ASSESSING THE VALIDITY OF SUBJECTIVE QOE DATA THROUGH RATING TIMES AND SELF-REPORTED CONFIDENCE

Werner Robitza
TU Berlin / Telekom Innovation Laboratories
Ernst Reuter Platz 7
10587 Berlin
Germany

Helmut Hlavacs
Universität Wien
Entertainment Computing Research Group
1090 Wien
Austria

ABSTRACT

When users give ratings in subjective Quality of Experience experiments, they may experience a lack of confidence in assigning a score to a stimulus. This results in invalid votes that might need to be removed from the acquired data. We show results from a subjective video quality experiment in which we measured the rating times and asked observers for their confidence, in addition to just the quality score. It is shown that the average confidence is higher for extremely good or bad MOS values. The results also demonstrate that collecting rating times is an effective way to detect issues with single votes caused by observers struggling to decide for a rating.

Index Terms— Quality of Experience, Subjective Quality Assessment, Rating Times, Human Factors

1. INTRODUCTION

The creation of objective models to predict the Quality of Experience (QoE) of a service requires the existence of ground truth. Subjective experiments that follow a well-defined procedure are the primary means to acquire this data. These psychophysical experiments traditionally aim at gathering data in a controlled environment, so as to produce repeatable results and eliminate the number of confounding variables between and during individual test sessions. The most frequently used methodologies to assess subjective (video) QoE are defined in ITU recommendations, such as P.910 [1] and BT-500.13 [2].

A common denominator in above-mentioned methodologies is that they require participants to give a rating to each stimulus (or pair of stimuli in the case of pair-comparison tests). The stimulus–response pattern is repeated for the full collection of stimuli, without the possibility for the observer to either actively select the stimulus they want to see next or to abstain from rating at all. In this regard, the response is forced, and might not represent the observer’s true perception when it was given under the pressure of having to complete the experiment session. Yet, it would be desirable to detect responses that could be invalid, in the sense of having knowingly being made in order to proceed to the next item without expressing an actual opinion in the rating.

In this paper we question the elicitation of forced ratings and their impact on the potential validity of the resulting data. Section 2 will discuss previous work done in this regard. In Section 3, we describe a subjective experiment in which we assessed the observers’ confidence and their rating times. The results from this experiment will be presented in Section 4. Finally, we discuss the implications of the findings and give recommendations on how experiment methodologies should be adapted to accommodate for potential validity issues.

2. MOTIVATION AND PREVIOUS WORK

Based on our experience with conducting subjective experiments in the video QoE domain, we found that experiment participants would sometimes struggle to decide on a rating for a specific stimulus. This would not occur throughout an entire session, but would still be noticeable when carefully observing the assessors. In the written instructions that are given to participants before an experiment with a one-dimensional rating scale, we usually inform them that they should vote quickly, without giving a lot of in-depth thought to what they had just seen, so as not to “over-think” in the process. Yet, the opposite seems to happen on occasion. We therefore hypothesize that users have troubles picking a rating on a presented scale when they are forced to. Also, we expect them to show a delay in rating when they cannot decide for a score. Therefore, it is important to be able to quantify how confident the observers were when providing a score. Ideally, it would be feasible to measure this confidence indirectly, without requiring additional time in an experiment session (e.g., for questionnaires).

ITU recommendation P.910 defines validity as the “agreement between (...) ratings obtained in a test and the true value
which the test purports to measure” [1]. Disagreement between ratings and the true value occur when the ratings do not express what researchers interpret them as, which is the case when observers pick ratings at random, or are unsure doing so.

Previous work in this regard includes the findings from Engelke et al. [3], who measured the confidence of human assessors in image quality experiments. The authors claimed that—together with visual detection thresholds and personal preference of certain error types—the internal confidence of observers would influence the final Mean Opinion Score (MOS) given to a stimulus. The subjective experiments conducted within [3] were based on wireless transmission errors and their effect on image quality. They followed a single-stimulus Absolute Category Rating (ACR) protocol [1]. Apart from asking for visual quality, subjects also had to report their confidence while rating on a five-point scale, similar to the ACR scale. The reaction time between the end of a stimulus presentation and the submission of the quality rating was measured with a stopwatch. The authors found a high correlation between the average quality ratings (MOS) and the average confidence ratings. However, due to only acquiring ratings from fifteen subjects, the experiment lacked enough data in order to be able to meaningfully address the impact of rating times on the confidence. Additionally, the use of a stopwatch may have resulted in inaccurate measurements of reaction time. The average reported rating time of observers was around 1.32–1.53 seconds. Human errors in measurement however may amount to 0.1–0.7 seconds [4], thus significantly impacting the results.

