Estimating the Effect of Context on QoE of Audiovisual Services: Laboratory vs. Public Places

Toni Mäki
VTT Technical Research Centre
of Finland
Email: toni.maki@vtt.fi

Martín Varela
VTT Technical Research Centre
of Finland
Email: martin.varela@vtt.fi

Jukka-Pekka Laulajainen
VTT Technical Research Centre
of Finland
Email: jukka-pekka.laulajainen@vtt.fi

***Abstract*—Multimedia service providers and operators need tools that can estimate the quality of the services delivered to their end-customers. To this end, a variety of quality prediction models has been developed. The models are typically built and trained based on data acquired from user studies conducted in laboratories with precisely defined conditions. While laboratory originated data is accurate and reliable, the performance of Quality of Experience (QoE) models could be possibly further improved by taking into account the real context of use. Working towards that goal, in this paper, we compare results of a laboratory audiovisual quality subjective study and two smaller scale user studies conducted in public place. We also discuss our approach of enhancing the prediction accuracy with context-specific bias functions.**

I. Introduction

Multimedia services offered through IP networks, such as IPTV, have gained popularity and increased competition between service providers. In order to keep customers satisfied, service providers and operators need to take care that the quality of their video streaming offerings is good enough for their users. In quality management it is crucial to monitor the realized QoE of services delivered to customers. To this end, many so called *objective* quality models — that is, those which can provide quality estimates without human intervention — have been developed that can estimate quality, in some cases in real-time. An extensive overview of quality models, including the audiovisual models relevant to this work can be found in [1].

Typically the objective quality models are developed using knowledge gained from subjective user studies, both directly and through the development of Human Visual System (HVS) models. The standardized procedures for performing subjective video assessment requires having strictly controlled conditions for the evaluation environment and context. Some of the most used definitions for psycho-perceptual approach are specified in standards [2] and [3]. These define viewing and listening conditions e.g. related to viewing distance, luminance of the screen and background and the room illumination. These constraints are critical for achieving consistency in the assessments, and obtaining results that are comparable with those performed by other laboratories. However, it can be argued that some benefits could be gained by relaxing the requirements for the subjective test environment. It seems likely that the standard evaluation conditions are too far from the practical use conditions of the applications in question (e.g. for TV-like services the tightly-controlled lab environment differs very much from a typical living-room, and even such thing as a "typical" living room seems like a dubious concept). Taking the assessments out of the lab also enables other assessment models, such as crowd-sourcing, where significant cost savings could be achieved. This would of course have an impact on the consistency of the assessments, and significantly reduce the reproducibility of the results, but would enable more tests to be carried out, and in more ecologically valid conditions.

The possible benefits of conducting subjective video tests in non-laboratory environment are not new, but these approaches have only recently begun taking some hold in the quality assessment domain. In other domains of research, such as Human Computer Interaction (HCI), the realism of the context in which the tests are performed is considered as very important, cf. [4] [5] for examples related the quality of mobile applications. This user-centered approach has been applied e.g. in [6] and [7], where the usability of applications in mobile and laboratory environment has been compared. In [8], the authors question the validity of laboratory-based quality evaluation and present a comparison of mobile television quality studies conducted in an actual use context and in a conventional laboratory environment. They conclude that there is a difference between the quality ratings

This work was supported in part by Tekes, the Finnish Funding Agency for Technology and Innovation, under the CELTIC IPNQSIS project.

derived from laboratory and context studies. Their results show that users tend to be more tolerant to problems in real context than the laboratory studies imply.

In this paper, we compare the results of a laboratory-based audiovisual assessment campaign with those of two separate (and smaller scale) campaigns carried out in public places, under a completely different context. Besides the explicit goal of comparing the results of subjective assessments in a lab vs. non-lab environment, this work provides a first step into developing context-specific bias functions to easily and cheaply adapt quality models, typically trained on laboratory-based data, to new contexts of use. These experiments are the first in the series of experiments with purpose of understanding the effect the context of use have on QoE.

