

How good Mobile TV needs to be?

Raquel Navarro-Prieto

Fundació Barcelona Media- Universitat Pompeu Fabra

Ocata 1, 08003 Barcelona, Spain

Raquel.Navarro@barcelonamedia.org

Abstract

Converged technologies are seen as the new leap forward in integrating the latest advances on various digital mobile radio networks. Nevertheless, the new levels of flexibility that these new technologies could introduce increases further the complexity confronted by users, and potentially lowers the usability and ease of use of, especially for services like mobile TV. Our research goal was to understand which an acceptable quality is for mobile TV. We concluded that the methods proposed by the existing standards did not match our goals because of the conditions in which our content would be shown and the qualities to be implanted are very far from the existing standards. Next, we described an innovative methodology introduced last year based on psychophysics, for sport sequences. We have adapted that methodology and carried out a study where we tested two situations, mobile TV sequences on a mobile phone screen (MPEG 4 compression) and PDA screen (MPEG 2 compression). We tested three types of content: news, sport and music video clip. In the final paper we would include a brief summary of the results.

Key words: User requirements, mobile TV, subjective evaluations, convergence services, experimental methodology.

1. Introduction

Converged technologies are seen as the new leap forward in integrating the latest advances on various digital mobile radio networks. Convergent technologies refer to the efforts of eliminating the barriers between mobile cellular communications networks and broadcast networks. The underlying idea is to be able to provide the user with a unique device and set of services that provide both rich personalised one-to-one communication and distribute mass media content and data. The INSTINCT project works towards this goal, to address how the combination of mobile telecommunications and digital broadcasting “can be used to successfully introduce new mobile services located at the crossroads of broadcast and Telco domains” [1].

For these reasons, INSTINCT defined a complete work package to provide the specification of the scenarios and the user needs analysis that must be performed to facilitate the massive adoption of new ways of accessing content and services. In addition, this work package also investigates business models and system dimensioning. This paper describes one of the studies that were carried out in the context of this work package over a two year period. This paper is organised as follows. First, we will describe the previous work that we did to gather general user requirements. Second, we will present and discuss critically the different methodologies used in the past to gather subjective evaluations. Finally we will present our study, the experimental design.

2. Previous work on mobile TV user requirements in INSTINCT

A scenario-based approach has been employed to gather user input for the development of converged technology. The users' feedback has been used to validate the scenarios and to inform the rest of the work packages about the user needs regarding the technology that is being developed. Towards this goal, each scenario was implemented as a slide show presentation to make it easier to share it with prospective users. Users were presented with these illustrations and during the interview we asked questions designed to help them place themselves in the proposed situation, in order to understand the usefulness of the services envisioned in INSTINCT.

The data collected confirmed that the participants would like to have a portal that integrates the available services and allows them to access the services in any context but with different devices for each context (mobile phone, PDA, tablet PC, laptop, TV, etc.). Likewise the participants provided examples of how they would use the alert services to fulfil some needs that are not currently catered for by the services available to them now. It was considered that the exact content of the alerts would be very dependent on the demographic characteristics of the users and the county. For this reason, a second round of studies in the U.K has been planned. Other services that have been found useful and attractive by the participants have been the use of profiles, the authentication before using some services, voting services, subscription to newspapers, forums, online shopping, and context based information. Not surprisingly, the participants expressed their concern about the way the information is presented to them. This is true for every system that a user interacts with, but particularly relevant in the mobile context. One particular concern of users is the information overload, -- "*not too much stuff*" --. For instance, an interesting solution proposed by the users was placing the alerts in a folder where the users could recover them when they want. In general, the test subjects proposed several alternatives that would allow the information in the screen to be arranged in an optimal way from the user perspective. In addition, the results also point out that personalised programme guides (EPG) would be a good way to enhance user experience. For instance, users suggested that they would like to receive the EPGs with the programmes they liked, highlighted or presented separately. Therefore, it was established which services, from the original envisions, are really needed by the user sample that were intended to. Based on our goals, our next research question regarding mobile TV was: Which is an acceptable video and sound quality for mobile TV watching? In order to be able to answer this question we needed to gather detail data about user real acceptance of watching a low resolution TV in a small screen.

3. Review of the methodological approaches for subjective evaluations

First, we will need to define what subjective evaluations are. Although all evaluations that take input from users could be considered in some sense subjective, we are excluding those evaluations where the subjects are asked to perform a certain task and/or reply to some questions based on the task. Subjective evaluations refer to the assessments about the perceived quality of a technology or a system. Subjective assessment evaluations could be divided into the following categories, which would describe in this section:

1. Psychoperceptual assessments
2. QoS assessments
3. Methods based on psychophysics

3.1 Psychoperceptual assessments

The goal of this type of assessments is to be able to have a measure of the human perception of the quality of the output of a given system.

Methods for the assessment of audiovisual communications are presented in P.920. The overall methodology described there, is based on conversation opinion tests. The 5-point quality scale is recommended for assessing the video quality, the audio quality and the overall audiovisual quality. A 5-point 'effort needed to interrupt' scale can also be used (Watson and Sasse, 98).

