Evaluating QoE by means of traditional and alternative subjective measures: an exploratory 'living room lab' study on IPTV

Katrien De Moor¹,², Miguel Rios Quintero³, Dominik Strohmeier³, Alexander Raake³

¹ITEM-Norwegian University of Science and Technology, Trondheim, Norway
²iMinds-MICT-Ghent University, Gent, Belgium
³AIPA-Telekom Innovation Laboratories /Technical University Berlin, Berlin, Germany

Abstract

In this paper, we explore the use of a set of potential ‘alternative’, emotional state and engagement-related measures of QoE and investigate how they relate to traditional QoE measures. To this end, we present results from a living room lab study (N=28) in which the impact of slicing errors on QoE was investigated. Our findings indicate significant differences in QoE between the three used error profiles, both when considering the traditional and alternative measures. Furthermore, we found that the link between the traditional and alternative measures is rather weak, indicating that the former may need to be reconsidered and extended with alternative measures, which allow to measure QoE in terms of ‘delight’ or ‘annoyance’.

Index Terms: QoE, emotional state, user engagement, delight, IPTV, living room lab

1. Introduction

Internet video is increasingly dominating consumer Internet traffic and the use of Internet Protocol Television (IPTV) and other real-time streaming services is ever growing [1]. In the context of such services, video content is streamed to the user at the receiving end and its play out takes place in parallel to the transport of the remaining parts of the video stream. As a result, real-time streaming services entail important challenges in terms of Quality of Experience (QoE). Previous studies have focused on the influence of network performance parameters (e.g., latency, loss distance, ...) in this respect. It is however still poorly understood how they influence QoE from a real user perspective and which other factors might play an important role.

A number of previous studies pointed to the (possible) influence of a range of non-technical factors on QoE in the context of IPTV services. In [2] it was shown that characteristics of the viewing context (i.e., natural setting vs. controlled lab) influence impairment visibility and tolerance. Other work investigated the impact of video quality on viewing-related metrics of user engagement [3]. Research on adoption and customer acceptance of IPTV on the other hand, has pointed to aspects such as enjoyment, expectations, goals and motivations [4]. In QoE-research in the context of IPTV services however, such non-technical aspects are usually not, or only to a limited extent, taken into account. This observation has been one of the motivations underlying the work presented in this paper.

A second motivation lies in the discrepancy between theory and practice in terms of how QoE itself is defined and evaluated. In 2012, a new broadly supported definition was proposed, defining QoE as 'the degree of delight or annoyance of the user of an application or service. It results from the fulfillment of his or her expectations with respect to the utility and / or enjoyment of the application or service in the light of the user's personality and current state' [5]. In this sense, the concept of delight plays a prominent role. This definition represents an important step to go beyond the instrumental and utilitarian approach to QoE, i.e., to go beyond the goal of merely providing user satisfaction by also explicitly including positive emotions such as enjoyment (and hence, the fulfillment of hedonic needs) as a desired outcome. At the same time, it has crucial implications for the way in which QoE is evaluated in practice today. Most constructs embedded in the new definition are not at all or insufficiently taken into account in the current, dominant practices. QoE is traditionally not evaluated in terms of emotional states such as delight or annoyance, but rather in terms of user satisfaction in relation to different quality levels. Moreover, although the QoE framework presented in [5] makes reference to influencing factors at the human, system and context level, standardized QoE assessment approaches do not allow to take these factors into account. The work presented here is therefore also motivated by the observation that there is a need to explore additional, alternative measures of QoE, to investigate how they relate to traditional QoE measures and to reflect on how the existing evaluation and measurement approaches could/should be adapted in order to put the more holistic conceptual understanding of QoE [5] into practice.

