Quality of Experience of Adaptive Video Streaming: Investigation in Service Parameters and Subjective Quality Assessment Methodology

Samira Tavakoli[1], Kjell Brunnström[2,3], Jesús Gutiérrez[1], Narciso García[1]

[1] *Universidad Politécnica de Madrid, Spain,* [2] *Acreo Swedish ICT AB, Sweden,* [3] *Mid Sweden University, Sweden*

Abstract

Usage of HTTP adaptive streaming (HAS) has become widely spread in multimedia services. Because it allows the service providers to improve resource utilization and user's Quality of Experience (QoE). Using this technology, the video playback interruption is reduced since the network and server status in addition to capability of user device, all are taken into account by HAS client to adapt the quality to the current condition. Adaptation can be done using different strategies. In order to provide optimal viewing experience, impact of different adaptation strategies from point of view of the user should be studied. However, the time-varying video quality due to the adaptation which usually takes place in a long interval introduces a new type of impairment making the subjective evaluation of HAS challenging. The contribution of this paper is two-fold: first, it investigates the testing methodology to evaluate the HAS QoE by comparing the subjective experimental outcomes obtained from a standardized method and a semi-continuous method developed to evaluate the long sequences. In addition, influence of using audiovisual stimuli to evaluate video-related impairment is inquired. Second, impact of some of the adaptation technical factors including the quality switching amplitude and chunk size in combination with high range of commercial video content type is investigated. The result of this study provides a good insight toward achieving appropriate test methodology to evaluate the HAS QoE, in addition to designing adaptation strategies with optimal visual quality.

Keywords: Adaptive video streaming, Quality of Experience, subjective evaluation

1. Introduction

In recent years a great increase of video services has taken place providing the ease of access for the consumer from almost any type of device and different location. This has made video streaming as the most dominant application in the Internet and this dominance is expected to grow even further within the near future. One of the main advances in this aspect is using HTTP streaming over Transmission Control Protocol (TCP) as delivery method which is used by most video content delivery services such as YouTube and Netflix. In contrast to the traditional streaming over User Datagram Protocol (UDP) where packet losses results in audiovisual distortions, the packet retransmission feature of TCP ensures reliable delivery of the video content. Buffering of the content at the client side further allows to overcome the network resource limitations in a short time scale and assures a continuous playout of the video content. However, this delivery method is prone to the temporal impairments such as long initial delay (in case where large playout buffers have to be filled initially) or stalling (i.e. playback interruption due to the empty playout buffer).

To overcome these problems, several service providers have implemented HTTP adaptive streaming (HAS). HAS makes it possible to switch the video quality during the playback in order to adapt to the current network conditions. In HAS, the video content is available in multiple bitrate (called adaptive streams or *representations*) on the server-side, which may differ in terms of spatial resolution, temporal resolution (framerate), encoding quantization, as well as combination thereof. Each representation consists of independently decodable segments of a few seconds of video, termed *chunk*. The characteristics of the representations are stored at server-side in a file (or *manifest*) providing to the client the required information for the adaptation process. After filling the buffer with the initial bitrate, the video playout is started. During the playback, the adaptation algorithm in the client-side measures the current bandwidth and/or buffer status in order to decide on the appropriate bitrate for the next video chunk request, such that the available bandwidth is utilized best and stalling is avoided. Nevertheless, in terms of user perceived quality, another dimension, i.e. time-varying quality switching is introduced.

Apart from the benefit of dynamically adapting the current video bitrate to the available bandwidth, employing HAS provides further advantages compared to the classical video streaming. For instance, offering multiple bitrates

of the video enables service providers to adapt the delivered video to users with different demands and network/device accessibility. Furthermore, based on the available video quality, different pricing schemes and service levels can be offered to the customers. All these advantages have made employing this technology more popular. Nevertheless, in regard to the adaptation behavior that should be decided in the client side, there is still no clear guideline about the performance of different scenarios in terms of the user's Quality of Experience (QoE). To achieve this, first the QoE influence factors of adaptation should be determined.

Up to now, several research works in the HAS area have been conducted which can be differentiated along technical and perceptual based quality assessment. Technical analyses such as [1] mainly focus on optimal switching strategies to optimize the bandwidth utilization and other network related parameters. Whilst the perceptual based analyses consider the QoE impact of adaptation related parameters. Up to now, research has been mostly concentrated on technical aspects of HAS, but in order to optimize the user's QoE it is crucial to pay more attention on the perceptual aspect.

Different possible QoE influence factors have been already addressed in previous studies such as *switching frequency* (i.e. number of switches per time interval), *switching amplitude* (i.e. quality level difference per switching event[1]), and the influence of *content characteristics* on the user's perception of quality switches. Considering the switching frequency, results presented in [2, 3, 4, 5, 6] show that frequency of the adaptation should be kept as low as possible. On the other hand, the study presented in [3] shows that if the duration spent on a high quality level is sufficiently long, higher switching frequencies do not significantly degrade the QoE. Considering the amplitude of the switch, most of the previous studies [3, 4, 5, 6, 7] conclude that gradual multiple variations are preferred over rapid variations. Nevertheless, as highlighted in [4] this conclusion may not be applied to the scenarios where the quality levels' difference is very small.

Another possible influence factor could be the *length of the chunk*. In certain scenarios like a live broadcast, using long chunk size may not be be suitable as switching granularity is more considerable. Employing small chunks improves the client reaction time to network bandwidth variations but also

[1]Adaptation event denotes the period of video playback when the quality switching from the current level to the target level occur.

increases the activity on the client side. To the best of our knowledge, the impact of chunk size on the HAS QoE has not received much attention yet. Concerning the dependency of adaptation QoE on the content characteristics, it has been found in [3, 8] that the effect of spatial and temporal switching varies depending on the content type. This was found to be even true for switches with the same amplitude so that it is difficult to spot quality oscillation when there are frequent scene changes while in steady shots they are more noticeable.

One of the common approaches to evaluate the impact of visual distortions on the user's perceived QoE is through subjective assessments. Extensive research related to subjective studies of audiovisual quality has brought up several testing methodologies to obtain reliable results for the development of multimedia technologies. There are different international recommendations provided by standardization organizations such as ITU-R BT.500 [9] and ITU-T P.910 [10], which give guidelines to assess the quality of television pictures (e.g. content encoding, viewing distance and viewing environment). However, the novelties of adaptive streaming technology could require research for new assessment methodologies that allow to obtain representative conclusions regarding the HAS users' visual experience.

In particular, the most common methodologies, like Absolute Category Rating (ACR) [10] recommend the use of short test video sequences of around 15 seconds after which the observers provide their ratings. However, in adaptive streaming there are switching behaviors whose effect takes longer time. Therefore, longer test sequences may be more appropriate to study these cases. Also, as presented in [11], traditional testing methods may not accurately predict the perceptual quality since the relative impact of impairment types would change with the setting of subjective test. It means it is not clear if the perceptual quality of adaptation event solely evaluated using the ACR method would be the same as when it is occurred in a longer sequence. Another standardized methodology that might be appeared more suitable for adaptive streaming evaluation is Single Stimulus Continuous Quality Evaluation (SSCQE) [9] where the observers provide instant ratings of the video quality in a continuous way during the sequence. However the recency and hysteresis effect of the human behavioral responses while evaluating the time-varying video quality would lead to an unreliable evaluation through this methodology [12].

