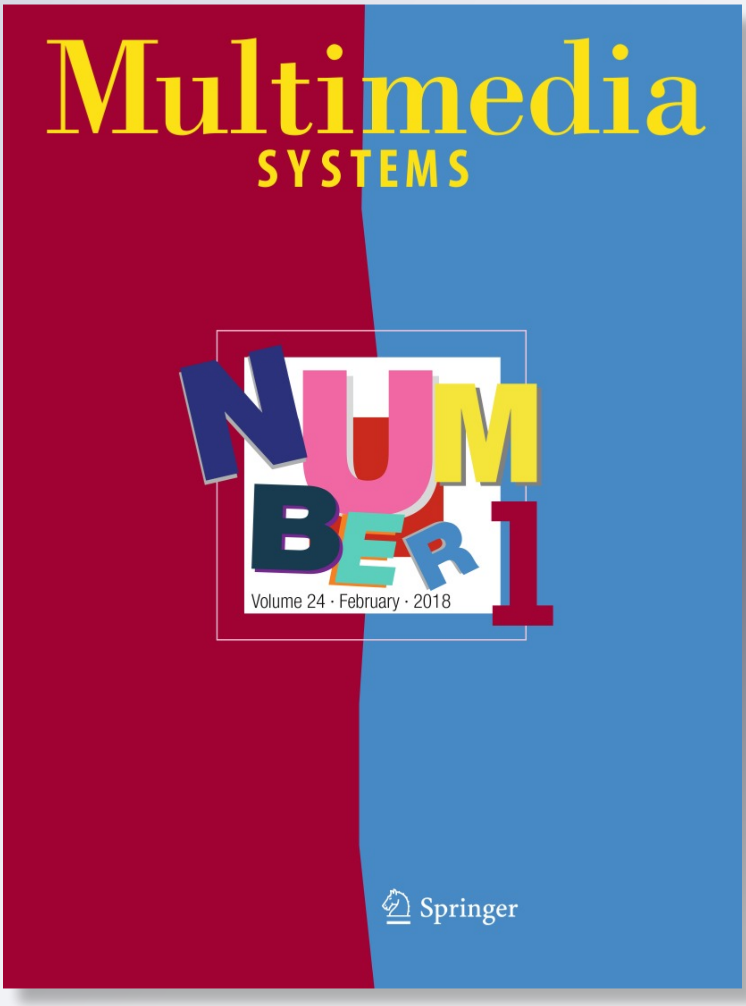
Impact of packet loss on the perceived quality of UDP-based multimedia streaming: a study of user quality of experience in real-life environments

Štefica Mrvelj & Marko Matulin

Multimedia Systems

ISSN 0942-4962
Volume 24
Number 1

Multimedia Systems (2018) 24:33-53
DOI 10.1007/s00530-016-0531-8



Multimedia Systems (2018) 24:33–53
https://doi.org/10.1007/s00530-016-0531-8

REGULAR PAPER

Impact of packet loss on the perceived quality of UDP-based multimedia streaming: a study of user quality of experience in real-life environments

Štefica Mrvelj[1] · Marko Matulin[1]

Received: 2 February 2016 / Accepted: 1 October 2016 / Published online: 12 October 2016

Abstract Multimedia content delivery has become one of the pillar services of modern day mobile and fixed networks. The variety of devices, platforms, and content providers together with increasing network capacity has impacted the popularity of this type of service. Considering this context, it is crucial to ensure end-to-end service quality that can fulfill users' expectations. The user quality of experience (QoE) for multimedia streaming is tempered by numerous objective and subjective parameters; therefore, it is important to understand the relationships among them. In this paper, we thoroughly examine the impact of packet loss on user QoE in cases when multimedia streaming service is based on underlying User Datagram Protocol. The dependencies between the chosen objective and subjective parameters and the user QoE were examined in a real-life environment by conducting a survey with 602 test subjects who rated the quality of a 1-h documentary film (72 different test sequences were prepared for the rating process). Based on the obtained results, we ranked the objective parameters by their order of importance in relation to their impact on user QoE as follows: (1) total duration of packet loss occurrences (PLOs), i.e., quality distortions in a video; (2) number of PLOs; (3) packet loss rate; and (4) duration of a single PLO. We also demonstrated how the overall user experience can be redeemed, despite the perceived quality distortions, if the content is entertaining to the viewer. The user experience was also found to be influenced by the existence/non-existence of video subtitles.

Keywords Quality of experience · Multimedia · Packet loss · User perception · Real life · Video subtitles

1 Introduction

Understanding the relationship between achieved network performance and user perception regarding the quality of a specific service remains a paramount objective for network operators and service providers. This knowledge is useful when attempting to improve network efficiency, reduce operating costs, and maintain certain levels of user satisfaction. In the research efforts undertaken after the advent of the quality of experience (QoE) concept, several authors found that real-life subjective evaluations of service quality produce different results when compared with evaluations conducted in a controlled environment. Primarily, results obtained in real life indicate that users are more forgiving to the temporal distortions of service quality. Considering these differences, it is not unexpected to find that certain authors question the usability of laboratory testing (see [1, 2]), since the user perception provides important information for determining the performance targets of different applications, i.e., quality of service (QoS) demands. From this perspective, it is important to highlight the findings of Kaikkonen et al. in [1] who conclude that the results obtained in the artificial, laboratory environments may suggest that a specific service needs higher QoS demands than it is actually the case. Therefore, it is worth conducting subjective tests of QoE in real-life environments and in situations that reflect everyday service usage scenarios [3].

Communicated by M. Claypool.

✉ Marko Matulin
mmatulin@fpz.hr

[1] Department of Information and Communications Traffic, Faculty of Transport and Traffic Sciences, University of Zagreb, Vukeliceva 4, 10000 Zagreb, Croatia

Reichl et al. in [4] discussed one of the first real-life subjective tests of QoE. The authors installed two cameras on the hat of a female participant. The cameras recorded the facial expressions of the woman as she used a mobile video streaming application during her daily routine. Later, the stored video was analyzed to determine the degree of enjoyment, frustration, boredom, etc. of the woman. In [5], the authors analyzed the same type of application for two groups of test subjects. The subjects in the first group used the application in real-life conditions (e.g., in a train station or on a bus). The second group of test subjects viewed the same short video clips under the same network conditions but in a controlled environment. The results showed that the first group of users did not notice as many impairments as the second group. These results were later confirmed by Staelens et al. in [6, 7], who tested user QoE using full-length movies. Their findings confirmed that the user's environment has a high potential to significantly affect the evaluation results.

The results presented in [8] indicate that different sets of mobile applications were used by users in the morning, in the evening, in the car, and outside of the office. In addition to the real-life environment, the authors concluded that the user rating is influenced by the importance of the mobile application to the task at hand. In addition, they showed that users who use a specific type of application on a computer tend to poorly rate the mobile version of that same application, i.e., previous experience significantly affects their QoE of mobile applications. Van den Broeck et al. in [9] analyzed the quality of the video stream of the Koksijde City Council meetings. Test subjects were asked to watch live streams of the meetings from their homes and rate the video quality. However, because the authors did not have information about the network performance during the multimedia streaming sessions, they were unable to correlate the QoS parameters experienced by the users with their QoE scores.

Subjective multimedia quality assessment procedures have certain limitations, because they are time-consuming and expensive to conduct [10]. This is especially relevant for real-life testing where an additional challenge arises: how to deliver content to the subjects and collect the rating data. As reported in [11], to overcome these limitations, several authors conducted real-life subjective service quality evaluations through remote assessors using the Internet. This approach is called *QoE crowdtesting*. For this purpose, commercial platforms [12] and social networks [13] can be used. In [14], the authors outlined several benefits of QoE crowdtesting, including (a) the reduction in costs and time needed for testing, (b) the ability to survey wide and diverse panels of test subjects, and (c) the use of real-life testing conditions. However, during QoE crowdtesting, the content first has to be downloaded on the subject's devices (e.g., while they are providing demographic information) and then, it is played locally. Hence, this study format is inadequate if the full-length videos are used for the evaluation of user QoE, because the content can be several gigabytes in size.

The alternative approach is to remotely obtain information about the network performance of end users, while they are watching the multimedia content that is streamed to their devices. An application capable of monitoring performance for end users can be developed for this purpose and installed on the devices used by the test subjects, as implemented in [8] or proposed in [15]. Nonetheless, it remains more difficult to conduct testing on a larger target group, because the test subjects must be convinced to install the application on their devices (the subjects must be assured about the harmless intentions of the researcher when any type of monitoring application is installed). Furthermore, similar to QoE crowdtesting, when the test is done using the full-length videos, it may be challenging to pursue the subjects to participate in the survey, since they have to stream several gigabytes of data to their devices.

Another possible solution is presented by Staelens et al. in [16], where the authors implemented a subjective quality assessment methodology into the application used for presenting video content on mobile devices. The authors were the owners of the mobile devices, and the prepared test sequences were stored locally on the devices. During the test period, the devices were provided to the test subjects who were instructed to watch the sequences in real life and rate its quality directly on the device via the application. Later, the authors collected the devices and the rating data for the analysis. This approach solves the majority of issues discussed above, because the content is not streamed or downloaded by the subjects and the rating data can be stored on the devices. However, this method reveals the purpose of the test to the subjects and is not a feasible solution for large-scale surveys such as this study.

In [17], Ickin et al. continued their previous work presented in [8] and developed a QoE evaluation methodology for Android-based smartphones. The methodology was used for user QoE analysis for mobile video streaming. The authors added functionalities to open source VLC Media Player, namely, the user interface of the player was upgraded to accommodate the QoE rating scale and a "freeze" button. During the streaming sessions, the button was pressed by the test subjects when they wanted to indicate noticing the frame freeze video artifact. Note that the study also assumed streaming of the content to the subjects' devices for rating and, similar to [16], the used methodology revealed the purpose of the test to the subjects.

