Quality of Experience, a very personal experience !

Antoine Lavignotte, Christophe Gravier, Julien Subercaze, Jacques Fayolle

Quality of Experience, a very personal experience !

Antoine Lavignotte, Christophe Gravier, Julien Subercaze, Jacques Fayolle
Université de Lyon, F-42023, Saint-Étienne, France;
Université de Saint-Étienne, Jean Monnet, F-42000, Saint-Étienne, France;
Télécom Saint-Étienne, école associée de l'Institut Télécom, F-42000, Saint-Étienne, France;
Laboratoire Télécom Claude Chappe (LT2C), F-42000, Saint-Étienne, France.
Email: {firstname.lastname}@telecom-st-etienne.fr

***Abstract*—At the Pervasive Computing area, end users expect to receive a multimedia service with an acceptable quality, anytime, and anywhere. Measuring this acceptability is usually referred to Quality of Experience (QoE). Unlike Quality of Service (QoS) which focuses on allocating expected systems and network resources, QoE is concerned with optimizing the perceived quality of a service by end-users. The appraisal of the user's acceptability thresholds is a key factor for service providers to perform adaptation decisions on their products (VoD, IPTV, online games, etc). While QoE is individualized, no study has yet examined to what extent. In this paper, we report an empirical study to understand if QoE should be managed globally, per cluster of users, or personally. We prove that every user has his very own vision of a same service, therefore that future QoE-based adaptive systems should take into account this property.**

***Keywords*-HDTV; TV Broadcasting; Human Factors; Human Behavior; Measuring; Monitoring**

I. Introduction

A lot of technologies and services which contain HCI (Human Computer Interface) can be deployed due to their ready-to-use technology. Most of them struggle to emerge because they fall short of users' expectations although they do satisfy some QoS (Quality of Service) thresholds. The reason is that QoS focuses on technical features and does not capture the end-user's expectations. The users' perception (QoE) is a key element to take into account during a new technology or service development, especially in Multimedia [1]. ITU-T defines QoE by: *"The overall acceptability of an application or service, as perceived subjectively by the end-user*" [2]. A more general notion of QoE is proposed by Alben to define all the different interaction aspects between a product and a person [3]. The most important fact to understand is that QoE contains an important degree of subjectivity, in a given context.

The paper is organized as follows. Section 2 defines the objectives and the experimental setup of this study. Results are reported and analyzed in section 3. Finally, section 4 concludes.

II. Experiments

A. Objectives

The main objective of the study is to define how to take into account the QoE in an adaptive system. To answer this question, we conducted a test on a users group. In a first approach, we tried to understand the perception difference between users by comparing grades associated on a video set by each subject. The next step, important to secure the results' consistency, was to ensure that user's perception did not change during the test. Finally, thanks to the acceptability threshold asked to each subject, we were able to compare the video quantity which met each user's threshold for a traditional multimedia session.

B. Experimental recommendations

We recreated a personal living room environment in order to carry out tests in a context close to real world situations. To match a great test environment, we decided to follow most of the ITU-R recommendations regarding subjective evaluation methods. All specifications can be found in [4].

This environment was composed of:

1) An HD TV 40" as the terminal,
2) A sofa 2.97 meters away in order to satisfy the PVD (Preferred Viewing Distance [4]),
3) And a room plunged into the dim light to cancel other room illumination.

This experiment is independent of a given optimal context. However, the context of the experience for each subject was strictly identical. That means: No light in the room during the test, the same TV configuration for each user (brightness, contrast, ratio, luminance,...) and no noise during the test.

C. Materials and methods

1) Selection of test materials: To emphasize QoE perception of the users in a certain context, we have chosen to base our test on a single movie encoded in different ways. The first parameter is the *image size*. Since each terminal has its own power of resolution, so the same service should be distributed with various parameters. The most common resolutions have been chosen: *480p, 720p,*

1080p. Keeping the same ratio size was important to respect the image proportions during the diffusion. Each video was broadcast in a full HD TV, that is to say 1080p. The second parameter is the *video encoding bitrate*. To determine the interesting encoding levels, we conducted preliminary tests to highlight the most significant bit rate levels. In a 3G connection (UMTS, HSDPA), the bandwidth can easily vary from 10Mbps to few kbps during hard handover or telco roaming. We chose a range between 10Mpbs and 250Kbps which fits with a real bandwidth variation on a cellular network. Different steps were also defined to propose multiple video modalities (in Kbps): {250, 500, 800, 1000, 1500, 2000, 9653}, where 9653 Kbps was the source encoding bitrate. These steps were chosen to be closed to different adaptive live streaming recommendations [5].

