What your phone makes you see

Investigation of the effect of end-user devices on the assessment of perceived multimedia quality

Péter András Kara, László Bokor
Department of Networked Systems and Services (HIT), BME
Budapest, Hungary
{kara, bokorl}@hit.bme.hu

Andreas Sackl
Forschungszentrum Telekommunikation Wien (FTW)
Vienna, Austria
sackl@ftw.at

Mariana Mourão
Universidade Federal de Minas Gerais (UFMG)
Belo Horizonte, Brazil
marismourao@ufmg.br

Abstract— The scientific researches performed in the past decade clearly highlighted that Quality of Experience (QoE) has become the ultimate measure of service value in this rapidly developing era of technology. Its continuous monitoring is now simply essential and is the key to service success. Subjectively measured QoE is considered to be somewhat unerring, however, several human factors may introduce distortions in the assessment of multimedia quality. Motivated by the phenomenon of cognitive bias, this paper attends the topic of end-user devices; how they may influence evaluation of video streaming quality. We performed an empirical analysis of mobile streaming multimedia quality with 100 participants, and utilized four different subjective measurement scales. Results show that the perception of streaming quality was significantly affected by the phenomenon through cognitive dissonance.

Keywords— Quality of Experience (QoE); perception of quality; cognitive dissonance; mobile handhelds; streaming multimedia

I. INTRODUCTION

Over the past 20 years, the academic community – especially in the field of telecommunications – has heavily explored the notion of Quality of Experience (QoE). The relevance of this issue is not only due to the accelerating and unbridled growth in both new and improving technologies, but it has great presence in the lives of ordinary people. The purpose of QoE-related research is to analyze the perceived quality of services based on the subjective assessments of end-users. Beside technical specifications, it is necessary to investigate which additional user aspects influence the subjective quality assessment. Previous works in this topic [1][2] showed that the user-factor “expectations” influence the perception of quality from a user’s point of view. For example, the “labeling” of an Internet connection evokes certain expectations, i.e. a wireless-labeled connection is evaluated more positively in contrast to a connection with identical technical properties, but labeled as wireline.

Other studies [3] attending the topic of user aspects in subjective assessment of service quality also detail the role of measurement scales. While discrete scales bound the evaluation task to a limited number of choices, continuous rating techniques allow a virtually infinite space for assessment. Continuous scales not only permit the expression of the lowest amounts of differences between the qualities of test stimuli, but psychologically also encourage these slight differences to be made. The choice of quantitative and qualitative scales [4] is also frequently investigated, since they might differ on several levels, i.e. subjective tags may introduce inconsistent distances between adjacent points of the scale [5], unexpected scoring variances can appear when tags are translated to other languages [6], scores might get compressed due to the avoidance of tags on the top and bottom end of the scale [7] etc.

Our work presented in this paper addresses the following research questions: first of all, we wish to investigate a distinct subtype of the labeling effect in which the brand of the end-user device is the only main difference between identical test cases; we examine how much the type of the smart mobile device solely affects the perception of multimedia quality. In addition, special attention is put to perception alteration itself in order to obtain information on the background of assessment decisions, to see what changes are achieved in human perception. Second, the primary types of measurement scales (discrete / continuous, quantitative / qualitative) shall be put to analysis to examine what specific scales may allow or even force on the subject regarding quality assessment. The novelty of this contribution is the analysis of frame rate alteration and artefact insertion in the perception of visual quality due to cognitive bias.

The structure of the paper is organized as follows: Section 2 provides an overview of similar approaches in the area, followed by the details of the chosen configuration of the performed measurement in Section 3, results of which are introduced in Section 4, and concluded in Section 5.

