Evaluating Feedback Devices for Time-Continuous Mobile Multimedia Quality Assessment

Shelley Buchinger, Werner Robitza, Matej Nezveda, Ewald Hotop, Patrik Hummelbrunner, Martijn C. Sack, Helmut Hlavacs

University of Vienna
Entertainment Computing research group
Währingerstr 29, 1090 Vienna, Austria

Abstract

In January 2014, the new ITU-T P.913 recommendation for measuring subjective video, audio and multimedia quality in any environment has been published. This document does not contain any time-continuous subjective method. However, environmental parameter values are changing continuously in a majority of outdoor and also most indoor environments. To be aware of their impact on the perceived quality, a time-continuous quality assessment methodology is necessary. In previous standards, targeting laboratory-based test settings, a desk-mounted slider of substantial size is recommended. Unfortunately, there are many environments where such a device cannot be used.

In this paper, new feedback tools for mobile time-continuous rating are presented and analysed. We developed several alternatives to the generally adopted desk-mounted slider as a rating device. In order to compare the tools, we defined a number of performance measures than can be used in further studies. The suitability and efficacy of the rating scheme based on measurable parameters as well as user opinions is compared. One method, the finger count, seems to outperform the others from all points of view. It was been judged to be easy to use with low potential for distractions. Furthermore, it reaches a similar precision level as the slider, while requiring lower user reaction and scoring times. Low reaction times are particularly important for time-continuous quality assessment, where the reliability of a mapping between impairments and user ratings plays an essential role.

Keywords: Subjective quality assessment, Quality of Experience,
1. Introduction

For a long time, subjective multimedia quality assessment has been performed in a fully controlled laboratory setting, following the guidelines described in ITU recommendations [1, 2, 3, 4]. Due to the increased number of mobile devices, this traditional approach does not seem adequate anymore. Alternatives, such as (semi) living-labs have been considered [5, 6, 7], but also measurements in realistic environments have been performed [8] in the recent past. Based on these considerations, a new ITU recommendation defining measurement settings for subjective video, audio and audiovisual quality in any environment was published in January 2014 [9]. In this specification, a set of five methods and acceptable as well as discouraged changes to these methods are presented. In contrast to previous recommendations, the new ITU-T P.913 document does not contain any time-continuous rating method.

At a first glance, collecting single ratings for each test sequence seems to have the advantage of being able to define a mapping between user ratings and media quality. But in a mobile testing scenario, this is only partially true: In general, the stimuli that are used have a duration of around 8 to 10 seconds. In [1] it is stated that “for still pictures, a 3-4 second sequence and five repetitions (voting during the last two) may be appropriate” and that “for moving pictures with time-varying artefacts, a 10 s sequence with two repetitions (voting during the second) may be appropriate”. In the new standard [9] it is specified that stimuli should range from 5 to 20 seconds and that “eight- to ten-second sequences are highly recommended”.

The length of 8 to 10 seconds has been defined to avoid uncertainties that might be caused by the primacy and recency effect. Those psychological effects explain why judgments are increasingly based on earlier or later parts of the sequences. However, within a time frame of 10 seconds, environmental conditions can change drastically and can seriously affect the experienced quality. Take for example a truck driving by, casting a shadow on the users’ device and creating a strong background noise. In this work, we did not analyze the effects on changing environments on the perceived quality. Here, we simply want to point out that the usage of a time-continuous quality assessment methodology seems to be unavoidable when allowing any environment for subjective testing as aimed at in ITU-T P.913 [9].
One of the reasons why time-continuous methods have not been included in ITU-T P.913 [9] might be that no adequate rating device is available for outdoor or even mobile quality assessment. According to [1, 2], time-continuous subjective tests have to be performed by using a desk-mounted slider. Such a slider cannot be used for measuring the quality of mobile multimedia. It is too large to be carried around and it is impossible to perform ratings while consuming mobile multimedia.

To overcome this problem, we developed several different tools for a time-continuous rating in mobile quality assessment. We estimated that some of these tools would be more applicable for mobile quality assessment than others. Therefore, we defined a set of measures that enabled us to compare the performance of the time-continuous rating schemes.

We decided to compare them to the currently used slider according to objective and subjective criteria. The precision of the rating methodology represented an important aspect of our research. We also computed the time needed to react and to perform the intended rating. We estimated the potential distraction that occurred due to the rating methodology and we anonymously collected biometrical data that might have impact on the efficacy of certain rating methodologies. Finally, we gathered the users’ opinions on each single method and asked the test persons to rank them in respect to their subjective preferences.

The paper is structured as follows. In Section 2 we present a selection of related work concerning rating devices and scales. In Section 3, the different test methodologies and their implementation are described. Three different user tests were carried out to assess the performance of the test methodologies. Those tests and their selected quality criteria are described in Section 4.

2. Related work

Research on the suitability and comparability of different subjective tests has been frequently carried out [10, 11, 12]. It was found that the choice of the rating scale is not very critical. For example, user ratings based on an 11-point scale can be translated into a 5-point scale [12] without loss of information. This is beneficial, since it seems to be convenient to use the simplest scale for mobile testing, due to the complexity of the situation. However, the issue of the appropriateness of the rating device itself has been rarely addressed. In [9] it is mentioned that “voting may be recorded with
paper ballots or software”. Neither paper ballots nor software running on the viewing screen can be used for assessing time-continuous user feedback.

