Lessons Learned during Real-life QoE Assessment

Nicolas Staelens
Ghent University - IBBT
G. Crommenlaan 8 bus 201
Ghent, Belgium
nicolas.staelens@ugent.be

Wendy Van den Broeck
VUB - IBBT
Pleinlaan 9
Brussels, Belgium
wvdbroec@vub.ac.be

Yohann Pitrey
University of Vienna
Dr.-Karl-Lueger-Ring 1
Vienna, Austria
yohann.pitrey@univie.ac.at

Brecht Vermeulen
Ghent University - IBBT
G. Crommenlaan 8 bus 201
Ghent, Belgium
brecht.vermeulen@ugent.be

Piet Demeester
Ghent University - IBBT
G. Crommenlaan 8 bus 201
Ghent, Belgium
piet.demeester@ugent.be

ABSTRACT

Subjective video quality experiments are usually conducted inside controlled lab environments, where stringent demands are imposed in terms of lighting conditions, screen calibration and position of the test subjects. However, these controlled lab environments might differ from a more natural setting such as watching television at home. This, in turn, will influence subjects' Quality of Experience. In previous research, we conducted a series of subjective experiments, mimicking the natural environment of the targeted use cases as much as possible. In this paper, we provide an overview of these conducted studies and summarize our main research findings. Our results show that the environment and overall experimental setup influence the primary focus of the test subjects. Consequently, a significant difference in impairment visibility, tolerance and annoyance is observed compared to conducting experiments in pristine lab environments. Our results highlight the importance of considering the targeted use case and assessing video quality in more natural environments.

Categories and Subject Descriptors

H.1.2 [**User/Machine Systems**]: Human factors

General Terms

Measurement, Human factors

Keywords

Quality of Experience (QoE), Subjective video quality assessment, Real-life

1. INTRODUCTION

According to the definition of Quality of Experience (QoE) [3], *"the overall acceptability of an application or service, as perceived subjectively by the end-user"*, might be influenced by user expectations and context. Subjective video quality assessment methodologies as specified in, for example, International Telecommunication Union (ITU)-R Recommendation BT.500 or ITU-T Recommendation P.910 describe in detail how to set up and conduct experiments for obtaining real human ratings on perceived quality of (degraded) video sequences. These experiments must be conducted in a controlled lab environment subject to stringent demands in terms of, amongst other, lighting conditions and distance between the test subjects and the screen. Furthermore, observers participating with the experiment usually receive specific instructions on how to assess and evaluate the quality of the different video sequences. As such, users' expectations and users' context might already be influenced prior to the start of the experiment and hence will impact their QoE.

Due to the instructions given to the test subjects, observers' primary focus is on (audio)visual quality evaluation. Also, most of the existing subjective assessment methodologies focus on the evaluation of short video sequences (typically between 10 and 15 seconds long)[1]. Moreover, the standardized test environment in which the experiments must be conducted is not necessarily representative for a more realistic setting such as watching television in a living room environment [8].

In previous research [7, 8], we conducted a series of subjective video quality experiments in more realistic environments targeting specific use case scenarios. Consequently, we considerably changed the context and expectations of our observers compared to those during more standardized quality assessment. The results of these studies were also compared with results obtained during quality assessment in controlled lab environments as specified by the ITU Recommendations.

In this paper, we briefly present our research on assessing video quality in more realistic environments. By summarizing the main research findings of our real-life experiments, we also discuss the benefits and importance of mimicking realistic environments/settings during subjective quality assessment. Last, we identify some of the pitfalls encountered

[1] Except for the Single Stimulus Continuous Quality Evaluation (SSCQE) methodology as specified in ITU-R Rec. BT.500.

during real-life video quality assessment.

The remainder of this paper is structured as follows. In section 2, we present some related work on experiments conducted in less stringent assessment environments and methodologies enabling more natural quality assessment.

Next, section 3 provides more information on two subjective studies we performed to assess the influence of primary focus on perceived quality when conducting experiments in a more natural setting. Section 4 summarizes the main research findings of these experiments and highlights the importance of mimicking realistic environments during quality assessment. Finally, we conclude this paper.

2. RELAXING THE STANDARDIZED ASSESSMENT METHODOLOGIES

As pointed out in the introduction, several stringent demands are imposed on the controlled lab environment in which the subjective experiments should be conducted. On the one hand, this can make subjective experiments hard to set up and expensive. However, on the other hand, the standardized assessment methodologies facilitate repeatability of experiments.