The paper is organized as follows. Section II describes the test content and how it was generated. Section III describes the laboratory-based assessment campaign. In Section IV we describe the subjective tests performed at two exhibition halls. Section V discusses the results of the public place tests and compares them to the laboratory tests. We also discuss the performance of the laboratory trained model in the public contexts, and that of the potential bias functions based on the field tests. Finally, Section VI concludes the paper.

II. Test Content

Both the sequences used in the laboratory campaign and those used in the exhibition were generated using the same instrumentation system. The original audiovisual material was acquired from The Consumer Digital Video Library [9]. The samples that contained speech were in English. The frame rates of the downloaded samples were either 25 fps or 30 fps. The samples were edited as per the P.911 [2] guidelines (e.g. length approximately 10 s, no interrupted words) and encoded using H.264. The encoding was done with a 2-pass method (baseline profile), and into different bit-rates depending on the samples' resolution (6 Mbps for 1920x1080, 3 Mbps for 1280x720 and 1 Mbps for 854x480). The slicing feature of H.264 was exploited in order to fit a single slice into a single UDP packet. The Intra frame interval was configured to a maximum of 1 s. AAC was used for encoding audio, at two different bit-rates (96 kbps for 1920x1080 and 1280x720 resolutions and 64 kbps for 854x480 resolution). A network emulator (Linux's Netem, expanded with the Gilbert model extensions from the NetemCLG project [10]) was used to create realistic losses in the test bed, by replaying previously-created loss traces (generated with in-house developed tool[1]) with accurate loss rates and mean loss burst sizes. The test sequences were prepared prior to the assessment by recording RTP-based video streams transmitted over the emulated network.

TABLE I
The influence factors tested

Dimension	Description	Values
RES	The resolution of video sample	854x480, 1280x720, 1920x1080
EC	Error concealment method in use	Enabled, Disabled
LR	Percentage of packets being lost during the transmission of the video sample	0 %, 0.3 %, 0.6 %, 1.2 %, 2.4 %, 4.8 %

TABLE II
Original video samples

Original sample	Description
NTIA snowy day in the city (1e)	Three scenes with snowfall.
NTIA Jump Rope	A man jumping a rope.
NTIA Elephant Crane	An elephant crane playing on a stick.
NTIA Highway Cuts	Several views of cars driving.

It was expected that both within-subject and between-subject variance would be high in the public place tests, because of the small sequence set size. Therefore, in order to increase the confidence on the examined conditions, we decided to concentrate only on three influence factors Resolution (RES), Error Concealment (EC) and Loss Rate (LR). Movement Quantity (MQ) and Mean Loss Burst Size (MLBS) that were part of the original laboratory test plan were fixed to "Moderate" and "2", respectively. Table I displays the varied influence factors with their respective value ranges. The laboratory tests did not consider error concealment (EC), because of a programmatic error. Nevertheless, as we were interested to see its effect size, we included error concealment as a varied influence factor. The sequences evaluated at exhibition were generated from four different contents listed in Table II. In laboratory setting 12 different contents were used. In all the discussion that compares the context, only the equal conditions and content are used.

III. Subjective Tests in Laboratory

We conducted a subjective quality assessment campaign in the VTT multimedia laboratory. The goal of the assessment campaign was to capture the effect that certain influence factors have on QoE. The studied factors were resolution (RES), quantity of movement (MQ), loss rate (LR) and mean loss burst size (MLBS). The participants of the laboratory assessment campaign

[1] https://github.com/mvarela/Gilbert-Loss-Trace-Generator

consisted of 24 VTT employees. With exception of four persons, all subjects were native Finnish speakers. Seven subjects were female and seventeen male. Three subjects considered themselves non-technical and seven considered themselves technical person. Four subjects were multimedia experts and ten persons had prior experience with multimedia quality assessment. The average age of subjects was 32.5 years (range of 24 - 46). The assessments were done using an evaluation tool developed in-house[2] in an environment conditioned as closely as possible according to the requirements of [2]. Closed headphones were used for listening to the audio. Each subject evaluated a set of sequences containing 125 video sample pairs. The set of sequences contained 3 repetition sequences and two anchor sequences with extreme conditions. The subjects went through a short training session before beginning the actual assessment.

