©2016 Society for Imaging Science and Technology
DOI: 10.2352/ISSN.2470-1173.2016.16HVEI-123

Are lab-based audiovisual quality tests reflecting what users experience at home?

Miguel Rios Quintero, Technische universität Berlin, Berlin/Germany; Alexander Raake, Audio Visual technology, Technische universität Ilmenau, Ilmenau/Germany

Abstract

This study aims to understand the influence of the environment over the perception of Audiovisual quality on IPTV services. Thus, the test participants were distributed between 2 groups. One group performed the test in a living room lab. This laboratory has the typical characteristics of a regular living room. The second group performed the test in a room which characteristics were based on the ITU Recommendation ITU-R Rec. BT. 500. To ensure the reciprocity of the conditions between both rooms, extreme care of environmental conditions were taken. Hence, factors such as ambient light, ambient noise, temperature, and relation between TV size and viewing distance were replicated in both rooms. Preliminary results reveals discrepancies between subjective quality evaluations performed in both rooms.

Introduction

Laboratory based conditions and objective tests have been until now, the most efficient approach to achieve reliable results. Unfortunately, this type of set-ups rarely replicate the real conditions in which the evaluated activity takes place. A clear example of those limitations can be seen in the results obtained in traditional laboratory quality test. Standardized quality test results in terms of MOS (Mean Opinion Score), in addition collected in a Lab environment, cannot readily be used to evaluate how users actually appreciate multimedia services under real-life conditions. The main concern is that every user has a different perception created by several influential factors such as own previous experience or cultural background. Hence, lab oriented subjective quality assessment, provide little or no information about perceived value of a service may have for the user while using in real conditions (e.g. home).

Thus, to ensure the proper development of a user model, it is imperative to understand the effects of the environment on user perception. Therefore,the development of test targeting specific factors of a user case should be the next step on the development of more reliable user model test.

Laboratory conditions are a good basis for understanding the relation between technology and user-perceived quality, but it is unclear if the results can be extrapolated to a realistic usage. Therefore, it is necessary to identify any significant user factors which can lead to a possible differentiation when evaluating quality and acceptability. The intention is to implement subjective assessment methodologies that take into account extrinsic influencing factor.

The paper is structured as follow: Section one illustrate the limitations encountered in the traditional testing paradigm and how researchers nowadays develop different approach towards an ecologically valid test scenario; Section two describe the entire project set-up, aim of the study and experimental conditions evaluated; Section three describes the final test set-up, and test subject selection. Section four presents obtained results; and in section 4 we provide a conclusion and present our future approach towards and ecologically valid model.

Previous Work

Laboratory conditions are good basis for understanding the relation between technology an user-perceived quality, but it is unclear if the results can be extrapolated to a realistic scenario. Unfortunately, the possibilities in which people enjoy of audiovisual content are every day more diverse and complex. This phenomenon can be observed in the context of adoption of mobile telephony. As a result of the constant evolution and improvement of mobile network. the consumption of mobile audiovisual content has been able to reach the levels in which it can compete with traditional fixed services such Television. These complex scenarios presents a new an bigger challenges for researcher who strive to develop close to realistic Quality of experiences model. On these days, context of usage which were previously considered irrelevant such as the environment can cause a significant influence in the final perception of quality of users. A clear example of this have been seen in a test performed by Jumisko-Pyykoö, and Hannuksela in [1]. Their test revealed that similar technical parameters can be perceived and evaluated differently depending of where is the measurement performed. The obtained results demonstrated the effect of extrinsic influencing factors not previously taken into account. When talking about extrinsic influencing factors, we are referring to all those properties or characteristics which are originated in the environment that surrounds the service user. Hence, it means that such factor could influence in the final intention of the user of adopting or dropping a service. unfortunately,). These factors are particularly difficult to analyse in real conditions, due to their singularity, which is dependent on the context.

The most effective approach to mitigate the effect of extrinsic influencing factors is to develop standardized tests. Good guidelines can be followed from standardized procedures such as the presented in ITU-T P910 [2] and BT.500-11 [3]. Regrettably, the gain of reliability and repeatability is traded of with its ability of resembling an ecologically valid scenario.