Sparling et al. [5] measured the impact of different rating scales on rating time and cognitive load for web-based product reviews (e.g., “like buttons” or star ratings). They found that the amount of items on a rating scale is positively correlated with the average rating time. Also, users would spend more time evaluating the central items of a scale rather than the extremes, which corresponds to the findings from Engelke et al.

Our goal was to improve the methodology by which confidence and rating time are measured, and acquire data from video (instead of image) quality experiments. We targeted a higher number of observers in order to collect more data points for regions in which [3] found no significant influences, and increased the precision of the acquired rating times.

3. EXPERIMENT

For gathering confidence scores and quality rating times, we conducted a subjective video quality experiment. 27 users participated in this experiment: 12 male and 15 female. Their average age was 32, ranging from 19 to 49 ($\sigma = 11.13$). We focused on the effects of error concealment techniques for H.264/MPEG-4 SVC transmissions. Scalable Video Coding (SVC) is an enhancement of H.264/MPEG-4 AVC, which allows for multimedia transmissions to contain multiple streams at different quality levels.

3.1. SVC Error Concealment

In SVC, scaling of quality can be achieved temporally (by reducing framerate), spatially (by reducing coded picture dimensions) or in terms of quality (by increasing the amount of lossy compression). In a simple setting, the SVC stream could consist of a base layer and an enhancement layer (with better quality). If there are sudden limitations in bandwidth or transmission errors, an SVC decoder may switch from one of the enhancement layers to the base layer—provided a good enough transmission rate. This comes at a cost of visual degradation of the complete stream shown to the viewer, even if only parts of the transmitted pictures might have been affected by the loss. Instead of switching to the base layer entirely, the decoder could also try and merely conceal impairments in the higher SVC layers by taking information from the base layer.

3.2. Source Material and Test Conditions

The test material for this experiment is based on the IRC-CyN/IFC SVC4QoE Replace Slice Video VGA database\(^1\), with 9 SRC videos in VGA resolution ($640 \times 480$) encoded at 30 Hz framerate: Aspen, BoxingBags, HalfTimeWide, Highway, MesaWalk, Powerdig, rbnews, ShadowBoxing and SkateFar. Each video had a total duration of 10 seconds. There was no audio stream.

To generate the conditions (14 in total, plus one reference condition) for this database, the authors of [6] simulated a transmission by removing slices from the source bitstreams based on a loss simulator. The loss was applied in such a fashion that only one slice out of four in a picture was removed, and that visually important regions of a scene would be affected. Hence, there was no fully randomized error pattern. The duration of the error was one second. Then, the base layer was encoded at either 15 or 30 Hz framerate, with bitrates of 120 and 200 kBit/s, respectively. In order to reconstruct the stream from the distorted transmission bitstream, the missing slices or frames were patched into the enhancement layer from the base layer.

A total of 135 stimuli (processed video sequences, PVS) were generated from these conditions.

3.3. Test Setting and Protocol

We conducted the experiment on a Samsung Galaxy Note 10.1 tablet with the videos being shown in their native resolution on a grey screen. We implemented the presentation of the stimuli with a custom interface for the Android operating system. After each PVS, observers could rate the quality on

---

\(^1\)http://www.irccyn.ec-nantes.fr/spip.php?article768
hypothesized that an overall index of self-esteem would be strongly correlated with the average confidence given during an experiment. Knowing about this effect could make it easier to interpret the confidence scores taken.

We therefore additionally administered the Rosenberg Self-Esteem Scale (RSES) to the participants [7]. It is a questionnaire-based test with ten items to be agreed or disagreed with (on a Likert-type scale). It includes statements such as “I am able to do things as well as most other people”. The RSES allows a maximum test score of 40 (indicating high self-esteem), with a typical population average of 30. The test takes no longer than a few minutes to complete.

The questionnaire was given to the participants before any introduction to the nature of the experiment they would be taking part in. To further check the impact of the actual video quality test, we administered the questionnaire at the very end of the experiment session.

4. RESULTS

4.1. Usage of Confidence Scale

We first look at the distribution of confidence scores in order to identify how the rating scale was used. Figure 2 shows that for most ratings, users felt confident (51%) or very confident (28%). Only in very few instances they felt inconfident (3.8%) or very unconfident (0.3%). We expected a distribution in this shape—otherwise the overall design of the experiment might have been questionable if it had only provoked unconfident responses.