Different types of rating equipment have already been compared in [13]. This study was published in 2005 when mobile multimedia usage was still very limited. A joystick, a sliding bar, a throttle and a mouse have been taken into consideration. Unfortunately they do not represent suitable alternatives for mobile use. The performance of the test equipment has been evaluated by means of a user test that focused on four main criteria: (1) the user friendliness, (2) the simplicity of use, (3) the visibility and clarity of the functional units, and (4) the feeling for the position within the evaluation scale. Important properties such as the rating precision and the potential amount of distraction was not considered. In this publication, the best result was obtained by the slider.

It seems that one reason why time continuous measurement cannot be considered in a mobile test scenario is the lack of a portable and reliable rating device. As mentioned above, we estimate that time continuous measurement is of important value for mobile testing in real environments. We decided to focus our research on two topics: (1) the device itself, its shape and the according rating procedure (we searched for possible alternatives that could be easily used everywhere – even while commuting) and (2) the design of a testing protocol that can be used in general to check the suitability of a rating device.

Experimental results can be found in Section 5 and finally, conclusions are drawn in Section 6.

3. Material and methods

Large or heavy devices are inappropriate for mobile multimedia quality assessment. Since they are uncomfortable to use, they contain a high potential for distractions. In order to find a suitable alternative to the slider for future testing, we developed four alternatives that will be presented below.

3.1. Slider

The slider is probably the most common tool of continuous quality assessment in user laboratories. The usage of a slider is defined in [1]. An example of such a device can be seen in Figure 1. According to [1], the slider should be mounted on a desk or another horizontal surface. Its travel range should be 10 cm long and the values from the slider should be recorded at least
twice a second. For our experiments a sliding bar using a 10 cm linear Alps potentiometer has been used. The values were converted by a 10-bit A/D converter (LTC 1090) and were serialized for a RS-232 connection. These values were read out by our own software and were then inserted into our presentation software called Subjective Player.

Using the slider is popular for several reasons. On the one hand, the parts necessary to assemble the device are very cheap. On the other hand, the flexibility of rating scales allows for many types of assessment: The slider could be labelled with an arbitrary number of points to simulate different rating scales, e.g. 5-point or 11-point scales. Moreover, the beginning and end of the scale is easy to reach for the observers. Although feasible, the ITU does not recommend locking the potentiometer in certain positions so as to facilitate categorical rating. This is probably due to possible distraction. However, during our experiments, test persons were told not to look at the slider’s position. Another advantage is that with the standardized use of sliders, test beds in laboratories can be set up easily.

However, the standardized use is also one of the sliders’ biggest disadvantages. Since the slider needs to be mounted on a desk and requires the
observer to use their hand for rating, this device can never be used for mobile quality assessment. Also, persons with too small hands might need to reposition their hand while rating. This can be seen in Figure 1, where the test person can not reach the full scale without repositioning.
Figure 2: Glove
3.2. Glove

An alternative to the slider is a digital glove that uses at least one sensor to capture the bending of the hand and/or several fingers. Such a glove is depicted in Figure 2. For our experiments, a glove manufactured by 5DT Technologies was used.\(^1\) The glove used for the experiment features a total of 14 sensors, each placed at a different position. There are two sensors per finger ("near" and "far") and additional sensors between each pair of fingers. Although there are several models with different sensor counts available, it was found that one sensor per finger should be enough to capture the position with good enough accuracy. When analyzing the data, only the sensory output of index and middle finger can be used.

Also, there are models available for left and right handed users. The outer material is elastic. Therefore, the glove should fit all hand sizes. In preliminary experiments [14], users with smaller hand sizes had to wear a second woolen glove underneath the digital glove. This measure improved the quality of the obtained values, however, test subjects found it rather straining because moving their fingers required higher muscular efforts and the hand easily started to sweat, especially during the summer period.

The glove’s main advantage is portability: Because it fits the hand almost perfectly and the rating gesture is very natural, there is only minimal distraction from the rating device. The USB connection would allow for a laptop to record the values while moving. There is also a wireless module available which would free the test subject from any necessary wiring.

Because of the implemented technology, the glove is very expensive. Even if the software can be easily programmed, higher costs should be taken into account. Some other disadvantages result from the lack of a distinct scale. In our preliminary experiments, it was found that a scale compromising more than five points can not be accurately represented when using the glove. Users have been instructed as follows on how to use the device properly:

Rating is performed by opening and closing the hand, i.e. opening the hand results in a good score, closing the hand to a fist is interpreted as a bad score. While opening their hand, they had to take care not to over-bend the fingers. This would result in inaccurate sensor data. We also considered an alternative rating procedure where five different hand positions were shown to the user on a clock model, but the representation of low values was found

\(^1\)http://www.5dt.com
to be tiring.

According to [1, 2] the corresponding sensor values should be recorded at least twice a second. In our experiment, the values are polled every few milliseconds by a software written in C++.

3.3. Finger distance

Another method tested for subjective scoring consists of measuring the distance between the thumb and index finger. A larger distance equals to a good score, the smallest possible distance results in a bad score. The gesture can be described to the users as “pinching” and can be seen in Figure 3.

To measure the distance, a web cam was positioned at a height of $\sim 50$ cm, recording the hand from above. To capture the position of each finger, two knitted caps had to be put on by the users, each in a different color. To find the corresponding position of the markers a software was implemented using C++ and OpenCV\(^2\) library. We decided to use the camshift algorithm for tracking of the markers. It performs well if the objects are single-hued and hue alone allows the objects to be distinguished from the background and other objects. These requirements were perfectly covered by our testing conditions. Nevertheless it can occur that the tracker is lost because of an

\(^2\)http://opencv.willowgarage.com/wiki/
abrupt motion. Therefore the test persons were asked to give their ratings by a smooth movement of their fingers.