II. RELATED WORKS

The influence of different scales on subjective quality ratings has been evaluated in various contexts, e.g. the investigation of the difference between discrete and continuous scales of quality assessment in 3D video experiments [8]. According to these findings, there are no relevant differences between continuous and discrete scales.
for the subjective evaluation of 3D video content. The analysis of various rating scales in regular 2D video quality assessment is also an attended topic [9], i.e. the comparison of a 5-point discrete scale, an 11-point continuous scale, a 5-point continuous scale etc. However, no statistically significant differences between the rating scales were found either.

Watson et al. [7] question the common ACR scale [10] with the standardized labels of “bad, poor, fair, good and excellent”. According to the purpose of this scale, the 5 labels should be distributed uniformly, but based on their findings the common ACR scale is not an interval scale. Therefore, in a later research [11] the authors compared an unlabeled continuous scale with the standard ACR scale in the context of audiovisual quality assessment studies. Surprisingly, both scales led to similar results.

In accordance with the related work, the design of rating scales play an important role in the context of subjective quality assessment, even though sometimes they might not appear to be statistically different. However, several other factors – especially human factors – may prove to be even more influencing in the shaping of service quality evaluation.

Cognitive bias can appear in several measureable forms, i.e. the contrast effect [12]. One user-factor that needs to be specially highlighted is expectation towards quality, since the definition of QoE [13] itself also approaches service value from among others the “fulfillment of expectations”. The psychological phenomenon behind the influence of expectations [1] is cognitive dissonance [14], which indeed plays a noteworthy role in the evaluation of multimedia services [15]. The way it affects assessment sturdily depends on the specific user [16], and it correlates with one’s prior experience and knowledge regarding service technology [17]. Results in the topic [18] show that most of this distortion is originated to environmental information about the service – since preconception is usually one of the conflicting cognitions – but some are simply hard to hide from measurement participants [3], thus bound to take effect if present.

The end-user device itself – among the many other variables – can also have a major impact on the experienced quality [19][20], especially when modern-day smart phones and tablets are considered [21]. Results of Schatz et al. [22] point out the effect of display resolution and size on QoE ratings, but device brand alone is worth investigating as well.

In our measurement, we selected end-user device brand as distinct, hard-to-hide information, and chose the combinations of discrete, continuous, quantitative and qualitative scales, as detailed in the measurement setup.

III. MEASUREMENT SETUP

The measurement was designed to investigate the impact of preconceptions related to various mobile end-user equipments on the perception of multimedia streaming video quality and also the influence of different rating scales. Three mobile phones were chosen for this purpose, namely an LG Optimus NET, an HTC Desire S and an Apple iPhone 5. Although there are significant technical differences between these models (see TABLE I), we aimed to display multimedia content with constant, unvarying quality in order to focus on perception alterations attained by preconceptional assessment distortion.

<table>
<thead>
<tr>
<th>Device</th>
<th>Resolution</th>
<th>Screen size</th>
<th>Color</th>
</tr>
</thead>
<tbody>
<tr>
<td>LG Optimus Net</td>
<td>320 x 480</td>
<td>3.2&quot;</td>
<td>18 bit</td>
</tr>
<tr>
<td>HTC Desire S</td>
<td>480 x 800</td>
<td>3.7&quot;</td>
<td>24 bit</td>
</tr>
<tr>
<td>Apple iPhone 5</td>
<td>640 x 1136</td>
<td>4&quot;</td>
<td>24 bit</td>
</tr>
</tbody>
</table>

To achieve our research goal, we designed a mock-up measurement in an isolated, controlled laboratory environment. From the participants’ point of view, the measurement was the assessment of streaming multimedia on three smart phones. However, the videos were completely identical, since they were not streamed but played locally on the devices. The true nature of the measurement remained entirely invisible to the participants.