A slightly modified version of the Degradation Category Rating (DCR) method described in P.911 was deployed. The subjects were presented first the original video sample (a sample transmitted over an error-less network) and then the distorted sample. After each sequence the subjects were presented with three questions (in two separate stages) instead of one. In the first stage, the subjects were asked to rate the difference in audiovisual quality of the pair of sequences, as in normal DCR. In the second stage, the subjects were asked to rate separately the difference in the audio quality and the difference in the video quality. The five-level impairment scale from P.911 [2] (Very Annoying, Annoying, Slightly Annoying, Perceptible but not annoying, Imperceptible) was used for all the voting. The results of the subjective assessment were found to be consistent under statistical analysis. This allows us to posit that the voting methods used in laboratory and in exhibition context themselves produce comparable results regarding audiovisual quality.

IV. Subjective Tests at Exhibition

The public place assessments were conducted in exhibition halls during Future Network and Mobile Summit 2012 in Berlin, Germany and during Celtic-Plus Event and Proposers' Day 2013 in Kayseri, Turkey. The sequences were presented and evaluated using laptop computers that were exactly similar to the one used in laboratory assessments (15,4" screen with 1920 x 1200 resolution), with the exception of having an external monitor in laboratory (25.5" screen with 1920 x 1200 resolution). Closed headphones were used for listening the audio. The assessment application was configured to use the normal (single-stage) DCR method described in P.911 [2] for voting. After each configuration (pair of samples) subjects gave their opinion of the audiovisual quality difference between the samples. The five-level impairment scale of P.911 was again used. Each subject evaluated a single set of sequences containing 13 video sample pairs. Within each set of sequences there were one repeated sample (as a consistency check) and two anchor sequences with extreme conditions. Ten different sets of rendered sequences were randomly drawn before the assessments. Each subject was assigned a set in round-robin fashion. Instructions were given in paper format in English. The instructions were based on Section II.2 of Appendix II of P.911. There was a short training session before the actual assessment also in the public place tests.

In Berlin, twenty people did the assessment. This is comparable in scale to the laboratory tests. The subjects reported eleven different languages as native tongue (a single person was a native English speaker). Three of the subjects were female and seventeen were male. The average age of subjects was 35.4 years (range of 24 - 59). Seventeen subjects evaluated themselves as technical people, a single person as multimedia expert and two subjects as multimedia assessment experts. A small gift was given to participants as a reward (a few participants did the test without a reward, as more subjects than expected participated in the campaign).

In Kayseri, the results of 9 subjects were usable, due to technical problems in the assessment process. This small sample size somewhat limits the strength of the conclusions we can draw from these results. The subjects reported six different languages as native language (no native English speakers were found in this group). One of the subjects was female, and 8 were male. The average age of subjects was 41 years (range of 27 - 60 years). Three of the subjects did not have previous assessment expertise. Two of the subjects were multimedia experts, and the rest considered themselves as technical people.

Moving from a laboratory environment to an exhibition hall implies major contextual differences in the assessment. The most remarkable difference is probably in the environment itself. While the laboratory environment was peaceful and properly lit, the exhibition hall was noisy, occasionally crowded and variable in illumination. Another notable difference between the environments was the presence of other people. In the laboratory, the subjects had full privacy, while in the exhibition they were often accompanied by other people (although only they had a direct view on the screen). There was a difference also in screen size as the assessment in exhibition were done using the laptops' native screen, while in laboratory an external 24-inch monitor was used. Main differences are summarised in Table III. The test content and the influence factor combinations used at exhibition was a subset of those used in laboratory.

[2] This tool is freely available for research purposes, please contact the authors for further information.

TABLE III
LABORATORY VS. EXHIBITION HALL CONTEXTS

Laboratory	Exhibition hall
Conditioned background	Disturbance on background (people discussing, working etc.)
Controlled illumination	Variable illumination
No audio disturbance	Audio disturbance caused by other people
Full privacy	Surrounded by other people
External monitor	Laptop's internal monitor

In the second campaign in Kayseri, participants were similarly rewarded with a small gift for their time, as in Berlin.