Even factors such as the length duration of a video sequence need to be scarified in order to gain reliability. In a typical standardized audiovisual test, an audiovisual sequence has a duration of between 10 and 15 seconds. Not only it is understood

DOI: 10.2352/ISSN.2470-1173.2016.16HVEI-123

that content of such duration cannot be compared with the experience of watching a complete movie or a tv series, but even when compared with sequences which are slightly longer it can be seen a difference in evaluation of a service. An example of this phenomenon is seen in the work of FrÖhlich et al [4]. Their study demonstrate that even a difference between sequences below 3 minutes, there can be a significant difference on quality rating. The study concluded with the premise that shorter videos are evaluated more critically than longer sequences.

Fortunately, in last years, researchers have been taken a more holistic approach on the development of more close to realistic scenarios. Prominent examples this new adopted paradigm can be seen in the study of mobile communication. Examples of this can be studied in [5] and [6]. These studies encounter the challenges of performing quality tests in scenarios systematically controlled. Their obtained results reported how acceptability is affected, not only by the technical parameters, but by the environment. Specifically speaking of Schatz et al. study [5]. Their results revealed that test participants were more critical in the evaluation process of the usage of web services in the laboratory than in an ecologically valid environment.
In the area of audiovisual quality, more prominent studies can be found. One of the most significant studies to the date is the one presented by Staelens et al. [7]. In their study, they attempted to perform a test simulating a entire home experience. Despite of the difficulties encountered in the complexity of the test set-up. The studies revealed the significant influence that the home environment can have over user quality of experience while watching a movie in the comfortability of their home places.

For our study, we developed a systematic approach in which we strive to analyse the effect of technical quality levels on a close to realistic scenario. The present study is the first project of a series of 4 projects towards a better understanding of acceptability and Quality of experience in IPTV services.

Audiovisual Quality Assessment Project

This section describes a series of audiovisual Subjective quality test designed for the purpose of understanding the possible influence from the visual appeal of the environment on perceived quality. to this end, we designed a series of test based on the single stimuli audio visual quality test performed according the cf. ITU-T Rec. P.910 [2] and P.800 [8].
The visual appeal differences between rooms can be seen in 2 and 1. 2resembles an standard ITU-T based laboratory room, and 1 a typical living room environment. However, both rooms were build taking into account the standardized requirements for room testing based on ITU-T standards. The project was divided into two phases. The first phase, or exploratory phase, consisted of two independent groups of test participants who performed separated audiovisual quality tests on different environment with similar controlled environmental conditions, but with different visual appeal characteristics. The second phase, or repeated-measures phase, consist on a repetition of the first phase. The difference between phases resides in the fact that test some participants from the first phase were invited to perform the test in the second environment.

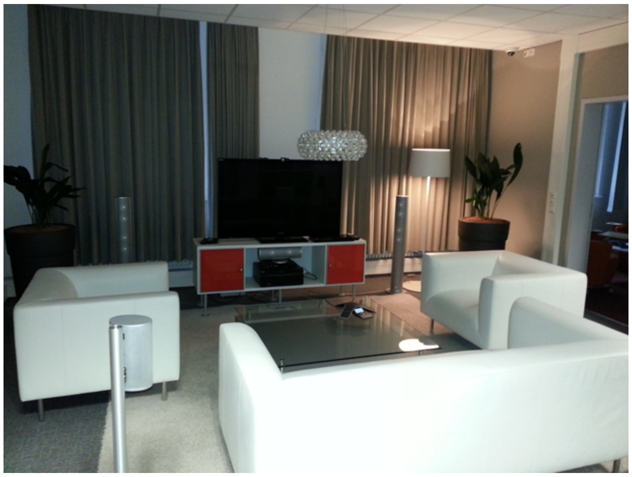


Figure 1. *Living Room test environment*

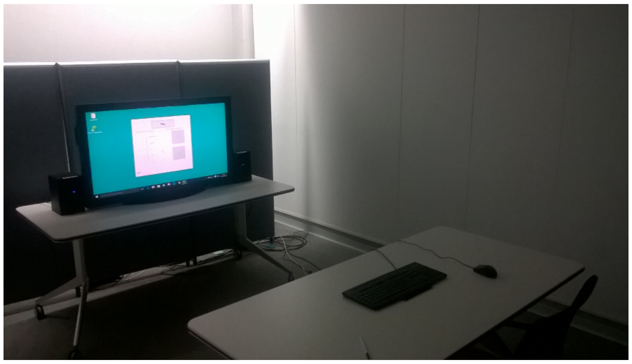


Figure 2. *Laboratory test environment*

In the table 1 are listed all the condition are all the condition which were controlled during the project.