In order to initialize the tracking, the markers were selected manually with a bounding box at the beginning of each test. The euclidean distance in pixels between the centers of the two markers was calculated every $\sim 50$ ms. In an off-line step these distances were mapped to a five point scale for every person separately.

An advantage of this method is the simple gesture. Furthermore the minimum and maximum of the scale are easy to reach. Unfortunately, it can be argued that this method is not fully suitable for mobile testing and in particular while commuting. The reliability of the finger distance approximation might be limited due to some involuntary camera motions and uncontrolled distance variations between the camera and the fingers. However, an example for a wearable gestural interface that performs well can be found in [15]. In this case a tiny projector and a camera were mounted on a hat.

3.4. Finger count

Showing a certain number of fingers is probably one of the simplest forms of representing a score. With possible discrete values from 1 to 5, the standard absolute category rating (ACR) scale can be represented. The advantage of this method is that it seems to be very intuitive. Moreover, counting or scoring by using hand gestures is very common. When trying to refine this scale, for example by showing half points, the method would probably be less intuitive and its precision would suffer.

3.4.1. Web Camera based implementation

As with the measurement of the finger distance, a web camera was used to record one hand of the test subjects from above. This setup can be seen in Figure 4.

The system was developed in C++ with the assistance of OpenCV for image and cvBlobsLib\(^3\) for blob processing. The main aim was to separate the user’s hand from the background in order to recognize the gesture. As background we used a black cardboard (see Figure 4) because it has a high contrast to the skin color of caucasian people and thus simple segmentation techniques can be applied. Moreover, all test persons were wearing a black armband around their wrists to separate the hand from the rest of the arm.

\(^3\)http://opencv.willowgarage.com/wiki/cvBlobLib
Figure 4: Finger count
First, each frame in RGB was blurred with a $7 \times 7$ kernel to eliminate noise. Next, the frame was converted to a binary image with a fixed threshold. These steps resulted in two blobs, namely the user’s hand and the rest of their arm. Due to the positioning of the camera, the hand blob always had a bigger area and was hence easily selectable.

To count the number of fingers, we used the method described in [16]. A circle was placed at the center of mass of each blob, passing through each outstretched finger. Then, the number of intersections between background and foreground was counted. The result was divided by 2 to consider the two edges of a finger.

The radius of the circle is of big importance. A too small or too big circle leads to wrong results. Therefore we introduced an initialization step for every test person at the start of the experiment. First, the area of the user’s hand showing five fingers was calculated. To estimate the radius for a test person, the measured area was divided by a constant. The constant depends on the hand size and was defined for every person individually.
3.4.2. **Glove based implementation**

The idea of expressing the quality by showing a certain finger count was strongly accepted by our users. However, similar to the finger distance method described in Section 3.3 the suitability of the proposed approach seems to be rather limited for mobile use. Again, camera motions and uncontrolled distance variations between the camera and the fingers might have some impact on the reliability of collected data. Therefore, we decided to implement the finger count method by using the data glove described in Section 3.2.

In Figure 5, a test person wearing the glove and showing the number 3 is depicted. The “far” and “near” sensors were used to check whether the finger is flexed. In order to investigate the status of each finger, i.e. bent or outstretched, fixed thresholds for the sensors were determined. When polling the current finger count, the thresholds are evaluated against each other to estimate the number of fingers shown. To identify these thresholds, a test person traced all sensor values while showing each of the five numbers. The selected sensor values for counting one finger have been determined to be:

- Thumb near < 0.3
- Index far < 0.6
- Middle far < 0.6
- Ring far < 0.6
- Little near < 0.2.

As mentioned before, persons with small hands were asked to wear a glove under the digital glove in order to receive comparable sensor data.

3.5. **Moving a mobile phone**

Another very intuitive method for quality assessment utilizes the acceleration and orientation sensors of a mobile phone. The test person moves the phone in a semicircle from portrait position to upside-down position according to the desired score (see Figure 6). Very similar to the thumb-up/thumb-down gesture, the “normal” position represents the best score, while the worst score is indicated by holding the phone upside-down. If desired, the continuous values of the angle can be transformed into discrete values (e.g. 1 to 5). This is achieved by dividing the semicircle into sectors
of equal size – in this case 36 degrees. Because of this, the mobile phone methodology is interval-based and it was only assessed whether users could reach the specified interval.

3.5.1. General approach

The mobile phone that was used for our experiments was a HTC Hero\textsuperscript{4}, which is based on the Android operating system already mentioned before. Our application on the mobile phone was implemented using Java in combination with the Android SDK to access the sensors. For synchronizing the start of the video with the mobile application, we sent a single packet over the WiFi connection to the mobile phone that would indicate the application to begin logging. From the raw sensor data of the magnetic field and acceleration sensors (in the device coordinate system), the rotated values for

\textsuperscript{4}http://www.htc.com
azimuth, pitch and roll (in the world’s coordinate system) were computed using built-in Android methods. For our purpose, pitch is the value of choice – it ranges from -90 degrees (portrait) to 90 degrees (upside-down). To assure comparability of results, we then converted the continuous values into five discrete scoring levels.