Therefore, research on new methodologies more appropriate for evaluating the adaptive streaming system is required. In fact, some approaches have

been already proposed. For instance, in our previous work [13] the evaluation of set of subsequent adaptation scenarios was made using long sequences (around 6 minutes) selected from the content that is usually watched by the users at home (e.g., movies, news, etc.). Assessment methodology used in this study was firstly presented in [14] and subsequently named Content-Immersive Evaluation of Transmission Impairments (CIETI). The idea behind designing this method was to simulate realistic viewing conditions by using longer sequences so the observers become more engaged to the content as they would be in real life, rather than focusing on detecting impairments, which can happen using traditional methodologies with the short and less entertaining test videos.

In this sense, other approaches have been also presented, such as the method proposed in [15] for immersive evaluation of audiovisual content, based on the use of long test stimuli to encourage the observers' engagement with the content and simulating real situations of using audiovisual applications. In this work, not only longer sequences are recommended for evaluating video quality, but also using test sequences with audio (in spite of traditional standard recommendations). This recommendation makes sense since video-only presentations poorly represent the users' experience of an audiovisual application, as people rarely watch videos without sound. Another interesting study was presented in [8] where the authors analyzed the effects of a single adaptation event in sequences of around 2 minutes.

Summarizing, there are a broad investigations on HAS. However, many research questions are still open. This is because of different issues such as limited number of test conducted addressing a specific factor, results obtained from non-statistical analysis, or in some cases contradictory outcomes from different studies. In addition, there are some completely new research questions such as HAS QoE testing methodology that have not been addressed so far [16].

In response to this state of research, the novelty brought by this paper is two-fold: first it takes the open research question in regard to HAS QoE evaluation method into account. Second, it investigates the perceptual influence of some of the important technical parameters for adaptive streaming in combination with high range of video content characteristics.

The reminder of the paper is organized as follows. Section 2 describes the study factors and the experimental setting. The experimental results are described in Section 3 and further discussed in section 4. Finally, in Section 5 the general conclusions are presented.

2. Study description

2.1. Study Factors

Following aspects in relation to HAS perceptual quality and respecting the home viewing condition were considered to study in this paper.

- *Switching strategy*: Among the possible technical switching parameters, the impact of switching amplitude and chunk size when decreasing and increasing the video quality were considered for this study. By employing short and long chunk size to change the quality in step-wise (gradually) and abrupt (rapid) way different adaptation scenarios were designed. In addition, the trade-off between the quality adapted video and constant quality video was investigated.

- *Content type*: Effect of the content characteristics on perceptual quality of the test stimuli was highlighted in several previous works. This dependency can be based on the psychovisual factors such as objective characteristics of the content (e.g. spatial and temporal complexities) as the rate distortion performance of encoded video depends largely on these factors [17], in addition to psychological factors like user expectations and desirability of the video [18] which could be because of the content genera. To investigate the impact of these factors on perception of adaptation, the switching scenarios were applied on various video sequences in different objective characterizations and genera which are usually watched by the actual viewers in the real-world condition.

- *Evaluation methodology and impact of audio presence on quality assessment*: Another study factor was to investigate subjective methodology to assess the video sequences with time-varying quality. To this aim, set of experiments were conducted in two laboratories. The stimuli and experimental setups were held constant across the labs but the evaluation methodology was varied: in one lab a standardized test method (ACR) was applied and in the other one a semi-continuous method developed to evaluate the perceptual quality in long test sequences (CIETI). Moreover, the influence of audio presence on evaluation of the video-related HRCs was investigated. For this purpose, the CIETI method was employed in two different experiments: one by showing the video-only stimulus and the other by showing the stimulus in the presence of audio.

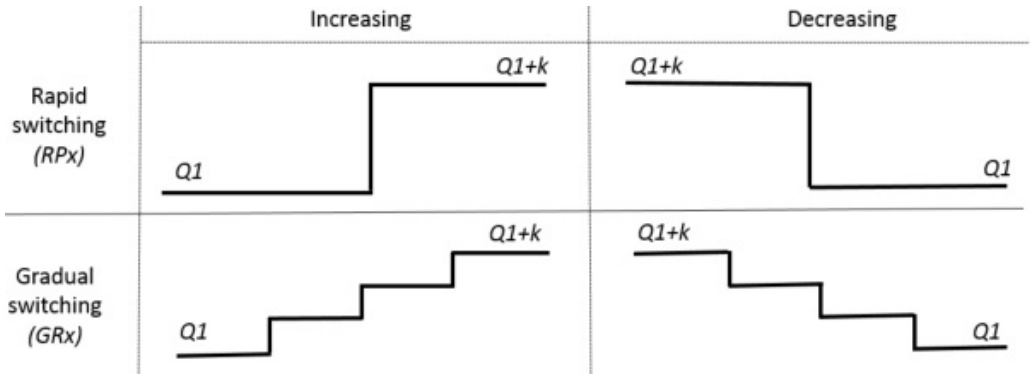


Figure 1: Quality switching pattern for rapid and gradual switching (RPx and GRx respecitvely where x stands on chunk size). *Q1* and *Q1+k* denote the quality levels (cf. Table 2).

2.2. Subjective Experiments

2.2.1. Test materials and conditions

Choosing among commercial content, seven source videos (SRC) of approximately six minutes long were chosen as listed in Table 1. They were originally 1080p picture size and 24 or 50 fps. The spatial and temporal information of the content (SI and TI in order) determined using the metric provided by [10] covered a large portion of SI-TI plane.

The adaptive streams were provided considering the encoding setting used in practice by the video streaming companies for the living-room platform. Following the recommendation of [3], the compression domain was considered as the switching dimension. For each SRC, four quality levels were produced using Rhozet Carbon Coder with the setting summarized in Table 2. It was assumed that the network bandwidth varies along these levels.

For each of the status when client should request from the server lower bitrate chunk (down-switching) or higher bitrate chunk (up-switching), the gradual and rapid way of switching were simulated following the patterns presented in Figure 1. For the chunk size, 2 sec and 10 sec length were considered to be inline with current HAS solutions. To study the perception of adaption streams in different content, four HRCs were considered representing the constant quality level. All HRCs are listed in Table 3.

To produce the test video, each SRC was segmented using Adobe Premiere Pro CS6. The segmented videos were subsequently considered as either Processed Video Sequences (PVS) or voting-segment (VS) . The PVS was made

Table 1: Source sequence description

Content Type	Description
Movie1	Action, adventure, fantasy; with some scene in smooth motion, some with group of walking people, some with camera panning
Movie2	Drama, music, romance; mostly with the smooth motion in the static background, some scene with group of dancing people
Movie3	Action, Si Fi, drama; with rapid changes in some sequences, cloudy atmosphere
Sport	Soccer, 2010 World cup final, average motion, wide angle camera sequences with uniform camera panning.
Documentary	Sport documentary; mostly with handheld shooting camera
Music	Music concert; high movement of the singer with some sudden scene change
News	Spanish news broadcast; some scenes with static shooting camera with one/two standing/sitting people, some outdoor scenes

Table 2: Trasncoding parameters of quality levels

Stream code	Framerate	Resolution	Target bitrate (kbps)
Q1	24	720p	600
Q2	24	720p	1000
Q3	24	720p	3000
Q4	24	720p	5000
Video: H.264, high profile, closed GOP, disabled scene change detection			
Ref. frame: 2, B frame: 2, CBR, adaptive QP			
Audio: AAC, 192 Kbps			

Table 3: List of the test adaptation strategies (HRCs)

