IMPACT OF VIDEO RESOLUTION CHANGES ON QoE FOR ADAPTIVE VIDEO STREAMING

Avşar Asan[*], *Werner Robitza*[§], *Is-haka Mkwawa*[*], *Lingfen Sun*[*], *Emmanuel Ifeachor*[*], *Alexander Raake*[†]

[*]Signal Processing and Multimedia Communications Lab, Plymouth University, Plymouth, U.K.
[§]Telekom Innovation Labs, Deutsche Telekom AG, Berlin, Germany
[†]Audiovisual Technology Group, Technical University of Ilmenau, Ilmenau, Germany

ABSTRACT

HTTP adaptive streaming (HAS) has become the de-facto standard for video streaming to ensure continuous multimedia service delivery under irregularly changing network conditions. Many studies already investigated the detrimental impact of various playback characteristics on the Quality of Experience of end users, such as initial loading, stalling or quality variations. However, dedicated studies tackling the impact of resolution adaptation are still missing. This paper presents the results of an immersive audiovisual quality assessment test comprising 84 test sequences from four different video content types, emulated with an HAS adaptation mechanism. We employed a novel approach based on systematic creation of adaptivity conditions which were assigned to source sequences based on their spatio-temporal characteristics. Our experiment investigates the *resolution switch effect* with respect to the degradations in MOS for certain adaptation patterns. We further demonstrate that the content type and resolution change patterns have a significant impact on the perception of resolution changes. These findings will help develop better QoE models and adaptation mechanisms for HAS systems in the future.

Index Terms— Quality of Experience, Video Quality, Resolution Switch, HTTP Adaptive Streaming

1. INTRODUCTION

Today, HTTP Adaptive Streaming (HAS) is the most popular method of streaming videos to end user devices over the web infrastructure. It is cost-effective and ensures multimedia service constancy and stability. HAS adapts the video playback according to the network characteristics. This is typically achieved by switching between representations of different bitrate and resolution of video. The impact of such resolution changes during the playout on the users' perceived quality is an important factor; previous work [1] has already shown how the Quality of Experience (QoE) can be influenced by buffering events or variations in quality over time. QoE also significantly affects decisions on the preference to use a service or not [2]. Negatively affected QoE due to unstable network conditions may trigger a chain reaction, starting from individual service abandonment up to users leaving their service/content providers (i.e., user churn) in the long term.

Video resolution switch phenomena and their effects on QoE have not yet been fully investigated. The main objective of this work is to provide a systematic analysis of resolution changes and their impact on QoE. We present the results of a quality assessment test which investigates *resolution switch* effects. In our work. the term *resolution switch* corresponds to the video player switching from one played resolution to another. We also define *adaptivity* as an overall effect, i.e. the sum of resolution switch events in a sequence. With the aid of our systematic approach, it is possible to analytically investigate adaptivity patterns with respect to their Mean Opinion Score (MOS).

We begin by describing related work in Section 2.We then propose a novel theoretical framework for the assessment of resolution adaptivity in Section 3. Our audiovisual test setup is explained in Section 4. In Section 5 we interpret the results of our assessment. Finally, in Section 6 we discuss our findings and list future work. The paper is concluded in Section 7.

2. RELATED WORK AND MOTIVATION

Although the detrimental effects of various playback impairments such as initial loading, stalling or quality variations have been widely investigated (e.g., comprehensive surveys are found in [1, 3]), there is still a need for dedicated and systematic studies to tackle the impact of resolution adaptation. However, it is not just the obvious effects on video player that contribute to user QoE: our literature analysis focuses on the key influencing factors on QoE for HAS adaptivity.

Human perception system characteristics play a key role in subjective quality assessment tasks. Cranley et al. [4] emphasize that human visual perception is able to adapt to a specific video quality only after a few seconds. The authors noted that impairment effects become more annoying if the quality changes happen frequently in a very short time period. Although HAS is technically capable of changing the quality

every few seconds while streaming, in practice, adaptation is carried out more slowly to prevent large quality variation periods or oscillations. The authors of [5] investigated up to six quality changes in a 20-second video, which in the light of the aforementioned considerations is beyond realistic.