Rooms Controlled Conditions (Ranges)		
Ambient Light	200 Lux at the viewing spot	
Ambient Noise	60 5 dB on Quiet Scenes	73 5 dB on Loud Scenes
Room Temperature	22C	
Viewing distance	2.5 Times the size of the television	

Table1 :Controlled environmental parameters

Audiovisual Material

The audiovisual material used in this test consist of 70 sequences of 15 seconds. 15 of the sequences belongs to the dataset used in the development and validation of the standardized model ITU-T Rec. P.1201 [8] and P.1202 [9]. The other 35 sources were selected under the criteria of bein highly engagin content. The selection of the highly engaging content was made using a search script which evaluates the most popular movie scenes in the area of Germany in 2012.
All sequences were process in FULL HD and with 3 different quality levels. In table 2 are described the details of the technical parameters from the original sources. the different levels of quality were defined based on the percentage of packet loss

©2016 Society for Imaging Science and Technology
DOI: 10.2352/ISSN.2470-1173.2016.16HVEI-123

(0%,0.625% and 1.25% packet loss). the errors were inserted directly into the stream files using a 4-state Markov model [4]. The impairment style used was slicing. At the end, 210 sequences with different quality levels were generated for this project. As a result of the large quantity of sequences used in the project, the total list was randomized, divided and presented in two session by an intermediate break of 5 to 10 minutes.

Technical Parameters	
Video	
Bandwidth	15 Mbits/s
Resolution	1920 x1080p
Codec	H.264/AVC
Frame Rate	24 fps
Audio	
Bandwidth	320 kbits/s
Channels	2
Codec	AAC
Sampling frequency	44100Hz

Table2 : Sources Processing Parameters

Evaluation Methods

In this Project, emotional state information was collected using a single dimensional seven point pictorial scale. The evaluation was repeated during the three phases: at the briefing, after the intermediate break, and at the end of the test. The goal of such evaluative schema aims to detect possible changes in the emotional state which could influence the evaluation criteria of test participants.

The audiovisual quality evaluation was performed in terms Absolute Category Rating (ACR) scheme on a 5-point scale, cf. ITU-T Recs. P.910 and P.800. During the briefing phase, the test participants received precise instructions on the usage of the evaluating system. Furthermore, a series of sequences with different quality levels were used in the training phase to indicate what kind of audiovisual impairments the test participant should expect. Additional to the quality evaluation, test participants were asked to provide ratings of acceptability on a binary yes/no scale. The test participants were encouraged to evaluate acceptability based on the hypothetical questio: " would you use a service under the current condition or not?"

Additionally, socio-demographic and television consumption habits information was collected using surveys during the briefing and post-test phases.

Project Development

AS previously mentioned, the project was divided into two phases, the exploratory phase and the repeated-measures phase. The first phase aims to identify any possible differences in the rating behaviour between the different environment and quality levels for two distinct groups. In the second phase, the study focused on a deeper understanding of test participants rating behaviour across the different environments. To this end, a sub group of participants from the first phase were invited to perform once more the test this time in a different environment in which they performed the test during the exploratory phase. Prior starting the repeated-measured phases, we established a pause of 2 week after the completion of the exploratory phase. With the pause we attempt to reduce any possible bias triggered by performing the same test in for a second time in a short period of time.

In both phases, the test set-up remained unchanged, and consisted of the following six phases:

1. *Briefing*: a visual acuity test, a socio-demographic and an initial emotional state information were evaluated using surveys
2. *Training*: Test participant were instructed on the usage of the rating system and evaluative criteria. Also they performed a mocking test with several sources showing the variety of quality conditions
3. *First audiovisual test*: Test participants evaluated half of the randomly presented sequences were evaluated (105 sources)
4. *Intermission*: during the pause an emotional state survey was answered by the test participant
5. *Second audiovisual test* : The second half of the randomly selected sequences were evaluated (105 sources)
6. *Post-test*: Here, a third emotional state and TV consumption behavior surveys were evaluated by the test subjects

The programmed test duration was of approximately 1 hour 45 minutes.