V. RESULTS

In this section we discuss the results acquired from the exhibitions. We start by looking into the reliability of the votes and the main effects caused by influence factors studied for the Berlin campaign. The Kayseri results are excluded from this analysis, because of the low number of assessments. They are considered in the discussion of the contextual differences. Finally, we compare the assessments of a subset of the sequences (those having equal configuration to that of laboratory ones) between laboratory and exhibition contexts.

The reliability of the conducted tests and the certainty of the subjects was examined with SOS (Standard Deviation of Opinion Scores) analysis according to [11]. According to the SOS hypothesis the MOS and the SOS^2 of well-behaving results should have a square relationship characterized by a parameter named a. Figure 1 displays the realized standard deviation of MOS values as a function of MOS values and the associated fit of the square function. The standard deviations of randomly drawn MOS values are shown for comparison. The a parameter calculated for the subjective data takes value of 0.1636. This value is close to those presented for video streaming user studies in [11] (0.1078 - 0.2116). This suggests good comprehension of the voting task and realistic variability in the voting by the subjects.

Figure 2 illustrates the main effects caused by the influence factors studied plus the effect of the content. The Loss Rate and the Error Concealment factors have a strong effect on the perceived quality. Resolution, on the other hand, has a weak effect. The content does not have a significant effect on the MOS (let us remind that we are discussing DCR voting). ANOVA results show that the effects of Loss Rate and Error Concealment are statistically significant (with alpha level 0.05), with p-values of 1.869E-14 and 0.042, respectively. No significant interactions were observed.

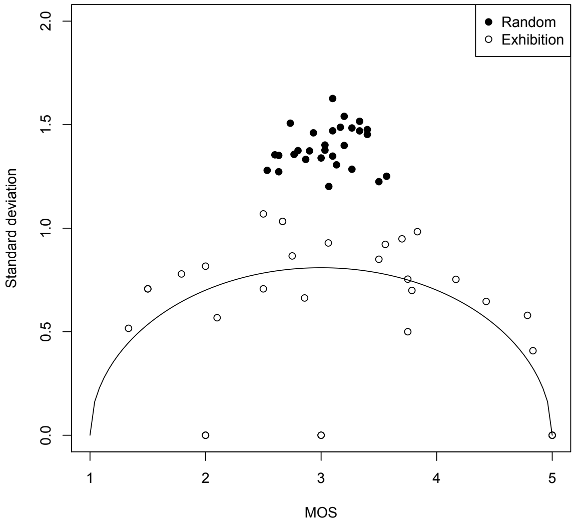

Fig. 1. Standard deviation of MOS as function of MOS

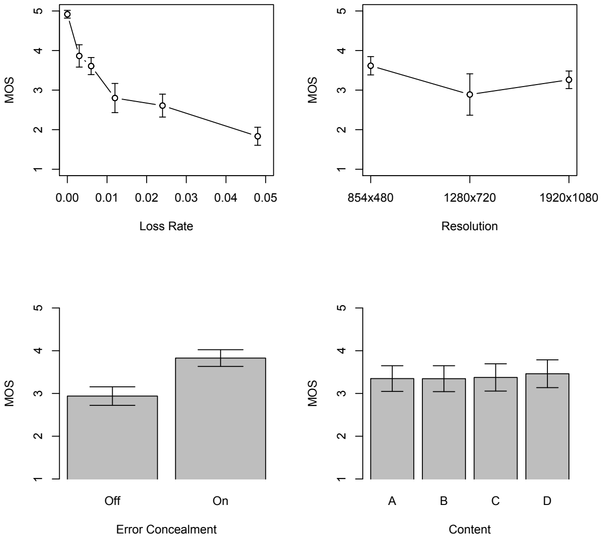

Fig. 2. The main effects of the influence factors and content

Figure 3 illustrates the MOS as function of LR for the laboratory and exhibition contexts (the confidence intervals of the Kayseri votes were left out for clarity. They are naturally wider than those for the Berlin campaign). Figure 4 illustrates the MOS (averaged over all loss conditions) for the smallest and the largest resolution. To make the data comparable only votes from the sequences with EC off are included. Both figures suggest the context could play a significant role in the composition of the whole QoE. Judging by the confidence intervals, the differences between laboratory originated and Berlin exhibition originated assessment results are statistically significant. The results from Kayseri should be interpreted very cautiously, because of the large confidence intervals.