Status	Possible behavior		HRC code
Increasing quality	Gradually	10 s chunk	IGR10
		2 s chunk	IGR2
	Rapidly	10 s chunk	IRP10
		2 s chunk	IRP2
Decreasing quality	Gradually	10 s chunk	DGR10
		2 s chunk	DGR2
	Rapidly	10 s chunk	DRP10
		2 s chunk	DRP2
Constant	Whole PVS at 5 Mbps		N5
	Whole PVS at 3 Mbps		N3
	Whole PVS at 1 Mbps		N1
	Whole PVS at 600 kbps		N600

by applying one of the HRCs on the segment (for the adaptation HRCs this was done by concatenating the chunks from different quality levels as presented in Figure 1). The subsequent segment with no degradation was considered as VS. In the same way, all HRCs were applied on the subsequent segments with the VS intervention. Because of the session time limitation and high number of the SRCs and HRCs, the full factorial design was not feasible[2]. To respect the ITU-T recommended test session length, four out of seven SRCs (Movie1, Movie2, Sport and Documentary) were used in two different variants. By this way, relevant amplitude degradations (i.e. comparing GRx and RPx, cf. Figure 1) and constant quality levels with potential non-perceivable difference (i.e. comparing N3 and N5, as well as N600 and N1) were compared in an individual video segment of the aforementioned content. As a result, 11 test sequences (TS), i.e. for each individual HRC 11 different video segment (4x2+3), and consequently the total of 132 PVSs (11x12) were generated for evaluation. Length of the PVSs were variable depending on the HRCs, 40 seconds for those considering the quality switching

[2]To have a full factorial designed experiment and understand the perceptual difference of the 12 HRCs, each of them had to be applied on every individual segment (cf. segments used to make PVS1, PVS2... in Figure 2) from all seven SRCs. This would lead providing 12 variants for each content and subsequently over 360 hours long test session.

with 10 seconds chunk, and 14 seconds for the rest of the HRCs.

2.2.2. Evaluation method and experimental setup

Three experiment were conducted evaluating the identical PVSs but through different approaches. The first experiment was conducted in Acreo Swedish ICT's lab (denoted as 'Acreo' experiment). The randomized order of the PVSs were presented to the test subjects following the ACR methodology adapted from ITU-T Rec. P.910. After presentation of each PVS, the subjects were asked to evaluate the PVS by answering two questions: the overall quality of the PVS (rating on the five graded ACR scale *Bad*(1), *Poor*(2), *Fair*(3), *Good*(4) and *Excellent*(5)), and if they perceived any change in the quality (options: *Increasing*, *No change*, *Decreasing*).

Two other independent experiments were carried out in Universidad Politecnica de Madrid's lab using CIETI method: one by presenting only the video stimulus (denoted as 'UPM-NoAudio' experiment), and the other one in the presence of audio[3] (denoted as 'UPM-Audio' experiment). Figure 2 shows an example of the TS used in UPM experiments. First segment in the TS had no degradation providing a reference of the video quality. '0' printed in the lower right corner of the frames indicated the start of the test. The following segments were the subsequence of PVS and VS. The VS frames had printed a number indicating the former PVS number which was also the box number in the paper questionnaire to rate by the test subject. In the test session, the 11 TSs (each including 12 sequential PVS-VS pairs) were presented in a random order. For the PVS evaluation, the test subjects were asked to answer the same questions as in Acreo experiment and using identical rating scales. As a new task, after evaluating the 12 PVSs of each TS there was another question in the questionnaire asking about the overall quality of the whole sequence. 40 sec after terminating the evaluation of each TS the next one was played.

In order to allow for cross-lab comparison, the ambient and all the hardware and software used in UPM were adjusted similar to Acreo both complying with the recommendation ITU-R BT.500-11 [9]. A 46" Hundai S465D display was used with the native resolution of 1920x1080 and 60 Hz refresh rate. The viewing distance was set to four times of display height. The TV's peak white luminance was $177 cd/m^2$ and the illumination level of the room was 20

[3]Quality of the audio stream was held constant during the playout.

Figure 2: Test sequence (TS) structure. In Acreo experiment, the randomized order of all PVSs was shown following ACR methodology. In UPM experiments, according to the CIETI methodology, the randomized order of TSs (including sequential PVS-VS) was presented such that the test subject was evaluating the PVS while continuously watching the 6 min video.

lux. The TSs (PVSs in Acreo experiment) were displayed in uncompressed format to assure that all observers were presented the same sequences. A computer connected to the TV was used to play the sequences using variable length codeword (VLC). In order to avoid any temporal distortion introduced by the player, the videos were preloaded into the computer's RAM. The TV resolution was set to the resolution of the test videos (720p) to avoid scaling when displaying the videos.

In all three experiments, prior to the test session the test subjects were screened for visual acuity and color vision. Later on, the test instruction and the rating scale provided in the observers' native language were given (mainly Spanish in UPM, Swedish in Acreo, and English for the international observers of each experiment). After reading the instruction, a training session was conducted by showing some TS/PVS samples specially prepared for training of each assessment method to familiarize the observers with the range of the qualities, quality variation and test procedure. The whole session was divided into three parts including two breaks of about 10 minutes when the subject was encouraged to leave the test room to minimize the fatigue effect on his evaluation. The total experiment lasted approximately 1 hr and 10 min.

After post-screening of the subjective data in accordance with the latest recommendations from Video Quality Experts Group [19], the scores of 23 observers from Acreo (7 female and 16 male, age from 18 to 68), 21 observers from UPM-Audio (6 female and 15 male, age from 27 to 50) and 22 observers from UPM-NoAudio experiment (5 female and 17 male, age from 24 to 54) were considered for the evaluation. In each experiment, about 80% of the subjects had telecommunication background (engineer, researcher, etc.) and 4 to 6 of them had subscription from a media service providers.

3. Results

Two sets of data including the scores for the 'perceptual quality' and the 'quality switching detection' were collected from each experiment and accordingly the Mean Opinion Scores (MOS) and 95% confidence interval (CI) of their statistical distribution were calculated. To investigate the impact of evaluation methodology on observers' assessment, first the MOS of the two UPM experiments were compared and later on the result was compared with Acreo experiment. Subsequently, the QoE of the adaptation strategies were investigated.

3.1. Cross-Experiment Comparison

3.1.1. UPM experiments: Impact of the audio presence

Figure 3 shows the difference between the MOS obtained from Audio and NoAudio experiments. It can be observed that in some PVSs (specially when increasing the quality in Sport, Music, and Movie2 which included scenes with dancing people) the presence of audio had positive influence up to 0.9 MOS value on evaluation of the observers. However, there are also few PVSs in other content (such as News) which perceived up to 0.6 MOS value better in NoAudio test. By computing the Pearson linear correlation coefficient between the MOS it was observed that the results of two experiments have a similar trend (correlation = 93%). Figure 4 shows this relationship having the results of NoAudio experiment on the x axis and the Audio experiment on the y axis. Considering the diagonal solid line as the main diagonal (indicating the ideal case in which both data sets would match to each other) and the dash-dash line as the regression mapping of Audio data to NoAudio, it is observed that the data have a small deviation upside of the reference which indicates the higher scores given in the Audio experiment compared to the other one. There was also slightly larger span in the MOS of NoAudio experiment (from 1.3 to 4.7) compared to Audio one (from 1,5 to 4,6).
To explore the significance of difference between the results of two experiments the repeated measure of ANalysis Of Variance (ANOVA) was performed on scores as the *dependent factor*, the experiment as one *between factor*, the 11 TSs and 12 HRCs as *within factors*, and $\alpha = 5\%$ as the *level of significance*. No significant difference was observed in the main effect of the experiments i.e. the audio presence ($p = 0.02$). The Tukey HSD post-hoc test also showed that there is no single PVS significantly perceived different

