkinen, "Evaluation of subjective video quality of mobile devices," in *Proceedings of the*

13th Annual ACM International Conference on Multimedia (MULTIMEDIA '05), pp. 535–538, Singapore, November 2005.

[20] S. Winkler and C. Faller, "Audiovisual quality evaluation of low-bitrate video," in *Human Vision and Electronic Imaging X*, vol. 5666 of *Proceedings of SPIE*, pp. 139–148, San Jose, Calif, USA, January 2005.

[21] H. Knoche, H. de Meer, and D. Kirsh, "Extremely economical: how key frames affect consonant perception under different audio-visual skews," in *Proceedings of the 16th World Congress on Ergonomics (IEA '06)*, Maastricht, The Netherlands, July 2006.

[22] M. M. Hannuksela, V. K. Malamal Vadakital, and S. Jumisko-Pyykkö, "Synchronized audio redundancy coding for improved error resilience in streaming over DVB-H," in *Proceedings of the 3rd International Mobile Multimedia Communications Conference (MobiMedia '07)*, Nafpaktos, Greece, August 2007.

[23] M. M. Hannuksela, V. K. Malamal Vadakital, and S. Jumisko-Pyykkö, "Comparison of error protection methods for audio-video broadcast over DVB-H," *EURASIP Journal on Advances in Signal Processing*, vol. 2007, Article ID 71801, 12 pages, 2007.

[24] N. Ravaja, K. Kallinen, T. Saari, and L. Keltikangas-J"arvinen, "Suboptimal exposure to facial expressions when viewing video messages from a small screen: effects on emotion, attention, and memory," *Journal of Experimental Psychology: Applied*, vol. 10, no. 2, pp. 120–113, 2004.

[25] H. O. Knoche, J. D. McCarthy, and M. A. Sasse, "Reading the fine print: the effect of text legibility on perceived video quality in mobile tv," in *Proceedings of the 14th Annual ACM International Conference on Multimedia (MULTIMEDIA '06)*, pp. 727–730, Santa Barbara, Calif, USA, October 2006.

[26] A. Köpke, A. Willig, and H. Karl, "Chaotic maps as parsimonious bit error models of wireless channels," in *Proceedings of the 22nd Annual Joint Conference of the IEEE Computer and Communications Societies (INFOCOM '03)*, vol. 1, pp. 513–523, San Francisco, Calif, USA, March-April 2003.

[27] A. Willig, M. Kubisch, C. Hoene, and A. Wolisz, "Measurements of a wireless link in an industrial environment using an IEEE 802.11-compliant physical layer," *IEEE Transactions on Industrial Electronics*, vol. 49, no. 6, pp. 1265–1282, 2002.

[28] I. S. Reed and G. Solomon, "Polynomial codes over certain finite fields," *SIAM Journal of Applied Mathematics*, vol. 8, no. 2, pp. 300–304, 1960.

[29] K. Grill-Spector and R. Malach, "The human visual cortex," *Annual Review of Neuroscience*, vol. 27, pp. 649–677, 2004.

[30] M. S. Lewicki, "Efficient coding of natural sounds," *Nature Neuroscience*, vol. 5, no. 4, pp. 356–363, 2002.

[31] S. T. Fiske and S. E. Taylor, *Social Cognition*, McGrow-Hill, Singapore, 1991.

[32] U. Neisser, *Cognition and Reality: Principles and Implications of Cognitive Psychology*, W.H. Freeman, San Francisco, Calif, USA, 1976.

[33] K. Oatley and J. M. Jenkins, *Understanding Emotions*, Blackwell, Oxford, UK, 2003.

[34] S. Jumisko-Pyykkö, J. H"akkinen, and G. Nyman, "Experienced quality factors: qualitative evaluation approach to audiovisual quality," in *Multimedia on Mobile Devices 2007*, vol. 6507 of *Proceedings of SPIE, 65070M*, pp. 1–12, San Jose, Calif, USA, January 2007.

[35] M. C. Meilgaard, G. V. Civille, and B. T. Carr, *Sensory Evaluation Techniques*, CRC Press, New York, NY, USA, 1999.

[36] R. G. Picard, "Mobile telephony and broadcasting: are they compatible for consumers," *International Journal of Mobile Communications*, vol. 3, no. 1, pp. 19–28, 2005.