The evaluation task was to rate the video quality of multimedia streaming. This means that the participant was not asked to evaluate the overall quality, only the streaming related aspects. The stimuli contained no audio since it was not targeted for quality assessment; the Human Visual System (HVS) was in focus. Evaluation was performed on four types of measurement scales (see Fig. 1): discrete quantitative (DN), discrete qualitative (DL), continuous quantitative (CN) and continuous qualitative (CL). The quantitative scales ranged from 1 to 5, and the qualitative scales ranged from “bad” to “excellent”, in alignment with the ITU-T recommendations [10]. Each participant used only one type of scale for assessment and scores were recorded digitally by the participant via PC. An operator was present, controlling the mobile devices, so the participant had no explicit interaction with them beyond observation (see Fig. 2) to ensure the integrity of the mock-up measurement.

---

**TABLE I. VIDEO CAPABILITIES OF THE SELECTED DEVICES**

![Fig. 1. Visual representations of the scales](image-url)
The stimuli were prerecorded HTTP video streams. The image quality was preserved to remain flawless, however, serious stuttering was intentionally introduced. The average frame rate was reduced to 10, which is rather low compared to what users encounter today. Image quality was untouched and no artefacts were included, because they might ring a bell after the third observation, yet this does not apply to the amount of stuttering. Since the same video file was copied onto each device, we had to keep in mind the capabilities of the weakest device, thus the video was rendered accordingly.

The stimulus consisted of 3 sections, which were to be evaluated separately. The first one was a fighting scene from X-Men: Days of Future Past [23], the second was a selection of high motion scenes from Appleseed Alpha [24], and the final one was a cut trailer of The Nut Job [25]. These shall be referred to as stimulus A, B and C later on. All had the duration of 30 seconds and afterwards 15 seconds were provided for the digital assessment of quality, right after the end of each stimulus. The complete duration of the measurement was nearly 7 minutes, excluding the approximately 1-minute introduction of the evaluation task at the beginning. The 3 stimuli were first displayed on the LG, then on the HTC, and finally on the iPhone (see Fig. 3).

We decided to complete all three videos on each device before continuing with the next device to prevent post-decision dissonance [26], which might be also interesting to investigate (stimulus A gets displayed on the 3 devices, and then comes stimulus B and C). Due to this phenomenon, the scoring pattern of the first stimulus is likely to be repeated during the evaluation of the other stimuli to support a prior decision. We chose the order displayed in Fig. 3 to leave more space for regular cognitive dissonance [14] by letting 1.5 minutes pass between identical stimulus sections; memories regarding perception were easier to be overwritten by preconceptions.

During such measurements [15], the two conflicting cognitions are perception and preconception; what we see and what we think we should see. If they do not match, a conflict occurs. Dissonance reduction favours one and alters the other or adds a new cognition to the picture. Such new cognition can be memory; as can be read in the results, sometimes memory is easier to amend then the “real time” perception of quality.

According to the schedule of the evaluation (see Fig. 3), the order of the devices was not randomized; it was the same for every participant. Even though randomization is a commonly used technique in such researches, we decided to keep a strict order for several reasons. First of all, we presumed that the majority would have the highest expectations towards the iPhone, so we wanted to “save the best for last”. We also wanted to have an increasing order in screen size and resolution, since bigger is not always better. In fact, in this case they were identical. The rating task was designed to provide equal conditions to all participants and thus avoid specific psychological effects affecting only a portion of the participants, e.g. contrast effect [12]; with a randomized order, the different orders could have induced different contrasts.

A total of 100 participants completed the evaluation task, each measurement scale utilized by 25 people. Each and every participant observed and rated the stimuli on all 3 devices, and each person used 1 specific rating scale to assess visual quality. The age of the participants ranged from 20 to 28, with the gender distribution of 40 men and 60 women. Nationality was diverse; we had participants from Algeria, Australia, Brazil, Egypt, France, Georgia, Hungary, Iran, Niger, Norway, Poland, Portugal, Serbia, Syria and South Korea, typically university students. No prior detail was provided to the participants, they only info they possessed was that they were about to participate in a measurement in the field of telecommunications. Information exchange between participants was not possible and no information was provided regarding the access network technology.