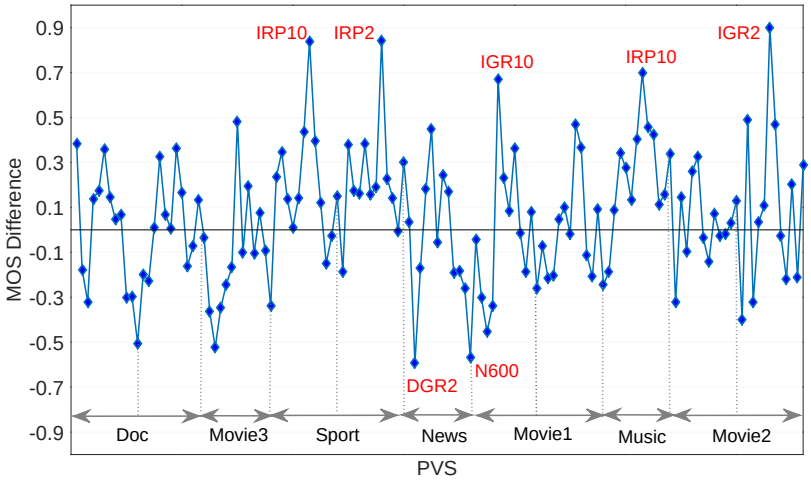


Figure 3: Difference between the MOS in Audio and NoAudio experiments (UPM)

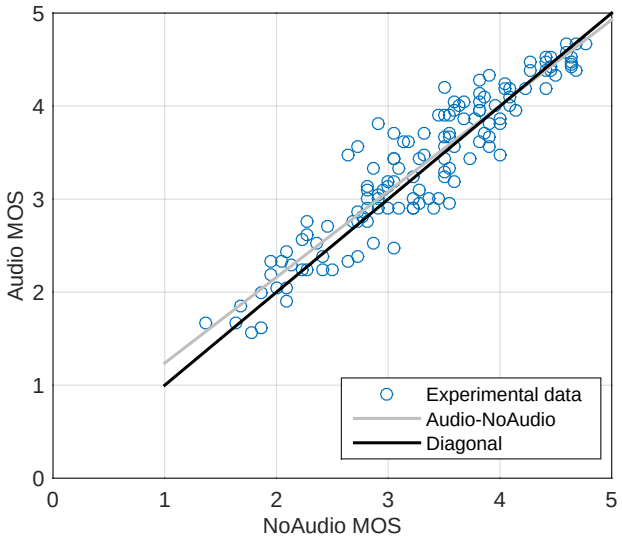


Figure 4: Relationship between Audio and NoAudio experiments

in two experiments. Therefore two data sets were combined by numerically averaging the MOS obtained (denoted as 'UPM' in the rest of the paper).

3.1.2. UPM vs. Acreo: Impact of the evaluation methodology

Figure 5 shows the difference between the MOS of Acreo and UPM experiments that in some of the PVSs is quite significant. From another side, the Pearson correlation between them shown in Figure 6 was 90%. The repeated

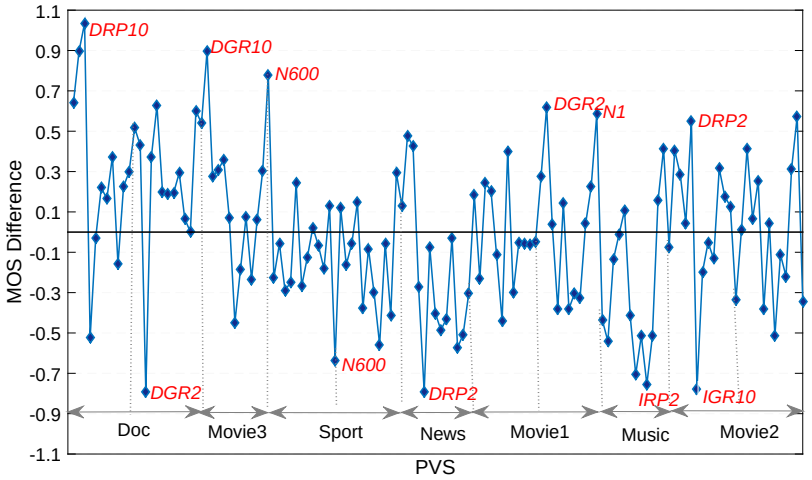


Figure 5: Difference between the MOS in Acreo and UPM experiments

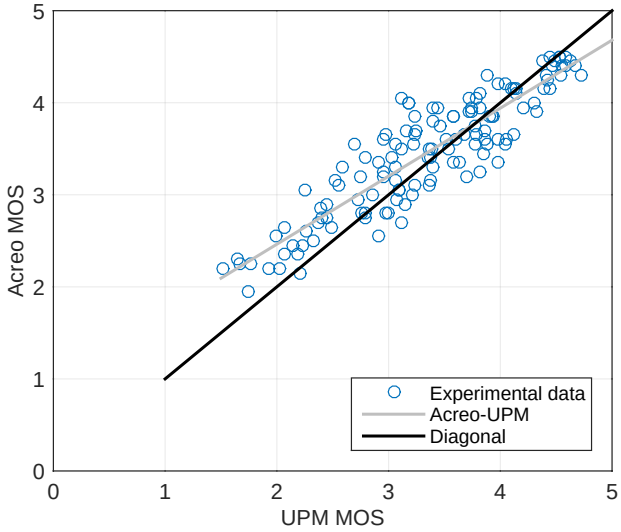


Figure 6: Relationship between Acreo and UPM experiments

measure of ANOVA applied on two data sets considering the same setting as the previous part revealed no statistically significant difference between the main effect of the evaluation methodology ($p = 0.02$). Considering all the pairwise Tukey test comparing the Acreo data once with the individual UPM experiments and later on with the combined UPM data showed no single PVS significantly perceived different in any of the cases. However,

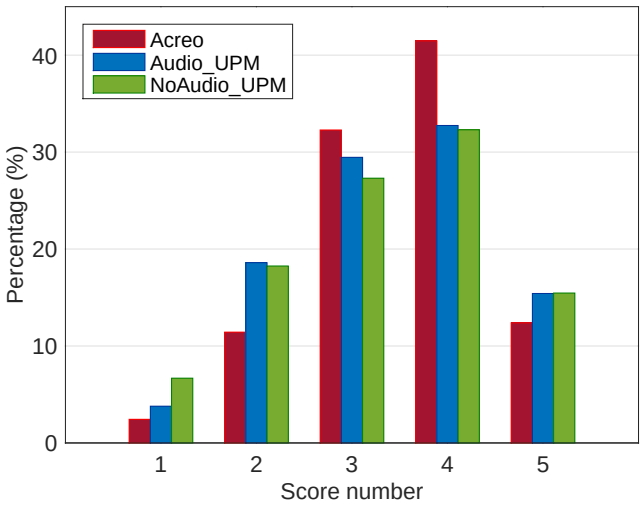


Figure 7: Observers' vote distribution

by comparing the distribution of the votes in two studies, it was observed that the usage of the voting scale was different in two labs (Figure 7). After applying the linear transformation of UPM's data to Acreo's using linear regression technique , the difference almost vanished and consequently the results of two studies were combined to be used as a single evaluation.