Ecological validity refers to how useful and valid results from a laboratory study are when they are applied in real life. Experiments with artificial settings or test scenarios based on an imaginary situation produce decontextualized results and may not be implemented in daily life, as recently discussed in [6]. In the domain of QoE, it is known that short-term video quality prediction models (e.g., as shown in [7, 8]) can obtain high performance, but the ecological validity of these models is questionable, especially for longer video durations and their applicability on HAS algorithms. Finally, a recent study [9] revealed that millennials (18–34) tend to spend around 14 hours per week on video streaming services and that longer video durations are more preferable.

Traditional testing methods show short, non-entertaining stimuli, with repeating contents, which is known to bore users. Having users be immersed and entertained is one of the key factors to get more ecologically valid results from a lab-based quality test. However, it seems that there is a lack of application of this paradigm. When they are immersed, users feel "sucked" into the media [10]. It may make them less aware of their surroundings, be more enjoyed, and help reduce stress during an experiment. Pinson et al. [11] first suggested a new test method: a source stimulus should be used only once so that the subjects can focus on the content rather than evaluating the same sequences over and over. In the same work, it is also proven that in an immersive test design, boredom and fatigue can be significantly reduced. Robitza et al. [12] successfully applied the immersive test design for HAS QoE. They note that stimuli should be entertaining and meaningfully complete for a more ecologically valid test.

Using different content types with various characteristics helps in developing more general statements about QoE. Content may differ in genre and enjoyability, but also in technical parameters such as spatiotemporal complexity, the latter having a significant impact on the quality of compressed video encodes. In [13, 14] the only content is computer-generated graphics. Also some studies such as [5] did not analyse the impact of content on their obtained results. In one study [15] investigating an adaptive streaming model, the authors chose seven content types having almost the same spatiotemporal complexity, although video stimuli are from different genres. Consequently the work failed to explain a logical relationship between spatiotemporal characteristics and the content type. Additionally the impact of quality switches is not investigated.

Rodriguez et al. [7] modelled the impact of video quality level switching. However, the authors assumed that the impact of a switch would be the same, no matter if it happened in good or bad quality regions. Our results however will show that the impact is more complex. Finally, Liu et al. [16] investigated the quality level variation factors depending on average level, number of quality changes and average change magnitude. However, the underlying quality metric they use does not model the impact of quality switches. Also, their work assumes a 1:1 relationship between bitrates and resolutions, which is not constant in practice.

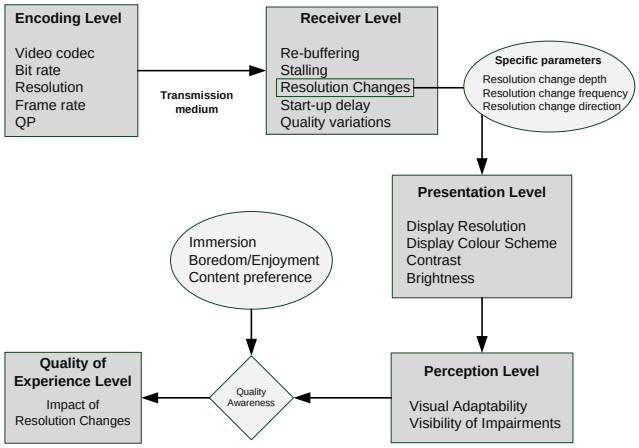

Fig. 1. Conceptual framework for QoE in HAS services.

3. CONCEPTUAL QoE FRAMEWORK

From the points insufficiently tackled in previous literature we developed a conceptual framework (see Figure 1). It comprises all steps in the transmission chain – from the source to the user – and highlights the factors that need to be investigated to fully understand the impact of adaptivity on QoE. It may guide in creating a study setup, interpreting our results and developing our future agenda. In our framework, we simplified concepts from [17], but added specific parameters for our investigation purpose.