IV. RESULTS

Analysis of the measurement results presents the collected data as the mean scores for the different devices. Quality assessments of the 3 sections of the stimulus are not detailed in this paper, i.e. stimuli of section A, B and C are not analyzed separately. The overall Mean Opinion Score values of 100 participants show a rising evaluation pattern, and the same can be said for the separate MOS results of the measurement scales (see Fig. 4). This means that in general the stimulus displayed on LG received the lowest assessment scores, on HTC it was given more and the multimedia quality...
on iPhone was evaluated to be the best. Although the order is the same for each scale, however, the measure of difference between devices is not quite the same in case of DL as the on the other scales (see Fig. 5). For all graphics, CI is 0.95.

While the differences between the mean score comparisons of adjacent devices fit into the interval from 0.1 to 0.2 in case of DN, CN and CL, the assessment difference between iPhone and HTC is 3 times bigger than between HTC and LG. On continuous scales, the participant has the freedom of choosing the scoring difference between test cases as small as he or she desires, within the boundaries of what the implementation of the scale allows. However, on discrete scales, the participant needs to make a choice whether the experienced dissimilarity in quality is big enough to assess it with an adjacent integer or tag, or not. Such a decision on a quantitative scale is slightly easier to make than on qualitative one, since for instance elevating evaluation from “fair” to “good” might not be as simple as from 3 to 4. Due to this, several participants of DL hesitated to differentiate the first two devices, but made a clear difference for the last one.

Of course not each and every participant supported the pattern of the overall mean scores. Even though the majority favored quality of the multimedia stream on the iPhone, some preferred the other devices (see Fig. 6). The distribution of this notion is quite similar regardless of measurement scale. Note that both counters are incremented in case of a tie; e.g., 4 HTC/iPhone ties in DN.

At this point of the study, some important questions may rise: If the stimulus on the devices were identical, then why did the participants make such scoring differences in their assessment of quality? Did their perception of quality differ as well? At first sight, it could be easily stated that all these results are due to the different capabilities of the devices; differences in screen size, brightness, colors etc. The situation is not that simple, since the evaluation task was to rate the video quality of multimedia streaming (i.e., image quality, motion fluency etc.), and not to assess the general user experience. Moreover, almost every participant experienced and reported differences in the unvarying elements of technically constant streaming quality.

Beside the assessment scores, verbal evaluation and comments were also collected from the participants. This of course did not happen in a mandatory manner; the participants could add comments before, during and after the measurement, but specific questions were not asked. Asking a set of questions prior to the measurement – especially regarding expectations – does provide a clearer view on the correlations between preconceptions and actual perception, however, it may also bias judgment.

A rather uncommon and surprising experience was the identification of visual artefacts in the videos. As it has been declared in the measurement setup, image quality itself was completely flawless and untouched. Nevertheless 11 participants saw artefacts or witnessed other degradations of visual quality. These experiences typically appeared in a posterior way; e.g., during the observation of the video stream on iPhone, a participant recalled some artefacts regarding the streams on LG and/or HTC. This is a clear case of dissonance reduction. The two opposing cognitions are preconception, stating the quality on iPhone has to be better, and perception, without having any actual proof to support the preconception. To reduce the dissonance, a third cognition, the memory regarding earlier experiences is altered, providing evidence of poor prior quality.

Evaluation performed by a specific test participant in CN was a notable example (see Fig. 7) for the alteration of perception, regarding the presence of artefacts. The so-called preconception overcompensation phenomenon [3] was encountered, in which the participant possesses a specific
preconception, but the lack of support inverts it. The prior idea concerning quality was in alignment with the increasing pattern of the mean scores. Everything went according to preconception until test #7 (first test case on the iPhone). Perception was completely missing the confirmation of a superior streaming quality. This immediately repositioned the iPhone from the top to the bottom of the list, being even slightly worse than the LG. Aside stuttering, the participant started seeing minor artefacts here and there in all 3 stimulus sections (A, B and C), including pixel error and blur. In this case, degradation of visual quality did not happen in a posterior way; instead of altering the memory of the experienced quality on the first device, the perception of streaming quality was modified on the last device, in a “real time” manner.