2) Methodology used: This test is based on a derivative of the Single Stimulus Methodology (SS method). It is the most widely used among visual quality researchers evaluating Image Quality Assessments (IQA) algorithms [6]. This method has been formalized in the ITU-R BT500-11 and ITU-T P.910 recommendations. To better understand the difference of perception and to discover the quality threshold for each user, some changes has been made on the traditional methodology. Unlike many studies where video clips duration varied from 8 to 30 seconds [7], we decided to use a full video trailer (Thor movie) lasting 2 minutes and 26 seconds. This choice allowed us to provide a movie with various types of scene (slow parts for discussions, fast parts for action,...). Although the time of each video is important, the user can switch from one video to another at any time. This is more closed of the real life where the user can stop the session if the quality is not good enough. He is therefore not forced to watch the entire video every time. Moreover, to respect the ITU-R BT.500-11 recommendations, a time limitation of 30 minutes has been done. To respect this limitation, the use of only one video was necessary. We chose focus on quality (number of modalities) compared to the quantity (number of videos). To test the influence of previously seen videos and the objectivity of each grade per video, we included two video repetitions into the playlist. A presentation of the playlist is presented in Table I. The videos from the playlist are all downloadable [1]. Repeated videos are respectively marked by an * and **.

D. Subjective assessment

1) User Panel: Fifty-two subjects took part in this study. They ranged between 18 and 58 years old. The average age is 33 and the median is 29. Concerning the gender distribution, a total of 65,38% men and 34,62% women took part in the tests.

2) Scoring Method: Subjects were asked to use the MOS (Mean of Opinion Score) scale to grade each video. This is a way to quantify numerically the outcomes of a subjective experiment. It has been defined by ITU [4]. The MOS scale contains 5 grades. Each grade is designed to reflect a judgement of users concerning video quality. No half grades were available during the test. Table II reproduces the MOS scale for the reader, as it was presented to the subjects. The MOS scale was chosen because it is already widely used in the scientific literature, therefore amenable to comparison with previously obtained results. This method yields a distribution of judgements across a scale of categories for the different videos proposed.

3) Test Environment: To ensure that users cannot be influenced by others, all tests were taken in isolation. The tests subjects were seated on a couch exactly in front of the 40" HD TV. Each user had first to give his/her name, age

[1] http://datasets-satin.telecom-st-etienne.fr/alavignotte/QoeExperience/

Table I
THE EXPERIMENTAL PLAYLIST

Playlist	Video encoding Bitrates (Kbps)	Image size (Pixels)
1	9653	1080p
2	250	480p
3	1000	720p
4*	2000	1080p
5	500	480p
6	1500	720p
7	1500	1080p
8	800	480p
9	2000	720p
10	1000	1080p
11	1000	480p
12	9653	720p
13**	800	720p
14	800	1080p
15	1500	480p
16**	800	720p
17	500	1080p
18*	2000	1080p
19	2000	480p
20	500	720p
21	250	1080p
22	9653	480p
23	250	720p

Table II
ITU-R QUALITY AND IMPAIRMENT SCALES

Five-grade scale			
Quality		**Impairment**	
5	Excellent	5	Imperceptible
4	Good	4	Perceptible, but not annoying
3	Fair	3	Slightly annoying
2	Poor	2	Annoying
1	Bad	1	Very annoying

