Design of a large-scale subjective test in the cinema

Katriina Kilpi, Wendy Van den Broeck, An Jacobs
iMinds-SMIT, Vrije Universiteit Brussel
Pleinlaan 2, 1050 Brussels, Belgium katriina.kiilpi@vub.ac.be; wvdbroec@vub.ac.be; an.jacobs@vub.ac.be

Abstract— **This paper discusses the design of a subjective quality test on speckle perception. The test was conducted in a movie theatre with 187 participants. The paper will discuss the goal and set-up of the test, practical issues on the selection profile and recruitment of respondents as well as the proceeding and the challenges of the actual experiment. We conclude with some reflections on the applied set-up and some first results.**

Keywords— subjective tests in commercial settings, quality of experience, real-life setting

I. INTRODUCTION

This paper reports on a large-scale subjective Quality of Experience experiment that was conducted in a movie theatre with 187 respondents. Quality of Experience can be defined as "the overall acceptability of an application or service, as perceived subjectively by the end-user". According to ITU (International Telecommunication Union), this "overall acceptability may be influenced by user expectations and context [1]. The aim of the subjective experiment was to investigate the user perception of laser illuminated projection in a real-life setting. Laser illuminated projection is the latest development in cinema projection, therefore our QoE experiment was designed to take place in a natural cinema setting with actual cinemagoers. In this paper we reflect on the design and planning process of the large-scale subjective test in the cinema, which called for careful planning. A laser projector prototype as developed by technology company and digital cinema projector provider Barco was placed in a cinema in the city of Antwerp (Belgium). The main goals of the test were: 1) to evaluate the general image quality of the laser projection, 2) to investigate the perception and evaluation of 'speckle' in the image after familiarizing the people with the concept of speckle and 3) to compare the general quality and color equality of a lamp and a laser projector in a real-life setting. Speckle can occur in laser projection as an observable granular pattern on the projection screen, thus masking the displayed content [2]. The cinema audience was familiarised with speckle by displaying a picture with and without speckle noise. Prior to this test in the cinema, a lab test was conducted. In this test, both objective and subjective measurements of color quality and speckle in still images were conducted, using a butterfly test set-up, showing two images simultaneously [2].

Subjective video quality tests usually take place in a standardized and controlled lab setting. This implies following strict guidelines on the specific setting, the number of impairments and the specific length of the tests (e.g. ITU guidelines BT.500 on the subjective assessment of the quality of television pictures [3] and ITU guidelines P.911 on the subjective audiovisual quality evaluation of multimedia services [4]). Previous experiments on video quality for IPTV streaming services have, however, taught us that there is a clear difference in Quality of Experience (QoE) of video quality depending on the focus of the viewer. Staelens et al (2010) conducted subjective experiments to measure QoE in Internet Protocol Television (IPTV) and Video on Demand (VoD) offerings by employing a subjective quality assessment methodology based on full length movies in a real-life setting (under the same conditions where users watch television) [5]. Their research showed that when people actively evaluate short video sequences in a lab, they are specifically focusing on the perceived quality. In contrast, when they watch video content in the natural setting of their living room, their primary focus is on the actual content (e.g. the plot or the dialogue in the movie). In addition, their quality expectations and appreciation is different than the QoE measured in a lab setting [5]. Borowiak et al. [6] also suggest new methodology of quality evaluation of long duration audiovisual content using a slider knob. This knob could be used by the research subjects for the continuous self-adjustment of the displayed quality instead of giving quality scores. Besides a tendency to develop methods for the evaluation of longer video sequences, there is also a tendency to take subjective experiments out of the lab. Recently, more and more research is using or mimicking the actual setting of the use case when conducting QoE subjective experiments [7, 8, 9]. Previous tests in cinema setting were conducted by Rahayu in Norway [9, 10]. In these tests, the digital cinema setting was used, but only a limited number of 5 participants was conducting the test at the same time, all seated at the same distance from the screen. The used data set for quality evaluation consisted of short fragments (10 seconds) as described in the ITU guidelines. These tests indicated for example that content type significantly influences the perceived visual quality of both image and motion picture quality [9, 10].

In our QoE experiment, the main aim was to test the speckle perception and acceptance of cinemagoers, while consuming natural cinema content in an actual cinema setting. Therefore, we designed an experiment in which we used real cinema content (trailers and a full movie) in a cinema with multiple subjects seated in different positions. The experiment was conducted only one time. In the experiment we used a combination of conscious testing (trailers and butterfly testing) and blind testing (full movie). In the next paragraphs we will discuss the design of the experiment and the specific choices we made.