Are these standards suitable for mobile TV acceptability evaluation?

In our case, our goal is to measure the 'Quality' of media (video and audio) previous to the development of the services, in order to know which would be the minimum quality required for the users to enjoy mobile TV. Although the usage of an established standard would be desirable for our study, when we look at the existing ones there are a number of issues that, either are not specified on them, or that do not much our test goals:

- The effect of using a particular sample with specific characteristics as we consider being the case for the user group for INSTINCT services. The standards assume that all the human being will rate quality in the same way.
- The duration of the videos and audio samples is 10 second, as it is argued by Sasse and Watson, "it is not clear that 10 seconds video sequence is long enough to experience the types of degradations common to multimedia communication" and in general for multimedia over packet networks. On the other hand, with longer test sequence problems may rise of load on the users memory and users being "distracted" in their judgments by the content of the video (Aldridge et. al.)
- The effect of the context of usage of the system. In our case the mobile context is critical and therefore the quality that is assume acceptable by the users would not necessarily be the same than when they are in front of their TV sets or PCs.
- The standards do not contemplate the circumstances in which the quality of the video is quite poor. They were developed for system with a much higher output quality than the one that would be provided by INSTINCT.
- The standards only specify accumulative table as method for analysis, we consider that in our case we will need a more sophisticated statistical analysis to satisfy our goals.
- These standards recommend a 5-point scale to gather user responses. The 5-point scale has also being criticized as not being as international as was intended. Moreover, this type of scales treats quality as a single measurable dimension, despite of the evidence to the contrary. On the other hand continues rating with Single Stimulus Continuous Quality Rating (SSQR) (De Ridder, & Hamberg, 97) have proven to be distracting for users (Bouch, & Sasse, 2000).
- Lastly, the vocabulary needs to be change since it seems to not be universally recognize and not suitable for multimedia.

3.2 QoS Assessments

This second type of assessment is not focus on the sensorial perception of quality but in the subjective view of the user of other parameters of a system, trying to evaluate the QoS of the output in a general sense. Among the numerous dimensions that have been studied in this area (Sasse, 1999), the more frequently used dimensions include: Easy of use; Accessibility; and Security.

The idea behind is to obtain a more holistic description of users' perception of QoS not only the quality of the video.

3.3 Psychophysics approach

This innovative approach has been introduced for the first time by McCarthy, Sasse and Miras (2004) for multimedia technologies. These authors proposed this methodology for comparing the effects of quantization versus frame rate in streamed video. In their study they concluded that the rule “high motion = high frame rate” does not apply to small screens for sport content. After reviewing the methods for absolute rating in 5-points scales, they concluded that these methods were not suitable for multimedia content. Their claim was that a new method to elicit continuous ratings of quality with minimal effort on the users part was needed. Therefore, they introduced a new methodology, adapted from classical psychophysics, to discover the functions relating physical quality to perceived quality. This method is based on gradually increasing and decreasing video quality within a single clip to identify the threshold level at which quality becomes acceptable or unacceptable to these users. According to McCarthy et. al., “this metric tackles many of the drawbacks of alternative approaches in that:

- It is easy for users to understand
- It is less disruptive to the user than other continues rating techniques (like slides)
- It can be used with variable video quality
- It is more relevant to service providers” (since it will allow to test the variability that would be found in real networks)

The task for the user was to say when the quality of a sequence starts to be acceptable or unacceptable. Their clips were 210 seconds in length and the quality was increase or decrease in discrete steps every 30 seconds. Users were not aware of this quality structure – the authors simply told them they would be watching films that “varied in quality”. In their case, because they were interested in understanding the relationship between frame rate, quantization and perceived quality, they examined three different quality gradients (i.e. varying Frame per Second (FPS), Quantization, and both FPS and Quantization).

Which approach better match our research goals?

Our research question was not test the quality of a video sequence or to test if on sequence has a better quality than another, but rather to test if the quality of the video is good enough, “acceptable”, for the users to watch it (especially in circumstances when they are mobile). Therefore, after reviewing the existing standards and the problems with their application to reach our goals, we have concluded that need to create our own experimental design and session planning. We will draw some information from the standards but only about the presentation methods in order to make our experiment easily replicable among the scientific community.

In the light of the different methodologies and standards, and how the measure multimedia material, we proposed to use an methodology that tries to solve the problems with existing ones, while measuring both the quality of the video and the overall impression of the users about the service. Because of this, we decided to use the method based on psychophysics, with some variations to fit the characteristics of INSTINCT project.

In our case, the critical variable would be the kilobits and megabits transmitted per second. Therefore we will have 4 different levels of the speed of transmission; we will call this variable thought this report Quality (although we know that Quality is integrated by many variables). We proposed 4 discrete steps would be also of 30 seconds, since they have been proven to be a good length for users to realise about the differences in quality. We will therefore, have sequences of 120 seconds each. In addition, we only manipulated the FPS, associated with the two different screen sizes that would be used in the experiment.