In general, the mechanism of the preconception overcompensation phenomenon is quite simple yet astonishing. Preconception contains a specific statement S – which can be deemed either true or false – and waits for the other cognition, perception to prove that S is true. Still, if perception proves that S is true, that does not exclude cognitive dissonance, since perception can be adjusted to prove the statement. However, if perception cannot be altered in one direction to support S, it will be in fact modified in the opposite direction. Basically, if the cognition cannot prove that statement S is true, then it is considered to be a proof that S is false. This results in an inverted preconception.

The most common cause of quality differentiation was the stuttering due to the low frame rate. This applied to nearly everyone to some extent, however, there were 26 participants who experienced major, significant differences in frame rate and evaluated accordingly (see Fig. 8). That means that while streaming on one device was experienced as smooth and fluent, on a different one was labeled as quote “serious stuttering”. The distribution of the scores of these participants regarding measurement scale is quite uniform; 6 of them belong to DN and CL, and 7 to DL and CN. Some of them pointed out rather notable dissimilarities in quality assessment. On the discrete scales, 6 of them had a score difference of 2 or greater. These are of course mean scores for the devices, the actual difference between stimulus sections can even surpass it; i.e., in DN, while a specific section was rated 1 on one device by an individual, it was given 4 on the other one.

If we divide the participant into two clusters (see Fig. 9), based on witnessing significant changes in frame rate, it is revealed that the primary concepts of the overall mean scores are unmodified, i.e., the assessment difference between iPhone and HTC is still 3 times bigger than between HTC and LG in the DL cluster of 18 participants. This separation of scores also demonstrates that the differences in quality assessment are indeed the smallest in case of the remaining group of CL; it already had the smallest differences (see Fig. 5), but the separation highlighted the finding.

All in all, it needs to be stated that not even 1 person out of the 100 made an unvarying evaluation; everyone distinguished streaming quality at some point at least to some extent. There were 13 participants using discrete scales who assessed quality the same on 2 out of 3 devices. On the continuous scales no such evaluation pair was made, since continuity allowed score differences to be as tiny as possible.
V. CONCLUSIONS

The paper has presented our empirical study regarding the role of end-user devices in subjective assessment of service quality. The results provided by the 100 people participating in our series of measurements have shown that perception and memory can be easily altered by preconception through the dissonance reduction method of cognitive dissonance, and that user devices can play an important role in the process. More than every fourth participant experienced significant differences between test stimuli which were in fact identical, and no unvarying evaluation was made at all. We also encountered some interesting phenomena, i.e., the preconception overcompensation phenomenon, in which an unproven preconception becomes its opposite. In alignment with previous findings, we also found that a continuous qualitative scale comes with the smallest scoring differences in such a measurement scenario, since it enables diminutive distances between evaluations and even though tags are indeed subjective, they are not as arbitrarily utilizable as quantitative numbers. In our future works, we shall continue to investigate objective assessment of multimedia quality in modern day scenarios, with the addition of involving objective and hybrid measurement methods. Modeling distortions of human evaluation may lead to higher levels of service quality prediction accuracy and can enhance our understanding of user satisfaction.

ACKNOWLEDGMENT

The research leading to these results has received funding from the European Union’s 7th Framework Programme ([FP7/2007-2013]) under grant agreement n° 288502 (CONCERTO project). At the time of final submission, P.A. Kara was supported by Kingston University through the European Union’s H2020 Framework Programme (H2020-MSCA-ITN-2014) contract n° 643072 (QoE-Net). The authors would like to thank all the organizations and the many individuals whose work and participation made this research possible.

REFERENCES