978-3-901882-50-0

II. SETTING UP A LARGE-SCALE EXPERIMENT

A. Designing the subjective experiment: test set-up

The experiment was a subjective user test in the natural context of the cinema with a large group (aimed at N=200) of users with different profiles (i.e. socio-demographics, frequency of movie visits, visual acuity). The respondents were asked to participate in a Quality of Experience test of a new projector. As a compensation for participating in the test the participants were offered the chance to watch a recent full movie immediately after the test for free, including some free soft drinks and popcorn during the break.

Figure 1. Subjective experiment in the movie-theatre

We opted for a digital voting device on which the participants could cast their vote on the QoE questions to make sure that everyone answered each question individually. The voting system illustrated in real time which devices and seat numbers had completed the vote, allowing us to control immediately that we received everyone's vote.

The first part of the test consisted of the speckle and image quality test. At the beginning of the test, we introduced the concept of speckle to the participants by showing them two images where speckle was clearly visible. By introducing the phenomenon of speckle, we wanted to make sure that the participants were clear on what they were supposed to pay attention to as they judged the visibility and tolerance in the video sequences. Next some general questions (both multiple choice and yes/no questions) were posed in order to make the participants acquainted with the voting devices before the actual rating of the video sequences started. Also at the start of the test a question on color blindness was added (are you colorblind: yes/no/I don't know) as was the Ishihara Test, i.e. a test that identifies not only colorblind participants, but also those with a vision defect (partial color blindness).

Secondly, there was a quality evaluation of short video fragments (average of 1.30 minutes). For this test we used six movie trailers with different characteristics (animation movies vs. regular movies, dark vs. bright colored trailers and fast vs. slow action scenes). The choice of showing movie trailers for the experiment fits within the aim of mimicking a real-life user experience in the cinema as also here trailers of upcoming movies are shown. Therefore we started with a full version of the trailer which was shown in a randomized order of filter settings. Two different settings of the laser projector were to be tested: the Neutral Density filter and the Polarizing filter. After watching the full trailer, respondents were asked about their appreciation of the content (i.e. would they go and see this movie based on the trailer?). Afterwards they gave a rating (on a 5 point scale) of their appreciation of the general picture quality, the perception of speckle and its perceived annoyance. Then 20 seconds of each trailer were shown in a double stimulus experiment (the same fragment of 20 seconds in each filter setting directly after each other in random order) and participants were asked to select the picture quality they liked the best ("I preferred the 1st clip/2nd clip/I don't know').

A third part of the QoE experiment was a butterfly test with 11 still images displaying different images (e.g. dark vs. bright landscapes, images varying in detail), comparing the lamp and the laser projector. The same image was projected simultaneously side by side with both the lamp and the laser projector. Here respondents were asked to give a quality score (1-10) for both sides. Finally, they were also asked to indicate how they perceived the color equality of the left and right pictures (completely different/a bit different/the same).

In the fourth part of the test (the full movie), our aim was to investigate the difference in speckle tolerance when people have a different primary focus, i.e. focusing on image quality versus focusing on the content. In this part the focus was on content, by performing a blind test with the full movie. To make it a blind test, at the end of the butterfly test the participants were thanked for their participation and asked to stay seated as the full movie would start immediately. The full movie was portrayed as a reward for taking part in the tests. This made the participants believe the test was already over. The whole movie was projected on the laser projector. The movie started with one filter setting and some 30 minutes into the movie, we had a break of 15 minutes during which the other filter was changed. While the filters were switched without the participants knowing, the participants could take a bathroom break and get some refreshments that were offered to them as a reward.

When the movie was over, the participants were asked to stay seated a moment longer and vote on a few more questions (such as whether they liked the movie and whether they noticed any speckle).

In the final and fifth part of the test, a selected group of respondents (N = 32) participated in short focus group sessions of 20 minutes. In these sessions we discussed their perception of the speckle in the short trailers and in the movie. To help them remember the content they had evaluated, we displayed images of the trailers and butterfly test images. We also discussed other contextual factors that influence their QoE when watching a movie in the movie theatre (e.g. the most annoying things about watching a movie in the cinema, the best things about going to the cinema etc.).