To overcome these challenges, in this study, the test sequences were prepared in an emulated network environment where different packet loss rates (PLRs) were set during streaming sessions. The sequences of different qualities were distributed on a DVD to test subjects who were unaware about the purpose of the test. They were only asked to watch the DVD in the environment where they usually watch TV programs and to open a sealed envelope (containing the questionnaire) after the screening. Hence, a methodology similar to that from [7] was used for the preparation of the test sequences and their distribution to the subjects. However, this study had several distinguishing features as follows: (a) a considerably larger number of test sequences with different properties were produced and evaluated; (b) in addition to the PLR, the impact of the number of packet loss occurrences (PLOs) and their duration on user QoE was analyzed; (c) a set of subjective factors, such as user annoyance, level of entertainment, social context and user fatigue, was analyzed; and (d) the impact of video subtitles on user QoE was investigated.

The objective of this study was to thoroughly examine the impact of packet loss related issues on user QoE in cases when 1-h multimedia content is streamed using underlying User Datagram Protocol (UDP). We strived to disclose how different PLRs, number of PLOs, and their total duration correlate with the level of user annoyance and QoE. This knowledge will be used in our future research when we will try to develop no reference objective video quality assessment model for assessing the user QoE.

The remainder of this study is structured as follows. Section 2 describes the process used to create the test sequences and the method used for the subjective evaluation of video quality. The evaluation results are presented and discussed in Sect. 3, and the conclusions and future work are outlined in Sect. 4.

2 Research method

2.1 Properties of test stimuli

In [6] and [7], the authors used full-length movies to evaluate user QoE for video streaming services, because the users of internet protocol television (IPTV) or video on demand (VoD) services typically watch videos that last longer than the test sequences used for experiments performed in controlled environments. Furthermore, several researchers have found that when using short video clips, the evaluation of the QoE does not often match the real-life quality perception, i.e., in real life, it is necessary to increase the duration of the test sequences [18, 19]. Hence, the video content used in this research was a 1-h documentary film about the solar system. The video was encoded using advanced video coding (H.264/AVC) and advanced audio coding (AAC). The video was coded at a bit rate of 9.8 Mbps and a frame rate of 50 fps. The resolution of the video was 1920 × 1080 pixels; the audio was coded at a bit rate of 256 kbps.

The video was streamed in an emulated network environment between two computers using the UDP on the transport layer. This type of streaming differs from the Hypertext Transfer Protocol (HTTP)-based streaming which uses the Transport Control Protocol (TCP) as the underlying transport protocol. Nowadays, the HTTP-based adaptive streaming is the relevant scenario in practice; however, UDP-based streaming is still used for delivering live multimedia content as well as IPTV, especially for those services that use set top boxes.

During the streaming sessions, PLRs of 0.05, 0.1, 0.5, 1, 1.5, and 2 % were introduced using the emulator client (the burst packet loss length was set to 1). Six incoming video signals, each completely affected by different PLRs and containing video artifacts (jerkiness, frame freeze, blurring, blocking, error blocks, object persistence, edge busyness, and mosquito noise), were stored in the same format as the original video. To test the impact of the decreases in network performance on the user experience, 1, 4, 7 or 10 short video clips from a degraded video signal were inserted into the original video signal. The duration of a single inserted clip, i.e., a single PLO, varied between 1, 4, and 7 s. Variations in these three objective parameters allowed for the creation of 72 different test sequences (Table 1), whose quality was rated by the test subjects.

Since we adopted the full-length movie quality assessment methodology defined in [7], the inserted clips, i.e., PLOs, were evenly distributed over the entire duration of all test sequences. We did not experiment with different distributions of PLOs; thus, their distribution in the sequences was somewhat deterministic. The main reason for even distribution of PLOs, when longer test sequences are used in the analysis, can be found in [20] where the results showed that, if the quality distortions are grouped into the first few minutes of the screening, the quality scores are observed to increase. Conversely, if the quality distortions are grouped into the last few minutes, the scores are observed to decrease. This is due to the humans' short-term memory and recency effect which will be discussed further in Sect. 3.4.

In contrast, the distribution of the quality distortions can be modeled with, for instance, two-state exponential model as it is done in [17]. The model assumes that a video stream can be in one of the two states (ON or OFF). In the ON state, the stream is unaffected by the quality degradations, while in the OFF state, the quality is degraded. In [17], the authors investigate the inter-picture time in cellular-based video stream and define that if the time is ≤100 ms then the stream is in the ON state. Otherwise, the stream enters the OFF state. While streaming a 10-min long test sequence,

Table 1 Properties of different test sequences (TSs) and the number of responses (NoR) for each TS (the TS properties are presented in brackets, where the numbers have the following meanings: PLR of the inserted video clips; number of PLOs; duration of a single PLO; total duration of all PLOs)

TS no.	TS properties	NoR	TS no.	TS properties	NoR	TS no.	TS properties	NoR
1	(0.05 %; 1; 1 s; 1 s)	9	25	(0.05 %; 4; 4 s; 16 s)	6	49	(0.05 %; 7; 7 s; 49 s)	8
2	(0.1 %; 1; 1 s; 1 s)	8	26	(0.1 %; 4; 4 s; 16 s)	10	50	(0.1 %; 7; 7 s; 49 s)	10
3	(0.5 %; 1; 1 s; 1 s)	9	27	(0.5 %; 4; 4 s; 16 s)	8	51	(0.5 %; 7; 7 s; 49 s)	7
4	(1 %; 1; 1 s; 1 s)	8	28	(1 %; 4; 4 s; 16 s)	7	52	(1 %; 7; 7 s; 49 s)	7
5	(1.5 %; 1; 1 s; 1 s)	7	29	(1.5 %; 4; 4 s; 16 s)	9	53	(1.5 %; 7; 7 s; 49 s)	10
6	(2 %; 1; 1 s; 1 s)	10	30	(2 %; 4; 4 s; 16 s)	8	54	(2 %; 7; 7 s; 49 s)	7
7	(0.05 %; 1; 4 s; 4 s)	9	31	(0.05 %; 4; 7 s; 28 s)	8	55	(0.05 %; 10; 1 s; 10 s)	7
8	(0.1 %; 1; 4 s; 4 s)	8	32	(0.1 %; 4; 7 s; 28 s)	8	56	(0.1 %; 10; 1 s; 10 s)	9
9	(0.5 %; 1; 4 s; 4 s)	8	33	(0.5 %; 4; 7 s; 28 s)	7	57	(0.5 %; 10; 1 s; 10 s)	9
10	(1 %; 1; 4 s; 4 s)	6	34	(1 %; 4; 7 s; 28 s)	8	58	(1 %; 10; 1 s; 10 s)	7
11	(1.5 %; 1; 4 s; 4 s)	8	35	(1.5 %; 4; 7 s; 28 s)	11	59	(1.5 %; 10; 1 s; 10 s)	8
12	(2 %; 1; 4 s; 4 s)	9	36	(2 %; 4; 7 s; 28 s)	11	60	(2 %; 10; 1 s; 10 s)	7
13	(0.05 %; 1; 7 s; 7 s)	9	37	(0.05 %; 7; 1 s; 7 s)	9	61	(0.05 %; 10; 4 s; 40 s)	6
14	(0.1 %; 1; 7 s; 7 s)	8	38	(0.1 %; 7; 1 s; 7 s)	7	62	(0.1 %; 10; 4 s; 40 s)	10
15	(0.5 %; 1; 7 s; 7 s)	10	39	(0.5 %; 7; 1 s; 7 s)	7	63	(0.5 %; 10; 4 s; 40 s)	9
16	(1 %; 1; 7 s; 7 s)	6	40	(1 %; 7; 1 s; 7 s)	8	64	(1 %; 10; 4 s; 40 s)	9
17	(1.5 %; 1; 7 s; 7 s)	8	41	(1.5 %; 7; 1 s; 7 s)	8	65	(1.5 %; 10; 4 s; 40 s)	7
18	(2 %; 1; 7 s; 7 s)	10	42	(2 %; 7; 1 s; 7 s)	10	66	(2 %; 10; 4 s; 40 s)	11
19	(0.05 %; 4; 1 s; 4 s)	9	43	(0.05 %; 7; 4 s; 28 s)	12	67	(0.05 %; 10; 7 s; 70 s)	9
20	(0.1 %; 4; 1 s; 4 s)	7	44	(0.1 %; 7; 4 s; 28 s)	8	68	(0.1 %; 10; 7 s; 70 s)	8
21	(0.5 %; 4; 1 s; 4 s)	10	45	(0.5 %; 7; 4 s; 28 s)	7	69	(0.5 %; 10; 7 s; 70 s)	9
22	(1 %; 4; 1 s; 4 s)	6	46	(1 %; 7; 4 s; 28 s)	8	70	(1 %; 10; 7 s; 70 s)	8
23	(1.5 %; 4; 1 s; 4 s)	7	47	(1.5 %; 7; 4 s; 28 s)	10	71	(1.5 %; 10; 7 s; 70 s)	10
24	(2 %; 4; 1 s; 4 s)	9	48	(2 %; 7; 4 s; 28 s)	7	72	(2 %; 10; 7 s; 70 s)	10

the authors showed that the ON and OFF durations were distributed exponentially.

In this study, the first and last 7 min and 17 s of the test sequences were unaffected by the quality distortions, which allowed the test subjects to immerse themselves into the video in the beginning of the screening and to contemplate what they had experienced toward the end.

The total duration of all PLOs in a test sequence (i.e., the total duration of the quality distortions) varied between 1, 4, 7, 10, 16, 28, 40, 49, and 70 s, depending on the number of PLOs in a video and the duration of a single PLO, as shown in Table 1. Several test sequences had the same PLR and the same total duration of all PLOs; however, the number of PLOs in each video and the duration of a single PLO differed. For instance, four inserted video clips of degraded quality, each lasting 7 s, equaled 28 s of quality distortions in a test sequence, which is the same as when seven video clips, each lasting 4 s, are inserted.