3.2. Analysis of the QoE of adaptation strategies

Figure 8 shows the overall quality of the adaptation strategies and constant quality levels, having the HRCs on the x axis and the average MOS of over all the content on the y axis. For the increasing and decreasing status, number 1 to 4 stands to GR10, GR2, RP10 and RP2, and for the constant status stands to N5, N3, N1 and N600 in order (cf. codes in Table 3). To statistically analyze the results, the repeated measure of ANOVA was applied on the data considering 11 TSs, 3 test status (increasing, decreasing, and constant), and 4 conditions considered for each status as within factors. The result showed that the perceptual quality of constant 5 Mbps (N5) and 3 Mbps (N3) encoded videos are not significantly different ($p = 1$) while both being significantly better compared to 1 Mbps (N1) and 600 Kbps (N600) encoded videos, and all increasing scenarios ($p < 0.001$). This outcome was also confirmed by the Tukey test ($p < 0.05$ for all pairwise-comparisons). Furthermore, the QoE of all increasing scenarios were significantly better than 1 Mbps and 600 kbps encoded videos ($p < 0.001$, also confirmed by the

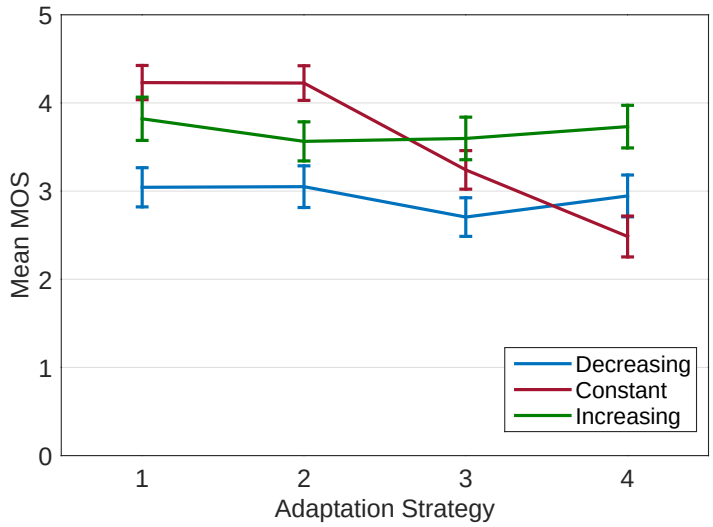


Figure 8: Overall behavior of the adaptation strategies. Number 1 to 4 stands to GR10, GR2, RP10, RP2 for the *decreasing* and *increasing* quality status, and to N5, N3, N1, and N600 for the *constant* quality state in order (cs. Table 3).

post-hoc test). By comparing the MOS of the increasing scenarios together, no difference was observed as the corresponded CIs were clearly overlapped. However, the post-hoc test revealed that the perceptual quality of IGR2 and IRP10 (interrelationship: $p = 0.99$) are significantly lower than IGR10 and IRP2 (interrelationship: $p = 0.5$) with $p < 0.001$. Regarding the decreasing scenarios, both ANOVA and post-hoc test showed the significantly lower QoE of DRP10 compared to the other scenarios ($p < 0.001$).

3.2.1. Perceptual quality and detection of adaptation in different content

Figure 9 presents the QoE of adaptation strategies in different content (for those SRCs examined in two variants the average MOS obtained from two variants is presented). It can be seen that in general the quality switching were perceived differently in different content (cf. *Mean* in the figures) and this is more significant in Sport content. Considering the impact of amplitude and period of the quality switching and their interaction, although in most of the content no significant difference can be observed, in some of them those observations presented in the previous part (i.e. low QoE in IGR2, IRP10 and DRP10; cf. Figure 8) are highlighted.
Figure 10 shows the QoE comparison of switching amplitude in the identical video segment of those content examined in two variants. The labels in the

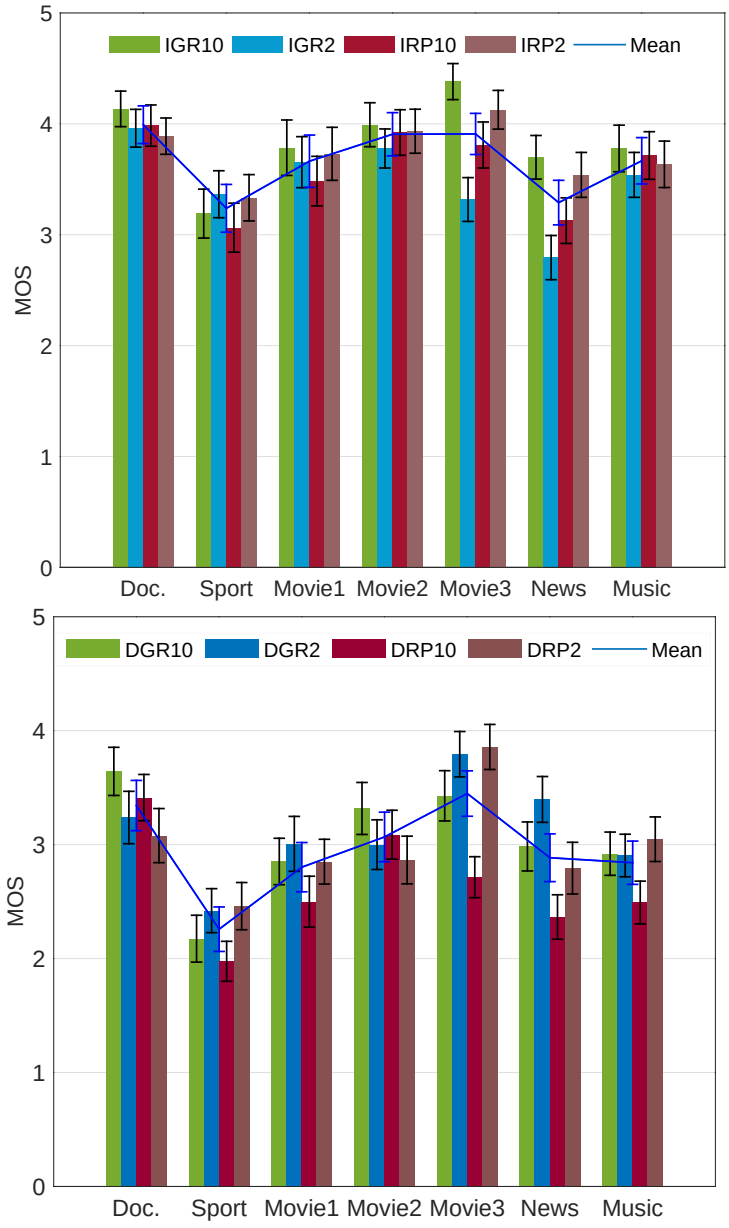


Figure 9: Perception of up-switching strategies (figure above) and down-switching strategies (figure below) in different content

figures (e.g. Doc-1, Doc-2, Doc-3 and Doc-4) represent distinct segments from the main source content. Right part of the figures shows the switching