This paper therefore aims to make a contribution in two ways. First of all, by investigating whether and how three different gradations of slicing - a common, packet-loss based error in the context of IPTV services - influence QoE. Secondly, by exploring the relevance of possible ‘alternative’, emotional state-and engagement-related measures of QoE and their relation to traditional quality evaluations. As the emphasis mainly lies on this second objective, it goes beyond the scope of this paper to give an extensive overview of related work on the impact of packet loss-based impairments on video quality perception.

In this work, we present results from a study in which 28 test subjects evaluated 10-minute movie sequences in terms of QoE. QoE was measured by means of traditional and alternative subjective measures. Since a preceding survey study on IPTV viewing experiences (N= 284 IPTV viewers) indicated that the living room is the most common place in the home environment where the TV set is located (97.2%), our test took place in a closer-to-realistic environment, i.e., a living room lab. The rest of the paper is organized as follows: in the next section we briefly link up to a number of related studies pointing to the
importance of non-technical factors in the context of IPTV-related services and take a first step to explore the literature on user engagement and delight. Section 3 describes the test setup in more detail. Section 4 discusses the analysis and results and is followed by a number of conclusions and suggestions for future research in Section 5.

2. Prior research and relevant concepts

In the introduction, we already referred to the study of Staelens et al. [2] which highlights the importance of real-life QoE assessment (i.e., longer duration content, natural viewing conditions) in the context of IPTV and Video on Demand (VoD) services. They conducted a series of subjective QoE assessments both in a lab setting (standardized approach) and in a real-life environment (using a new test methodology based on full length movies). The results indicated a significant difference between both test settings when it comes to the visibility of impairments and test subjects’ tolerance levels. Moreover, the results pointed to the importance of the ‘flow experience’ of movie content, which has an impact on how (un)acceptable playout interruptions are. In this respect, Staelens et al. found that there is a preference for solutions that keep the flow of the content and the flow experience itself intact [2]. These findings fueled the discussion on ecological validity in/of standardized QoE testing. Moreover, this study pointed to the influence of human influencing factors such as users’ expectations, next to a set of system and context influencing factors.

Other related studies on video quality and IPTV have looked at specific human factors, such as user engagement. Dobrian et al. [3] investigated whether video quality of live and VoD services (both short and longer duration content) has an impact on user engagement when considering different types of video content. User engagement was quantitatively measured: playtime (i.e., duration of a viewing session) was used as metric of engagement at the view level; number of views and total playtime across all videos were the included engagement metrics at the viewer level. The results pointed to the relation between specific quality metrics and user engagement at both levels. At the view level, buffering ratio was found to be a crucial metric, both for long and short VoD content. For live content, the average bitrate turned out to be important as well, next to buffering ratio. Moreover, the results indicated that ‘join time’ is likely to influence the customer retention, yet when compared to the other investigated metrics, the impact on view-level engagement was lower. When considering engagement at viewer level, it was found that join time is very important. In addition, the investigated quality metrics had an influence on both the number of views and the total play time. Finally, Dobrian et al. found that the most important metrics on view level engagement also impact engagement at the viewer level [3].

The concept of user engagement itself has also been extensively discussed in the User Experience (UX) literature. It plays a prominent role in UX research, in which the promotion of ‘rich, engaging interactions between users and systems’ is of crucial importance [7]. Engagement has been defined in different ways (for a thorough and critical literature review, see [6]). There is however a consensus that it has several components, at the behavioral (as incorporated in [3]), affective and cognitive level. Moreover, it is understood as a multifaceted, key quality of users’ experiences, which is characterized by a range of attributes [7]. In [6], an interesting process-based framework of engagement (including stages such as dis- and re-engagement) was proposed. Moreover, a set of attributes of engagement (including e.g., focused attention, affect, motivation, …) have been proposed [6, 7] and linked to motivations in recent UX studies, such as [7]. Based on this related work from the UX-field and since QoE also is and aims to be about people interacting and engaging with applications and services, we believe that user engagement is an important concept to explore more profoundly and to integrate in QoE research. In our study, we also wanted to get an indication of user engagement and two attributes were therefore included as possible alternative measures (see 3.3).