[37] R. G. Picard, "Interacting forces in the development of communication technologies," *European Media Management Review*, vol. 1, no. 1, pp. 18–24, 1998.

[38] F. D. Davis, "Perceived usefulness, perceived ease of use, and user acceptance of information technology," *MIS Quarterly*, vol. 13, no. 3, pp. 319–340, 1989.

[39] V. Venkatesh, M. G. Morris, G. B. Davis, and F. D. Davis, "User acceptance of information technology: toward a unified view," *MIS Quarterly*, vol. 27, no. 3, pp. 425–478, 2003.

[40] M. Amberg, M. Hirschmeier, and J. Wehrmann, "The compass acceptance model for the analysis and evaluation of mobile services," *International Journal of Mobile Communications*, vol. 2, no. 3, pp. 248–259, 2004.

[41] E. Kaasinen, *User acceptance of mobile services—value, ease of use, trust and ease of adoption*, Doctoral thesis, VTT Information Technology, Helsinki, Finland, 2005, VTT publications 566.

[42] M. Pagani, "Determinants of adoption of third generation mobile multimedia services," *Journal of Interactive Marketing*, vol. 18, no. 3, pp. 46–59, 2004.

[43] G. C. Bruner II and A. Kumar, "Explaining consumer acceptance of handheld Internet devices," *Journal of Business Research*, vol. 58, no. 5, pp. 553–558, 2005.

[44] S. Sarker and J. D. Wells, "Understanding mobile handheld device use and adoption," *Communications of the ACM*, vol. 46, no. 12, pp. 35–40, 2003.

[45] R. Aldridge, J. Davidoff, M. Ghanbari, D. Hands, and D. Pearson, "Regency effect in the subjective assessment of digitally-coded television pictures," in *Proceedings of the 5th International Conference on Image Processing and Its Applications (ICIP '95)*, pp. 336–339, Edinburgh, UK, July 1995.

[46] A. D. Baddley, *Working Memory*, Oxford University Press, New York, NY, USA, 1998.

[47] M. Ries, R. Puglia, T. Tebaldi, O. Nemethova, and M. Rupp, "Audiovisual quality estimation for mobile streaming services," in *Proceedings of the 2nd International Symposium on Wireless Communications Systems (ISWCS '05)*, pp. 173–177, Siena, Italy, September 2005.

[48] H. Knoche and J. D. McCarthy, "Good news for mobile TV," in *Proceedings of the 14th Wireless World Research Forum Meeting (WWRF14)*, San Diego, Calif, USA, July 2005.

[49] R. P. Aidridge, D. S. Hands, D. E. Pearson, and N. K. Lodge, "Continuous quality assessment of digitally-coded television pictures," *IEE Proceedings: Vision, Image and Signal Processing*, vol. 145, no. 2, pp. 116–123, 1998.

[50] H. de Ridder and R. Hamberg, "Continuous assessment of image quality," *SMPTE Journal*, vol. 106, no. 2, pp. 123–128, 1997.

[51] R. Hamberg and H. de Ridder, "Time-varying image quality: modeling the relation between instantaneous and overall quality," *SMPTE Journal*, vol. 108, no. 11, pp. 802–811, 1999.

[52] A. Bouch and M. A. Sasse, "Case for predictable media quality in networked multimedia applications," in *Multimedia Computing and Networking 2000*, K. Nahrstedt and W. Feng, Eds., vol. 3969 of *Proceedings of SPIE*, pp. 188–195, San Jose, Calif, USA, January 2000.

[53] D. S. Hands and S. E. Avons, "Recency and duration neglect in subjective assessment of television picture quality," *Applied Cognitive Psychology*, vol. 15, no. 6, pp. 639–657, 2001.

[54] R. T. Apteker, J. A. Fisher, V. S. Kisimov, and H. Neishlos, "Video acceptability and frame rate," *IEEE Multimedia*, vol. 2, no. 3, pp. 32–40, 1995.

[55] D. Wijesekera, J. Srivastava, A. Nerode, and M. Foresti, "Experimental evaluation of loss perception in continuous media," *Multimedia Systems*, vol. 7, no. 6, pp. 486–499, 1999.