Project Results

This section presents the results obtained during both phases of the project. Firstly, the results from the socio-demographic and audiovisual content consumption surveys performed by test participant are described. Later, results from emotional state survey in term of mood evaluated on different time lapses of the test will presented. Finally, the findings from both phases of the project are exposed and discussed.

Socio-Demographic Results

In the entire project, 48 test participants took part. 22 were men, and 26 women. The average age of the participants was 31.9 years old being the youngest participant 18 years old, and the oldest participant 51 years old with an standard deviation of 9.03 years. The average consumption of audiovisual content per service is described in the table 3.

Values Represented in Minutes			
TV	*Online-Media*	*Live-TV*	*Video On demand*
137,54	50,18	23,090	64,63

Table3: Participants Average Multimedia Consumption in Minutes

Emotional state results

A mayor concern in the development of the test was a possible negative effect in quality evaluation caused by the duration of the test. As a control measure, we performed an exploratory comparison between mean ratings across the different time periods. The results illustrated in figure 3 presents mean ratings for each time period on each test environment. At first glance, figure 3 shows that tests participants who took part the living room test have slightly higher mood that those who took part in the test in the laboratory, as well as, their rating behavior is similar in both cases. This insights reveals that the test duration have a slight effect on users mood. However, the differences between evaluations on all different periods of times and the environment is minor and can be consider negligible.

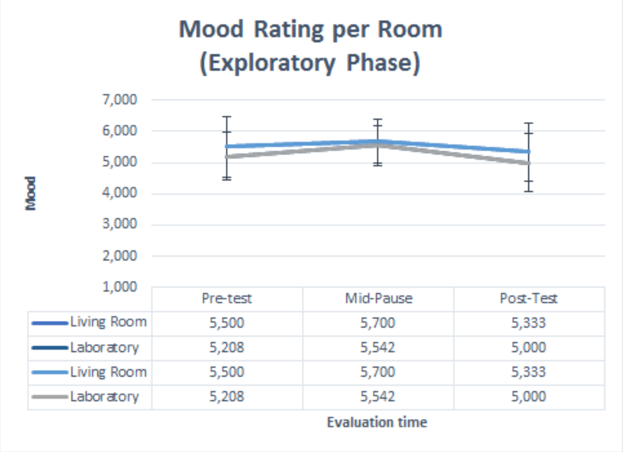


Figure 3. *Mood evaluation per group*

Exploratory Phase

During the exploratory phase, 48 test participants were evaluated. Unfortunately, data from 2 test participants had to be excluded due technical problems. In this phase, test participants were divided in 2 groups and each group performed the test in one of the environments. An exploratory analysis revealed that rating behaviour was slightly similar across conditions for both groups . Based on this observation, an ANOVA was performed to assess possible differences in perceived audiovisual quality influenced by the environment between the to groups. Ours findings presented in table 4 reveals that there is significant effect of the room on quality behavior when the sequences has no impairments. However, in cases where the sequences presents impairments, the effect of the room seems not significant.

The relation between acceptability and Quality rating was evaluated through a point-biserial correlation. The results presented in figure 5 revealed and a significant correlation between acceptability and Perceived audio visual quality in all cases.

Packet loss Percentage	**F**	*p*
0,000%	(1,7488)= 126,768	,000
0,625%	(1,7488)= 23,98	,117
1,250%	(1,7488)=12,53	,411