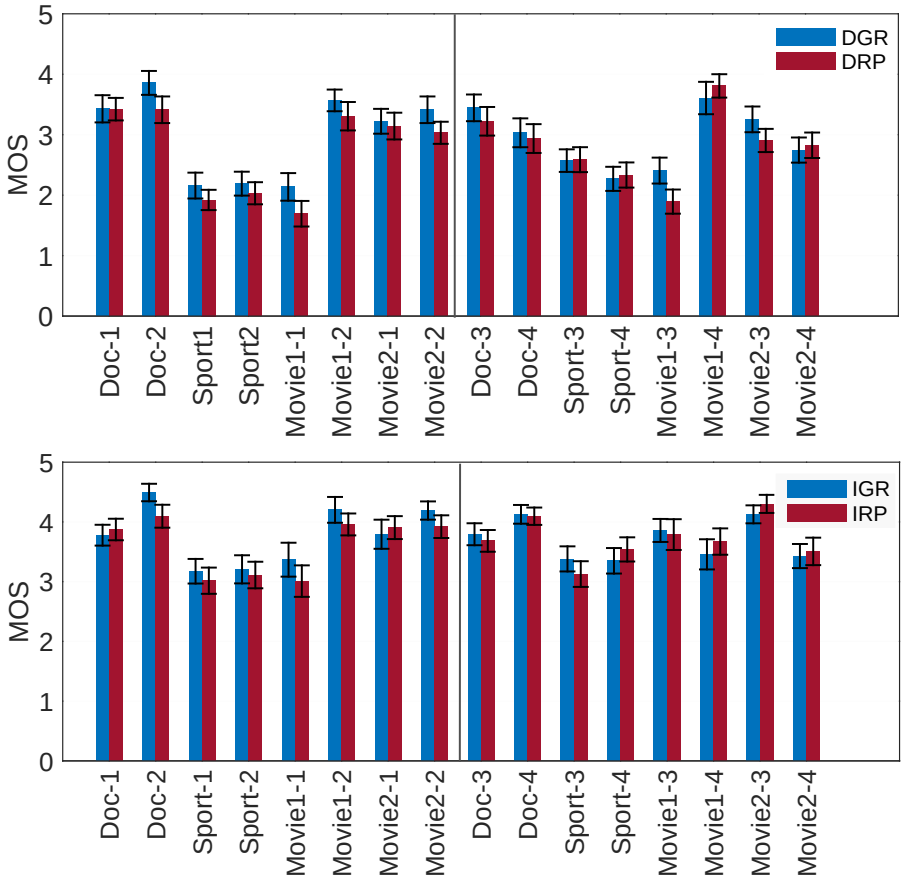


Figure 10: Perceptual impact of the switching amplitude (gradual,vs. rapid) when decreasing (figure above) and increasing (figure below) the quality using small chunk (right side) and large chunk (left side) in different content

scenarios using 2 sec chunk while the left part shows those using 10 sec chunk. In both cases of up-switching and down-switching it can be observed that the switching amplitude does not have significant influence on users perceptual quality. However, in most of the content we can find a trend showing the better quality of the gradual switching when using the 10 sec chunk, but no specific trend can be found when using the 2 sec chunk. One possible reason could be that step-wise changing the quality every 10 sec lets users to get used to the presented quality, and because of the human short-term memory effect as defined in [20], they get less annoyed by changing the quality from two consecutive level. However, this cannot be the case

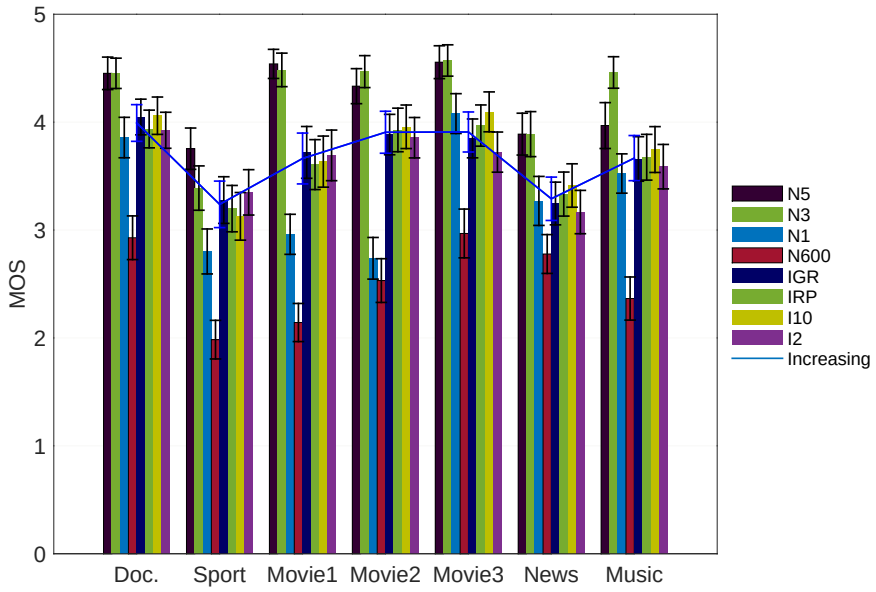


Figure 11: Perception of increasing strategies vs. constant quality. The data labeled as 'Increasing' represents the average of all increasing scenarios.

when the quality changes every 2 sec. On the other hand, characteristic of the content could be another reason of obtaining different observations from the two aforementioned test scenarios. Because the 'video segment' utilized for evaluating the quality switching using 10 sec chunk were different than the one using 2 sec chunk (cf. Figure 2). Figure 11 shows the MOS in up-switching strategies and those PVSs encoded in constant quality. First of all, it can be observed that the PVSs with the constant quality were perceived differently so that in some content the quality gets extremely detracted when encoded in N1 and N600 (e.g. extreme MOS reduction in N1 perceived in Movie1 and Movie2). Another interesting observation was the significantly better QoE of N3 compared to N5 in Music. By exploring in the corresponded results achieved from three experiments it was observed that the QoE of N5 obtained from Acreo experiment was about 0.9 MOS value lower than Audio and 0.25 value lower than NoAudio experiments. One possible reason could be the difference in the context of the video segment in which N5 and N3 were applied (the former video segment mostly shows the audience in the dark scenes with the smoke on the air which could be perceived as an impairment, while the later one showing the singing group in more bright ambient) which

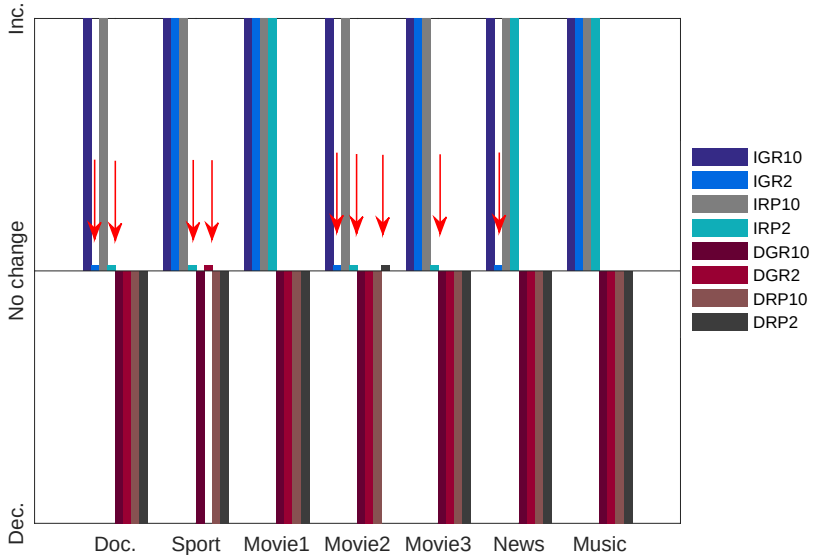
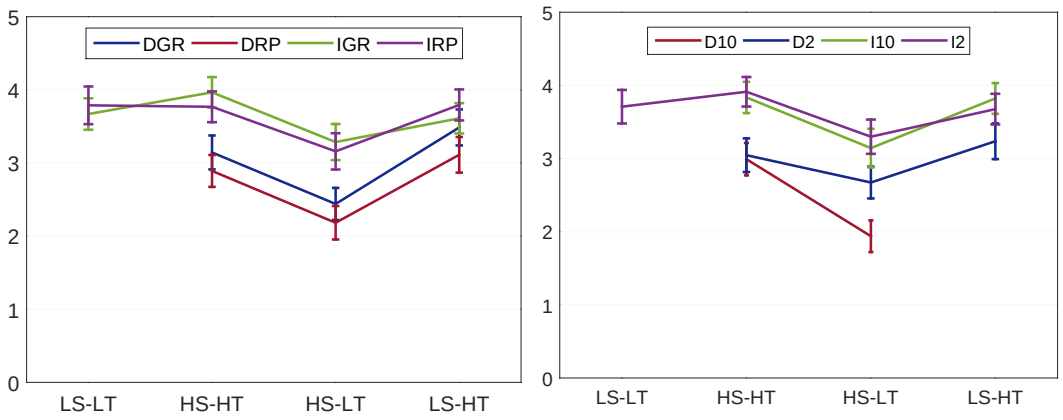


Figure 12: Detection of the quality change in the videos including quality variation. Red arrows show those switching scenarios voted as 'no change'.

can affect more in the single stimulus presentation of the PVS (Acreo). From this figure it can be also observed that in some of the content (Documentary, Movie3, News and Music) increasing the quality does not provide better perceptual quality compared to the constant 1 Mbps encoded video (and of course to the constant 600 Kbps). The MOS related to detection of quality change in the PVSs including quality switching scenarios is presented in Figure 12. It can be seen that finding the quality switching is also different in different content. Another interesting observation was found about the switching scenarios using 2 sec chunk where in some of the content were voted as 'no change'. In an overall view (considering the Mean MOS for all the content) however, the test participants accurately identified the sort of quality variation/stability in all the HRCs.