The sequences were distributed to the test subjects on a DVD, thereby enabling them to view the documentary film in a real-life environment. Because the original video was in high-definition resolution, it was necessary to convert the prepared test sequences into the DVD format. The conversion into the DVD format was performed using CyberLink PowerDirector 11 with settings that maintained the best possible video quality. The PAL system and the MPEG-2 video encoding format were used for the conversion. During the conversion, all video enhancement features of the software were turned off, and the encoding codec did not use any error concealment methods. The DVD format was chosen for the following reasons.

1. As explained in Sect. 2.2, the test subjects in this study were students. Thus, it was necessary to use a format that would enable the majority of the test subjects to watch the video in real-life conditions. Note that students often live in student dormitories, campuses or rented apartments where they usually do not have access to, e.g., Blu-ray players. Because DVDs can be played on most personal and laptop computers and the availability of DVD players to the student population is higher than that of Blu-ray players, it was decided that the DVD format was the most appropriate for conducting research among this population.

2. It allowed experimenting with the video quality distortions in a controlled environment under known network conditions, while enabling a survey to be conducted among the test subjects in real-life environments.
3. It ensured easy distribution of the test sequences.

2.2 Targeted population and response rate

In this research, the test subjects were students of the University of Zagreb. This population was targeted because (a) according to Datta et al. [21], video streaming services are generally used by users between the ages of 18 and 24, which corresponds with the age group of a typical student population, and (b) this population was easy accessible for conducting such a survey (i.e., the convenience sampling method [22] was used).

Initially, 864 students received one DVD with the sequence to be rated. Apart from the sequence, the students also received a sealed envelope containing a short summary of the purpose of the research, a questionnaire (the questionnaire can be found in Appendix 1 of this paper) and instructions on how to complete the exercise. They were instructed not to open the envelope before the end of the screening and to complete the questionnaire immediately after the screening. Hence, the students were unaware about the purpose of the test prior to watching the video.

Each test subject watched only the video that was given to him or her, and they were instructed to watch it once before completing the questionnaire. The questionnaire contained multiple choice questions. However, for the questions related to the subjective perception of the video quality and the extent of the perceived video quality distortions, an 11-point numerical scale was used, as designed in ITU-T Rec. P910 [23].

Given the possibility that some of the distributed DVD disks may have been damaged and/or a user's equipment was defective (e.g., DVD players), the questionnaire also contained questions with the purpose of detecting such instances. For example, by answering question A3.1, the subjects provided information about the types of video artifacts that appeared during the screening. If the answer(s) indicated that the subjects experienced degradations that were not related to the specific test sequence, and that questionnaire was excluded from further analysis (e.g., a response of "d" for a test sequence with a PLR of 0.05 % led to the exclusion of that questionnaire because it was known that those test sequences contained no frame freeze video artifacts). Removing these questionnaires from the analysis was important, because in those instances, the subjects experienced video quality distortions that were unrelated to our experiment. Furthermore, responses of "a" and/or "b" to question B4 also served as rejection criteria, because this was a direct indication of malfunctioning of the user's equipment.

The questionnaires also contained questions that were used to detect and exclude outliers from further analysis (questions A1 and A3.3–A3.5) as well as to identify test subjects with impaired sight and/or hearing (question B8). Farrokhi and Mahmoudi-Hamidabad stressed that the exclusion of outliers is especially relevant when non-probability sampling methods are used, of which convenience sampling is notorious [22]. In this study, if the subject's answers to questions A1 and A3.5 differed by more than 8 points, that questionnaire was excluded from the analysis. These questions were aimed at discovering the user's perceptions of the overall quality of the video and the viewing experience, respectively. Thus, it was considered that a difference of greater than 8 points indicated inconsistent and abnormal rating, because it is unlikely that a subject would perceive the video as being of *Bad quality* while also having an *Excellent experience* of watching it (or experience a similar quality disparity in the opposite direction). A questionnaire was also excluded from the analysis if an overly stringent rating was applied by the subject in questions A3.3 and A3.4. For instance, if the subject indicated that the total duration of all video quality distortions was 1 s and rated that duration as more than 6 on the annoyance scale (question A3.4), this was considered to be stringent, unrealistic rating behavior that should be treated as abnormal.

Finally, questionnaires were also rejected if they were not fully completed and in the following cases:

- If response "c" was provided to question B2 (the reason: the subject did not watch the complete video).
- If response "d" was provided to question B3 (the reason: the noise level in the subject's environment may have interfered with his or her perception).
- If response "b" was provided to question B6 (the reason: the subject did not notice the quality distortions on their own; instead, the person(s) in their company suggested that the quality was degraded, ergo, the subject was unable to correctly evaluate the type of degradation, its duration and frequency as well as the level of annoyance toward something which remained hidden to him or her).
- If response "c" was provided to question B6 when response "a" was given for question A2 (the reason: inconsistent responses).
- If response "b" was provided to question B10 (the reason: the subject did not complete the questionnaire immediately after the screening; instead they have completed it after one, two or more days and thus might have forgotten the quality distortions they experienced, potentially leading to false ratings).

- If response “a” was provided to question B11 (the reason: the subjects were familiar with the topic of the research prior to the screening and thus may have been focused on noticing and memorizing the distortions, which is unlike real-life conditions).

According to ITU-T Rec. P.910 [23], at least four test subjects are needed when conducting a subjective evaluation of the quality of video sequences. To ensure the minimum required sample size, 12 DVD copies of each test sequence were created, and two questionnaires were inserted into the envelopes. Given the possibility that the subjects might watch the video in the company of someone else, subjects were asked to pass the second questionnaire to the person in their company. After a period of 2 weeks, 830 questionnaires were collected. Over 27 % of the collected questionnaires were rejected (27.47 %, or 228) for the reasons discussed in this chapter.[1] Appendix 2 contains a table which shows the number of rejected questionnaires for each specific criterion. The user QoE analysis was performed using a sample consisting of 602 test subjects (Table 1 indicates the number of accepted questionnaires per test sequence).

In [14], the authors discuss the methods for excluding unreliable responses in crowdtesting. From our list of rejection criteria, it can be observed that we (a) used *consistency questions* to identify unreliable, abnormal responses; (b) investigated the *hardware environment* to detect the malfunctioning of user equipment; and (c) examined *hidden influence factors* such as the level of noise in the test subject’s surroundings during the screening.

3 Evaluation results and discussion

3.1 Analysis of user QoE

The average QoE rating and margin of error (with confidence level of 95 %) was calculated for each test sequence, and the results are presented in Fig. 1 (the four subplots correspond to different numbers of PLOs in a video). The QoE rating for each test subject was calculated as the average of the ratings given in questions A1 and A3.5. Note that when discussing the results, we use the linguistic meanings defined for 11-point quality scale in question A1 (i.e., 0–2 *Bad quality*, 2–4 *Poor quality*, 4–6 *Fair quality*, 6–8 *Good quality*, 8–10 *Excellent quality*). The boundaries between these five sets are not firmly determined because the linguistic meanings are given to assist test subjects during rating. This feature makes the scale suitable for exploring user opinions, which are usually fuzzy in nature.

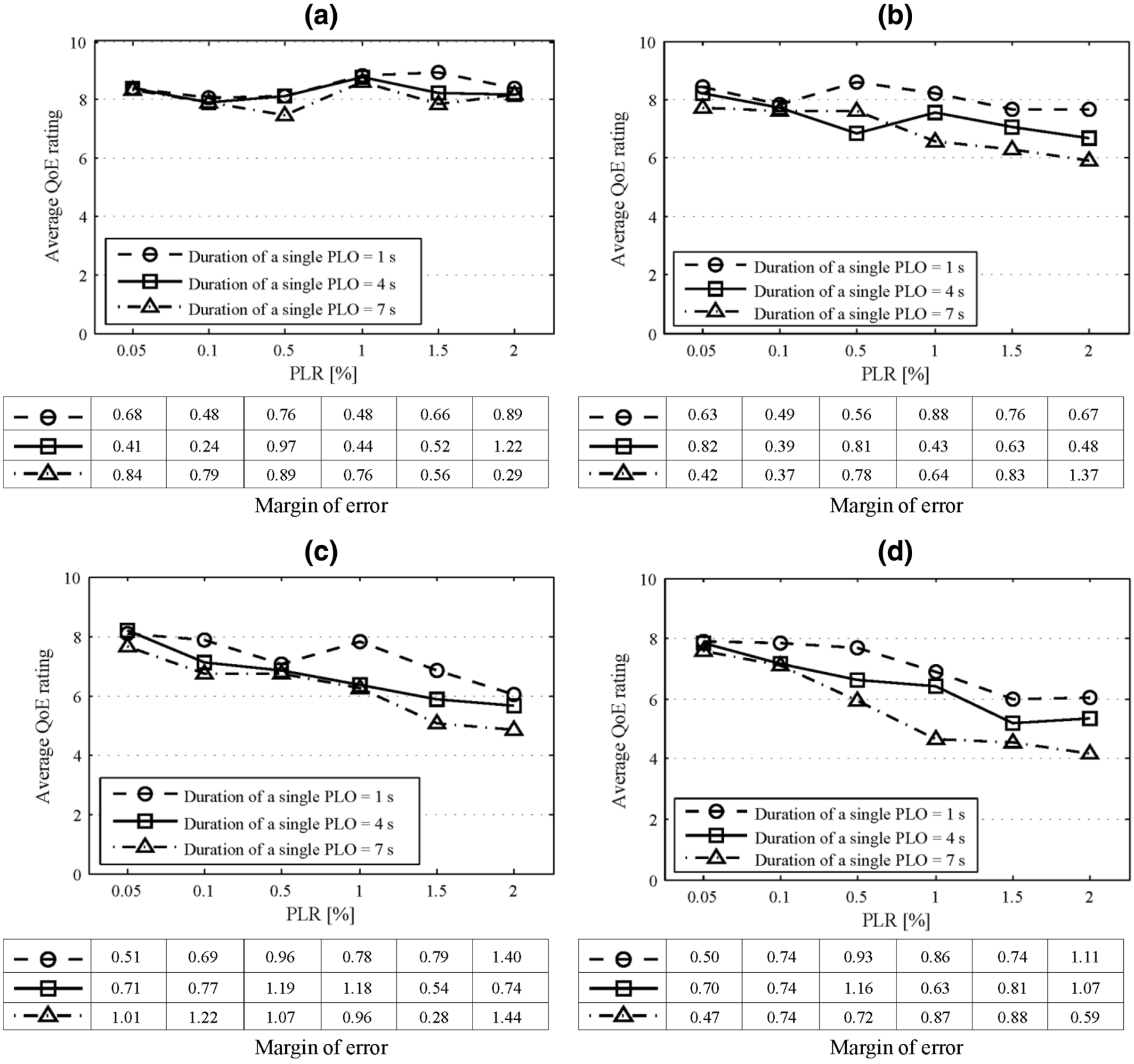
The statistical significance of the obtained results was tested in three ways: (1) the Kruskal–Wallis (KW) test was used to test whether the ratings constituting a particular curve (corresponding to a particular value of the duration of a single PLO) originated from the same distribution (Table 2); (2) if the p values calculated from the KW test were lower than the significance level $\alpha = 0.05$ (i.e., if significant differences were observed within a set of QoE ratings), then the Mann–Whitney (MW) U test was used to determine between which two independent ratings the significant difference existed (Table 3); and (3) the KW test was again used to test whether there were any significant differences between the sets of QoE ratings presented in different subplots of the figure (Table 4). We chose to use the KW and MW U tests rather than the traditional Student’s t test or ANOVA because the observations did not follow a normal distribution. Note that the significant values, for which $p < \alpha = 0.05$, are marked with numbers written in bold text format.