In comparison to other methods, the mobile phone has the great advantage of mobility: it is light and has at least one built-in wireless connection. The test person is usually accustomed to the handling of a mobile phone and it does not matter if the user is left or right handed. A smart phone may not be as cheap as the slider, but often, there already is one available. Finally, since a modern mobile phone is very flexible, one could think of other gestures for the scoring – this would be just a minor adaptation of the application. When designing such an alternative gesture or rating mechanism using the mobile phone one has to keep in mind that there are some restrictions caused by the rating scenario. The assessor might only one hand available for the rating device because he needs the other hand for holding the viewing device. Secondly, the evaluators attention has to be focused on the content to be viewed and not on the rating procedure. Summarizing, only single-handed gestures that are easy to perform are possible. For example, turning the mobile phone in one hand might be risky: the device could easily fall on the floor or the user could lose orientation for example by forgetting the position of the point of origin. We preferred to use the sensors over a touch screen as input mechanism because people might be tempted to look at the rating device. The potential for distractions seems to be rather high. For the same reason, we decided not to use the trackball or single buttons.

3.5.2. Haptic feedback

The described method can be extended by adding haptic feedback. Just like the markings on the slider, the borders between the five discrete scoring intervals can be indicated using the vibration functionality of the mobile phone. This is very convenient because the test person is not forced to look away from the video to ensure the chosen level is correct. Since the sensor values often fluctuate around the actual position, we introduced an interval of +/- 2.5 degrees around the borders of the discrete scores. Here, the scores remained unchanged (see Figure 7).
Figure 7: Five segments with interval
4. Theory

4.1. Experimental settings

Since we aimed at studying the performance of the rating methodology itself, we chose to set up a specific test scenario that differs from the typical multimedia quality assessment settings. Instead of presenting stimuli containing artificially introduced impairments, we used videos where every five seconds a new number or category was shown to the users. The numbers were randomly generated between 1 and 5 and the categories were randomly selected from the five ACR categories “excellent, good, fair, poor, bad” [1, 2]. The content was displayed in white bold letters in front of a black background. Each time a new number appeared on the screen, the user had the task to find the rating value displayed with the targeted rating methodology.

Three user tests were performed to identify the necessary requirements of a mobile rating scheme and to understand the users preferences and reasons for refusal.

4.1.1. Test 1: Slider versus glove

First, we performed a subjective test described in [14] where the glove was compared (1) to a visible and (2) to a hidden slider, which has been simply covered with a towel. The experiment showed that the glove represents an attractive alternative to the slider because the user satisfaction is high (average MOS for glove and slider: “good” and the open slider was preferred), the rating delay is short (glove: 500 - 1000 ms and slider: 1000 - 1500 ms). It is not likely that test persons are distracted by the equipment (below one second for a simpler task). Even if the precision of assessed data revealed to be within acceptable bounds, we noted that this aspect needed to be improved. Furthermore, it was not clear whether other completely different approaches might not be even more appealing to the users providing similar rating performance.

4.1.2. Test 2: Several alternatives

Because of this, we started to develop several alternative rating methodologies and performed a second user study. The methods (1) slider (visible and hidden) (2) glove (3) finger distance (4) finger count with camera (5) mobile phone (with and without haptic feedback) were evaluated by more than 15 test persons each. Results of this test revealed that the finger count method is appreciated by the users. We observed that the time to react and
to perform a score is low. It is also rather similar for different scores (e.g. “bad” and “excellent”) and users in contrast to the other methods. More detailed results on this study will be provided in Section 5. The major drawback of the users’ favorite approach – the finger count – consisted in its static implementation by using a camera.

4.1.3. Test 3: Improved finger count

Hence, we decided to conduct a third user study where we compared the rating methodologies (1) slider (2) glove (3) finger distance and (4) finger count with glove. The mobile phone was excluded this time because the users seemed to reject it in previous studies: the device was perceived to be too big and the potential for distractions too high. Since we aim at presenting our final results, the test settings of the third test are presented in more detail below. However, a very similar setup had been adopted for the other two studies.

Before taking part in the third experiment, 12 male and 3 female participants with an average age of 29 years were asked to fill out a questionnaire, where some of their physical qualities were assessed (sex and age, hand size). In Section 5 the impact of hand sizes on the rating precision will be analysed and presented.

The experiment consisted of three main phases with a total duration of \( \approx 40 \text{ to } 45 \) minutes: introduction (\( \approx 3 \text{ min} \)), familiarization and accuracy task including a satisfaction questionnaire for each methodology (\( \approx 5 \text{ to } 6 \) min each) and finally a mobile rating task followed by a satisfaction questionnaire (\( \approx 6 \text{ min} \)).

Dedicated stations for the different rating methods were set up. At the beginning, during the introduction phase, the experiment and the different methodologies were presented to the users in general. Then, test persons moved to one of these stations. There, they received more explanation on the method, in order to get accustomed to the device. Then, users performed two tests in random order: one time using videos showing random numbers between 1 to 5 and another time using videos containing the five ACR categories. All videos were exactly one minute long, including an introduction and a message at the end, thus leading to 10 ratings each user had to perform.

All the tests started with a shorter training session, i.e. a familiarization task. Only when the test persons were accustomed to the entire test setting, the accuracy task was performed. After each test, the users were asked to complete a satisfaction questionnaire where they had to rate the methodol-
ogy using the five ACR categories and to justify their choice by providing some comments. This procedure was repeated for each rating methodology described above. The test persons visited the rating stations in random order to avoid bias by always testing one method first.

Finally, users performed a mobile test in order to understand which rating methodology fits best for our major aim, the mobile quality assessment, in terms of user satisfaction. Test persons were asked to use the rating methodologies described above while walking through our building. This time, a normal movie containing impairments was shown, in order to reproduce a real subjective test. The test was concluded by compiling a satisfaction questionnaire, in which the users had to judge the methods according to the criteria defined in [13] that are relevant also for mobile use: the user friendliness, the simplicity of use and the feeling for the position within the evaluation scale. Furthermore, they were asked how distracting and tiring the applied methods were perceived. These opinions were provided by using a 5-point Likert scale. At last, the users had to select one method that they judged to perform best.