3.2.2. Impact of spatiotemporal characteristics of the content in QoE of adaptation

Previously in [21] the impact of content characteristics on the user's perception of our test switching scenarios was studied. This was done by analyzing the MOS of adaptation scenarios in the PVSs which were classified by a combination of their amount of spatial-temporal (ST) complexities as formulated in [10]. Subsequently four PVS classes were resulted: low spatial-



(a) Increasing scenarios (b) Decreasing scenarios

Figure 13: Impact of the spatiotemporal characteristics of the content on perception of switching amplitude (left) and chunk size (right). The content classes and MOS are presented in x- and y-axis in order.

low temporal (LS-LT), low spatial- high temporal (LS-HT), high spatial- low temporal (HS-LT), or high spatial- high temporal activity (HS-HT). The statistical analysis of MOS showed that the QoE of both up- and down-switching in the content from HS-LT class is significantly lower compared to other classes. In the current study we extended this idea by analyzing the impact of content characteristics on perception of different amplitude and period of quality switching. The results obtained from existing content classes presented in Figure 13 showed that no matter about the amplitude and period of quality switching, the impact of adaptation in aforementioned content class is significantly lower than the content from other classes.

3.2.3. Overall quality of the adapted test sequences

Table 4 shows the correlation between the MOS about the whole TSs voted by the subjects at the end of evaluating the 12 PVSs in the UPM experiments, and the mean MOS of the last 5 and also all 12 PVSs in each of the sequences. It can be seen that the MOS about the quality of entire sequence was highly correlated with the average of 12 PVSs' scores in the sequence but far less correlated with the average of last 5 PVSs' scores. For both cases, the correlation is lower in Audio experiment. On the other hand, it was observed that the scores gave for the quality of whole sequences in Audio experiment were up to 0.45 MOS value more than the ones obtained

from NoAudio experiment (in 8 out of 11 sequences).

Table 4: Correlation between the MOS of whole seq. and mean MOS of the PVS, last 5 PVS in each sequence (UPM experiments)

Experiment	Whole seq. vs. all PVS	Whole seq. vs. last 5 PVS
Audio	0,95	0,79
NoAudio	0,99	0,89

4. Discussion

One of the goals of this study was to investigate subject testing methodologies to evaluate sequences with time-varying quality. To this aim, we ran two experiments using the CIETI methodology developed for evaluating the long sequences to make mimicking the attribute of mentioned degradation possible. To study the impact of audio on evaluation of the test stimuli, one of the experiments was done in the presence of audio. It was observed that in some of the content (such as Sport, Music and one of the movie content including scenes with dancing people) the audio presence had positive effect up to 0.9 MOS value on the test subjects evaluation. Whilst a negative impact of audio presence up to 0.6 MOS value (in News) was also observed. One possible reason of lower subjective rating for this video could be the context of the presented news. As it has been addressed by previous studies (e.g. [22]), the video context appeals to different psychological process (understanding, desire, engagement) which results in complex interaction with the users' perception. Here the presented news was about the usage of drugs by the young generation that perhaps in the presence of audio was more effectual for the observers.

In spite of the observed differences in two experiments, the high Pearson correlation (0.93) between the experimental results as well as the no significant difference resulted from ANOVA and post-hoc tests verified that two experiments can be combined. In addition, no large difference in the range of MOS values of two studies was observed (Figure 7). This finding was different than what was previously discussed in [15] in regard to the impact of changing from video-only stimuli to audiovisual stimuli on our ability to distinguish between HRCs.

Comparing the UPM combined results with the results of Acreo, our previous study using ACR method, also indicated a relatively high correlation (0.9) and their non-significant difference was verified by ANOVA and this time also all the post-hoc tests of the same PVS in both cases. Because of difference in the usage of voting scale in two studies, the Acreo data was linearly transformed to UPM's. The average of UPM and transformed Acreos data were used to evaluate the switching strategies.

As mentioned, in spite of some large differences between MOS obtained from different methodologies, the result of statistical analysis did not show any significant difference between three experiments. This does not rule out that the different methods can lead to different results, but with the number of subjects involved and the statistical variance in the data the differences in the MOS values of the same PVSs were not big enough to show any statistically significant difference.

By statistically analyzing the combined results, it was observed that the perception of adaptation scenarios and also constant encoded test videos are different in different content. As it was observed in Music content, where the significantly better QoE of N3 was shown in comparison to N5 (cf. 11), the context of the video sequence could be an influence factor in this regard. From another side, this content dependency was also observed in detection of quality switching by the test subjects.

In respect to content dependency, since subjective testing methods as investigated here and in Pinson et al. [15] are not full matrix, this impact becomes more difficult to analyze. In principal studying the impact of the content is not the goal in these types of investigations, but rather gets the impact of degradations in general. However, it may still be interesting to get some indication of the impact of the content on the obtained quality ratings. The influence of spatial-temporal complexities of the content in perceptual quality of the adaptation was analyzed in [21]. By studying the QoE of the PVSs which were classified based on their ST complexities, it was observed that the perceptual quality of adaptation in the content with high spatial and low temporal complexities is significantly lower compared to the other content. Since ST classification of the videos were done after the experiments, the number of content in each class was not homogenous and in some cases not enough to elaborate our analysis beyond all the test conditions. Even so, by investigate among those existing samples which were the two variables of identical SCR but from distinct content classes (i.e. the variable 1 and variable 2 of the SRCi, where HRCj was applied on, had different ST

characteristics) it was observed that the amount of temporal complexity can cause up to 1.6 point improvement in MOS. This observation was mostly highlighted in the up- and down-switching using 10 sec chunks.
The result showed that the MOS of the entire sequence was highly correlated with the average of individual PVSs MOS in the sequence but far less correlated with the average of last 5 PVSs MOS. For both cases, the correlation was lower in the Audio experiment. Furthermore, it was observed that the scores gave for the quality of whole sequences in Audio experiment were up to 0.45 MOS value more than the ones obtained from NoAudio experiment. However, it may not be ruled out the subjects were affected by scoring the individual cases when giving their overall score.
It is worth to remind that the duration of PVSs was either 14 sec or 40 sec. Such a considerable difference on PVSs duration could affect the evaluation of the subjects, especially in Acreo study where the subjects watched the PVSs in single stimulus event, while in UPM study because of the continuous presentation of the video sequences, this issue might have had less impact. We believe, although the study could not show it, to assess the quality of adaptation events which in practice may last differently in different conditions, it is better to apply a continuous evaluation method rather than a clip by clip way like the ACR where the time difference easily noted by the subject.