In the light of the new definition of QoE as a ‘degree of delight or annoyance’, we are also particularly interested in the concept of delight. Although it goes beyond the scope of this paper to provide an extensive overview of the literature in this respect, we briefly point to a number of interesting contributions that might stimulate the debate on the delight concept and its implications for QoE research. Over the last decades, the notion of ‘delight’ has received growing attention, both from practitioner-oriented as well as academic perspectives. As a result, a lot of interesting research on delight, its foundations and characteristics can be found in the literature [8-12]. In fields focusing on e.g., service quality, customer behavior/loyalty, marketing, it is usually addressed by the concept of ‘customer delight’ as. Oliver et al. [10] argue that - from a managerial perspective - the underlying interest in providing customer delight lies in the possible behavioral outcomes and consequences (see e.g. [13]) and the potential for loyalty-driven profit.

In [11], a distinction is made between two views on customer delight, namely the ‘zone of delight’ view (i.e., customer delight is seen as a specific ‘zone of nonlinearity’ in the relation between customers satisfaction levels and their loyalty) and the ‘direct response’-view of customer delight. In the latter, delight is seen as an ‘emotional response’ to a provided value or service performance that was not as anticipated or expected and positively surprising [9]. In [10], different theoretical foundations of delight are discussed (such as the disconfirmation theory, the satisfaction paradigm and theories of affect and emotion). In this paper, we approach delight from the latter, as a positive emotional response. This perspective of delight as a pleasant surprise has been strongly influenced by the work of Plutchik [14], who considered delight to be a secondary dyad, an outcome of the interaction between two basic emotions, namely joy and surprise. Measures for both emotions were therefore incorporated in our study (see 3.3).

3. Test setup

3.1. Sample description

28 people participated in the study (average age 28.5, Standard Deviation (S.D.)= 5.75). The youngest participant was 22, the oldest 45. The majority of the test subjects (60.7%) were female (versus 39.3% male). All test subjects had normal or corrected-to-normal vision and reported normal hearing. More than two out of three participants (71.4%) indicated to watch TV at least several times per week and on an average weekday, 44.4% watches at least one hour. In terms of location and social context, 85.7% watches at home. Half of the participants usually watches alone, one out of three usually watches with others.
3.2. Test procedure
First, the test subject was welcomed by the test leader and given a short introduction. Next, every participant was tested for color vision deficiency (Ishihara’s test) and visual acuity (Wenzel’s plate). Thereupon, he or she was asked to fill in a short pre-test questionnaire aimed at collecting some basic socio-demographical and additional contextual information about the natural TV viewing behavior. The participant was then asked to choose between two content options and guided to the living room lab. During the second part, each test subject was shown three video sequences of 10 minutes (extracts from the content chosen in part 1), which were followed by a questionnaire containing a set of subjective measures (see 3.3). After the test, every participant received a gift voucher as incentive.

3.3. Subjective measures
After every sequence, participants were asked rate the overall quality of the video (five-point Absolute Category Rating scale, as specified in ITU-T Rec. P.910) and its acceptability. The main reason for not using a continuous quality evaluation technique, is that we wanted to avoid biasing the participants to a pure quality evaluation task during watching, rather than evaluating the experience as such and related affective states. For acceptability, a binary scale was used and the participants were asked to indicate whether they found the overall quality of the video acceptable (operationalized as: ‘good or good enough to watch the video’) or not.