III. PRACTICAL CHALLENGES IN ORGANISING THE EXPERIMENT: PRIOR TO THE EXPERIMENT

Designing a subjective experiment of this scale in a real-life setting calls for many practical preparations. In the following sections we will reflect on the different practical challenges of organizing the experiment as we walk the reader through the steps from planning to some first results that were gained.

A. Recruiting the respondents

As mentioned, we aimed at attracting a variety of cinemagoers; both frequent goers as well as non-frequent goers in combination with a variation in gender, age and visual acuity (wearing glasses, colour blindness). In order to mimic the natural setting of watching a movie in the cinema, people were encouraged to come together with friends or family, as this is usually the natural context of the movie going experience. For the recruitment, we made use of an informational webpage which was linked to an online subscription form with 24 questions on socio-demographic profile, cinema going behavior and visual impairments. Participants could also mention the names of the other people they were coming with, but each of them also had to complete the registration individually. This information was used in multiple ways: as useful contextual information; as variables in the posttest analysis of QoE; and to gain insight into the demographics of the participants as well as their seating preferences. To recruit the participants, we made use of multiple channels: mailing list of the cinema in which the experiment took place (Kinepolis) addressing daily, weekly and monthly cinemagoers; the Kinepolis website; the participating university's student mailing list; posts via the social network profiles of colleagues, family and friends; and by handing out flyers in railway stations. Most effective way of recruiting turned out to be was the snowball method, since after we had gained the attention of one participant (mostly via the movie theater mailing list), they recruited the company in which they wanted to participate in the experiment and consequently watch the full movie.

To combine the quantitative subjective measurement with more in depth qualitative data about the motivations and the overall experience, we planned to organize 4 focus groups immediately after the test. To recruit people for the focus groups, we sent the invitation to participate in a focus group to all registered participants. Rather than narrowing down the specific inclusion profile of the focus group in advance, we decided to create the pool of candidates from those who positively replied to the invitation, and decide on the criteria for each focus group afterwards. This way we avoided the risk of having to convince non-interested people to take part in focus groups and overload them with mailing, thus creating a potential drop in the overall test participation beforehand.

A total of 50 people indicated being interested in the focus groups. The four groups of 8 people each were chosen based on the following criteria: a) people sitting in the front, b) people sitting in the middle, c) people sitting in the back and d) most frequent cinemagoers (weekly, as nobody indicated to be a daily cinemagoer). These categories were selected based on the previous butterfly lab tests which suggested that there could be a difference in quality perception according to seating position.

B. Timing

The tests took place at the end of October 2012. We planned the test deliberately during the autumn break, to attract more people to take part in the test. The test took place in the morning in the middle of the week from 9:30 to 13:00, since this was the only occasion that was available for a movie theater room of that size.

The recruitment started in the first week of October, leaving us with one month to recruit the aimed 200 participants. To be able to keep people interested and motivated in the test they had signed up for, it was important to have the opening of the registration and the actual experiment placed relatively close to each other.

C. Seating in the movie theatre

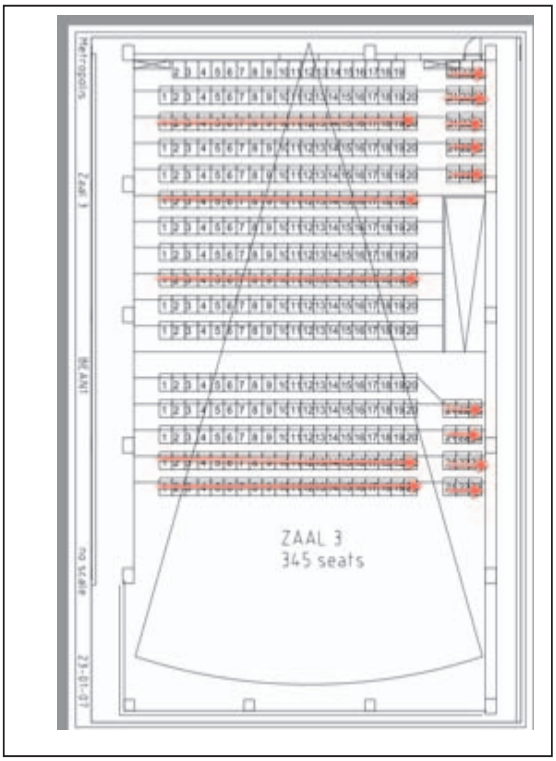

Figure 2. The movie theater room used in the subjective test

The room, in Kinepolis Antwerp, in which the experiment took place, has the capacity to seat 354 people. As we aimed to recruit only 200 participants, we first wanted to define where these participants would be seated. We chose to seat the participants equally in the front, middle and back on all together 11 rows (see Figure 2). This specific way of seating (front-middle-back) was also used in the analysis of the results of the QoE test, in order to investigate if there was a significant difference in speckle perception according to the seating position.