5. Conclusions

In this study, we have compared two subjective testing methodologies, the standardized ACR method and CIETI semi-continuous method suggested by Gutierrez et al. [14] using video-only and audiovisual stimuli in two separated experiments; so in total three subjective experiments were performed. Through these experiments we analyzed the impact of some important technical parameters for adaptive streaming services.
On the comparison between the subjective testing methodologies, no significant effect was found between the experimental methods, that is the main effect was not significant and not any PVS was significantly differently rated by the two methods. The correlation was high (0.9), which is, however, slightly lower than expected when repeating the same experiment twice with the same method and condition. Based on this, the subjective data was merged into one set for facilitating the analysis of the technical parameters.
The key findings of the analysis of the technical parameters were:
- Overall there was a statistically significantly higher QoE for the two in-

creasing scenarios, IGR10 and IRP2, and significantly lower QoE for the decreasing scenario DRP10.
- The constant bitrate cases N1 and N600 had significantly lower QoE compared to increasing the quality by adaptation.
- We could not find any significant difference of the quality between the high constant bitrate cases, N5 and N3, which were both better than N1 and N600 as well as the increasing scenarios.
- Considering different results above, no main effect (from ANOVA and Tukey HSD post-hoc test) was obtained for any of the adaptation-related study factors. In other word, in contrary to many of previous studies, using short chunk or abrupt quality switching do not necessarily degrade the QoE.
- The content had an influence on the absolute MOS results especially in comparison between N1, N600, increasing scenarios, and switching frequency. However regarding switching amplitude (smooth vs. abrupt) no significant differences between different test contents could be found.
- Very strong correlation was found in the no-audio test: 99% between the overall quality of a whole 6 min video clip and mean of all the individually scored PVSs MOS, that the video clip consisted of. This correlation for the audio test was slightly lower (95%).
We believe that this study give useful results for optimizing adaptive streaming services. Although we did not find any statistical difference between the experimental methodologies, we think this would need to be investigated further involving a larger number of test subjects in order to be able to statistically infer about any differences between the methods.

6. Acknowledgment

Many thanks go to Kun Wang and Börje Andrén from Acreo Swedish ICT, and Muhammad Shahid from Blekinge Institute of Technology involved in design, execution and evaluation of the subjective tests. The work at UPM has been partially supported by the Ministerio de Economía y Competitividad of the Spanish Government under projects TEC2010-20412 (Enhanced 3DTV) and TEC2013-48453 (MR-UHDTV). The work at Acreo Swedish ICT AB was supported by VINNOVA (Sweden's innovation agency), which is hereby gratefully acknowledged.

References

[1] T. Arsan, Review of Bandwidth Estimation Tools and Application to Bandwidth Adaptive Video Streaming, in: Proc. 9th International Conference on High-Capacity Optical Networks and Emerging/Enabling Technologies (HONET 2012), Istanbul, Turkey, 2012.

[2] M. Zink, J. Schmitt, R. Steinmetz, Layer-encoded Video in Scalable Adaptive Streaming, IEEE Transactions on Multimedia 7 (1) (2005) 75–84.

[3] P. Ni, R. Eg, A. Eichhorn, C. Griwodz, P. Halvorsen, Flicker Effects in Adaptive Video Streaming to Handheld Devices, in: Proc. 19th ACM International Conference on Multimedia, MM '11, Scottsdale, AZ, USA, 2011, pp. 463–472.

[4] A. K. Moorthy, L. Choi, A. C. Bovik, G. de Veciana, Video Quality Assessment on Mobile Devices: Subjective, Behavioral and Objective Studies, IEEE Journal of Selected Topics in Signal Processing 6 (6) (2012) 652–671.

[5] R. Mok, X. Luo, E. Chan, R. Chang, QDASH: A QoE-aware DASH System, Proc. 3rd Multimedia Systems Conference (2012) 11–22.

[6] L. Yitong, S. Yun, M. Yinian, L. Jing, L. Qi, Y. Dacheng, A study on Quality of Experience for adaptive streaming service, in: IEEE International Conference on Communications Workshops, 2013, pp. 682–686.

[7] M. Zink, O. Künzel, J. Schmitt, R. Steinmetz, Subjective Impression of Variations in Layer Encoded Videos, in: Proceedings 11th International Conference on Quality of Service, 2003, pp. 137–154.

[8] D. C. Robinson, Y. Jutras, V. Craciun, Subjective video quality assessment of http adaptive streaming technologies, Bell Labs Technical Journal 16 (4) (2012) 5–23.

[9] ITU-R, Methodology for the Subjective Assessment of the Quality of Television Pictures, ITU-R Recommendation BT. 500 (Jan 2012).

[10] ITU-T, Subjective Video Quality Assessment Methods for Multimedia Applications, ITU-T Recommendation P.910 (Apr 2008).

[11] N. Staelens, S. Moens, W. Van den Broeck, I. Marien, B. Vermeulen, P. Lambert, R. Van de Walle, P. Demeester, Assessing Quality of Experience of IPTV and Video on Demand Services in Real-Life Environments, Broadcasting, IEEE Transactions on 56 (4) (2010) 458–466.

[12] C. Chen, L. Choi, G. de Veciana, C. Caramanis, R. Heath, A. Bovik, Modeling the Time Varying Subjective Quality of HTTP Video Streams with Rate Adaptations, IEEE Transactions on Image Processing 23 (5) (2014) 2206–2221.

[13] S. Tavakoli, J. Gutierrez, N. Garcia, Subjective Quality Study of Adaptive Streaming of Monoscopic and Stereoscopic Video, IEEE Journal on Selected Areas in Communications 32 (4) (2014) 684–692.

[14] J. Gutierrez, P. Perez, F. Jaureguizar, J. Cabrera, N. Garcia, Subjective Assessment of the Impact of Transmission Errors in 3DTV Compared to HDTV, in: 3DTV Conference: The True Vision - Capture, Transmission and Display of 3D Video (3DTV-CON), 2011, pp. 1–4.

[15] M. Pinson, M. Sullivan, 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.

[16] M.-N. Garcia, F. De Simone, S. Tavakoli, N. Staelens, S. Egger, K. Brunnström, A. Raake, Quality of Experience and HTTP Adaptive Streaming: a Review of Subjective Studies, in: 6th International Workshop on Quality of Multimedia Experience (QoMEX), 2014.

[17] L. Janowski, P. Romaniak, Qoe as a function of frame rate and resolution changes, in: Proceedings of the 3rd International Conference on Future Multimedia Networking, FMN'10, Springer-Verlag, Berlin, 2010, pp. 34–45.

[18] P. Kortum, M. Sullivan, The effect of content desirability on subjective video quality ratings, Human Factors: The Journal of the Human Factors and Ergonomics Society 52 (1) (2010) 105–118.

[19] VQEG, Report on the Validation of Video Quality Models for High Definition Video Content, Video Quality Expert Group, Available: www.vqeg.org (June 2010).

[20] T. Hosfeld, S. Biedermann, R. Schatz, A. Platzer, S. Egger, M. Fiedler, The memory effect and its implications on web qoe modeling, in: Teletraffic Congress (ITC), 2011 23rd International, 2011, pp. 103–110.

[21] S. Tavakoli, M. Shahid, K. Brunnström, N. Garcia, Effect of Content Characteristics on Quality of Experience of Adaptive Streaming, in: 6th International Workshop on Quality of Multimedia Experience (QoMEX), 2014.

[22] G. Ghinea, J. P. Thomas, Qos impact on user perception and understanding of multimedia video clips, in: Proceedings of the Sixth ACM International Conference on Multimedia, MULTIMEDIA '98, ACM, New York, NY, USA, 1998, pp. 49–54. doi:10.1145/290747.290754.
URL http://doi.acm.org/10.1145/290747.290754