The *Encoding Level* is the primary phase for the preparation of videos. It emphasizes the non-linear relationship between encoding parameters and QoE. By encoding parameters we mean the combined configuration of video bitrate, framerate, resolution and the codec chosen for different HAS video representations. On the *Receiver Level*, we look at effects of unstable network conditions, shown as typical HAS impairments. Specifically, we focus on adaptivity – together with its parameters. The *Presentation Level* is about how multimedia is presented on the user side. It denotes viewing conditions and types of devices. Our subjective test has been designed with a focus on that level. The *Perception Level* is about the way humans consume multimedia services. Influence factors in this level are subjective: for example, they depend on whether people can visually adapt to the impairments, perceive any impairments at all or really pay attention to what is happening on the screen. Those are aspects

that seem intangible at first, but are very important for consideration in future work. Finally, the *Quality of Experience Level* is where the sense of quality is formed after perception. Quality awareness is a cognitive gate component between the *Perception* and *Quality of Experience* level. It relates to user anticipation, content preference, enjoyment and immersion as a function of the subjects' desired quality features [10]. In order to have an understanding of the users perceived quality, the preceding levels in our framework should be well understood first. This is where our test comes into play.

4. EXPERIMENT SETUP

In this section we will present the technical setup of our test, which systematically addresses adaptivity as a function of resolution switches. We took special care to include the above-mentioned considerations on ecologically valid conditions, immersive source video selection and assignment of conditions based on spatiotemporal characteristics.

4.1. Source Stimuli

Many existing databases do not take into account factors such as immersiveness and enjoyment. Our 43 original videos were obtained from various online portals, choosing the popular genres sports, cooking, sightseeing and music videos. Due to the fact that such sequences have already been compressed by the content provider, only 4K (3840×2160) sources were chosen to ensure a high enough bitrate. We additionally checked every video for its quality to be pristine (e.g., presence of camera noise or shakiness, compression artifacts). The 43 videos were then cut to logically complete test scenes of 45 seconds length, resulting in 84 source clips (from here on: *SRCs*).

4.2. Conditions (Adaptivity Patterns)

Our test conditions (i.e. the way the resolution switches occur over time) were based on three resolution levels: 240p, 480p and 1080p. We first defined three reference conditions with those constant resolutions; all other conditions had one or two resolution switch in them. For our systematic design we considered the following adaptivity patterns, 21 in total:

- Reference conditions: 1080, 480, 240
- Single drop or increase: 1080 → 240, 1080 → 480, 240 → 1080, 480 → 1080, 240 → 480, 480 → 240
- Symmetrical drop: 1080 → 240 → 1080, 1080 → 480 → 1080, 480 → 240 → 480
- Fluctuations: 1080 → 240 → 480, 480 → 240 → 1080, 240 → 1080 → 480, 480 → 1080 → 240
- Constant drop or increase: 1080 → 480 → 240, 240 → 480 → 1080

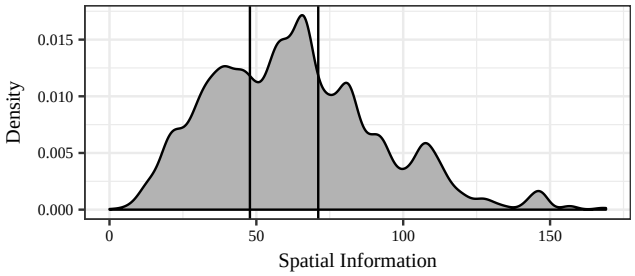

Fig. 2. Spatial Information values for all source sequences.

- Symmetrical increase: 480 → 1080 → 480, 240 → 480 → 240, 240 → 1080 → 240

As can be seen, the conditions allow comparisons against each other, which we will detail in Section 5.

4.3. SRC–Condition Assignment

Spatiotemporal characteristics play a key role in comparing different contents in terms of how much spatial detail and motion there is. Especially in our test design, every SRC is shown only once, hence, it needs to be made sure that their characteristics are equally spread out. While they may originate from the same original content (e.g. a longer sports sequence), individual portions of the clip may differ in their characteristics. We hypothesize that these characteristics have a direct impact on the visibility of resolution switches.