By comparing the average ratings for the first 18 test sequences shown in Fig. 1a, we can observe that the ratings varied on the interval [7.46, 8.96]; however, the rating differences determined using the KW test are insignificant (the p values are provided in Table 2). Therefore, different PLRs and different values of the duration of a single PLO do not result in degradation of the user QoE when the number of PLOs is equal to 1, i.e., when there is only one packet loss affecting the video.

When the number of PLOs was equal to 4 (Fig. 1b), the ratings decreased to below 6 for the first time (falling into the *Fair quality* set), but this occurred only when the PLR was 2 % and the total duration of all PLOs was equal to 28 s (i.e., duration of a single PLO = 7 s). For this subplot, the p values from Table 2 indicate statistically significant differences between the ratings constituting the curves marked with rectangles and triangles (duration of a single PLO = 4 and 7 s, respectively). For these curves, the results of the MW U test (Table 3) reveal significant differences between the QoE ratings (a) for PLRs of $\leq$0.1 and $\geq$1.5 % with a single-PLO duration of 4 s and (b) for PLRs of $\leq$0.5 and $\geq$1 % with a single-PLO duration of 7 s (with one exception: the ratings for PLR = 2 % are significantly different only from those for PLRs of $\leq$0.1 %). This shows that four quality distortions (lasting 16 s or longer) in a video lasting 1 h can degrade user experience in cases when a higher PLR occurs.

Further increasing the number of PLOs to 7 (Fig. 1c) caused greater user dissatisfaction, as seen from the fact that the average QoE rating decreased to less than 5; however, this was true only for the most degraded sequence

[1] It is noteworthy to mention that it was expected to have relatively large share of rejected responses, since it was recognized that not all students will take their participation in a survey seriously.



(a) Margin of error

– ⊖ –	0.68	0.48	0.76	0.48	0.66	0.89
–□–	0.41	0.24	0.97	0.44	0.52	1.22
·–△–·	0.84	0.79	0.89	0.76	0.56	0.29

(b) Margin of error

– ⊖ –	0.63	0.49	0.56	0.88	0.76	0.67
–□–	0.82	0.39	0.81	0.43	0.63	0.48
·–△–·	0.42	0.37	0.78	0.64	0.83	1.37

(c) Margin of error

– ⊖ –	0.51	0.69	0.96	0.78	0.79	1.40
–□–	0.71	0.77	1.19	1.18	0.54	0.74
·–△–·	1.01	1.22	1.07	0.96	0.28	1.44

(d) Margin of error

– ⊖ –	0.50	0.74	0.93	0.86	0.74	1.11
–□–	0.70	0.74	1.16	0.63	0.81	1.07
·–△–·	0.47	0.74	0.72	0.87	0.88	0.59

Fig. 1 Average QoE ratings as a function of PLR; the subplots of the figure are referring to the number of PLOs in a sequence: **a** 1, **b** 4, **c** 7, and **d** 10

Table 2 *p* values calculated using the KW test to test whether the QoE ratings constituting a particular curve (corresponding to a particular value of the duration of a single PLO) originated from the same distribution

Subplot of Fig. 1	*p* value calculated for the set of QoE ratings corresponding to a particular curve		
	Circle	Square	Triangle
(a)	0.34	0.24	0.38
(b)	0.05	**0.02**	**0.004**
(c)	0.09	**0.003**	**0.003**
(d)	**0.013**	**0.005**	**5.99e−6**

in that series (PLR = 2 % and duration of a single PLO = 7 s). In this subplot, significant rating differences are also recorded for the curves corresponding to single-PLO durations of 4 and 7 s (Table 2). From Table 3, it can be observed that the differences between the ratings in this subplot are significant for PLRs exceeding 1 %, especially for a single-PLO duration of 7 (for this curve, the differences between ratings are significant for PLRs of ≤0.5 and ≥1.5 %).

The lowest average QoE ratings were recorded for PLR ≥ 1 % in the sequences that contained 10 PLOs that lasted a total of 70 s (Fig. 1d, duration of a single

Table 3 *p* values calculated using the MW *U* test to compare two independent QoE ratings

Comparison between specific PLRs [%]	*p* values for Fig. 1b		*p* values for Fig. 1c		*p* values for Fig. 1d		
	Square	Triangle	Square	Triangle	Circle	Square	Triangle
0.05 ↔ 0.1	0.3548	0.8336	0.0759	0.2459	0.6718	0.3848	0.3117
0.05 ↔ 0.5	0.0525	0.6849	0.1082	0.2712	0.7102	0.1242	**0.0023**
0.05 ↔ 1	0.3146	**0.0074**	**0.0308**	0.1049	**0.0476**	**0.0157**	**0.0015**
0.05 ↔ 1.5	**0.0291**	**0.005**	**0.0004**	**0.0022**	**0.0092**	**0.0066**	**0.0008**
0.05 ↔ 2	**0.0137**	**0.0258**	**0.0007**	**0.0038**	**0.0152**	**0.0088**	**0.0002**
0.1 ↔ 0.5	0.0682	0.6025	0.9079	0.9223	0.7896	0.4875	0.1121
0.1 ↔ 1	0.732	**0.0136**	0.2480	0.5912	0.1243	0.1205	**0.0046**
0.1 ↔ 1.5	**0.0451**	**0.0132**	**0.0163**	**0.0376**	**0.0161**	**0.0084**	**0.0016**
0.1 ↔ 2	**0.0183**	**0.0258**	**0.0151**	**0.0248**	**0.0299**	**0.0183**	**0.0004**
0.5 ↔ 1	0.1828	**0.0427**	0.7282	0.4433	0.3611	0.3536	**0.0485**
0.5 ↔ 1.5	0.6648	**0.0297**	0.3768	**0.0112**	**0.0234**	0.1248	**0.0305**
0.5 ↔ 2	0.713	0.0699	0.3379	**0.0262**	0.0699	0.1281	**0.0055**
1 ↔ 1.5	0.204	0.6496	0.7893	**0.0314**	0.1052	**0.0111**	0.9646
1 ↔ 2	**0.0273**	0.7102	0.6852	0.0639	0.2769	0.1597	0.5338
1.5 ↔ 2	0.5317	0.9476	0.8447	0.4642	0.9539	0.9639	0.5706

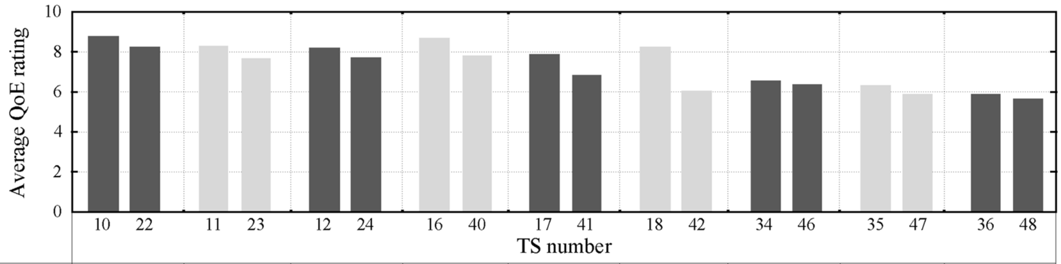
Table 4 *p* values calculated using the KW test to compare the sets of QoE ratings (for particular values of the PLR) between different subplots of Fig. 1

Comparison between specific subplots of Fig. 1	*p* values calculated for the two QoE ratings corresponding to a particular PLR [%]					
	0.05	0.1	0.5	1	1.5	2
(a) ↔ (b)	0.647	0.739	**0.037**	**0.000391**	**0.000333**	**0.000458**
(a) ↔ (c)	0.768	0.314	0.194	**0.000879**	**5.48e−7**	**0.000106**
(a) ↔ (d)	0.263	0.367	**0.011**	**5.7e−6**	**1.16e−6**	**3.42e−6**
(b) ↔ (c)	0.748	0.493	**0.018**	**0.021**	**0.000091**	**0.009384**
(b) ↔ (d)	0.414	0.627	**0.001156**	**0.000255**	**0.000169**	**0.000473**
(c) ↔ (d)	0.828	0.523	0.277	**0.003975**	**0.000419**	0.094

PLO = 7 s). The worst average QoE rating was recorded for the most degraded test sequence (PLR = 2 %, number of PLOs = 10 and duration of a single PLO = 7 s). However, that rating equaled 4.16, still within the *Fair quality* set, indicating that the subjects either forgot some video quality distortions that were experienced after watching a 1-h documentary film or they thought that the perceived video artifacts did not completely interfere with the seamless reproduction of the video. This type of rating behavior confirmed the results presented in [24] where it is shown that user quality requirements decrease over time. This finding encourages the investigation, in future, of the extent of video quality distortions that could lead to the worst possible quality ratings. For this purpose, the *method of limits* could be applied as by Menkovski et al. in [25]. After the KW test was applied to the data presented in Fig. 1d, the results indicated significant differences between ratings even when the duration of a single PLO was 1 s (Table 2). The MW *U* test indicated that an increase of the single-PLO duration to 7 s significantly degraded the user QoE despite a lesser increase in PLR.