Table 4: ** values significant at p <.001

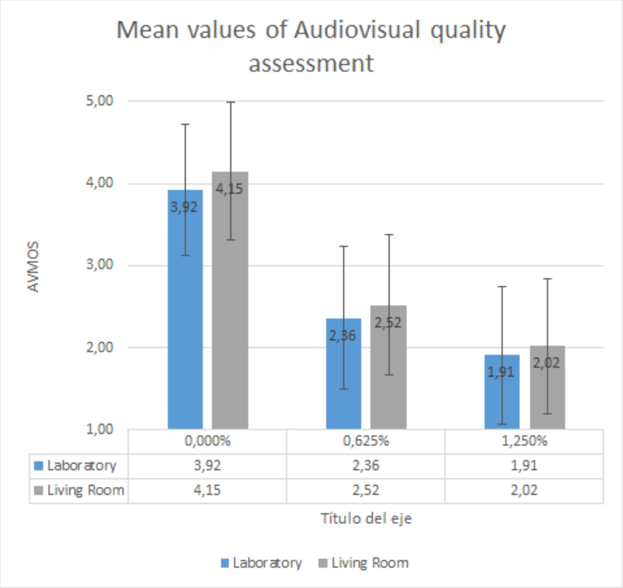


Figure 4. *Audiovisual Quality Assessment from the Exploratory Phase*

Correlation Acceptance-Audio Visual Quality				
			95% BcA CI	
	N	***r***	**Lower**	**Upper**
Laboratory				
0,000%	490	.415**	.316	.498
0,625%	490	.641**	,591	.692
1,250%	490	.640**	.581	.696
Living Room				
0,000%	490	.433**	.339	.519
0,625%	490	.602**	.554	.649
1,250%	490	.562**	.498	.621

Table5: ** values significant at p <.001

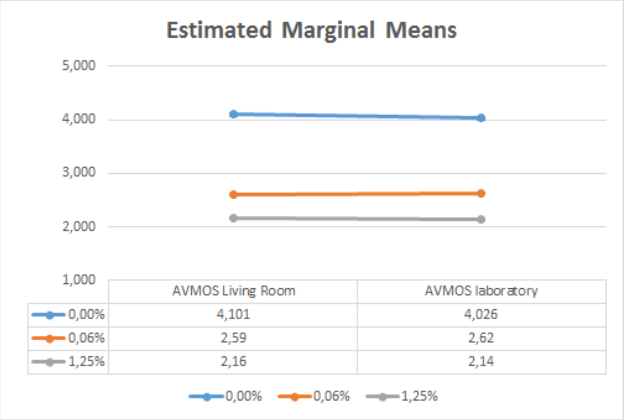
Paired Comparison test

In this phase we had the participation of 24 tests participants from the exploratory phase. This time 11 men, and 13 women participated. By repeating the test with the same participants, we strive to understand wether the test subjects evaluative behaviour remains constant when the visual appeal of the environment is manipulated. To analyse the data, we performed a repeated measure audiovisual quality test having the environment as the independent variable with two levels(laboratory and living room). Before performing the analysis, we split the data-set per quality condition. This approach allows us to have a better understanding of the effect of the environment, which could, otherwise, would be difficult to interpret on a three by two repeated measure ANOVA. Also, we support such approach based on the differences on rating behaviour by quality level observed during the exploratory phase.

Thus, the assumption of specificity is not taken into account for this test because there are only 2 repeated-measures conditions. The results of the study indicated a significant environment effect in cases were there is no visual impairment present on the sequences. However, in cases were visual impairments are present, test participants tended to reveal a constant rating behaviour independently of the environment. The results are presented in detail in . Follow up comparisons indicated that the pair wise difference was significant (p <.001). Thus the illustrated results in 5 illustrate that quality is perceived better in the living room when no errors are present, but in presence of impairments, the evaluative criteria from test participants becomes constant.

Packet loss Percentage	F	*p*
0,000%	(1,1679)= 7,979	,005
0,625%	(1,1679)= ,700	,403
1,250%	(1,1679)= 1,256	,263

Table : ** values significant at p <.005



	AVMOS Living Room	AVMOS laboratory
0,00%	4,101	4,026
0,06%	2,59	2,62
1,25%	2,16	2,14

Figure 5. *Repeated measured estimated marginal means*

The last analysis of our study explores the relation between acceptance and perceived quality. To this end, we conducted a point-biserial correlation as performed in the exploratory phase. The results are presented in 7, revealing a significant correlation in all cases.