A set of alternative subjective measures were also included. More concretely, we used the Self-Assessment Manikin (9-point scales), a pictorial self-assessment technique for assessing specific emotions (Pleasure – Arousal – Dominance) in relation to a stimulus [15]. We also included three constructs from the Differential Emotions Scale (DES) developed by Izard [16]. DES is a self-report scale which considers discrete categories of emotions to measure subjective emotional experiences. The underlying assumption is that fundamental emotions can be differentiated and distinguished from each other by self-report. We included the constructs Interest, Joy/Enjoyment (containing a ‘delight’-item) and Surprise, all measured based on 3 items, using a 5-point scale ranging from 0 (‘not at all’) to 4 (‘extremely’). Internal consistency reliability was checked (Cronbach’s alpha, threshold of .70) and new variables were computed for Enjoyment (α= .85), Surprise (α= .80) and Interest (one item had to be eliminated, Spearman ρ= .63). In addition, two adapted multiple item constructs from the self-report User Engagement-scale developed by O’Brien and Toms [7] were included: Focused attention (5 items, α= .85) and Felt involvement (3 items, α= .78).

The last part included an evaluation of Expectations (5-point scale, based on the Expectations Continuum discussed in [17]), Content likeability (i.e., of the specific sequence) and possible distortions. Test subjects were asked to describe impairments in case they perceived any, to link them to the content and to indicate their Degree of annoyance (on 5-point scale). In this paper, we discuss the results for a set of these measures.

3.4. Test material and test environment
The appeal of the video was an important factor for this test. Therefore, we made an initial analysis of most popular movie scenes on YouTube, limiting the scope to the area of Germany. From the initial selection, 2 different movies were chosen based on their popularity (action movie and comedy/romantic movie). Then, the movies were divided in 30 minute long sequences, and analyzed in terms of temporal and spatial complexity. After the analysis, one section of 30 minutes per movie was selected based on their similarities in terms of temporal and spatial complexity, and subsequently divided in 3 parts of 10 minutes. Each part of 10 minutes was processed in order to create the different video profiles used during the test. Table 1 indicates the characteristics of the three used error profiles, which were randomized across the sequences. The 4-state Markov model was used for inserting the errors, audio and video were multiplexed. The resultant sequences were encoded in H.264 and AAC in full HD 1080p with an average bitrate of 25mbits/s for video and 384kbits/s audio. The study itself took place in a ‘living room lab’, using a 55-inch Dynamic Edge LED-backlit HDTV with Full HD 1080p resolution at recommended viewing distance of 2.5 meters, and with a room illumination of ~25 lux.

4. Analysis and results

4.1. Evaluation of QoE in relation to the error profiles

4.1.1. Traditional QoE measures
As can be observed in Table 2 and Figure 1, the overall quality was – as expected – evaluated as best in the clips without errors. The clips with low but continuous errors were evaluated as worst. To investigate whether these observed differences between the error profiles are significant, we conducted a Kruskall-Wallis test using the overall quality ratings as dependent and the different error profiles as independents. The results indicate significant differences (H(2)= 55.78, p< .000) between the three error profiles.

<table>
<thead>
<tr>
<th>Error Profile</th>
<th>No errors</th>
<th>Low errors</th>
<th>High errors</th>
</tr>
</thead>
<tbody>
<tr>
<td>M</td>
<td>S.D.</td>
<td>M</td>
<td>S.D.</td>
</tr>
<tr>
<td>Overall quality</td>
<td>4.36</td>
<td>0.56</td>
<td>4.79</td>
</tr>
<tr>
<td>Pleasure</td>
<td>6.83</td>
<td>1.09</td>
<td>4.86</td>
</tr>
<tr>
<td>Arousal</td>
<td>4.86</td>
<td>2.09</td>
<td>4.46</td>
</tr>
<tr>
<td>Dominance</td>
<td>6.86</td>
<td>1.82</td>
<td>6.57</td>
</tr>
<tr>
<td>Expectations</td>
<td>0.70</td>
<td>1.17</td>
<td>-1.41</td>
</tr>
<tr>
<td>Interest</td>
<td>2.98</td>
<td>0.84</td>
<td>2.70</td>
</tr>
<tr>
<td>Enjoyment</td>
<td>2.35</td>
<td>1.05</td>
<td>1.72</td>
</tr>
<tr>
<td>Surprise</td>
<td>1.81</td>
<td>0.89</td>
<td>1.44</td>
</tr>
<tr>
<td>Focused attention</td>
<td>2.91</td>
<td>0.98</td>
<td>2.25</td>
</tr>
<tr>
<td>Felt involvement</td>
<td>3.20</td>
<td>0.88</td>
<td>2.48</td>
</tr>
<tr>
<td>Content likeability</td>
<td>3.04</td>
<td>0.85</td>
<td>2.26</td>
</tr>
</tbody>
</table>