The results presented above can be summarized as follows: (a) the PLR and the duration of a single PLO cannot affect user QoE if there is only one PLO in a 1-h video; (b) for PLRs of ≥1 %, a quality degradation that lasts ≥16 s can be negatively perceived by users; (c) the duration of a single PLO becomes an important factor as the PLR increases (≥1.5 %) if the video contains 7 or more PLOs; (d) the number of significant differences between two independent ratings in a particular subplot increased with an increase in the PLR and in the duration of a single PLO; and (e) based on the results presented in Table 4, an increase in the number of PLOs significantly affects user QoE for PLRs of ≥0.5 %.

For two test sequences with the same PLR (≥1 %) and the same total duration of all PLOs, a higher average QoE rating was recorded for the sequence with the lower number of PLOs (Fig. 2). However, the differences between the



PLR [%]	1		1.5		2		1		1.5		2		1		1.5		2	
Number of PLOs	1	4	1	4	1	4	1	7	1	7	1	7	4	7	4	7	4	7
Duration of a single PLO [s]	4	1	4	1	4	1	7	1	7	1	7	1	7	4	7	4	7	4
Total duration of all PLOs [s]	4		4		4		7		7		7		28		28		28	
Margin of error	0.44	0.88	0.52	0.76	1.22	0.67	0.76	0.78	0.56	0.79	0.29	1.40	0.64	1.18	0.83	0.54	1.37	0.74
p-values	0.2615		0.3248		0.0701		0.1752		0.0273		0.0272		0.7527		0.1389		0.1891	

Fig. 2 Comparison of the average QoE ratings for test sequences (TSs) with the same PLR and the same total PLO duration

two ratings (according to the MW U test with $\alpha = 0.05$) were significant in only two cases (between TSs 17 and 41 and between TSs 18 and 42, indicated by the shaded p values in the attached table). It can be concluded that the total duration of quality distortions more strongly affects the user QoE than do the number of PLOs and the duration of a single PLO, as it is evident that user QoE decreases with an increase in the total duration of quality degradations (from 4 to 28 s).

3.2 The relationship between the stimulus and user response

The p values presented in Tables 2, 3 and 4 indicate that certain transitions between the various levels of video quality do not evoke significant changes in the subjects' perception. This finding urged us to further investigate the relationship between the stimulus (i.e., the values of the objective parameters in the test sequences) and the QoE of the subjects. In [26] the relationships between two network parameters (bit rate and PLR) and Mean Opinion Scores of the subjects are described using the logarithmic functions. Apart from the logarithmic mapping, Fiedler et al. in [27] experimented with exponential relationship between QoE and QoS parameters, called IQX hypothesis. The incentive to use these relationships originates from a cognition that the user awareness of the QoE is more pronounced when the experienced quality is high. Specifically, when the QoE is very high, a small quality degradation will strongly decrease the QoE. Conversely, if the QoE is already low, a further disturbance is not perceived significantly [27, 28]. The two approaches are compared in [28] where it is show that the IQX hypothesis (i.e., the exponential relationship between QoE and QoS) outperforms the logarithmic relationship.

In this chapter the IQX hypothesis is used to show the exponential interdependency between the total duration of PLOs, PLR, number of PLOs, duration of a single PLO and user QoE. The results are presented in Fig. 3. Note that each subplot of the figure depicts the average QoE rating for all test sequences with a given value of only one parameter. For instance, the average QoE rating in Fig. 3a was calculated for all test sequences with specific total duration of quality degradations (without considering the differences in the other three parameters). The subplots also depict the minimum and the maximum recorded QoE ratings. The dispersion of the measurements, i.e., the range of these min/max intervals gives a clear indication how all four parameters create an affiliated effect on user perception.

The legend of the figure shows that two types of the exponential data fitting was applied. First, the fitting is done for the average QoE ratings (indicated by the full lines). It can be observed that the obtained exponential functions yielded high coefficients of determination ($R^2 > 0.91$ in all cases). Secondly, the fitting was conducted for all measurements (indicated by the dashed lines). In this case the calculated coefficients are remaining relatively low, due to the abovementioned dispersion of the ratings. However, we wanted to depict both fittings to show the similarities between the two curves. These results confirmed those from [26–28], i.e., a given amount of change of the objective parameters has a different impact on resulting change of QoE, depending on the current level of QoE.

Furthermore, the values of the four parameters are normalized with the purpose of discovering which parameter can degrade the user QoE the most (Fig. 4). When the values of the first derivative of the obtained functions, for any specific point of the functions, are compared, it can be observed that:

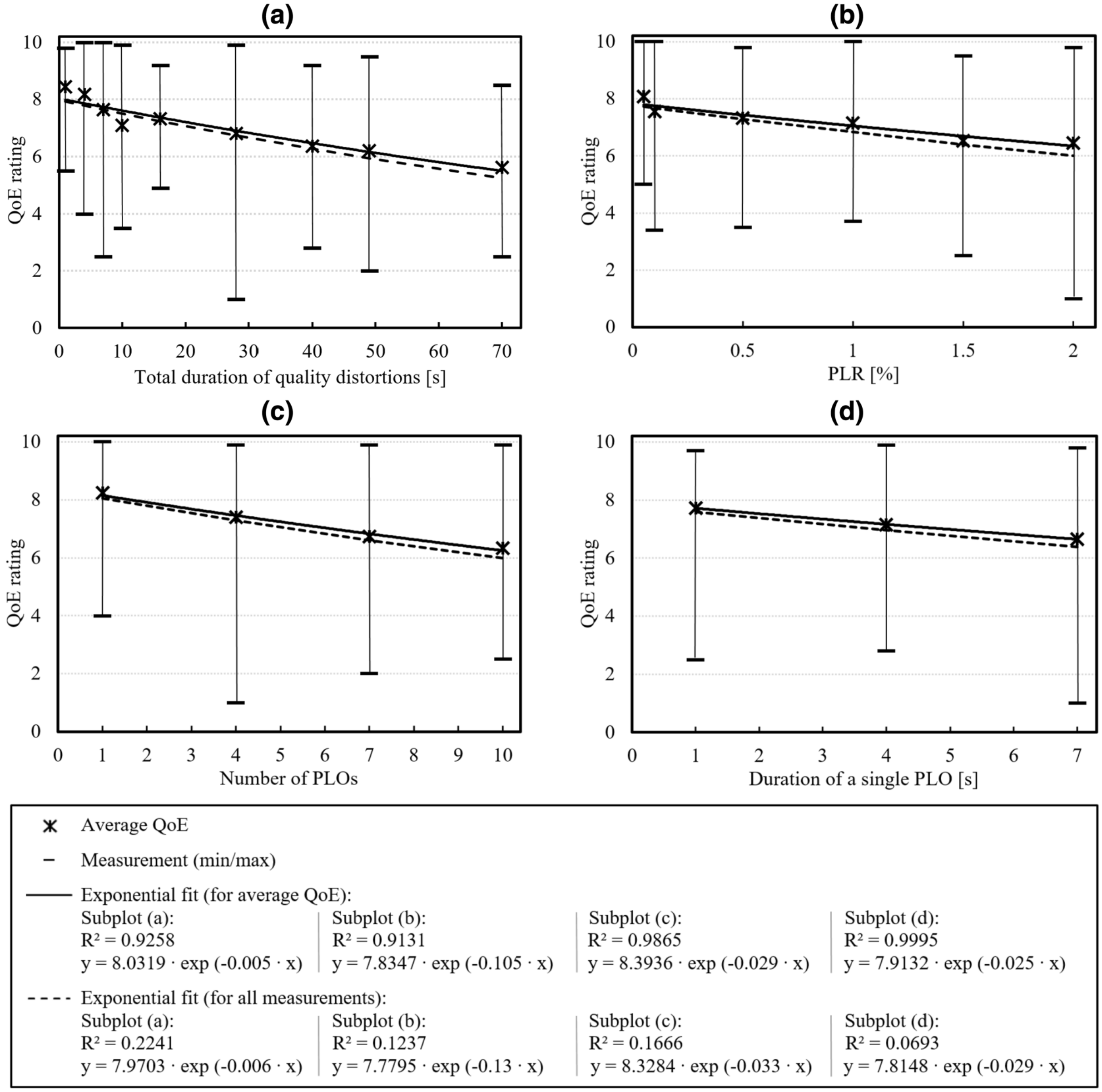


Fig. 3 QoE ratings for different parameters: **a** total duration of quality distortions, **b** PLR, **c** number of PLOs, and **d** duration of a single PLO

$$\mathrm{QoE}'_{\text{Total duration}}(x) < \mathrm{QoE}'_{\text{Number of PLOs}}(x) < \mathrm{QoE}'_{\text{PLR}}(x) < \mathrm{QoE}'_{\text{Duration of a single PLO}}(x), \qquad (1)$$

where x represents the normalized values of different parameters. Based on these results, the objective parameters can be ranked by their order of importance in relation to their impact on user QoE as follows: (1) total duration of quality distortions in a video, (2) number of PLOs, (3) PLR, and (4) duration of a single PLO.