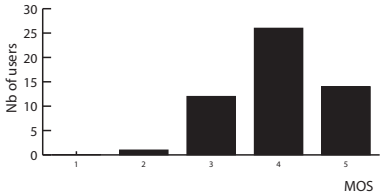

Figure 1. Video 15: Users distribution per MOS Value

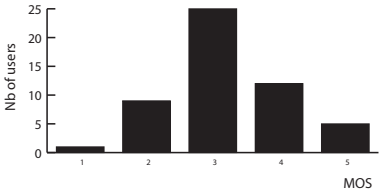

Figure 2. Video 3. Users distribution per MOS Value

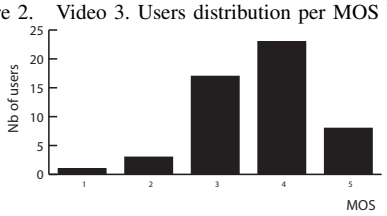

Figure 3. Video 18. Users distribution per MOS Value

and gender. Then, the first video started to show. At any moment, as soon as the user had chosen a grade, he/she was able to select it by pressing the appropriate key on the keyboard. Automatically, the next video was started and the previous one stopped. A security of 10 seconds was added to prevent any keyboard manipulation error. This continued until the end of the playlist. Before stopping the tests, one important thing to do was to ascertain the users' threshold of acceptability and more specifically, at what quality he/she was ready to watch the movie for a normal showing lasting an hour and a half. It was really important, in terms of analysis, to know how many videos had an acceptable level of quality for the user within this context.

III. Results & Analysis

A. Mark distribution per video examples

1) Exemple for a 480p video: To begin with, we propose to study the users' perception of the fifteenth video of the playlist. As a reminder, this video is encoded with a 1500 Kbps bitrate for a size of 480p. The Figure 1 presents a graph detailing the results for the same fifty-two views. It integrates the MOS value assigned by each user on this video. An initial quick scan of this graph shows that the grade distribution is important. Approximately 23% of the tested people chose grade 3. 48% 4 and 27% 5. Only one of them chose grade 2 and nobody grade 1. These results show that clients do not really have the same perception of a service. A quarter of the people think they are watching a perfect image movie with no distortion. Another quarter are watching a movie with only a fair level of quality! If we were in a context where an operator provides a video on demand to his client, a general adaptation decision based on a global users acceptability threshold could push 25% of the users aside because the quality is not good enough.

2) Example for a 720p video: For illustration, we have chosen the third video which is encoded with a 1000Kbps bitrate for a size of 720p. Results can be found in Figure 2. The video still differentiates the users. 10% of the people have chosen grade 5, 23% 4,48% preferred 3 and 17% the 2. Only one person has chosen the mark 1. Compared to the previous video, we noticed that almost half of the users agree on the mark 3.

3) Example for a 1080p video: Finally, results for a 1080p video are represented in Figure 3. We have chosen the eighteenth video which is encoded with a 2000Kbps bit rate for a size of 1080p. 6% of the users have chosen the grade 2, 33% grade 3, 44% grade 4 and 15% grade 5.

4) Results comment: Figure 1, 2 & 3 show that MOS is distributed along the quality axis, with a rather big standard deviation. We observe here Thurstones's law on comparative judgements [8], which is based on the assumption that subjective scores are distributed around the true value. But on the use case of video on demand (VoD) or video streaming, this results are very important. A bad quality reception could lead to stop in the use of the video service. That also implies that everybody does not have a same perception of a service. But we do not actually know if the service has an enough good quality to be seen by each user.

B. Mark distribution analysis

To differentiate user expectations and understand the above results, further analysis is needed than just a superficial reading.

1) Global Mark Evaluation: To have a better view of the distribution mark for each video, we have represented the results in a box plot graph which is displayed in Figure 4. We recall that the reading mode for column graphs is: On each box, the central point is the median. The edges of the box are the 25th and 75th percentiles. The whiskers extend to the most extreme data points. Outliers are not considered by the whiskers, but they are plotted individually. What we saw before in the example presentation of results (480p, 720p and 1080p) is repeated. Almost all the results are the same. The disparity of acceptability is very important in each video. 50% of videos have a distribution over 4

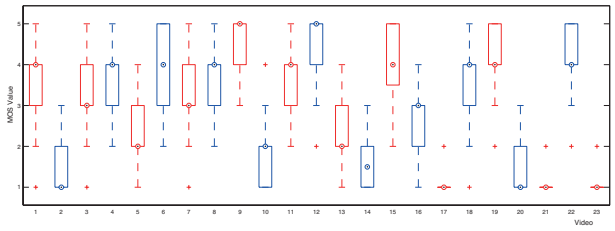

Figure 4. Box plots summary for MOS values obtained per video

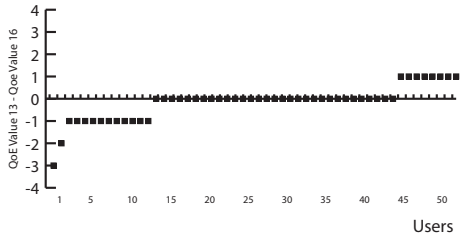

Figure 5. Difference between video 13 & 16 (ordered per user)