[56] R. Steinmetz, "Human perception of jitter and media synchronization," *IEEE Journal on Selected Areas in Communications*, vol. 14, no. 1, pp. 61–72, 1996.

[57] N. Kitawaki, Y. Arayama, and T. Yamada, "Multimedia opinion model based on media interaction of audiovisual communications," in *Proceedings of the 4th International Conference on Measurement of Speech and Audio Quality in Networks (MESAQIN '05)*, pp. 5–10, Prague, Czech Republic, June 2005.

[58] A. Watson and M. A. Sasse, "The good, the bad, and the muffled: the impact of different degradations on Internet speech," in *Proceedings of the 8th ACM International Conference on Multimedia (MULTIMEDIA '00)*, pp. 269–276, Los Angeles, Calif, USA, October-November 2000.

[59] R. Pastrana, J. Gicquel, C. Colomes, and H. Cherifi, "Sporadic Signal Loss Impact on Auditory Quality Perception," 2004, http://wireless.feld.cvut.cz/mesaqin2004/contributions.html.

[60] R. R. Pastrana-Vidal, J. C. Gicquel, C. Colomes, and H. Cherifi, "Sporadic frame dropping impact on quality perception," in *Human Vision and Electronic Imaging IX*, vol. 5292 of *Proceedings of SPIE*, pp. 182–193, San Jose, Calif, USA, January 2004.

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

[62] E. M. Rogers, *Diffusion of Innovations*, Free Press, New York, NY, USA, 5th edition, 2003.

[63] A. Watson and M. A. Sasse, "Measuring perceived quality of speech and video in multimedia conferencing applications," in *Proceedings of the 6th ACM International Conference on Multimedia (MULTIMEDIA '98)*, pp. 55–60, Bristol, UK, September 1998.

[64] A. Watson and M. A. Sasse, "Evaluating audio and video quality in low-cost multimedia conferencing systems," *Interacting with Computers*, vol. 8, no. 3, pp. 255–275, 1996.

[65] C. Carlsson and P. Walden, "Mobile TV—to live or die by content," in *Proceedings of the 40th Annual Hawaii International Conference on System Sciences (HICSS '07)*, p. 51, Waikoloa, Hawaii, USA, January 2007.

[66] C. Södergård, Ed., "Mobile television—technology and user experiences," VTT Publications 506, VTT Information Technology, Espoo, Finland, 2003.

[67] 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.

[68] G. Faria, J. A. Henriksson, E. Stare, and P. Talmola, "DVB-H: digital broadcast services to handheld devices," *Proceedings of the IEEE*, vol. 94, no. 1, pp. 194–209, 2006.

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

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

[71] ITU-T P.920, "Interactive test methods for audiovisual communications," International Telecommunications Union - Telecommunication sector, 2002.

[72] H. Coolican, *Research Methods and Statistics in Psychology*, J. W. Arrowsmith, London, UK, 4th edition, 2004.

[73] E. A. Styles, *The Psychology of Attention*, Psychology Press, Hove, UK, 1997.

[74] U. Reiter and S. Jumisko-Pyykkö, "Watch, press, and catch—impact of divided attention on requirements of audiovisual quality," in *Proceedings of the 12th International Conference on Human-Computer Interaction (HCI '07)*, pp. 943–952, Beijing, China, July 2007.

[75] V. K. Malamal Vadakital, M. M. Hannuksela, M. Rezaei, and M. Gabbouj, "Method for unequal error protection in DVB-H for mobile television," in *Proceedings of the 17th IEEE International Symposium on Personal, Indoor and Mobile Radio Communications (PIMRC '06)*, pp. 1–5, Helsinki, Finland, September 2006.

[76] S. Jumisko-Pyykkö, M. V. Vinod Kumar, and J. Korhonen, "Unacceptability of instantaneous errors in mobile television: from annoying audio to video," in *Proceedings of the 8th International Conference on Human-Computer Interaction with Mobile Devices and Services (MobileHCI '06)*, pp. 1–8, Helsinki, Finland, September 2006.

[77] R. M. Hogarth and H. J. Einhorn, "Order effects in belief updating: the belief-adjustment model," *Cognitive Psychology*, vol. 24, no. 1, pp. 1–55, 1992.