3.3 User annoyance caused by packet loss

Higher PLRs can damage the image and hamper screening for a longer period of time, thus increasing the level of user annoyance. Figure 5a depicts the Average Annoyance Level (AAL) of the test subjects as a function of the PLR as well as the AAL ± margin of error (MoE) with confidence level of 95 %. The subjects provided their ratings on an 11-point numerical scale, which can be found in the appendix (question A3.2). The differences between the AAL ratings were

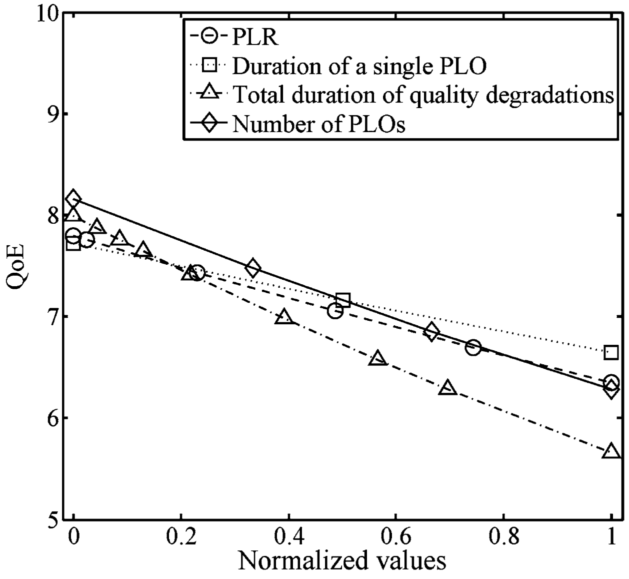


Fig. 4 Normalized values of the objective parameters and modeled QoE

evaluated using the MW *U* test (the results of this test are presented in Table 5). Higher PLRs caused the subjects to report higher AALs; however, even the highest score (4.38 for a PLR of 1.5 %) was well below the worst ratings (*Annoying* and *Very annoying quality distortion*). In general, quality distortions in the sequences with PLR < 1 % were mostly *Imperceptible* to the subjects, and the distortions caused by PLR ≥ 1 % were usually perceived only as *Slightly annoying*. The *p* values reported in Table 5 indicate that the differences between ratings in this subplot are predominantly significant.

However, Fig. 5a shows the AALs for all test sequences with a given value of the PLR, without considering the differences in the number of PLOs and the duration of a single PLO. The results of a detailed analysis (Fig. 5b) reveal that the highest AALs were recorded for test sequence number 72. For this sequence, the AAL is classified in the *Very annoying quality distortion* category and results in an overall assessment of *Fair quality*. For the data presented in Fig. 5b, the MW *U* test reveals that the differences are significant between the ratings for PLR = 0.05 % and for PLRs of ≥0.1 %; however, it can be argued that for these sequences, the total duration of quality degradations (70 s) annoyed the subjects more than did the changes in the PLR (as discussed in the previous chapter).

For the sequences in which the duration of a single PLO was 7 s, the AAL and average QoE ratings reported by the subjects were compared as a function of the PLR (Fig. 6). Note that the margin of error is again calculated for the confidence level of 95 %.

Because the results of the significance tests for the QoE ratings can be found in Tables 2, 3 and 4, for the data depicted in Fig. 6, the testing was conducted only between the different AALs presented in subplots a, b and c (Table 5 already contains the *p* values referring to the AAL curve presented in subplot d).

As previously stated, when the number of PLOs in the entire streaming session was equal to 1, the other two objective parameters had a limited impact on the QoE of the subjects (Fig. 6a). The quality distortions in these sequences remained *Imperceptible* to the subjects; thus, the QoE ratings were high (Table 6 shows mostly insignificant rating differences between the annoyance levels). An increase in the total duration of the quality distortions to 28 or 49 s (Fig. 6b, c, respectively) adversely affected the perception of the subjects, resulting in an increased AAL.

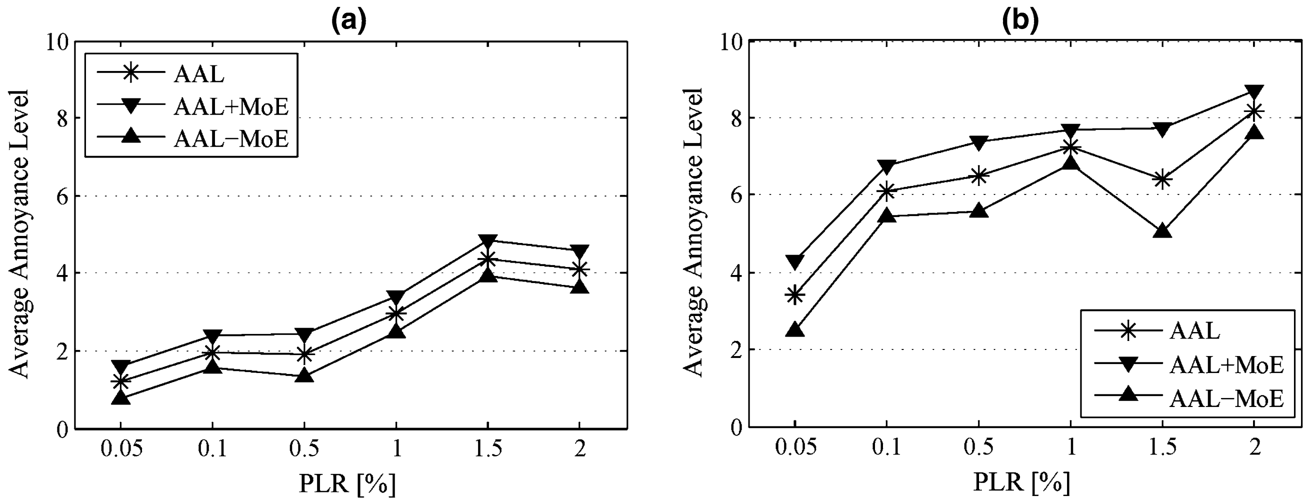


Fig. 5 Annoyance level (average ± margin of error) as a function of PLR: **a** results for all test sequences, and **b** results for test sequences with 70 s of quality distortions

Table 5 p values calculated using the MW U test to compare two independent AAL ratings (the significant values, for which $p < \alpha = 0.05$, are marked with numbers written in bold text format)

Comparison between specific PLRs [%]	p values for Fig. 5a	p values for Fig. 5b
0.05 ↔ 0.1	**0.0022**	**0.0005**
0.05 ↔ 0.5	**0.0147**	**0.0003**
0.05 ↔ 1	**<0.0001**	**0.0005**
0.05 ↔ 1.5	**<0.0001**	**0.007**
0.05 ↔ 2	**<0.0001**	**0.0002**
0.1 ↔ 0.5	0.4413	0.665
0.1 ↔ 1	**0.0009**	**0.0274**
0.1 ↔ 1.5	**<0.0001**	0.6569
0.1 ↔ 2	**<0.0001**	**0.0014**
0.5 ↔ 1	**0.0002**	0.2107
0.5 ↔ 1.5	**<0.0001**	0.9025
0.5 ↔ 2	**<0.0001**	**0.008**
1 ↔ 1.5	**0.0001**	0.534
1 ↔ 2	**0.0017**	**0.0164**
1.5 ↔ 2	0.3393	0.0696

For these subplots, the differences are predominantly significant between the ratings corresponding to PLRs of $\leq$0.1 and $\geq$1.5 %. Note that the increase in the AAL curve is somewhat steeper than the decline of the QoE curve; thus, the subjects were more annoyed by the increase in packet loss intensity. However, this factor was not entirely reflected in the overall experience.

The last subplot (Fig. 6d) shows that 70 s of quality distortions was sufficient to cause higher AAL (*Annoying* and *Very annoying*) ratings in practically all test sequences of that series, and the increased packet loss intensity caused the most distinctive decrease in the average QoE ratings, from 7.59 (PLR = 0.05 %) to 4.16 (PLR = 2 %).

A scatter plot of the user QoE ratings and their corresponding annoyance levels confirms the existence of a correlation between these two parameters (Fig. 7). The correlation coefficient of −0.69 indicates a moderately negative linear relationship. The results from this figure support previous claims by demonstrating that an increase in the level of user annoyance is not entirely reflected in the QoE.

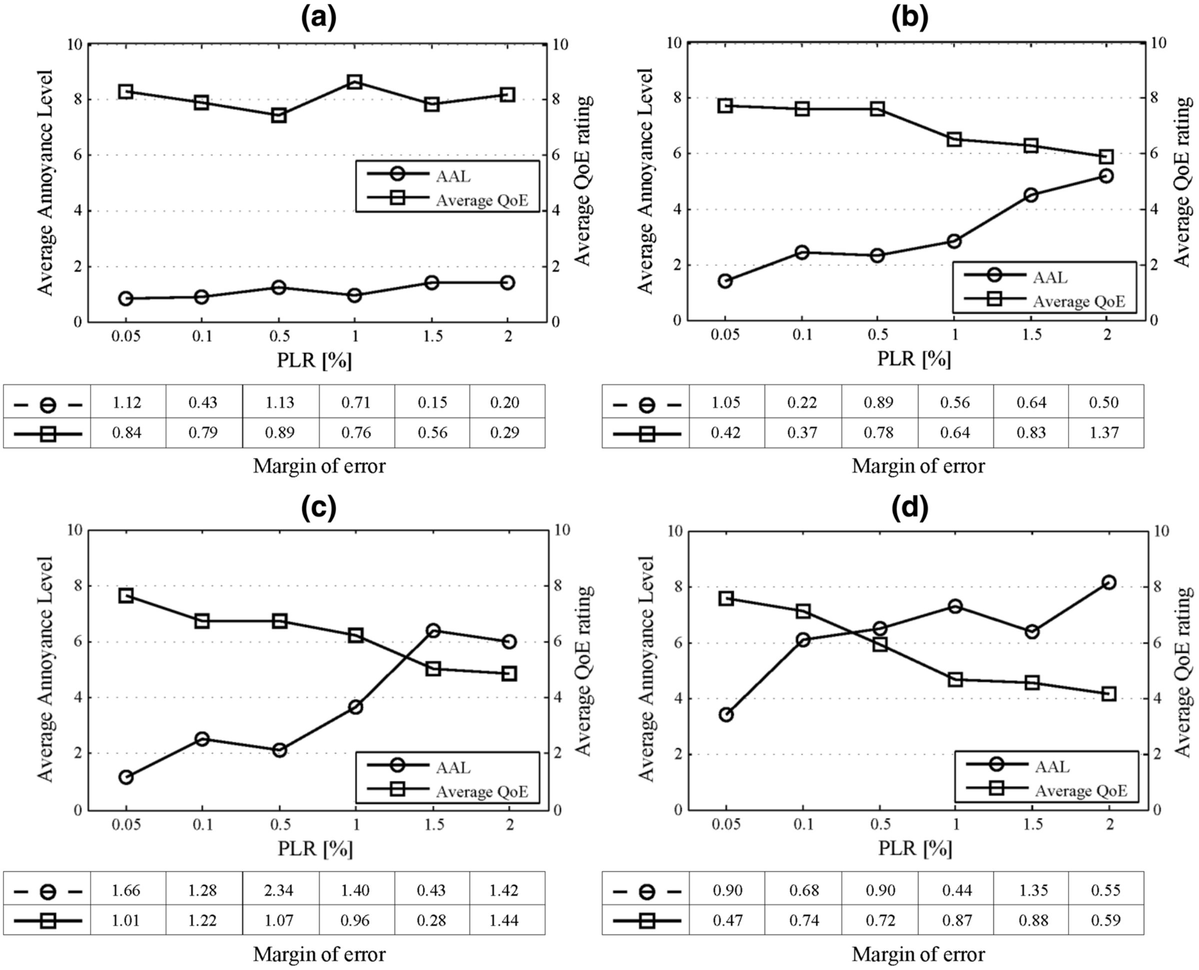
3.4 The impact of humans' short-term memory and recency effect