4.2. Criteria for comparison

In this investigation we aim at finding a new suitable alternative to the currently used slider for continuous video quality assessment. Therefore it is necessary to show that the results obtained by using a new method are as satisfying as the results obtained from the slider. On the one hand, the major criteria is the rating precision. A new method needs to be precise enough to adequately capture the user opinion. Errors should be as minimal as possible.

On the other hand it is also necessary to estimate how intuitive the new method might be. For this purpose we calculated two different rating delay times: the reaction time and the scoring time. The reaction time represents the time that has elapsed until the test person starts to rate while the scoring time is the time needed to perform a certain rating. Furthermore, the test users should not be distracted by the rating methodology itself because they need to concentrate on the presented content.

Finally, we asked the users about their opinion on the suitability of each method. There might be a very precise and intuitive approach for mobile quality assessment that needs to be discarded because it requires too much physical effort in the worst case leading to cramps after a certain test duration.
4.2.1. Precision

As described in Section 4.1, a video showing different numbers or categories was presented to the users. These numbers are denoted by \( n_1, \ldots, n_5 \) and the categories by \( c_1, \ldots, c_5 \). In our case, the rating scales comprised \( C = 5 \) elements each. Several repetitions \( R \) of each element have been shown to the user in random order. During the experiment, test persons had to match presented scores shown on the screen with their device for exactly \( N = C \times R \) times. In the following, this score is denoted as \( x^j \) where \( j \in [1, 2, \ldots, N] \). The mean rating value over time (a number or category is shown to the test person for 5 seconds) performed with one of our tools is denoted by \( y^j_i \) where \( j \) represents the same index used for \( x^j \) and \( i \in [1, 2, \ldots, M] \) identifies the test person that has performed the rating. \( M \) is the total number of users involved in the experiment.

In order to estimate the precision of the performed ratings, we computed the difference between the number or equivalent category shown to the user on the video and the rating that has been performed by the user. This difference was computed in several ways.

**Statistical person.** In this approach, the average rating score over all test persons has been computed for each video score. The result can be viewed as the ratings performed by one fictive person. Each average score is then subtracted from the intended score that was shown on the screen. For each task \( j \) the mean error \( P_{SP}(j) \) is obtained by considering the absolute value of the computed difference.

\[
P_{SP}(j) = \left| x^j - \frac{1}{M} \sum_{i=1}^{M} y^j_i \right|
\]

In this way, the rating precision of a statistical person is computed. This approach follows the data evaluation procedure applied in previously published research results [17, 18].

**Statistical error.** The average score over all test persons probably hides some rating error that is committed by each single person. When adding ratings of lower and higher values, the average rating might be correct – even if each single rating was not precise enough. Therefore, we additionally decided to compute \( P_{SE} \) where the error committed by each test person is calculated
first. Then, the mean error is computed.

\[ P_{SE}(j) = \frac{1}{M} \sum_{i} |x^i - y^i| \]

Of course, results are less appealing, but they represent a more realistic view for this type of investigation.

**Detailed analysis.** In most rating methodologies, the lowest and highest scores are much easier to represent than the intermediate values. For example, the values 1 and 5 are easy to rate when using a slider. However, in many cases, especially the intermediate values might be of high interest for video quality evaluation. Therefore we decided to compute the precision of both types (statistical person and statistical error) for each possible score. This can be expressed by the following formulas:

For each \( k \in n_1, \ldots, n_5, c_1, \ldots, c_5 \) the number of occurrences \( N_k \) of \( k \) is computed and can be written as \( N_k = \#\{j : x^j = k\}, 1 \leq j \leq N \). Now, for each \( k \) the results from the statistical person \( P_{SP}^k \) as well as from the statistical error \( P_{SE}^k \) can be computed according to:

\[ P_{SP}^k = \frac{1}{N_k} \sum_{j : x^j = k} P_{SP}(j) \]

and

\[ P_{SE}^k = \frac{1}{N_k} \sum_{j : x^j = k} P_{SE}(j). \]

The statistical error or person can now be used for ranking the rating methodologies in terms of precision. They express the deviation between the assessed and the intended user rating. In the following, methodologies with significant error will be judged to be inappropriate for subjective quality assessment because their use affects the correctness of collected data.

**4.2.2. Reaction and scoring times**

Since different methods require different efforts from the test users, we also decided to consider the reaction and scoring times, as already described in Section 4.2.

An example of these rating delays can be seen in Figure 8, which is a plot taken from a slider measurement. At time \( T_1 \), the number 2 was shown on
the screen. The user started to react at time $T_2$, less than one second later. Then, he tried to reach the displayed score with the sliding mechanism. After a certain adjustment period the slider is settled for a final position at time $T_3$. Finally, at time $T_4$, a new number appears on the screen. The reaction time is computed by $T_2 - T_1$ and the scoring time by $T_3 - T_1$.

Reaction time. The reaction time does not only provide information on the single test persons’ rating ability, but it also represents an important parameter that can be used to determine the rating difficulty of each methodology.

For example, the translation from the category “good” shown to the user on the screen to a specific distance between the thumb and index finger might require a certain amount of time. However, the translation of the number “4” to the related finger count might be more spontaneous for the test users. By computing the reaction time required for each single rating task, the translation difficulty can be quantified.
The length and the variance of reaction times might be crucial in multimedia quality estimation because they might affect the validity of test results for two reasons:

- Users might still be concentrated on rating an error while the next error occurs. They are distracted and the rating of the following error might not be valid.