As part of aiming at an authentic movie theater experience, we paid attention to the participant seating preferences as well as allowing participants to sit next to the company of their choice.

1) Organising seating according to participant preferences

When signing up for the experiment online, the participants had been asked to indicate where they usually like to sit in a movie theater. Although we attempted to seat all participants according to their indicated preferences (e.g. those who

preferred sitting in the back, got a seat in the back), we had to move some participants to other seats, as, for example, rather few (N= 2) indicated their preference to be in the front. We chose to give a seat for those with a preference for the middle, in the front, as the difference of sitting in the middle is smaller than seating those with a preference for the back to the front. People did indicate to have a clear preference for the middle (N=97) and the back (N=83) and only a few were indifferent about the location of their seat (N= 4).

2) Organising seating next to friends or family

As going to the movies is a social occasion for most people, we made sure to allow participants to sit next to their friends. To accommodate this and to seat 200 people in a meaningful way (i.e. to cover the chosen regions of the room), we asked people to indicate the names of people they were coming with as they were filling in the survey. As soon as all of the people in the indicated group had signed up, we were able to assign a seat for them on our seating plan.

3) Registration & seating arrangement planning

As the registration list was changing until the last minute (people dropping out and additional people joining existing groups), it became a constant labour intensive administrative effort to maintain and update the registration list. The list was organized on an excel file downloaded from the online survey software. The list included the participants who had registered until then and their indicated friends (registered/non-registered). This file was continually updated by the status of the participants (confirmed/cancelled/cannot be contacted etc.), and by adding more participants along with their registration survey data. At this point, we had to move fast and contact some participants individually by phone or by email to clarify whether or not they and their group were coming in the end.

D. Participant ID's

Once all parties in the indicated group had registered, we sent out the participation ID numbers in the confirmation mail. This number consisted of the row and seat number. Closer to the actual date of the event, we had to start allocating seats to all participants, even the groups that were not yet complete. The respondents who listed names of friends that had not registered at this time, were asked to encourage their friends to complete the registration process.

When the participant list was almost complete, i.e. we had provided the participant ID to most people by email, we were able to deliver the remote voting system provider the name and seating number of the participants. The company in charge of the voting system then produced a list where each participant's ID was paired with a number of a voting unit. Each unit had a number on its backside, which enabled easy checking from the list of unit and seats. The synchronization of the voting unit list and the seating plan was done only a week before the test, when most seats had been filled. This synchronization was important, as it allowed us to couple the voting data to the seating and to the specific profile of the participants.

IV. PRACTICAL CHALLENGES IN ORGANISING THE EXPERIMENT: DURING THE EXPERIMENT

On the day of the experiment, we had a team of seven researchers present. Four coworkers were active at the registration, while the other three were active in the cinema room to prepare it for the test. Participants were greeted at the entrance of the cinema in four different registration lines. We had printed out the name and ID lists of the participants and the researchers at the entrance ticked off the names of the participants as they were located on the list. The participants were then given a name badge along with their ID number as well as the voting system number. The participants who had been selected for the focus groups, were also marked separately (color coding) so that calling them out after the test would be easier.

The participants who had not registered but came in the company of others, were asked to register at the front desk. The registration survey had been printed out for the unregistered participants to complete. In the end, five people had to be added to the participant list afterwards. When participants had received their badges, they were offered a light breakfast just before moving into the cinema room. The participants were seated according to their ID's. As the room included a number of empty seats and rows, these were indicated with notes ("do not sit").