In [3], Engelke et al. report confidence scores for their subjective image quality experiment. There, observers showed an average confidence of 4.252 (n = 1200, σ = 0.773) for distorted images. In our experiment though, the average confidence was 3.977 (n = 2080, σ = 0.768), which is significantly lower (p < 0.0001). The reason for this difference can not be fully explained without reproducing the original experiment. Potential influence factors include the test environment and viewing device. The social background of the test participants (in fact, even the country of the experiment) may have also impacted this score. We however generally attribute the lower confidence to the usage of video as test material instead of images, which due to their increased temporal complexity—also in the form of temporally propagating errors—may be harder to rate.

4.2. Distribution of Rating Times

Of particular interest to us was the distribution of the quality rating times t_q (in contrast to the overall rating time which includes the confidence rating). How long would participants take to decide for a score? Figure 3 shows a long-tailed shape, with very short rating times predominating. Notable outliers exist for t_q, where there are six instances in which observers had taken more than 20 seconds to give a rating.
The average rating time was 1.86 seconds ($\sigma = 2.45$). 75% of all ratings were performed within 2.11 seconds, and the 95% quantile is at 4.73 seconds. We conclude that high rating times (e.g., above five seconds) are a strong indicator of possible validity issues with the actual quality score given, since they constitute outliers and therefore hint at the observer struggling to make a decision or being otherwise distracted.

4.3. Relation between Quality and Confidence

We first give an overview of mean confidence ratings compared with MOS. The results can be seen in Figure 4. Each dot corresponds to a PVS. The overall relation can be described well by a second order polynomial fit ($R^2 = 0.67$):

$$\text{Mean Confidence} = 4.65 - 0.99 \cdot \text{MOS} + 0.16 \cdot \text{MOS}^2$$

We expected a drop in confidence for mediocre quality ratings. The presence of strong visual degradations might trigger a reaction where subjects judge the quality as absolutely unacceptable—there would be no doubt about the score, since it could not be worse. Likewise, a stimulus where observers were unable to perceive any impairments will be assigned a perfect score. For medium quality levels however, subjects have to translate their internal rating into the specific item on the ACR scale, which in turn may result in indecisiveness and thus lower confidence.

4.4. Influences between Rating Time and Confidence/Quality

To further inspect the user behavior, we examined the average time to rate the quality ($t_q$) for each specific quality and confidence score. Figure 5 shows these results. On the x-axis we differentiate the quality scores 1–5 (“Bad” through “Excellent”) and confidence scores (“very unconfident” to “very confident”). The y-axis shows the average $t_q$ for the respective quality or confidence score.

We can observe a large average $t_q$ of 4.89 seconds ($\pm 3.31$ at 95% CI) for ratings where the observers were “very unconfident”. However, since there are only 11 ratings at this data point, a very large CI is expected. Nevertheless, there is a strong tendency of rating times increasing as confidence decreases, which confirms our statement that high rating times may be a sign of problematic votes, since they seem directly correlated to low confidence. It also refutes the findings from [3], where the rating time had surprisingly seemed to decrease for lower confidence ratings (however, in this case, only one single rating was given for the combination of “Bad” and “very unconfident”, and the authors declared it as
Average confidence scores would be predicted.

**4.5. Predicting Confidence from Rating Times**

To go a step further, it would be desirable to infer the observer confidence (and thus, the validity of ratings as indicated in Section 2) from the rating time. To do so, we model the confidence score as an ordinal dependent variable of the continuous independent variable $t_q$.

We performed a multinomial logistic regression [8] using the nnet package in the R software for statistics. The regression determines the likelihood of choosing a specific confidence level based on the time needed to rate the quality. This contrasts to a linear model in which continuous confidence scores would be predicted.

Before calculating the model, we removed extreme outliers where rating times were over 10 seconds. To visualize the results, we show the probabilities of a certain confidence being chosen at specified values of $t_q$ in Figure 6. As clearly visible, the odds to find a very high confidence drop significantly after several seconds. After roughly 5 seconds of rating, it is already more likely that an observer was unconfident. The model coefficients are given in Table 1 and allow to calculate the odds of choosing a confidence other than “very confident”.