Recently, a subjective experiment has been conducted to assess the influence of subjects' country and native language, playback/display device and overall test room conditions on audiovisual quality perception [6]. The audiovisual experiment was conducted several times in different environments and labs located in different countries. Furthermore, the experiments were conducted inside a controlled laboratory and inside a public area (e.g. cafetaria). Prior to the start of the experiment, subjects still received specific instructions and were thus focused on audiovisual quality evaluation. The results of this study showed that the environment in which the experiment is conducted does not greatly influence quality perception nor quality ratings. However, in the case of quality evaluation in a public area, a higher number of subjects is required for gathering stable results. This is a first indication that the stringent demands posed on the assessment environments can be relaxed to some extent.

The overall experiment duration and length of the video sequences are also two limiting factors of existing subjective quality assessment methodologies. In order to avoid viewer fatigue, experiment duration should not exceed 30 minutes. In [1], the authors propose a novel subjective methodology enabling quality evaluation of long duration audiovisual sequences. Instead of providing an actual quality rating in case of a degradation, subjects are asked to tune the quality of the sequence, by means of a knob, to the desired level. Based on feedback received from the test subjects, the proposed methodology requires less attention from the participants which allows them to focus more on the actual content.

A test bed for augmenting user experience by simulating sensory effects is presented in [10]. Using this test bed, the influence of, for example, wind, light or any other sensory effect on QoE can be evaluated. This also enables to simulate more natural environment conditions during quality assessment.

Research has also been conducted towards replacing the traditional quality rating scales by other scoring techniques such as providing feedback using a glove [2] or a gaming steering wheel [4]. In [5], a comparison was also made between different devices (mouse, joystick, throttle and sliding bar) for providing feedback in the case of time-variant video quality.

As such, active research is being conducted on relaxing the stringent demands imposed on the assessment environment and on creating more realistic environments during quality assessment.

3. MIMICKING REAL-LIFE CONDITIONS DURING SUBJECTIVE QUALITY EVALUATION

As user expectations and context influence QoE, we wanted to investigate the impact of conducting experiments in more realistic/natural environments. In particular, we are interested in discovering the influence of quality degradations when test subjects are not primarily focused on active (audio)visual quality evaluation.

In a first study [8], we worked out a novel subjective quality assessment methodology enabling quality evaluation in real-life environments. Our proposed methodology is based on injecting impairments in full length DVD movies. Next, test subjects were asked to take the DVD home and watch it in their most natural environment. In order to collect feedback concerning the perceived quality and the detection of degradations during playback, a questionnaire was provided to the test subjects in a sealed envelope containing, amongst other, the following questions:

- Did you perceive any kind of visual degradation during movie playback? If yes, which kind?
- Describe the scenes in which these degradations occurred.
- Indicate, on a scale from 1 to 5, impairment annoyance.
- Rate, on a scale from 1 to 5, the overall visual quality of the movie.

The participants were not instructed about the possible occurrence of visual degradations and they were asked not to open the envelope before having watched the entire movie. This way, we encouraged subjects to watch the DVD for its content without actively evaluating perceived quality. An additional questionnaire was used to obtain information on the environment in which they watched the DVD.

In a second experiment [7], we recently assessed the influence of audio/video (A/V) synchronization issues in the case of simultaneous translation of video sequences. Therefore, we mimicked the typical environment used by expert interpreters as shown in Figure 1. In this case, the experiment was conducted using both real interpreters (expert users) and non-experts. The main difference between the evaluations performed by the expert and the non-experts is the fact that the experts users were specifically instructed to

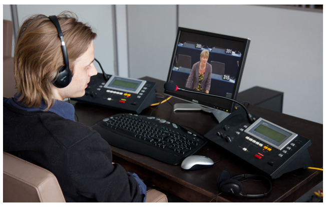

Figure 1: Environmental setup used for assessing A/V sync issues during our subjective quality assessment experiment.

perform a simultaneous translation of the video sequences (as they would do in a real-life scenario). By doing so, the interpreters were primarily focused on the translating part instead of detecting A/V issues.

4. LESSONS LEARNED WHILE MIMICKING REAL-LIFE CONDITIONS

By analysing the results obtained during the two experiments described in the previous section, we found some significant differences compared to results gathered using a standardized quality assessment methodology.

4.1 Influence of assessment environment and task on primary focus

Using our two subjective experiments, we changed the typical assessment environment used during subjective video quality assessment and tried to influence the focus of our test subjects by giving different instructions prior to the start of the experiment.

Concerning the experiment with the full length DVD movies, subjects were not aware of the fact that quality degradations could occur during video playback. Furthermore, the test subjects watched the complete movie in their preferred environment. Also, no impairments were injected in the first or last 15 minutes of the movie. Considering all these factors, participants were stimulated to watch the DVD primarily for its content. Based on the feedback collected from the subjects, we found that our novel methodology based on full length movies is capable of mimicking the typical lean-backward TV-viewing experience. As opposed to the standardized assessment methodologies using short video sequences, we are now also able to increase the immersion of our test subjects.