The subjects were asked to quantify the number of times that they noticed that the quality of a test sequence was distorted (question A3.3). Figure 8 shows the results of this analysis. The numbers on the x-axis represent the difference between the number of inserted PLOs and the number of quality distortions observed by the subjects. The line marked with circles indicates how many test subjects failed to notice a certain number of PLOs. Conversely, the line marked with rectangles indicates the instances in which the subjects thought that a greater number of quality distortions occurred in a sequence than was the case. For example, 100 test subjects failed to notice one PLO in the video (e.g., instead of 4, they noticed only 3, with a difference of 1), and 48 test subjects reported one additional, nonexistent PLO (e.g., instead of 4, they thought that they saw 5, with a difference of −1).

A considerable number of test subjects failed to notice some or all PLOs in the video (the sum of the answers comprising the line marked with circles is equal to 408). As emphasized by Jelassi et al. in [29], one of the factors influencing subject reasoning is humans' short-term memory. This factor clearly impacted the panel of test subjects in this study because after watching the 1-h documentary film, certain individuals simply forgot about the quality distortions that they may have noticed during screening. Furthermore, the test subjects were not focused on counting and memorizing the distortions because the experiment was conducted in a real-life environment where the subjects could focus their attention on the content.

In addition to humans' short-term memory, these results are also influenced by the psychological effect of recency. This effect is increasingly being referenced in related work when researchers attempt to explain how humans can more thoroughly recall ending scenes compared with the scenes shown in the middle of a test sequence [30]. In this study, the PLOs were evenly distributed over the entire duration of all test sequences (as discussed in Sect. 2.1). This means that in the sequences with one PLO, the quality distortion occurred in the middle of the video. Further analysis of the results reveals that 37.3 % of the test subjects who evaluated the sequences with one PLO failed to notice that PLO, thereby confirming the impact of the recency effect. Because the subjects were not asked to describe the scenes in which they noticed PLOs,[2] we are unable to analyze which PLOs were the most noticeable in other test sequences that contained 4, 7 or 10 PLOs. Nevertheless, from our knowledge of the results for the sequences with one PLO, we can infer that in other test sequences, PLOs that were placed near the middle of the video were not always recalled by the subjects.

[2] Since some test sequences contained 7 or 10 PLOs, it was considered that it would be difficult and time-consuming task for the subjects to exactly recall and describe in the questionnaire all the scenes with quality degradations (after watching the 1-h video). The alternative was to ask a multiple choice question in the questionnaire, so that the subjects would only indicate the scenes. However, then, the available answers would make them remember something what they forgot during the screening which, in turn, could affect their QoE rating.



(a)

−ϴ−	1.12	0.43	1.13	0.71	0.15	0.20
−⊟−	0.84	0.79	0.89	0.76	0.56	0.29

Margin of error

(b)

−ϴ−	1.05	0.22	0.89	0.56	0.64	0.50
−⊟−	0.42	0.37	0.78	0.64	0.83	1.37

Margin of error

(c)

−ϴ−	1.66	1.28	2.34	1.40	0.43	1.42
−⊟−	1.01	1.22	1.07	0.96	0.28	1.44

Margin of error

(d)

−ϴ−	0.90	0.68	0.90	0.44	1.35	0.55
−⊟−	0.47	0.74	0.72	0.87	0.88	0.59

Margin of error

Fig. 6 AAL and average QoE ratings as a function of the PLR for test sequences in which the duration of a single PLO was 7 s; the subplots of the figure are referring to the number of PLOs in a sequence: **a** 1, **b** 4, **c** 7, and **d** 10

Table 6 *p* values calculated using the MW *U* test to compare two independent AAL ratings (the significant values, for which $p < \alpha = 0.05$, are marked with numbers written in bold text format)

Comparison between specific PLRs [%]	*p* values for Fig. 6a	*p* values for Fig. 6b	*p* values for Fig. 6c
0.05 ↔ 0.1	0.2189	0.5624	0.1746
0.05 ↔ 0.5	0.4452	0.2673	0.8548
0.05 ↔ 1	0.2076	0.1025	**0.0366**
0.05 ↔ 1.5	**0.0135**	**0.0011**	**0.0044**
0.05 ↔ 2	**0.0155**	**0.0003**	**0.0053**
0.1 ↔ 0.5	0.6214	0.9079	0.5495
0.1 ↔ 1	0.9485	0.4942	0.2043
0.1 ↔ 1.5	**0.0457**	**0.0007**	**0.0002**
0.1 ↔ 2	0.0747	**0.0003**	**0.0084**
0.5 ↔ 1	0.662	0.385	0.2228
0.5 ↔ 1.5	0.055	**0.0043**	0.0564
0.5 ↔ 2	0.0634	**0.0005**	**0.0342**
1 ↔ 1.5	0.0810	**0.005**	**0.0006**
1 ↔ 2	0.1142	**0.0004**	0.1102
1.5 ↔ 2	0.7882	0.1479	0.4945

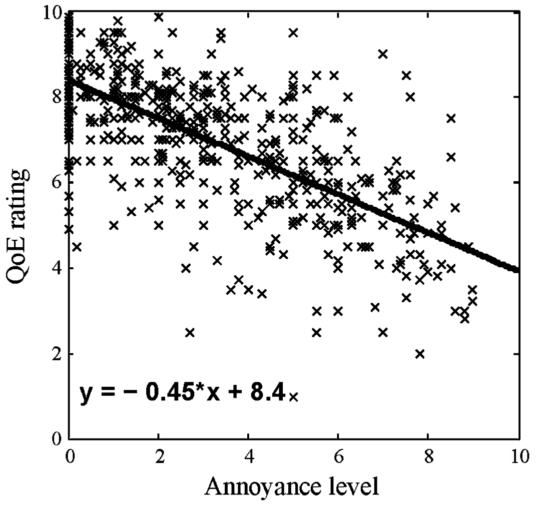


Fig. 7 User QoE and annoyance level

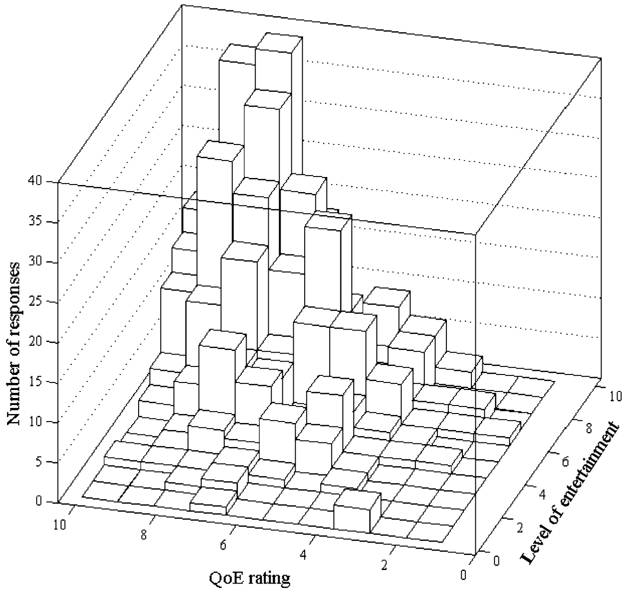


Fig. 9 Bivariate histogram of user QoE and level of entertainment

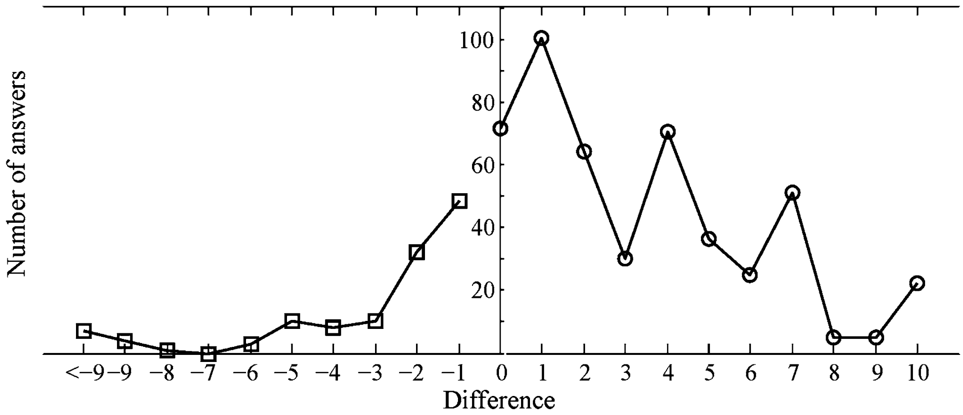
3.5 User QoE and their level of entertainment