We first calculated the Spatial Information (SI) values (according to ITU-T Rec. P.910) for every frame in our SRCs. From those, we obtained an SI density function, as shown in Figure 2, giving us the entire range of SI values in our test. By looking at the thresholds of the 33% and 66% quantile (47.8 and 71.1), we can then classify any SI values as *high*, *middle* or *low*.

For the allocation of our SRCs to the conditions, we first calculated the average SI for each third of every SRC, i.e. from 0–15, 15–30 and 30–45 s. We then assigned the average SI to the above-mentioned classes *high*, *middle* or *low*. For example, if one SRC was split into three parts and had the average SI values 33.5, 50.9, 79.2, it was classified as $L \rightarrow M \rightarrow H$. The same procedure was repeated with the SRCs split in half (i.e., from 0–22.5 and 22.5–45 s).

These SI characteristics were then systematically paired with the conditions under the following rules: 1) Each condition should have a SRC with an SI characteristic matching that condition. For example for a condition with 240 → 480, there would be a SRC with the SI characteristic $L \rightarrow M$. 2) There should be the inverse condition for that characteristic, e.g. $M \rightarrow L$ should be assigned to the SRC in the previous case. 3) Two SRCs with constant-high and constant-low SI would be mapped to the condition, too, respectively. This lead to four SRCs being applied to every

condition (21 conditions $\times$ 4 SRCs $=$ 84 sequences).

4.4. Video Encoding and Test Sequence Generation

For encoding the final sequences, we chose to simulate HAS offline. We first divided the SRC clips into two or three equally sized parts, depending on the condition assigned to them. These parts were then encoded with *ffmpeg* and *x264* and downscaled (if necessary) to match the condition pattern. x264 was set to use a Constant Rate Factor of 23 to ensure constant quality across the encode, with a one-pass encoding mode. The maximum bitrate was constrained for different resolutions – similar to what popular video streaming services implement: 400 kbps for 240p, 1.5 Mbps for 480p and 5.5 Mbps for 1080p. After that, the parts were upscaled to 1080p and concatenated to form the final processed video sequences (PVSes). Audio was not compressed during the PVS generation and played throughout the whole sequence.

4.5. Test Environment and Protocol

Our test was conducted in a standards-compliant environment (according to ITU-T Rec. P.910). The sequences were shown on a 42" LCD display with 1920×1080 resolution. Subjects were seated 3H (three times the height of the display) from the monitor.

First, subjects were introduced to the topic. They were then checked for visual acuity and colour blindness and had to fill out a simple demographic questionnaire. For the main experiment part, each PVS was presented after another, with a randomized playlist for each subject in order to minimize ordering effects. Before the actual PVSes were shown, we displayed five "training" clips whose ratings were not taken into consideration later. Subjects were asked to rate the visual quality of the stimuli. The ratings themselves were given on a standard Absolute Category Rating (ACR) scale with labels from *Bad* to *Excellent* (see ITU-T P.910), using the open source *AVRate* software. Finally, subjects filled a post-test questionnaire on what they had seen.

5. RESULTS

In our test, 30 subjects took part, 20 of which female. Their age ranged from 19 to 51 (average: 30). In order to eliminate unreliable viewers, we used the following procedure [12]: We first calculated the Pearson correlation between each subject's vote and the overall MOS for every PVS. Then, once a subject's correlation was below 0.70, they were removed from the pool and the procedure repeated. This lead to the exclusion of three subjects, meaning that our shown results are based on 27 assessors,

Our overall range of MOS is 1.48–4.70 (average: 3.13), showing a good use of the rating scale, which stems from a balanced test. In the following, we will explain the impact of certain experimental factors on the subjective ratings.

Table 1. MOS impact for one resolution switch, averaged over all content types.