To identify where exactly the significant differences are situated, we performed separate Mann-Whitney tests. These indicated that the overall quality is perceived as significantly better in the clips...
without errors than in the clips containing errors (p = 0.000). We have however no evidence to state that the differences between the low and the high error profile in terms of evaluated overall quality, are significant.

Figure 1: Absolute ratings for overall quality (per error profile)

When considering the acceptability evaluations, we also found a significant association between the three error profiles and whether or not the overall quality of the clip was found to be acceptable (using Pearson’s chi square test, \( \chi^2(2) = 46.48, p = 0.000 \)). As expected, the clips containing errors are much more likely to be evaluated as not acceptable (see also Figure 4). In total, 52.4% of the videos were rated as not acceptable. The percentage of videos with the high error profile evaluated as not acceptable is slightly higher than the one for low error profile videos (82.1% vs. 75.0%). This however also means that one fourth of the low error videos and 17.9% of the high error videos were evaluated as ‘acceptable’ (despite the inserted errors and despite the low quality evaluation scores).

4.1.2. Alternative measures

We also investigated possible differences in QoE when considering the alternative measures. Again, we conducted a Kruskall-Wallis test, using the alternative measures as dependents and the error profiles as independents. As is indicated in Table 3, significant differences between the error profiles were found for self-reported pleasure, for the degree to which the overall video quality meets the expectations, for enjoyment and for the user engagement-related constructs focused attention and felt involvement.

Table 3: Kruskall-Wallis test statistics (alternative measures)

<table>
<thead>
<tr>
<th></th>
<th>H</th>
<th>df</th>
<th>Sig.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pleasure</td>
<td>11.11</td>
<td>2</td>
<td>&lt; 0.05</td>
</tr>
<tr>
<td>Arousal</td>
<td>0.69</td>
<td>2</td>
<td>N.S.</td>
</tr>
<tr>
<td>Dominance</td>
<td>1.32</td>
<td>2</td>
<td>N.S.</td>
</tr>
<tr>
<td>Expectations</td>
<td>8.31</td>
<td>2</td>
<td>&lt; 0.001</td>
</tr>
<tr>
<td>Interest</td>
<td>1.44</td>
<td>2</td>
<td>N.S.</td>
</tr>
<tr>
<td>Enjoyment</td>
<td>6.83</td>
<td>2</td>
<td>&lt; 0.05</td>
</tr>
<tr>
<td>Surprise</td>
<td>2.51</td>
<td>2</td>
<td>N.S.</td>
</tr>
<tr>
<td>Focused attention</td>
<td>7.32</td>
<td>2</td>
<td>&lt; 0.05</td>
</tr>
<tr>
<td>Felt involvement</td>
<td>8.87</td>
<td>2</td>
<td>&lt; 0.05</td>
</tr>
<tr>
<td>Content likeability</td>
<td>8.31</td>
<td>2</td>
<td>&lt; 0.05</td>
</tr>
</tbody>
</table>

Additional post-hoc tests (using separate Mann-Whitney’s) indicated that the clips without errors correspond with significantly higher pleasure and enjoyment than those with a low error profile. Felt involvement and focused attention - constructs which we used to get an indication of the user’s engagement in the viewing sessions – are also significantly higher in the no error profile than in the clips with low error intensity but continuous error frequency. This indicates that – as in [2] – errors interrupting the flow seem to interfere with the engagement process in a negative way. There is also a significant difference between both profiles in terms of the indicated likeability of the content (i.e., at the semantic level, not in terms of quality). When there are no errors, the content is significantly better appreciated than when there are errors in the videos (this applies to both profiles containing errors in comparison to the error-free profile). The technical quality thus seems to interfere with the evaluation of the content in terms of liking or disliking (as was also observed in previous studies, such as [18]).