- If the variation of rating times is high, it might be impossible to relate performed ratings to the impairments contained in the stimulus.

The rating time was computed as follows. Since every new number or category appeared on the screen at a defined moment of the test (i.e. every five seconds), it was possible to determine the reaction time for each displayed score. As shown in Figure 8, the reaction time was defined to be the period of time that has elapsed between $T_1$ and $T_2$. In this period, ratings are nearly constant and start to increase significantly after $T_2$.

Hence, $T_2$ represents the moment when ratings start to vary noticeably. Since ratings are often slightly varying over time, e.g. because of trembling hands or fingers, we decided to determine $T_2$ as the first moment in which the ratings differ from the previous rating in more than 5% of the available rating range $R$. Equally to the method described in [14], the reaction time can be computed by using the following property:

If

$$|y_i - y_{i+1}| > \frac{5}{100} \cdot R$$

and

$$|y_j - y_{j+1}| \leq \frac{5}{100} \cdot R \quad \forall j : T_1 < T_j \leq T_i$$

then

$$T_2 = T_{i+1},$$

where $y_i$ denotes the rating recorded at time $T_i$.

If no reaction time could be detected (e.g., when values are not varying enough), then the mean rating used to compute the precision error is computed, but the section is neither used for the reaction time nor for the scoring time analysis. This error can for example occur when test persons with small hands use the glove. The sensors might not identify the changes when shifting from score 2 to score 1 if the position assigned to the value 2 is already too small. Another scenario could be that a user needs to shift
from value 2 to 1 on the slider but he has already reached the lowest possible position without knowing. Then, it is impossible to move the slider further and no reaction time can be calculated.

**Scoring time.** After a certain reaction time, the test person eventually starts to rate. At this moment, the user needs some time to reach the intended score. This period strongly depends on the manipulation difficulty of the considered rating methodology. For example, test users might be able to show three fingers nearly instantaneously but they might need some adjustment time to reach a certain arm position when using the mobile phone for performing a rating. Equally, the positioning of the slider needs a certain amount of time. The scoring time is completed at the moment the rating is constant again. In Figure 8, this moment is reached at time $T_3$.

For the scoring time, the length of the constant rating interval is additionally taken into account in order to be certain that the correct final rating value is identified. As it can be observed in Figure 8, it can happen that small constant rating intervals occur due to user hesitations. This might happen in moments of indecision, because of lack of concentration or for example when the user checks the correctness of the current position and then readjusts the rating to the desired score.

Therefore, we decided to judge a constant interval to be the final scoring interval when at least one fourth of all tracked rating points for each rating (including the ratings performed during the reaction time) are consecutively differing less than 5% (or 10% with the phone) from its predecessor. The tracking intervals depend on the software and hardware used, but are generally below 200 ms.

Hence, during one rating session, we have $K$ scoring points $y_i$ tracked at time $T_i$ where $1 \leq i \leq K$ and $0 \leq T_i \leq 5000$.

$$T_3 = \min_{1 \leq i \leq K} \{ T_i : |y_j - y_{j+1}| \leq \frac{5}{100} \cdot R, \ \forall j \in [i; i + \frac{K}{4}] \}$$

The scoring time is defined to be the time interval that elapses between $T_1$ and $T_3$. It represents the entire time required to perform a rating. If no constant interval could be identified, e.g., because the rating varies over the entire available time frame, then the scoring time equals to the reaction time.

4.2.3. User opinion

As already mentioned in Section 4.1 the user opinion for each methodology was assessed and test persons were asked to chose their favourite method.
Furthermore, test persons had to judge the user friendliness, the simplicity of use and the feeling for the position within the evaluation scale, the potential for distractions and the perceived tiredness by using a 5-point Likert-scale.

5. Results and discussion

5.1. User opinion

As described in Section 4.1, users were asked to state their opinion on the different methodologies. Every person had to select one method they preferred over the others. The results of the second test, where a camera based implementation of the finger count method was used, can be observed in Figure 9. Generally, it can be deduced that users prefer the finger count, the finger distance and the glove over the other methods.

Judging from user’s comments, the finger count seems to be the most intuitive method, followed by the glove. Both movements seem to be very natural since they don’t require too much concentration. Especially the glove proved to be the least distracting when testing all methods outside
of a building. The finger distance measurement was also judged to be very intuitive, but many users complained about the lack of orientation.

After improving the finger count method by adopting the data glove during test 3, other 15 users were asked the same question. This time they were explicitly told to imply the suitability for mobile use in their considerations.

In Figure 10 it can be observed that 85% of the users prefer a glove-based method. A majority of 58% prefer the new finger count method. The slider reaches only the third place with 17%. This result is in accordance with the statements collected regarding the user friendliness, the simplicity of use and the feeling for the position within the evaluation scale. In all cases, the finger count method performs best with a result ranging between “excellent” and “good” whereas the other methods are judged between “good” and “fair”.

For example, the results for the simplicity of use are provided in Figure 11, showing Mean Opinion Scores with their corresponding 95% confidence interval. It can be observed that the performance of the slider exceeds the acceptability limit provided by the rating “fair” only by a small amount.
Hence, from a user perspective, it seems to be clear that the finger count method should be used and that there is an urgent need to replace the slider. This suggestion is supported by the fact that users have a significantly better feeling for the position within the evaluation scale when using the finger count method. The detailed result can be observed in Figure 12, where all methods are rated to have a “fair” performance while the finger count is estimated to be “excellent” in this respect (with a notably small confidence interval).

Also, the evaluation of the tiredness caused by the single methodologies confirms the users’ choice for the finger count method. Equal to the slider, a number of seven persons strongly disagreed that the method was tiring when evaluating the Likert scale questions.