When all participants had found their places, the participants were welcomed and given some instructions on the big screen. They were reminded of the importance of staying seated in their original seats, even after the break, and asked to take good care of the voting devices. In addition, they were reminded about using only their personal voting device and not to switch it with a fellow participant's voting system. As the cinema room had more capacity than the number of recruited participants, the challenge was to keep the participants seated in the seats they were assigned to. It was important for them to hold on to their voting units that were also coupled with the seat numbers and seat locations. Therefore, we also started with the full movie immediately after the test sequences and had a break only when the movie had been running for some 30 minutes to increase the chances of people returning to their original seats.

A. Substitutions/cancellations

We had anticipated changes in the actual numbers of those participants who had signed up and those finally taking part in the experiment. Therefore, we planned to be flexible and prepared handwritten badges for the non-registered people and assigned extra seats at the registration counter. When it came to cancellations, most participants (N=23) let us know in advance either by email or by phone - some even on the day itself. People were quite conscientious about cancellations: on the day of the test, two participants called to check whether they could still take part in the test, although they would be running late.

We also made sure to have a plan B. This in case registered participants did not make it and on the other hand, if non-registered participants would arrive (e.g. as part of a group where a friend who had registered had cancelled). We therefore used the voting units of those registered participants who did

not arrive to the new non-registered participants. Since cancellation at the last minute was possible, it would be easier to keep moving the assigned voting units and keep track of this rather than change seats, as people were fixed to their ID's (row + seating number). In the end, 243 people had registered. A total of 62 who had registered, did not make it after all. Of these, 23 cancelled beforehand. At the day of the test, 6 new people registered. This meant a total of 187 people who took part in the test. Due to the flexibility we had planned in, it was possible to maximize the number of people using the voting system and participating in a uniform way in the test.

B. Testing the voting system

Even though we had run a test of the voting system a day before, our plan B also included a paper copy of all questions, in case of unexpected malfunctioning of the voting system.

Before the actual experiment started, the participants were asked to vote on a few test questions (how did you come here?/ do you have a TV in your bedroom?). This enabled us to check immediately whether everyone was technically able to vote, and allow participants to get used to the different response alternatives (yes/no, scoring on a scale of 1-10) and to perform a correct vote during the test.

V. FIRST RESULTS

In the analysis of the results, there were some clear differences between the conscious tests with the trailers and the blind test with the full movie. In the end 88.6% of the respondents did not note any difference in picture quality in the movie before or after the break (when filters were switched) and as such did not have a clear preference. In the trailers people usually did have a clear preference for one of the filters, depending on the fragment and the specific genre. For example in one of the animation movies, 28% did not have a clear preference for a specific filter while in the darker movie trailers, up to 65% did not have a clear preference for a filter.

The mean opinion score on general quality evaluation in the movie was 4.32 on the 5 point scale. 69.4% of respondents indicated the speckle as being imperceptible while 29% indicated it as being perceptible but not annoying. In the end, those who did not see speckle, rated the movie a bit higher, while those who saw much or very little speckle, rated it a little lower (p =.002). The MOS score for speckle annoyance for the full movie was 4.67.

In the trailer tests where people focused on the quality evaluation, MOS scores for the regular (non-animated) movies varied between 3.66 and 4.42. The location of the participant had an effect on speckle detection: in three fragments (2 with dark content, one with light and grey scenes), those sitting in the front noticed more speckle compared to those in the back (fragment 1p =.002, fragment 2 p =.011, fragment 3 p =.034 respectively). This indicates that the seating position in the movie theatre does influence the perceived Quality of Experience.

VI. DISCUSSION

The subjective experiment took manpower and effort (time and cost) and there was a high risk of failure (we could only run the test once). However, we do evaluate the subjective experiment in a positive way. We were able to retrieve QoE subjective measurements of a large number of respondents (N=187) and could link this data to the contextual information of the respondents (e.g. socio demographics, visual acuity, seating, likes and dislikes about going to the movie theatre). The combination of a conscious QoE test in which people focused on quality evaluation and the blind test with the full movie also provided us with interesting points of comparison between the results.

The most challenging and work intensive part of the experiment was coupling each participant's profile to a seat and a voting device upfront. This was complicated by multiple people that needed to be seated next to each other, however, of whom we had no certainty that they would register on time or show up at all. Opting for free choice of seating would have lead to the voting results not being coupled to any specific seat, therefore leaving one important parameter, namely, the participant's location in the room out of the experiment. The communication with respondents and the follow-up of the registration procedure took a lot of effort, which should be taken into account when designing such a large-scale subjective experiment. We made sure to underline the importance of registering on time and even more so, the importance of alerting the organizers in case there would be changes in the registered data (e.g. change of accompanying persons or not being able to participate). We had foreseen an additional 20% of respondents in the recruitment (N=243), but in the end, only 187 respondents showed up. In future experiments, the need to recruit up to 40-50% additional respondents should be considered.