When comparing the error-free profile with the high error profile (higher intensity, but lower frequency of errors), we find that the former is associated with higher pleasure and that the test subjects felt significantly more involved. Finally, the differences between the three error profiles concerning the degree to which the overall quality of the clips was meeting the test subjects’ expectations, which are shown in Figure 2 are all significant. Whereas the error-free clips are associated with an overall quality that is as expected or even better than expected, the opposite holds true for both profiles containing errors. The low error profile was evaluated as most below expectations in this respect.

Figure 2: Degree to which expectations are met (averages), compared for the three error profiles.

4.2. How do the traditional and alternative subjective measures relate to each other?

The second aim of our study was to explore how the different subjective measures might relate to each other. First of all, we have a look at how the acceptability and overall quality evaluations correspond to each other. To this end, we conducted a Pearson’s chi-square test which indicated a significant correlation (\( \chi^2(2) = 5.529, p < 0.001 \)): evaluations of overall quality as being poor or bad correspond with non-acceptability and vice versa, when the overall quality is evaluated as good or excellent, the odds that the overall quality was found to be acceptable, are much higher. These tendencies can also be observed in Figure 3, which show how the overall quality ratings and acceptability evaluations (in percentage) relate to each other, for the different error profiles.

We also take a closer look at the correlations between overall quality and the other measures (shown in Figure 4). Our findings indicate a relatively low but significant positive correlation.
between overall quality and the engagement measures (focused attention and felt involvement). Similarly, weak yet significant positive correlations can be observed between overall quality and the self-reported pleasure and enjoyment.

Figure 3: Acceptability (in %) in relation to overall quality scores, for the three error profiles.

Not surprisingly, the degree up to which the overall quality of the clip was in line with the expectations on the other hand, is strongly positively correlated with the overall quality ratings: when the perceived overall quality is worse than expected, this is reflected in worse quality ratings and vice versa ($\rho = .694$, $p = .01$). Finally, there is a negative and significant correlation between the reported annoyance due to the slicing errors and the overall quality ratings ($\rho = -.423$, $p = .01$): low(er) annoyance ratings (meaning that the slicing errors were not experienced as annoying and high(er) overall quality ratings seem to go hand in hand).

![Correlation with overall quality (Spearman's rho)](image)

Figure 4: Significant correlations between 'overall quality' and the other self-report measures, showing the values for Spearman’s $\rho$ (*= significant at the .05 level, **= significant at the .01 level).

Because of our interest in gaining a better understanding of “delight”, we are also particularly interested in the relation between the enjoyment construct and the other included measures of QoE. As is shown in Figure 5, there are several subjective measures that significantly correlate to enjoyment. A first interesting observation is that the (positive) correlation with overall quality is the lowest one of all significant correlations. An important implication of this finding is that the traditional overall quality measure of QoE is thus – in our study - not a good measure for QoE, understood in terms of delight.

As expected, enjoyment is strongly correlated to pleasure, which indicates that our measure of enjoyment can be considered as rather robust. In addition, there is a clear and significant positive correlation between enjoyment and the user engagements attributes included in the study, namely felt involvement and focused attention. Moreover, enjoyment also correlates with the likeability of the content ($\rho = .471$, $p = .01$): higher enjoyment and higher appreciation of the content tend to go hand in hand. Although enjoyment is also significantly and positively correlated to the surprise construct and to the degree to which the overall quality corresponds to expectations, these correlations are rather weak. Finally, there is a significant but also weak negative correlation between the enjoyment and the reported annoyance due to slicing ($\rho = -.271$, $p = .05$): lower enjoyment corresponds with indications of higher annoyance.