In our second experiment on assessing the influence of A/V synchronization issues, we explicitly instructed our expert users to perform a simultaneous translation of the audiovisual sequences. On top of that, the interpreters were aware of the fact the synchronization problems could occur during playback. After each sequence, the experts were asked to indicate whether they observed any synchronization issue and how they would rate its annoyance. The results clearly showed that performing a translation of the video sequences significantly changes the ability of the subjects to detect quality degradations. The non-experts participating with the experiment were only instructed to detect the A/V synchronization problems.

As such, by mimicking more realistic conditions during subjective video quality evaluation or by giving slightly different instructions to the test subjects, we are able to significantly change the primary focus of our observers. In turn, this greatly influences their judgements concerning impairment visibility and tolerance as will be explained in the next section.

4.2 Influence of primary focus on impairment visibility and tolerance

During both experiments, we changed the context of quality assessment and, hence, influenced QoE. Afterwards, we also compared these results with results obtained using a standard subjective quality assessment methodology. Due to the change in primary focus of our observers, we see some significant changes concerning impairment visibility and tolerance when mimicking more realistic environments.

While the test subjects watched the DVDs, their focus was mainly on the actual content of the movie. This greatly influenced tolerance towards certain types impairments. As depicted in Figure 2, frame freeze impairments are significantly less detected during real-life QoE assessment. However, when questioning the observers towards the tolerance of frame freezes or blockiness impairments, the majority of the subjects do not tolerate frame freezes. This is caused by the fact that freezes break the natural flow of the movie and have a greater impact on the immersive experience. As stated before, this immersive experience cannot be achieved using a standardized subjective methodology.

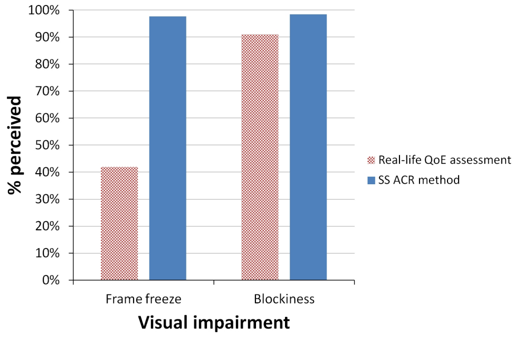

Figure 2: Impairment visibility of frame freezes and blockiness while conducting experiments in real-life and controlled lab environments.

In the case of detecting A/V synchronization problems, we see that the expert users detect these sync issues much less compared to the non-experts as depicted in Figure 3. In general, the interpreters do not detect lipsync problems except in the case delay between audio and video increases up to -240ms[2]. As such, due to a change of primary focus going from active video quality evaluation to simultaneous translation of video content, observers' attention is taken away from the quality aspect of the content.

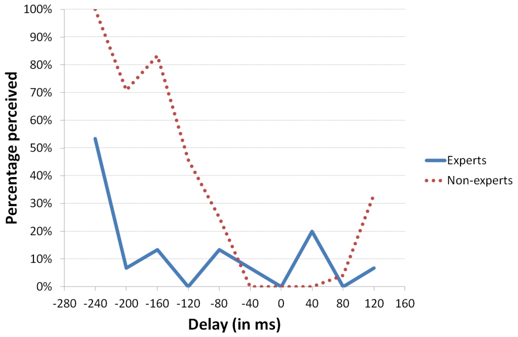

Figure 3: Percentage of the experts and non-experts who detected the A/V synchronization delay.

These results indicate that the specific task, immersion and

[2] A negative offset implies that audio is delayed with respect to the video.

primary focus significantly impact quality perception and detection.

4.3 Points of special interest to consider

By running subjective experiments in more natural environments/settings, we are able to mimic realistic conditions which influence subjects' QoE. These studies revealed some important differences compared to performing quality assessment inside a controlled lab environment. However, conducting experiments by mimicking natural environments requires some points of special attention.

From nature, subjective quality experiments are known to be time-consuming and expensive. In our case, we also conducted face-to-face interviews after the experiment with each of the participants. This aids in contextualizing the experiments and can provide more insights on how quality is assessed in natural environments. Unfortunately, these interviews complicate and further increases the time needed for conducting such experiments.