The administrative task of keeping the registered participants' list up to date was challenging with the tools we used. A more suitable tool was not readily available, which is why the data was collected into an excel sheet. The list of registered participants had to be manually updated every so often. The file was alive and being constantly updated so that it included the most accurate information. It was therefore easier to assign the handling of the registrations to a single person as delegating the task at this point of the experiment planning would have been more time consuming as it would have called for reorganizing the registration system so that it is clear to more people than one. This naturally also posed a great risk to the experiment, in the case that the person in charge of the registrations would have suddenly not been able to fulfill the task. Therefore, a more suitable system would make this task easier and open up its handling to more people in the research team.

During the experiment a large team was needed to limit the risk of unforeseen behavior of the respondents. Though we had anticipated this behavior (unregistered respondents showing up, people switching seats and people in need of help with their voting device), we were prepared with a sizable group who knew the risk factors to the experiement. As a

result, we were able to swiftly cope with these practical issues on the day itself.

The participants were subjected to a large amount of data they needed to evaluate. The test took longer than 30 minutes, which meant that the participants were tired and had possibly already lost their focus by the end of the experiment. This could have had an impact on the reliability of the results. Since we tried to embed test in the natural setting of cinema going, this may have been better tolerated (reward of the full movie). Trying to limit the test time in future experiments is however advisable. The most suitable part to limit test time is in the number of fragments or omitting the butterfly test. Also the learning curve of answering the different types of question categories could be limited by using less variation in answering formats. One should be aware of too much repetition in answering categories, as it may lead to less valid answers due to automatic responding.

ACKNOWLEDGMENT

The research described in this paper was funded by the Agency for Innovation by Science and Technology in Flanders (IWT) in the framework of an R&D project BRAVO (Barco Platform for Laser Video Projection) in cooperation with Barco N.V. The authors wish to thank Barco N.V. and Kinepolis for the support in the set-up and the proceeding of the subjective experiment and Nicolas Staelens and Shirley Elprama for their advice.

REFERENCES

[1] ITU-T Recommendation P.10/G.100 Amd 2, ““Vocabulary for performance and quality of service,”” International Telecommunication Union (ITU), 2008.

[2] Roelandt,S.; Meuret, Y.; Craggs, G.; Verschaffelt, G.; Janssens, P.; Thienpont, H. “Standardized speckle measurement method matched to human speckle perception in laser projection systems”. Optics Express, vol. 20, issue 8, pp. 8770-8783, 2012.

[3] ITU-R Recommendation BT. 500-11, ““Methodology for the subjective assessment of the quality of television pictures””, Geneva, Switzerland, 2002 (www.itu.org)

[4] ITU-T Recommendation P.911, “Subjective audiovisual quality assessment methods for multimedia applications,” International Telecommunication Union, Geneva, Switzerland, 1998.

[5] Staelens, N., Moens, S., Van den Broeck, W., Mariën, I., Vermeulen, B., Lambert, P., Van de Walle, R., De Meester, P., Quality of experience of IPTV and Video on Demand services in real- life environments, IEEE Transactions on Broadcasting, 2010.

[6] 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.

[7] Van den Broeck, W., Jacobs, A. & Staelens, N., Integrating the everyday-life context in subjective video quality experiments., workshop on Quality Metrics for Consumer Electronics (Qomex), 2012.

[8] Strohmeier, D., Jumisko-Pyykko !, S., Eulenberg, K., Open profiling of Quality: probing the method in the context of use, Workshop on Quality of Multimedia Experience, 2011

[9] Rahayu, F. N., Reiter, U., You, J., Perkis, A., & Ebrahimi, T. (2011). Subjective visual quality assessment in the presence of audio for Digital Cinema. 2011 Third International Workshop on Quality of Multimedia Experience. IEEE.

[10] Rahauy, F.N, Reiter, U., Ebrahimi, T. and Perkis, A. A study of quality of experience in D-cinema. Submitted to signal processing: image communicaion, Theory, techniques & appliactons, a publicatons of the european assosication for signal processing (EURASIP), ISSN: 0923-5965, 2011