**4.6. Self Esteem and Overall Confidence**

In Figure 7 we can see the scores from the pre- and post-test RSES (Rosenberg Self-Esteem) evaluation compared to each observers’ average confidence. For the pre-test RSES, we measured a significant correlation between RSES and average confidence of $r = 0.621 \ (p < 0.001)$. The post-test RSES, while still significant, does not correlate as well with the average confidence ($r = 0.477, p = 0.014$), which indicates an influence of the actual experiment on the self-esteem, and diminishing the usefulness of the post-test RSES value. The third order polynomial mapping function ($R^2 = 0.56$) is:

$$c = 0.001r^3 - 0.107r^2 + 3.465r - 33.86$$

where $c$ is the average confidence per user and $r$ their RSES pre-test score.

**5. DISCUSSION**

**5.1. What causes the lack of confidence?**

The distinct shape of confidence vs. quality scores shown in Figure 4 raises a question of causality: On the one hand, we may ask whether lower confidence scores for mediocre quality are caused by the content being harder to rate. In fact, as already proposed in Section 4.3, the presence of strong degradations or the absolute lack thereof may be easier to judge, whereas observers might find it harder to judge a subtle impairment as “good” or “poor”. On the other hand, observers who are generally unsure about their rating (e.g., because they did not pay attention to the video playing) may settle for middle scale items rather than the extremes, so as to reduce the possible error they make. As an example, consider the case for a strongly degraded sequence: an observer who did not see the video would rate it as “fair” with the absolute error being smaller than if they had rated it as “excellent”.

**Table 1: Multinomial logistic regression coefficients.**

<table>
<thead>
<tr>
<th>Response</th>
<th>Intercept</th>
<th>$t_q$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Very unconfident</td>
<td>-5.972</td>
<td>0.698</td>
</tr>
<tr>
<td>Unconfident</td>
<td>-3.120</td>
<td>0.619</td>
</tr>
<tr>
<td>Neutral</td>
<td>-1.386</td>
<td>0.529</td>
</tr>
<tr>
<td>Confident</td>
<td>0.139</td>
<td>0.307</td>
</tr>
</tbody>
</table>

**Fig. 6:** Predicted confidence levels as a function of quality rating time.

**Fig. 7:** Pre- and post-test RSES results compared with average confidence and third-order polynomial fit.
5.2. How should confidence be measured?

We believe that the ordinal confidence scale as used both in the previous literature as well as in our tests is problematic in its use. While it aligns with the ACR scale in terms of items, the overall confidence of an observer is not described very well by a Likert scale with a middle “undecided” item. In fact, we do not expect a person to be undecided about whether they are sure about a statement or not. At least in the context of this experiment, the responses to the question about confidence should be considered valid.

In addition to the semantic problems with the scale, the (semi-)ordinal responses in the form of Likert items require more complex data analysis. More generally, while assigning a number to the items and averaging the score is typically done for MOS, it is still a statistically questionable practice [9], because the underlying assumption of the items being equally spaced is not guaranteed. A visual analog scale (VAS) may be more suited for this task: the level of confidence would be reported on a slider from 0 to 100%. Further investigation has to be performed in this regard, but we expect a VAS to show more “honest” responses with less inherent bias than an explicitly labeled scale.

Lastly, it has to be considered that the rating methodologies we tested may in fact not be as prone to confidence drops than other experiment protocols. In fact, the impact of the rating methodology and the chosen test design may be a critical factor: a very narrow range of quality could increase the confusion among participants, and thus lower their confidence [10]. Still, despite rare occurrences of low confidence votes and the general mental demand of subjective experiments, observers seem very optimistic and determined in their ratings, which results from our qualitative surveys support.

6. CONCLUSION

In this paper, we presented the results of a subjective video quality experiment that aimed at investigating the relationship between rating times and self-reported confidence. Previous work in this direction only focused on image quality, and the acquired rating times might have been too imprecise due to the use of a stopwatch instead of computer-based measurements. The aim of our experiment was to acquire data for video quality experiments, and increase the precision of the measured rating times.

Our data shows that measuring rating time is not only beneficial to understanding whether users felt unconfident during a specific judgement, but also an efficient means to check the experiment data for invalid responses. Lack of confidence occurs either because the observers were distracted during the stimulus presentation, or because they had problems judging the impairment itself. We suggest that rating times above a certain threshold should result in the ratings being pruned from the dataset (unless a large number of votes for a stimulus makes the impact of invalid responses negligible). This threshold is suggested to be around 5 seconds, above which it is much more likely to find a subject being “unconfident” than “very confident”.

Naturally, since unconfident responses only happen rarely, more data would be needed in order to increase the explanatory power of existing models. Further experiments focusing on audio-only stimuli would help in determining whether different sensory experiences have noticeable impact on the observers’ confidence.

7. REFERENCES