4. Experimental Design for subjective evaluation of QoS

We used a 2X2X3X4 experimental design, with all the variables manipulated within subjects design (MPEG compression X Gradient of the quality X Clip content X Video Quality) with repeated measures. The descriptions of the values of the variables manipulated in this experiment are:

- MPEG encoding: with two values MPEG 4 and MPEG 2 for small size (average mobile phone size) and medium size (average PDA size). As specified previously, the small screen condition had a size and a resolution following the standards for a mobile phone screen, 160x128 pixels. The medium screen reassembled the size and number of pixels of an average PDA screen, namely 352x288.
- Gradient of quality: with two values, increasing and decreasing quality. We did that to follow the methodology defined by McCarthy, Sasse and Miras (2004).
- According to previous literature one of the most important variable regarding the video content for quality assessments, is the amount of movement in the sequence. Therefore, we manipulated the video content in order to be able to test if there is any effect of the content of the video on the users' judgments. In addition, sport, news and video clips are important contents for the project. We used three source clips with the following characteristics.
 - Source clip A: Music clip
 - Source clip B: Football sequence
 - Source clip C: News from a TV news program
- Quality: with 4 levels that were different for the MPEG4 and MPEG4 compressed sequences.
 - MPEG4 sequences' values: 0.45-0.8-1.15-1.45 Mbps
 - MPEG2 sequences: 1.45-1.15-0.8-0.45 Mbps

In total, 12 experimental conditions, as shown in Table 1 will be tested. In order to make sure that all the condition are presented the same number of times in all the possible positions of the conditions, a counterbalance across all the subjects a Greco Latin Square Design was used.

This way all the users saw all the 12 conditions, grouped in blocks of 3 clips each. Each block grouped either the increasing and decreasing gradients, and both types of compression conditions.

As we mentioned before, we used a different length per clip as McCarthy et. al. (2004), i.e. 120 seconds per clip, which will allow us to introduce 4 different levels of quality per clip. Each of the level of quality will last 30 seconds. Because the time to present each clip was 120 seconds them the total time of experiment, excluding the reading the instructions and the time filling questionnaires. was 1440 seconds or 24 minutes,

The user task was to say aloud when s/he thinks the quality is acceptable or unacceptable.

This binary acceptability ratings (e.g. 1= acceptable, 0= unacceptable) were transformed to a ratio measure by calculating the proportion of time during each 30 second period that quality was rated as acceptable, as McC. McCarthy, Sasse and Miras (2004). This way we were able to do statistical analysis (i.e. ANOVA) with the data.

4.1 Subjects

The sample should be similar to the sample used in the first part of our project, namely, young users but not teens, white collar workers, commuters. There were 24 participants.

4.2 Measures

Our dependent variables were the assessment ratings provided by the subjects about the perceived video quality, when it is acceptable and when is not acceptable.

4.3 Procedure

The subjects will be conducted through the following steps:

1. Introduction, explaining the procedure to the subjects.
2. Pre-test questionnaire to get more detailed information about their TV watching habits and mobile phone usage.
3. Presentation of the material. We will use the method that has been described above for the presentation of the videos and the users would provide their responses verbally.
4. Debrief. A short post-test questionnaire with questions about mobile TV usage.

5. Conclusions

We would like to highlight two mayor conclusions, the first one regarding the methodology and the second regarding the results found in the study. First, we can conclude that the methodology based on psychophysics introduced by McCarthy et al. (2004) for multimedia assessments of sport sequences, has proven to be very useful for mobile TV quality assessment with a variety of content.

In our presentation we will present a summary of the main results found with this methodology.

6. Acknowledgments

The authors acknowledge the support for this work that is funded by the EU under the IST program as the project INSTINCT (IP-Based Networks, Services and TermINals for Converged SysTems), IST-507014. The authors also want to acknowledge all partners in INSTINCT, especially in the ECO-System Aspects work package.

7. References

Sicre, J.L.; Duffy, A.; Navarro-Prieto, R.; et al (2004) "Three user scenarios on the joint usage of mobile telco and TV services for customers on the move." *WWRF12 meeting – WG1*. Toronto, 3 - 4 November, 2004

De Ridder, H. & Hamberg, R. Continuous Assessment of Image Quality. *SMPTE Journal*. 106(2): 123-128, Feb 1997.

McCarthy, J., Sasse, M.A., & Miras, D. (2004) Sharp or Smooth? Comparing the effects of quantization vs. frame rate for streamed video. *Proceedings of CHI 2004*, Vienna, Austria, April 20-24.

ITU-T Recommendation P.80 <http://www.itu.int/>.

Watson, A., Sasse, M. "Measuring Perceived Quality of Speech and Video in Multimedia Conferencing Applications", *Proc. ACM Multimedia '98*, Sept. 1998, Bristol, England

Bouch A., & Sasse, M. A. (2001): A user-centered approach to Internet Quality of Service: why value is everything. *Proceedings of IT-COM'2001*. Denver, Colorado, August 20-24, 2001.

ITU-T P.920 Interactive test methods for audiovisual communications. 1996. <http://www.itu.int/>.

ITU-T Recommendation P.910, “Subjective video quality assessment methods for multimedia applications”, August 1996. <http://www.itu.int/>.