Ref. Resolution	Switching Pattern	MOS Impact
1080	1080 $\rightarrow$ 240	-1.66
1080	1080 $\rightarrow$ 480	-0.49
480	480 $\rightarrow$ 1080	0.45
480	480 $\rightarrow$ 240	-1.12
240	240 $\rightarrow$ 1080	1.31
240	240 $\rightarrow$ 480	1.36

5.1. Impact of Conditions

First we want to verify that the chosen conditions have the expected impact on the overall ratings. Figure 3 shows all MOS values, grouped by condition, with the different content types highlighted. We can see that the pattern indeed has a strong effect on the ratings over the entire range of conditions.

Our systematic approach makes it possible to directly compare one pattern against another. This is especially useful when comparing, for example, a reference condition (e.g., continuous 1080P) against a condition with a resolution change (e.g. 1080-480). Here we can directly formulate the impact of the switch on the MOS, by subtracting the MOS of the switching condition from the MOS for the reference. In the above case, this is $4.50 - 4.01 = -0.49$. Thus, we can say that generally, when switching from 1080 to 480, this incurs a MOS impact of -0.49.

Table 1 shows an overview of the different MOS impacts identified for the patterns with one switch. It shall be noted that these values correspond to averages over all content types, however, we believe that those will be more useful for a content-independent modelling of QoE. From this table it can be seen that the depth (as measured in vertical pixels, i.e., 600 and 840) of the switch has a significant impact, as shown by an ANOVA between the depth and ratings ($p < 0.02$). Generally, we also observe that a change in resolution is worse when it occurs at a lower level.

The direction of the change appears to have an impact too: when users start with low resolution, they score any upwards resolution more positively than if they had experienced a drop in resolution. This could also be explained by a positional effect: when low quality is played at the last few seconds of the sequence, subjects may have given more weight to these portions. This is also called a "recency effect". It is visible in other conditions, too. For example, 480 $\rightarrow$ 1080, 240 $\rightarrow$ 480 and 240 $\rightarrow$ 1080 are scored significantly higher than their reversed counterparts that started well, but end at low quality. The same holds true for conditions with only one switch. We further conducted an ANOVA between the change directions

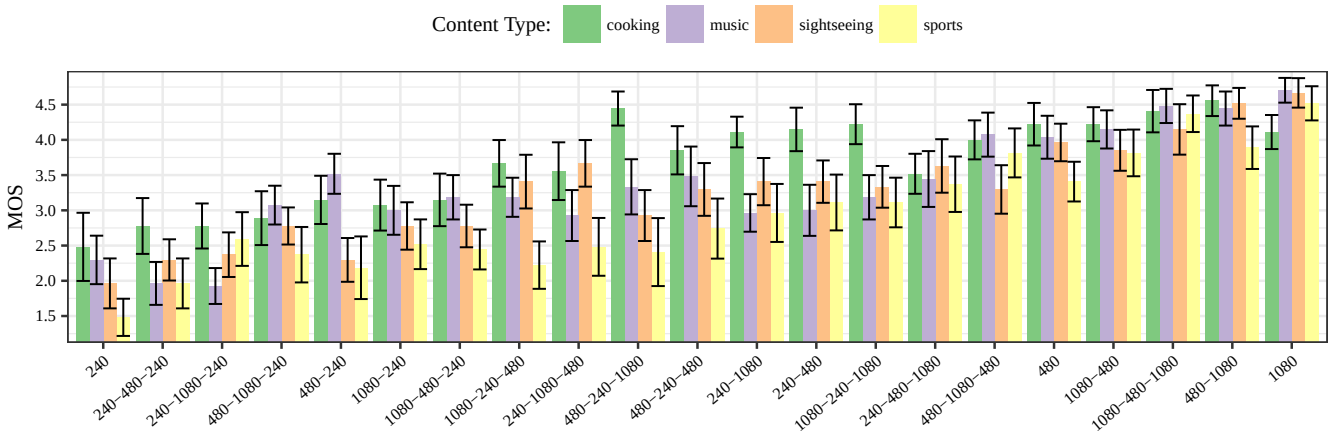

Fig. 3. MOS for each PVS, sorted by condition. Content types are highlighted in different colors.