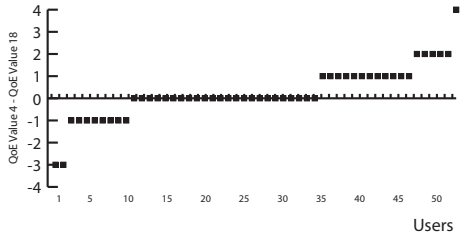

Figure 6. Difference between video 4 & 18 (ordered per user)

grades, 30% over 3 grades. All the others have a distribution over 2 grades at least. All users agree with mark 1 for videos with very low quality. (videos 17, 21 and 23). It tends to show that poor QoE is globally perceived, while enhanced QoE is perceived differently by users, and the difference is significant. These results suggest that everybody has their own personal scale. But is this personal scale, in the same context, fixed or does it change over time ?

2) Repeated videos, gap rating: To know if the users' opinion can change over the time for the same video, we repeated two trailers during the test. As a reminder, the first repeated video is 720p size encoded with a 800Kbps bitrate. They were shown 13th and 16th in the sequence. Their broadcast is just separated by a trailer with a low quality (1080p & 500Kbps bitrate). Results are shown in Figure 5. The graph represents the difference between the two grades awarded by the users. It is ordered by users. 60% have exactly the same opinion of a given video. If we work in absolute values, 96% find a difference less to 1 between the two videos. These results mean that the personal scale is quite stable. But we could say that the difference is small due to the proximity of the two broadcasted videos. To be sure of the first results, a second video was repeated. It was a 1080p size encoded with a 2000Kbps bitrate. The first display had been broadcast in the first quarter of videos (video 4). Then the replay was played in the fourth quarter (video 18). The results are shown in Figure 6. 46% have exactly the same opinion of the video. Working in absolute values, 85% discern a difference less to 1between the two videos.
Comparing with the first pair of videos, results are in good agreement. Only 10% of people are scattered in relation to the previous test. This tends to illustrate that users are not very affected by previously streamed views when they are introduced to a new multimedia stream. This is an interesting result when studying QoE during zapping sessions, although further tests may have to be conducted to draw stronger conclusions.

3) 3D Representation: The results can be represented using 3D modeling. This representation allows a quick overview of user's expectations in terms of QoE during this lab tests. Due to the space limitations, only three graphs are depicted in the paper but all the results can be found at http://datasets-satin.telecom-st-etienne.fr/alavignotte/QoeExperience/. Figure 7, shows the QoE modeling of User 31. Two thresholds appear: 800 & 1200Kbps. It can be observed that the user perception is not very sensitive with the resolution but much more with the bitrate. On Figure 8, results are drastically different. Just one threshold appears in high resolution (1.5Mbps). Compared to the previous user, the resolution is more important. For example, we have a 2 MOS gap with a same bitrate but with different resolution. In this context, the adaptation choice could be decisive on the user decision to continue the viewing or not. Another user shows another result (Figure 9). Only one threshold appear at 800Kbps and later, the grade changes proportionally to the bitrate growth. Sometimes, playing with the resolution can up the MOS score and could better satisfy the user (800Kbps / 480p & 1500Kbps / 1080p). Finally, Figure 10 shows the user's average representation. With a such result's disparity, we cannot base our adaptation's choices on an average grade which is not representative of all users.

4) Accepted videos: After all these graphs, we have yet to answer the main question: Does the QoE need to be addressed personally ?
A good way of answering this question would be to differentiate the number of accepted videos for each user[2] and to compare them. These results are shown in Figure 11. Never more than 13% people agree on the same number of accepted videos. There is a range from 5 to 19 accepted

[2]Based on the MOS threshold for an acceptable video streaming as declared at the end of the session by each subject.