![Significant correlations between ‘enjoyment’ and the other self-report measures, showing the values for Spearman’s $\rho$ (*= significant at the .05 level, **= significant at the .01 level).](image)

Figure 5: Significant correlations between ‘enjoyment’ and the other self-report measures, showing the values for Spearman’s $\rho$ (*= significant at the .05 level, **= significant at the .01 level).

Finally, we also explored how acceptability and some of the alternative measures relate to each other. Figure 6 shows how the values for enjoyment correspond to acceptability. We can observe that the overall video quality is evaluated as acceptable, the enjoyment tends to be higher than when it is not acceptable. A similar observation can be made when considering self-reported pleasure in relation to acceptability.

![Overall enjoyment ratings, split by acceptability](image)

Figure 6: Overall enjoyment ratings, split by acceptability

However, in our study, we found no evidence to state that acceptability one the one hand and pleasure, enjoyment, felt involvement and focused attention on the other, are significantly linked.
5. Conclusions and future work

In this paper, we explored a set of ‘alternative’, emotional state and engagement-related measures of QoE and investigated how they relate to traditional measures. To this end, we presented results from a living room lab study in which the impact of slicing errors on QoE was investigated. Our findings indicate significant differences in QoE between the three used error profiles, both when considering the traditional and alternative measures. Overall, the clips with low but continuous errors were experienced as worst in terms of overall quality and significantly linked to lower pleasure and enjoyment. The test subjects’ engagement was also significantly lower in the clips with errors than in those without. This finding can be linked to previous studies pointing to the importance of the ‘flow experience’ and implications of interrupted flow when watching movies. With respect to the relation between the included traditional and alternative measures, we found clear, significant correlations between enjoyment and the user engagement-related measures, the content likeability and the degree to which expectations were met. Enjoyment was also significantly correlated with the surprise construct, as well as with the overall experience’ and implications of interrupted flow when watching movies. With respect to the relation between the included traditional and alternative measures, we found clear, significant correlations between enjoyment and the user engagement-related measures, the content likeability and the degree to which expectations were met. Enjoyment was also significantly correlated with the surprise construct, as well as with the overall quality evaluations. The latter correlation was however very low. Although further research is needed to validate our findings, they point to the need to critically re-assess traditional measures of QoE and to further explore the value and implications of alternative measures, in particular if the goal is to really measure QoE in terms of delight as in [5] and to use this as a basis for optimization and differentiation.

As we did not conduct a comparative test in a standard test environment, a limitation of this study is that we cannot draw conclusions with respect to the influence of the test environment. Moreover, as we only used self-report measures, it would be valuable to complement the included alternative measures with autonomc (e.g., behavioral and physiological) measures that do not rely on introspection. It would also be interesting to compare our findings with a set up using a continuous quality evaluation approach (e.g., SSCQE). In our ongoing work, we try to address some of these limitations. In a recent, more standardized lab test, we used the same experimental stimuli. In this study (currently being analyzed), the traditional and alternative self-report measures were complemented with EEG, eye tracking and facial expression capturing to gather additional indications of engagement, enjoyment and attention. Future research should seek to better understand which factors and features play a vital role for fostering delight and for limiting annoyance. The classification discussed in [12] in terms of basic, performance and excitement factors could be an interesting starting point.

Based on a thorough review of the literature on delight, it would also be valuable to focus more explicitly on possible antecedents and implications of delight in the context of QoE.

6. Acknowledgements

This work was partially supported by the COST IC1003 European Network on Quality of Experience in Multimedia Systems and Services – QUALINET. Part of this work was carried out during the tenure of an ERCIM “Alain Bensoussan” Fellowship Programme (Katrien De Moor) and has received funding from the European Union Seventh Framework Programme (FP7/2007-2013) under grant agreement n°246016.

7. References