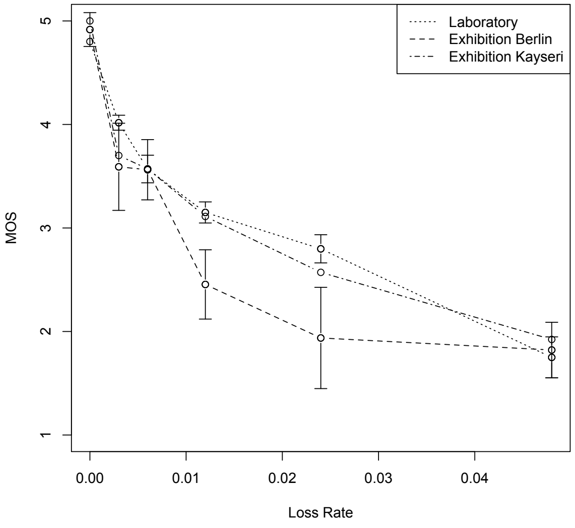

Fig. 3. MOS as function of Loss rate in different contexts

We then tested the accuracy of a Random Neural Network based PSQA [12] model trained with the laboratory-obtained data against the exhibition-obtained subjective data. Only the matching conditions from laboratory and exhibition user studies were included (i.e. the conditions with no error concealment on). The conditions were then used as input to the model and corresponding MOS estimations were calculated. The correlations and the errors of the QoE estimations to the actual MOS values are presented in Table IV. In both cases the model performed noticeably worse in the exhibition context, which, while expected, likely indicates a need for *post-hoc* calibration of models trained with laboratory data with data from realistic usage contexts.

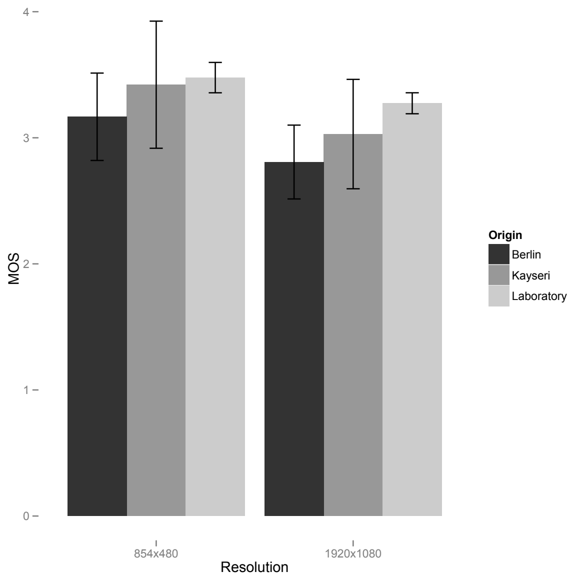

Fig. 4. MOS as function of Resolution in different contexts (averaged over all introduced loss rates). The results indicate that HD content is more susceptible to noticeable degradations in the presence of losses

TABLE IV
ACCURACY OF THE AUDIOVISUAL PSQA MODEL

Context	R_{spearman}	R_{pearson}	RMSE
Laboratory	0.802	0.846	0.485
Berlin	0.785	0.789	0.756
Kayseri	0.825	0.796	0.641

The differences observed between laboratory and exhibition results can be approximated by a quadratic function. Applying this context specific first-order correction function could be used to bring the model's estimations closer to the observed MOS values. We fitted the quadratic approximations with data from both Berlin and Kayseri. Then we used the data from Kayseri to validate the correction function derived from Berlin data (and *vice-versa*). There was a moderate improvement in the estimation accuracy regarding the Berlin votes, when Kayseri originated correction function was applied. In opposite case the estimation accuracy suffered a bit, because of overly strong correction. We plan to continue developing this approach by collecting more user data in order to cover wider range of contexts and thus make the estimations more reliable and generalizable.