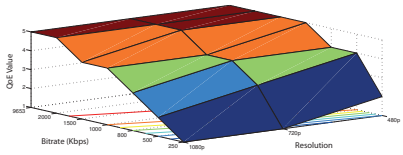

Figure 7. User 31-3D Representation

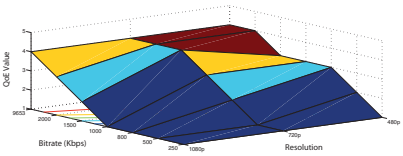

Figure 8. User 15-3D Representation

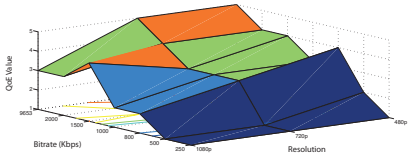

Figure 9. User 30-3D Representation

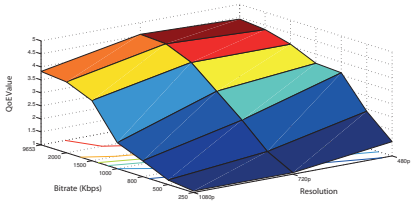

Figure 10. Average-3D Representation

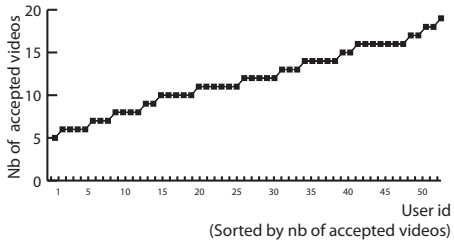

Figure 11. Number of accepted videos ordered per user

videos for the same context. How is it possible, therefore, to base the provision of multimedia on the results from a group of users with only two or three quality levels without actually knowing the individual user's expectations ? At least it illustrates the fact that treating QoE globally (for instance with a MOS threshold shared by all users) may not be the best way to maximize the personal multimedia experience. We can say that QoE based adaptation would be better performed in a personal manner and a clustering method can not be a solution to maximize the number of satisfied users.

IV. Conclusion

In this article, we have provided a new point of view on the uses of QoE. To conduct this study, we designed an experiment to address the video broadcast problem. The results lead to different findings.

The first one is that each subject has its own vision of a particular service, except for poor QoE multimedia, which are globally perceived as poor. If a service operator needs to adapt his video due to network congestion, he needs to consider that his clients will not have the same perception of the delivery content. This could result in the broadcast cancellation for a certain proportion of his clients. The second one shows that people's perception of a service is quite stable and will not significantly change during the test. This means that if you are aware of the users' acceptability levels, you would have all the necessary information for continuous adaptation to this user for a long time. You may not have to recover the personal QoE scale too regularly.

The third and final one supports the results of the first one. It shows that users do not have the same scale of expectations. We can just prepare a few different quality levels for a service and make decisions according to the network vagaries. But we have to take in consideration the real personal scale of the users to make the right decision to maximize QoE.

References

[1] R. Jain, "Quality of experience," *Multimedia, IEEE*, vol. 11, no. 1, p. 96, 2005.

[2] I.-T. S. G. 12, "Itu-t rec. p.10/g.100 amendment 2 (07/2008) vocabulary for performance and quality of service amendment 2: New definitions for inclusion in recommendation itu-t p.10/g.100," pp. 1–10, Mar 2009.

[3] L. Alben, "Defining the criteria for effective interaction design," *Interactions*, Jan 1996. [Online]. Available: http://www.albendesign.com/albenfaris/downloads/pdf/quality.pdf

[4] ITU-R, "Recommendation itu-r bt.500-11," pp. 1–48, Sep 2002.

[5] Apple, "Technical note tn2224," 2012. [Online]. Available: https://developer.apple.com/library/ios/#technotes/tn2224/_index.html

[6] J. Redi, H. Liu, H. Alers, R. Zunino, and I. Heynderickx, "Comparing subjective image quality measurement methods for the creation of public databases," pp. 752 903–752 903–11, 2010.

[7] S. Winkler and C. Faller, "Maximizing audiovisual quality at low bitrates," 2005.

[8] L. L. Thurstone, "A law of comparative judgment," *Psychological Review*, vol. 34, no. 4, pp. 273–286, 1927. [Online]. Available: http://content.apa.org/journals/rev/34/4/273