5.2. Distraction

During the first experiment [14], the distraction time was only roughly measured with a stopwatch. In these experiments, the time was estimated to about one second. We observed that users were generally distracted by the slider when it was visible for them. Nearly all test persons were tempted to glance to the rating device instead of watching the presented video. Using the glove instead resulted in very few distraction periods.
A more detailed analysis was performed during the second user study where the test persons were videotaped while providing ratings with the finger count (camera-based), the finger distance (camera-based) and the mobile phone (with and without haptic feedback) methods. Not all test runs could be filmed because there was only one camera available for several tests running in parallel.

The following relevant parameters are depicted in Table 1: (1) The number of test persons that were distracted, (2) the average number of distraction for the identified persons and (3) the average distraction period for one distraction occurrence.

<table>
<thead>
<tr>
<th>Analysis Parameters</th>
<th>Finger Count</th>
<th>Distance</th>
<th>Phone (camera-based)</th>
<th>Phone (mobile phone)</th>
</tr>
</thead>
<tbody>
<tr>
<td># users</td>
<td>3</td>
<td>6</td>
<td>8</td>
<td>7</td>
</tr>
<tr>
<td># distracted users</td>
<td>1</td>
<td>5</td>
<td>2</td>
<td>4</td>
</tr>
<tr>
<td>⊙ # distr. per person</td>
<td>3</td>
<td>3.17</td>
<td>15</td>
<td>3.5</td>
</tr>
<tr>
<td>⊙ distraction time (s)</td>
<td>0.78</td>
<td>0.74</td>
<td>0.59</td>
<td>0.78</td>
</tr>
</tbody>
</table>

Table 1: Frequency and duration of distractions
In Table 1 it can be observed that the mobile phone and the finger count method reveal to have a good performance. For both, the number of distracted users is small. The mobile phone results in a shorter distraction period whereas the number of distractions per person is significantly smaller for the finger count method. When computing the overall distraction period for the mobile phone, namely $\frac{1}{4} \times 15 \times 0.59 \approx 2.2$ seconds and for the finger count, being $\frac{1}{3} \times 3 \times 0.78 \approx 0.78$ seconds, it can be argued that the finger count method outperforms the other methods in terms of distractions.

During the third experiment, 14 out of 15 test persons were asked using a five point Likert scale (“strongly agree”, “agree”, “neither agree nor disagree”, “disagree” and “strongly disagree”) to indicate whether a method had distracted them. Seven persons agreed that the slider distracted them. The results for the finger distance method were ambiguous: Six persons agreed and six persons disagreed that the method was distracting them. The glove resulted in low perceived distractions: Six persons disagreed that the scheme was distracting.

Again, the best result was obtained by the finger count method where five persons strongly disagreed and four persons disagreed that the approach was distracting them.

5.3. Precision

From the results of the first experiment we understand that the slider provides more precise results than the glove. However, it is known that the good precision results of the slider can be misleading because they are partly due to the fact that test persons are checking the current position by glancing at the device. When showing the same number for five seconds on the screen, these distractions do not have any impact. In any real scenario, however, they would strongly affect the reliability of the results. Not surprisingly, a covered slider returned a lower precision, but still the precision of the glove does not seem to be satisfying in respect to the currently applied slider.

This issue represented one of the main drivers for developing several other rating possibilities. During the second and third experiment, increased analysis effort was put on this aspect. Therefore, the error performed by a statistical person described in Section 4.2.1 as well as the statistical error (as described in Section 4.2.1) were computed for each test methodology.

Obtained results for the second user test are depicted in the Figures 13 and 14. On the $x$-axis, the rating methods and the used scales are indi-
Figure 13: Test 2 - Error of a statistical person
cated (numbers 1 through 5 or the ACR categories), while the rating error is represented on the $y$-axis.

It can be observed that the rating precision obtained by using a mobile phone is increased in respect to the slider, also influenced by the usage of interval-based rating instead of continuous evaluation. Unfortunately, this method had not been accepted by the users and was therefore not taken into consideration for the next user test. Also, the finger count methodology returns good results, whereas the finger distance methods needs to be improved in terms of precision to compete with the slider. The glove returns around twice as much error as the slider. Furthermore, it can be observed that the amount of measured error differs less for different scales than for different rating methodologies. Hence, it is clear that the choice of the method represents the critical aspect, and the selection of the scale is of minor relevance.

To summarize, it can be said that the method preferred by the users – the finger count – reveals to have the same precision as the currently adopted
slider. Hence, it seems to be advisable to adopt this new feature for mobile subjective testing in a scenario where categorial rating is preferred. In the following, we refer to precision by using the definition of statistical error only. This is because both the statistical person and the statistical error provide similar results with a higher precision of the latter. Figure 15 shows the precision results of third user test, where the finger count method has been improved by using the glove. It can be observed that the precision of the slider still outperforms the other methods. However, the error for the finger count method is also rather low, in particular when using the numerical scale. In fact, users repeatedly stated to prefer this scale because they did not need to translate the categories to the number of fingers to be shown, which seems to be a difficult task.

In the course of the experiment, we noticed that the selected thresholds described in Section 3.4 do not fit for all hand sizes. This is due to the fact that the hand size of the person who defined the thresholds probably was
larger than average. Results for test persons with smaller hands seem to have a higher error than for persons with larger hand size.