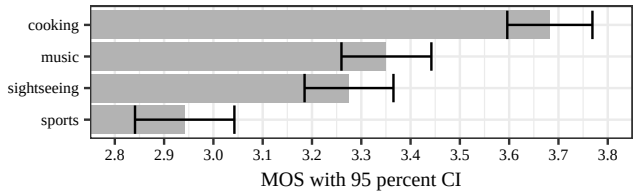

Fig. 4. MOS for each content type, averaged over all patterns.

("down" or "up") and the ratings, which showed a significant effect ($p = 0.04$) and confirms our results.

5.2. Impact of Content and other experimental Factors

As mentioned in Section 4, we included different content types in our experiment to get a more balanced MOS estimation for a given pattern, under the hypothesis that any systematic content effect could be averaged out. This is especially relevant in practical quality monitoring applications, where the content type may not be known.

However, for correctly analyse the subjective test results, the content type cannot be neglected: Figure 4 shows the average ratings for a given content type, considering all conditions, as clearly visible, there is a significant effect of the content type on the MOS ratings. We conducted a one-way ANOVA between content type and ratings ($p < 0.02$) to prove this effect. A post-hoc test (Tukey HSD) revealed that only the difference between sports and cooking videos was significant ($p < 0.01$). As we will later see in the questionnaire results, this is due to the visual characteristics of the content itself, not because of its enjoyability.

5.3. Questionnaire

From our pre- and post-assessment questionnaires we gathered more insight into the MOS ratings: subjects were asked to rank the four content types according to their liking. Only 13% of the subjects marked "cooking" as first priority. This contrasts with our quality ratings, where cooking content was judged significantly higher than others. This leads us to conclude that subjects found resolution switches less disturbing for this content, and that content preference and quality ratings are not necessarily correlated. We attribute these findings to the visual characteristics of the chosen cooking sequences which may have made the switches less visible. However, further analysis and tests are needed to confirm that hypothesis, which would require the inclusion of more content types with varying spatiotemporal characteristics.

6. DISCUSSION

As can be seen from the MOS results, our test is well-balanced in terms of the range of conditions and content types. The obtained MOS degradation values in Table 1 can be used as a component in QoE models, when it is necessary to quantify the impact of a single switch. Of course – as always the case for subjective studies – the factors shown here are just a small part in the big picture of HAS QoE, and we will conduct further test series in the future. In other words, it is impossible to design a test in which one can investigate all factors reliably. However, previous research has rarely been that systematic: the design of conditions should be done in such a way that they can be compared against each other. For future tests we can re-use some of the shown clips as anchor points, which will allow us to create a bigger database of adaptivity conditions that can also be systematically compared.

We could successfully apply an "immersive" paradigm in

our test, meaning that entertaining sequences were used, without repeating the same source. Our results indicate a strong impact of the content characteristics on the perceived quality.

At this stage, we believe that an attempt to model the impact of resolution switches would result in a too narrow view. In fact, it would require at least another study to serve as a database for validating any created model. Hence, our focus lies on producing a series of complementary tests, in order to be able to create more robust models in the end. For example, this process has also been successfully used for the models standardised by ITU-T.

7. CONCLUSION

In this paper we presented the results of a first study on the impact of resolution changes on user-perceived QoE. We were motivated by current literature being still inconclusive about investigating single impairment factors that are typical for HAS services, with resolution switches being one of them. Our novel, systematic test design – in which we could compare reference against adaptivity conditions – allowed us to predict the effect of a specific switch in terms of MOS degradation.

The experiment shown here is just one of a series of tests that we will conduct, in order to give a full picture on resolution switches. Our conceptual framework lists all these points as a guideline for future research: it includes factors such as switch visibility, positional effects of switches in longer sequences, impact of different device types, and – among others – socio-economic factors related to user demographics and pricing.

Acknowledgements
The work presented in this paper is funded by the European Union in the context of Horizon 2020 Research and Innovation Programme under Marie Sklodowska-Curie Innovative Training Networks (MSCA-ITN-2014-ETN) Grant Agreement No.643072, Network QoE-NET.