Correlation Acceptance-Audio Visual Quality				
			95% BcA Cl	
	N	***r***	**Lower**	**Upper**
Laboratory				
0,000%	490	,518**	,472	,559
0,625%	490	,723**	,703	,743
1,250%	490	,703**	,678	,726
Living Room				
0,000%	1680	,521**	,469	,566
0,625%	1680	,718**	,699	,738
1,250%	1680	,680**	,653	,704

Table 7: ** values significant at p <.001

Conclusion and further work

This paper present two different approach towards a better understanding of the influence of the environment visual appeal on evaluative behaviour of audiovisual content. The results specially remark that in cases when the conditions of the service are optional, the environment presents a strong influencing factor on the audiovisual experience. However, changes in the quality level can greatly impact the user leaving aside the influence of the environment. In addition, we want to highlight that this test resembles a typical audiovisual quality test in which sequences have a duration of 15 seconds. As a consequence of the short duration of the sequences, the evaluative process could be consider more critical than in longer sequences as seen in [10].
We considered that the information provided in this work represents a relevant move towards the development of more ecologically valid test setups. Thus, or next steps will be to address the influence of environment over evaluative behavior using longer and more engaging sequences. More precisely, during our next project, we expect to extend the already used content duration to an adequate length in which the level of engagement could be measured based on the content of the evaluated sources.

Acknowledgement

This project was funded by German Bundesministeriums fr Bildung und Forschung (01IS12056) and performed within the project AcToBeTVEx from Software Campus. The responsibility for the content of this publication lies with the authors.

References

[1] Satu Jumisko-Pyykkö and Miska 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.

[2] P ITU-T RECOMMENDATION, "Subjective video quality assessment methods for multimedia applications," 1999.

[3] ITUR Rec, "Bt. 500-11,," *Methodology for the subjective assessment of the quality of television pictures*, vol. 22, pp. 25–34, 2002.

[4] Oliver Hohlfeld, Rüdiger Geib, and Gerhard Haßlinger, "Packet loss in real-time services: Markovian models generating qoe impairments," in *Quality of Service, 2008. IWQoS 2008. 16th International Workshop on*. IEEE, 2008, pp. 239–248.

[5] Raimund Schatz, Sebastian Egger, and Alexander Platzer, "Poor, good enough or even better? bridging the gap between acceptability and qoe of mobile broadband data services," in *Communications (ICC), 2011 IEEE International Conference on*. IEEE, 2011, pp. 1–6.

[6] Toon De Pessemier, Katrien De Moor, A Juan, Wout Joseph, Lieven De Marez, and Luc Martens, "Quantifying qoe of mobile video consumption in a real-life setting drawing on objective and subjective parameters," in *Broadband Multimedia Systems and Broadcasting (BMSB), 2011 IEEE International Symposium on*. IEEE, 2011, pp. 1–6.

[7] Nicolas Staelens, Stefaan Moens, Wendy Van den Broeck, Ilse Marien, Brecht Vermeulen, Peter Lambert, Rik Van de Walle, and Piet Demeester, "Assessing quality of experience of iptv and video

DOI: 10.2352/ISSN.2470-1173.2016.16HVEI-123

on demand services in real-life environments," *Broadcasting, IEEE Transactions on*, vol. 56, no. 4, pp. 458–466, 2010.

[8] M-N Garcia, Phillip List, Savvas Argyropoulos, David Lindegren, Martin Pettersson, Bernhard Feiten, Jonas Gustafsson, and Alexander Raake, "Parametric model for audiovisual quality assessment in iptv: Itu-t rec. p. 1201.2," in *Multimedia Signal Processing (MMSP), 2013 IEEE 15th International Workshop on*. IEEE, 2013, pp. 482–487.

[9] P ITU-T RECOMMENDATION, "Parametric non-intrusive bitstream assessment of video media streaming quality," 2012.

[10] P. Frohlich, S. Egger, R. Schatz, M. Muhlegger, K. Masuch, and B. Gardlo, "Qoe in 10 seconds: Are short video clip lengths sufficient for quality of experience assessment?," in *Quality of Multimedia Experience (QoMEX), 2012 Fourth International Workshop on*, July 2012, pp. 242–247.

Author Biography

Miguel Rios joined the T-labs task force in 2010 researcher assistant (Wissenschaftlicher Mitarbeiter). He graduated as an informatics systems engineer at the Universidad Santa Maria La Antigua of Panama in 2003. During this time he worked in the area of IT management and project management of online enterprise solutions. In 2008 he acquired his master in the area of Management of the IT department and security of all the informatics resources in the Laureate international university in Panama. His actual research in the Technical University of Berlin focus in the area of QoS, QoE and business models.