VI. CONCLUSIONS AND FURTHER WORK

In this work we compared the results of formal subjective audiovisual assessment with more informal assessments done in actual usage contexts (in this case two public exhibition halls). We observed significant differences in the results, both in terms of the MOS values and on the impact of the different quality-affecting factors. Interestingly the results show that the subjects at public crowd were less tolerant to the quality degradations than the subjects in the laboratory. This is contradictory to the findings of [8]. Whether this is a result of using the DCR voting method or an effect of some other contextual factor or factors is a question that requires further studies. Specifically, tests separating the effects of contextual factors on a) voting behaviour and b) actual experience should be conducted.

We also demonstrated the viability and limitations of an audiovisual model trained on the laboratory-obtained data, when used in a different context, namely in crowded public places. The performance of the model in the exhibition context was inferior to the performance in laboratory context. However, the estimations could still provide usable estimations for quality monitoring purposes e.g. in public displays.

We are currently working on a model calibration method that uses information derived from lightweight user tests done in the specific context. The idea is to test and model the effects of the dominating influence factors in order to formulate a context specific correction function. To this end and in order to understand different contexts of use and devices generally, user tests outside laboratory shall be continued.

REFERENCES

[1] A. Raake, J. Gustafsson, S. Argyropoulos, M. Garcia, D. Lindegren, G. Heikkila, M. Pettersson, P. List, and B. Feiten, "IP-Based mobile and fixed network audiovisual media services," *IEEE Signal Processing Magazine*, vol. 28, no. 6, pp. 68 –79, Nov. 2011.

[2] *Subjective Audiovisual Quality Assessment Methods for Multimedia Applications*, ITU-T Std. P.911, 1998.

[3] *Methodology for the subjective assessment of the quality of television pictures*, ITU-R Std. BT.500-13, 2012.

[4] M. A. Sasse and H. Knoche, "Quality in context-an ecological approach to assessing qos for mobile tv," in *Proc ISCA/DEGA Tutor & Res Workshop Percept Quality of Systems*, 2006.

[5] A. Gotchev, A. Smolic, S. Jumisko-Pyykkö, D. Strohmeier, G. Bozdagi Akar, P. Merkle, and N. Daskalov, "Mobile 3d television: development of core technological elements and user-centered evaluation methods toward an optimized system," in *Society of Photo-Optical Instrumentation Engineers (SPIE) Conference Series*, vol. 7256, 2009, p. 15.

[6] T. Kallio and A. Kaikkonen, "Usability testing of mobile applications: A comparison between laboratory and field testing," *Journal of Usability studies*, vol. 1, no. 4-16, pp. 23–28, 2005.

[7] J. Kjeldskov and J. Stage, "New techniques for usability evaluation of mobile systems," *International Journal of Human-Computer Studies*, vol. 60, no. 5, pp. 599–620, 2004.

[8] S. Jumisko-Pyykkö and M. M. Hannuksela, "Does context matter in quality evaluation of mobile television?" in *Proceedings of the 10th international conference on Human computer interaction with mobile devices and services*. ACM, 2008, pp. 63–72.

[9] Intel Corporation, National Telecommunications and Information Administration's Institute for Telecommunication Sciences and University of California at Santa Barbara, "The consumer digital video library." [Online]. Available: http://www.cdvl.org

[10] S. Salsano, F. Ludovici, A. Ordine, and D. Giannuzzi, "Definition Of A General And Intuitive Loss Model For Packet Networks And Its Implementation In The Netem Module In The Linux Kernel," University of Rome - Tor Vergata, Tech. Rep., Aug. 2012.

[11] T. Hoßfeld, R. Schatz, and S. Egger, "SOS: the MOS is not enough!" in *2011 Third International Workshop on Quality of Multimedia Experience (QoMEX)*, Sep. 2011, pp. 131 –136.

[12] M. Varela, "Pseudo-subjective quality assessment of multimedia streams and its applications in control," Ph.D. dissertation, INRIA/IRISA, univ. Rennes I, Rennes, France, Nov. 2005.