During our experiments, we measured the hand size of all test persons. We therefore formulated the null hypothesis that the hand size does not affect the precision of the results. An analysis of variance [19, 20] performed with the statistical software R\(^5\) lead to a \(p\)-value of 0.08, therefore only some indication for a rejection could be reached. However, when having a closer look at the user data, it can be argued that two test persons seem to have experienced serious difficulties with the test setting. If they are considered as outliers, the \(p\)-value is reduced to 0.04 – we can then conclude that the hand size \emph{does} matter in terms of rating precision. In average, the error of users with a hand size greater than 20.5 cm is 0.17, 0.35 for hand sizes ranging between 19 and 20 cm and 0.42 for hand sizes smaller than 18 cm – all when using the numeric scale.

Summarizing, it can be stated that the error of the finger count method is comparable to the one of the slider. It is necessary to calibrate the glove sensor values to the hand size of each test person at the beginning of the experiment instead of using one fixed threshold for all hand sizes. Moreover, a combination of near and far sensors may lead to more accurate results when evaluating the number of presented fingers.

5.4. Reaction and scoring time

The reaction and scoring times were computed for all different methods and scales under test in the user studies 2 and 3. In general, when observing the results depicted in Figure 16 it can be noticed that all rating methods, on average, lead to reaction times from 800 ms to 1200 ms, except for the finger count and the mobile phone, which seem to have a better performance, and the glove when using categories that results to have a worse performance. In fact, results of the third user study seem to confirm this observation. Here, the measured reaction time of the finger count method was 611 ms in average. 906 ms were required when using the slider for representing the numbers shown on the screen. Since the mapping between impairments and user ratings seems to play a crucial role for the acceptability of a time-continuous measurement technique, the measured reduction of the users’ reaction time could represent a significant benefit provided by the finger count method.

\(^5\)http://www.r-project.org/
Figure 16: Reaction time
Furthermore, it can be noticed in Figure 16 that the reaction time is always smaller when numbers between 1 and 5 were shown on the screen instead of the ACR categories. Hence, it seems to be convenient to train the users in thinking in numbers before performing a test. Similar to the computed reactions times, the scoring times are comparable for all methods (around 1200 ms). During the second user test it could be noticed that the haptic feedback when using the mobile phone seems to add some further delay during the scoring procedure. Again, the finger count method seems to produce the lowest delay: around 700 ms in the second and 633 ms in average during the third user test. The slider instead required more than 980 ms in both cases.

For all methodologies, the gap between scoring times needed to represent the ACR categories is even more accentuated compared to the reaction times. Hence, the scoring procedure seems to add some additional delay when thinking in categories instead of numbers.

6. Conclusion

In this paper we presented an extensive study on alternative rating methodologies to the slider, which is the commonly used device for time-continuous subjective quality assessment. During our research activities, we started to develop and test several new methods that are suitable for mobile use. Results from three different user studies were used to understand the important issues of such a rating technology and successively led to important refinements. Finally, we were able to identify one method that seems to satisfy both the users’ needs and the required precision for providing reliable and comparable results.

Test persons performed ratings, for example by showing a certain number of fingers, by showing a distance between the thumb and the index finger, by wearing a data glove that contains sensors, and finally by using the accelerometer integrated in a mobile phone used for detecting gestures with and without haptic feedback.

We analyzed the new methods and compared them to the slider according to objectively measurable criteria such as precision, required reaction and scoring times needed to perform a rating and the amount of distractions. Furthermore, subjective user opinions were collected in order to be able to identify potential annoyance caused by the new methodologies. To summarize, the experiments lead to the following results:
General: The simplest method, the finger count, outperforms the other approaches regarding all considered quality parameters. In its current state of development, it seems to be convenient to replace the recommended slider by this technique, in particular when aiming at assessing mobile multimedia quality.

User opinion for mobile assessment: Undeniably, the finger count is preferred by the users. Only a small amount of users opted for the slider. It was stated that the finger count is the most user-friendly method, being simple to use with the best feeling for the position within the evaluation scale.

Perceived distraction and measured distraction: The finger count was estimated to be the least distracting method. This subjective perception has been confirmed by measurements where the finger count revealed to have the shortest distraction period.

Rating precision: Test results of the experiment revealed that the rating precision obtained with the finger count is similar to the one of the slider. Its reliability of results would be comparable when applying the new method, with room for improvement in the tracking software.

Reaction time: The measured reaction time when using the finger count is reduced compared to the slider. This result implies that the quality of a mapping between perceived impairments and performed ratings can be improved by using the finger count method. It is interesting to note that for all methodologies the reaction times needed for representing the categories “excellent, good, fair, poor, bad” are higher than for the numbers 5, 4, 3, 2, 1. This represents a general result that is independent of the selected rating device. Hence, when performing subjective tests, it seems necessary to train the users to think in numbers instead of categories.

Scoring times: Haptic feedback seems to be responsible for a significant increase in scoring time. As a consequence, feedback should be avoided if it does not lead to any other important performance gain.

When studying these results it can be concluded that for mobile subjective testing the finger count method needs to be adopted. It performed best with respect to the subjective and objective criteria. Its main advantage,
the simplicity of use, also represents its main drawback: it provides only a discrete rating scale. However, in mobile test scenarios, it seems to be convenient to adopt the simplest possible design as graduations between the single scale elements might not even be perceivable by the users.

7. Acknowledgement

We thank the reviewers for their valuable comments on a previous version of the paper, as well as our students who have acted as evaluators, and gratefully acknowledge grant ICT08-005 from the Wiener Wissenschafts-, Forschungs-, und Technologiefonds (WWTF, Vienna Science and Technology Fund) to Helmut Hlavacs and grant CS11-009 from WWTF to Ulrich Ansorge, Otmar Scherzer and Shelley Buchinger.

8. References