Another point of special interest to consider is the (highly) uncontrolled environment. During real-life QoE assessment, a lot of environmental factors such as lighting conditions, screen contrast ratio, distance to the screen, etc. cannot be controlled and can differ significantly from one subject to another. Therefore, special attention should be given to collecting as much parameters as possible characterizing the environment by means of, for example, an additional questionnaire. Further research is needed to assess the influence of different environmental conditions on quality perception.

In case of our full length DVD methodology, subjects were only questioned after having watched the entire movie about perceived quality and the number of detected visual degradations. As such, some bias can be introduced. For example, it can happen that certain impairments are forgotten, due to recency or primacy effects. Therefore, other means might be investigated for providing instantaneous feedback on perceived quality. However, this immediate feedback should not change the focus from movie content to active quality evaluation.

5. CONCLUSIONS

In this paper, we presented two subjective studies conducted to assess the importance of mimicking more natural environments during subjective video quality evaluation. By summarizing the main research findings of these experiments, we highlighted the importance of considering the targeted use case when conducting subjective experiments as this significantly influences subjects' expectations and context.

Our results show significant differences concerning impairment visibility, tolerance and annoyance compared to results obtained using one of the standardized quality assessment methodologies. This shows that results gathered during quality assessment in a controlled lab environment can be relaxed when measuring quality in more realistic environments. This calls for additional research on mapping controlled lab results to more real-life scenarios.

6. ACKNOWLEDGMENTS

Part of the research activities described in this paper were funded by Ghent University, the Interdisciplinary Institute for Broadband Technology (IBBT) and the Institute for the Promotion of Innovation by Science and Technology in Flanders (IWT). This paper is the result of research carried out as part of the IBBT projects *Video Q-Sac* (consortium of industrial partners Alcatel-Lucent, Telindus, Televic and Fifthplay in cooperation with the IBBT research groups: IBCN & MultimediaLab (UGent), SMIT (VUB) and IMEC) and *OMUS* (consortium of the industrial partners Thomson, Televic, Streamovations, Excentris in cooperation with the IBBT research groups: IBCN, MultimediaLab & WiCa (UGent), SMIT (VUB), COSIC (KULeuven) and PATS (UA)).

7. REFERENCES

[1] A. Borowiak, U. Reiter, and U. P. Svensson. Quality evaluation of long duration audiovisual content. In *IEEE Consumer Communications and Networking Conference (CCNC)*, pages 337 –341, January 2012.

[2] S. Buchinger, W. Robitza, M. Nezveda, M. Sack, P. Hummelbrunner, and H. Hlavacs. Slider or glove? Proposing an alternative quality rating methodology. *Fifth International Workshop on Video Processing and Quality Metrics for Consumer Electronics (VPQM)*, January 2010.

[3] ITU-T Rec. P.10/G.100 Amd 2. Vocabulary for performance and quality of service, 2008.

[4] T. Liu, G. Cash, N. Narvekar, and J. Bloom. Continuous mobile video subjective quality assessment using gaming steering wheel. *Sixth International Workshop on Video Processing and Quality Metrics for Consumer Electronics (VPQM)*, January 2012.

[5] O. Nemethova, M. Ries, A. Dantcheva, and S. Fikar. Test equipment of time-variant subjective perceptual video quality in mobile terminals. In *International Conference on Human Computer Interaction (HCI)*, 2005.

[6] M. H. Pinson. The influence of environment on audiovisual subjective tests. VQEG_MM2_2011_044_audiovisual lab comparison, Hillsboro, US, December 2011.

[7] N. Staelens, J. De Meulenaere, L. Bleumers, G. Van Wallendael, J. De Cock, K. Geeraert, N. Vercammen, W. Van den Broeck, B. Vermeulen, R. Van de Walle, and P. Demeester. Assessing the Importance of Audio/Video Synchronization for Simultaneous Translation of Video Sequences. *Springer Multimedia Systems*. Accepted for publication.

[8] N. Staelens, S. Moens, W. Van den Broeck, I. Mariën and, B. Vermeulen, P. Lambert, R. Van de Walle, and P. Demeester. Assessing quality of experience of IPTV and Video on Demand services in real-life environments. *IEEE Transactions on Broadcasting*, 56(4):458–466, December 2010.

[9] N. Staelens, B. Vermeulen, S. Moens, J.-F. Macq, P. Lambert, R. Van de Walle, and P. Demeester. Assessing the influence of packet loss and frame freezes on the perceptual quality of full length movies. *Fourth International Workshop on Video Processing and Quality Metrics for Consumer Electronics (VPQM-09)*, January 2009.

[10] M. Waltl, C. Timmerer, and H. Hellwagner. A test-bed for quality of multimedia experience evaluation of sensory effects. In *International Workshop on Quality of Multimedia Experience (QoMEX)*, pages 145 –150, July 2009.