Signal Processing: *Image Communication* 29 (2014) 921–934

Contents lists available at ScienceDirect

Signal Processing: *Image Communication*

journal homepage: www.elsevier.com/locate/image

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

ARTICLE INFO

Article history:
Received 25 January 2014
Received in revised form
3 July 2014
Accepted 9 July 2014
Available online 19 July 2014

Keywords:
Subjective quality assessment
Quality of experience
Experimentation
Human factors
Quality measurement

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 that 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.

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 [1–4]. Due to the increased number of mobile devices, this traditional approach does not seem to be adequate anymore. Alternatives, such as (semi) living-labs, have been considered [5–7], 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 audio-visual 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 8–10 s. In [1] it is stated that "for still pictures, a 3–4 second sequence and five repetitions (voting during the last two)

* Corresponding author.

http://dx.doi.org/10.1016/j.image.2014.07.001
0923-5965/

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 s and that "eight- to ten-second sequences are highly recommended".

The length of 8–10 s 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 s, 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 with 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 [10–12]. 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 [space]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 Fig. 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