8. REFERENCES

[1] M. Seufert, S. Egger, M. Slanina, T. Zinner, T. Hoßfeld, and P. Tran-Gia, "A survey on Quality of Experience of HTTP adaptive streaming," *IEEE Communications Surveys & Tutorials*, vol. 17, no. 1, pp. 469–492, 2015.

[2] A. Sackl, P. Zwickl, S. Egger, and P. Reichl, "The role of cognitive dissonance for QoE evaluation of multimedia services," in *2012 IEEE Globecom Workshops*. IEEE, 2012, pp. 1352–1356.

[3] M.-N. Garcia, F. De Simone, S. Tavakoli, N. Staelens, S. Egger, K. Brunnström, and A. Raake, "Quality of Experience and HTTP adaptive streaming: A review of subjective studies," in *QoMEX 2014*. IEEE, 2014, pp. 141–146.

[4] N. Cranley, P. Perry, and L. Murphy, "User perception of adapting video quality," *International Journal of Human-Computer Studies*, vol. 64, no. 8, pp. 637–647, 2006.

[5] S. Egger, B. Gardlo, M. Seufert, and R. Schatz, "The impact of adaptation strategies on perceived quality of HTTP adaptive streaming," in *Workshop on Design, Quality and Deployment of Adaptive Video Streaming*. ACM, 2014, pp. 31–36.

[6] M. Peeters, C. Megens, C. Hummels, A. Brombacher, and W. Ijsselsteijn, "Experiential Design Landscapes: Design Research in the Wild," in *Nordic Design Research Conference 2013*, 2013.

[7] D. Z. Rodríguez, Z. Wang, R. L. Rosa, and G. Bressan, "The impact of video-quality-level switching on user Quality of Experience in dynamic adaptive streaming over HTTP," *EURASIP Journal on Wireless Communications and Networking*, vol. 2014, no. 1, pp. 1–15, 2014.

[8] M.-N. Garcia, P. List, S. Argyropoulos, D. Lindegren, M. Pettersson, B. Feiten, J. Gustafsson, and A. 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] C. International, "Video streaming survey 2016," Tech. Rep.

[10] W. Robitza, S. Schönfellner, and A. Raake, "A Theoretical Approach to the Formation of Quality of Experience and User Behavior in Multimedia Services," in *5th ISCA/DEGA Workshop on Perceptual Quality of Systems (PQS 2016)*, 2016, pp. 39–43.

[11] M. Pinson, M. Sullivan, and A. Catellier, "A new method for immersive audiovisual subjective testing," in *8th International Workshop on Video Processing and Quality Metrics for Consumer Electronics (VPQM)*, 2014.

[12] W. Robitza, M. N. Garcia, and A. Raake, "At home in the lab: Assessing audiovisual quality of HTTP-based adaptive streaming with an immersive test paradigm," in *Seventh International Workshop on Quality of Multimedia Experience (QoMEX)*. IEEE, 2015, pp. 1–6.

[13] M. Grafl and C. Timmerer, "Representation switch smoothing for adaptive HTTP streaming," in *4th International Workshop on Perceptual Quality of Systems (PQS 2013)*, 2013, pp. 178–183.

[14] J. De Vriendt, D. De Vleeschauwer, and D. Robinson, "Model for estimating QoE of video delivered using HTTP adaptive streaming," in *2013 IFIP/IEEE International Symposium on Integrated Network Management (IM 2013)*. IEEE, 2013, pp. 1288–1293.

[15] S. Tavakoli, K. Brunnström, K. Wang, B. Andrén, M. Shahid, and N. Garcia, "Subjective quality assessment of an adaptive video streaming model," in *IS&T/SPIE Electronic Imaging*. International Society for Optics and Photonics, 2014, pp. 90160K–90160K.

[16] Y. Liu, S. Dey, F. Ulupinar, M. Luby, and Y. Mao, "Deriving and validating user experience model for dash video streaming," *IEEE Transactions on Broadcasting*, vol. 61, no. 4, pp. 651–665, 2015.

[17] A. Raake and S. Egger, "Quality and Quality of Experience," in *Quality of Experience*, pp. 11–33. Springer, 2014.