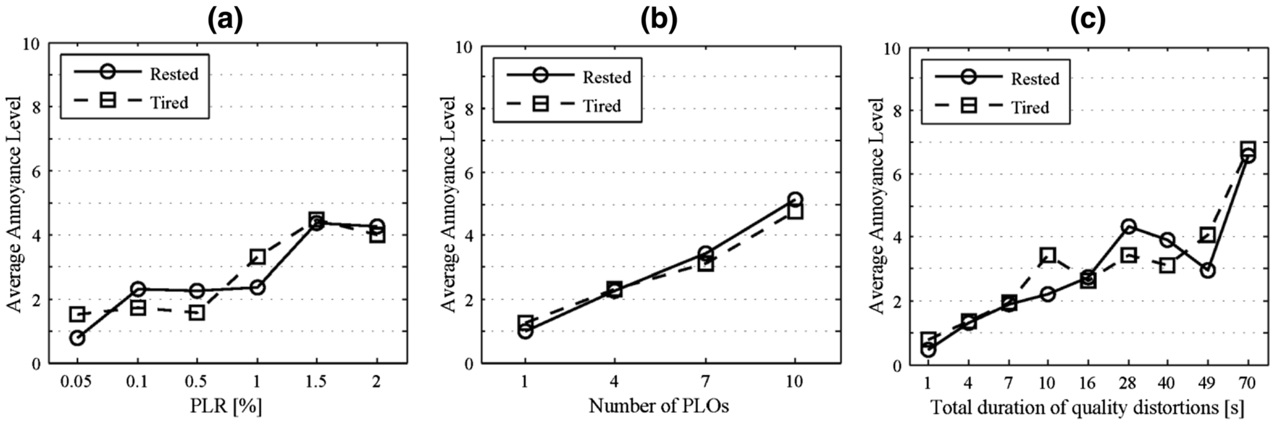
It is reasonable to assume that in everyday life, users watch video content that interests them. Acceptance of this assumption greatly affected the choice of multimedia content used in this research. One of the requirements that the content had to fulfill during the real-life experiment was to entertain the majority of the targeted population. In the research preparation phase, the use of music videos, sports matches, and movies of various genres (drama, comedy, thriller, etc.) was considered. However, due to the relatively large target group and the individual preferences, a documentary film about the solar system was chosen as the subject with the highest potential to awaken the subjects' curiosity and to entertain the majority of the subjects. Hence, entertainment-oriented content selection [31] was used, but one type of content was provided to the subjects because of the sample size and number of required test sequences for each video type.

The level of entertainment of the subjects was evaluated on an 11-point numerical scale (question B1). Figure 9 shows a bivariate histogram of user QoE and level of their entertainment. Note that the multimedia content was entertaining to the majority of test subjects. The average level of entertainment was 7.62 (margin of error equals 0.15 with confidence level of 95 %). The figure shows that a better user experience is obtained when the content is entertaining to the subjects, as previously reported in [32].

In light of the previously presented results, it may be argued that this type of content, which is mostly entertaining to the subjects, softened their criticism level, making them less annoyed and more forgiving of the quality distortions that they experienced during the screening sessions. This claim can be related to previous observations made while discussing the results shown in Fig. 6 that the increase in AAL was steeper compared to the decrease in QoE ratings, perhaps because the content was sufficiently entertaining to redeem the overall user experience despite the perceived distortions.

Fig. 8 Difference between the actual and noticed number of video quality distortions





—○—	0.40	0.58	0.69	0.70	0.70	0.73
– □ –	0.46	0.51	0.54	0.59	0.66	0.61

Margin of error

—○—	0.29	0.37	0.64	0.66
– □ –	0.33	0.38	0.45	0.58

Margin of error

—○—	0.33	0.35	0.48	0.95	0.75	0.77	1.13	1.28	0.72
– □ –	0.48	0.38	0.39	0.59	0.79	0.55	1.06	0.89	0.83

Margin of error

Fig. 10 AALs of the *Rested* and *Tired* test subjects as functions of: **a** PLR, **b** number of PLOs, and **c** total duration of quality distortions

Table 7 Summary of the MW *U* test results for the data shown in Fig. 10

Subplot	Level of user fatigue	Comparison between the ratings	p value ($\alpha = 0.05$)	Exceptions (i.e., insignificant differences)
(a)	Rested	PLRs of ≤1 % and ≥1.5 %	≤0.0016	No exceptions
	Tired		≤0.0001	1 % ↔ 2 % ($p = 0.1294$)
(b)	Rested	No. of PLOs = 1 and of ≥4	≤0.0054	No exceptions
	Tired		≤0.0119	No exceptions
(c)	Rested	Total PLO durations of ≤10 s and ≥28 s	≤0.0013	4 s ↔ 49 s ($p = 0.0877$) 7 s ↔ 49 s ($p = 0.2485$) 10 s ↔ 40 s ($p = 0.0909$) 10 s ↔ 49 s ($p = 0.3839$)
	Tired	Total PLO durations of ≤10 s and ≥16 s	≤0.0391	7 s ↔ 16 s ($p = 0.0848$) 7 s ↔ 40 s ($p = 0.2535$) 10 s ↔ 28 s ($p = 0.6413$) 10 s ↔ 40 s ($p = 0.3291$) 10 s ↔ 49 s ($p = 0.1617$)

3.6 The impact of user fatigue and social context

We asked the subjects to self-evaluate their level of fatigue after the screening (question B9), using the following two ratings: *Rested* or *Tired*. Earlier, we discussed the results of the AAL analysis in relation to the packet loss intensity (Figs. 5, 6); however, the questionnaire also contained questions about subject annoyance level caused by the number of PLOs and the total duration of the quality distortions in the video (questions A3.3 and A3.4, respectively). The subjects rated these annoyance levels also on an 11-point numerical scale. Thus, we were able to evaluate the AALs of the subjects in relation to these three variables individually, and a separate analysis was conducted depending on the level of user fatigue (Fig. 10). Note that the margin of error is again calculated for the confidence level of 95 %.

The figure shows that both groups of test subjects exhibited similar adverse reactions to the experienced quality distortions. The MW *U* test results reveal no significant rating difference between the *Rested* and *Tired* test subjects in all three subplots with three exceptions: (a) in Fig. 10a, the difference is significant for the ratings corresponding to a PLR of 0.05 % ($p = 0.0126$), and (b) in Fig. 10c, the differences are significant for the ratings corresponding to durations of 10 and 28 s (with p values of 0.0019 and 0.0346, respectively). Because of the large number of ratings, Table 7 summarizes the MW *U* test results for the data shown in Fig. 10.

This experiment was conducted in real life; therefore, it could be argued that tired test subjects were also resting and relaxing during the screening of the video. Thus, they showed similar attitudes toward the perceived quality degradations as the rested test subjects, which may explain why the differences in ratings between these two groups of test subjects were predominantly insignificant. However, because the authors do not have the required expertise to interpret the dependencies between human physical and psychological conditions and human reasoning, more accurate observations are necessary from relevant experts in the field to help interpret the obtained results.

When comparing the ratings and the results of the MW *U* test for each of the two groups individually, it can be observed that (a) higher PLRs significantly affected user annoyance in both test groups; (b) the number of PLOs in the video significantly affected the users, especially when that number exceeded 4; and (c) as revealed by Table 7, tired test subjects exhibited a somewhat lower tolerance of degradations that lasted 16 s or more.

Nearly two-thirds, or 64 %, of the test subjects watched the video in someone's company. The average number of persons in the company of the test subjects was 1.18. The subjects who had company during the screening noticed more quality distortions compared to the subjects who watched the video alone (the average number of noticed distortions was 3.84 compared to 3.32, respectively). Consequently, the subjects who had company gave lower average QoE ratings (7.02) compared with subjects who were alone during the screenings (7.22), although the difference between these ratings was insignificant (the *p* value obtained using the MW *U* test is 0.4526). However, the subjects with company found the videos to be more entertaining (their average level of entertainment was 7.82) compared with the subjects without company (the average level of entertainment for this group was 7.51). For these ratings, $p = 0.0383$, which indicates a significant rating difference.

We can assume that during and after the screening of the video, the subjects who had company discussed what they had experienced. They exchanged opinions about the multimedia content and its quality as they are normally discussing everyday TV program, and by doing so, they increased the probability of memorizing the quality distortions. However, this was not significantly reflected in their QoE ratings.

3.7 The role of video subtitles

When aired in Croatia, foreign TV programs (TV shows, movies, talk shows, documentary films, etc.) include subtitles on the bottom of the screen. Because the narrator of the video that we used in this research narrates in English, we decided that our test sequences had to have Croatian subtitles as well. The Arial font was used for the subtitles. The text appeared on the bottom of the screen (maximum two rows of text), and each subtitle line was active for between 3 and 7 s (depending on the number of words on the screen). The text did not contain any grammatical or typographical errors and was correctly synchronized with the video. When evaluating the quality of our subtitles (question B7), we found that only 4 test subjects (or 0.66 %) thought that the subtitles were poorly made, which allows us to conclude that the quality of our subtitles did not negatively affect the subjects' experiences.

Because we conducted a large-scale study of user QoE, we used the opportunity to test the impact of video subtitles on user QoE as well. Our intention was to investigate whether video subtitles conceal quality distortions that are appearing on the screen by drawing the viewer's attention to the text at the bottom of the screen. To the best of our knowledge, no prior work has been performed on this issue. For this purpose, we kindly asked our colleagues (English teachers and assistants at our university) to watch the video in a real-life environment without video subtitles. Test sequence number 57 (see Table 1) was chosen for this test, because it contained 10 PLOs of moderate intensity for this type of service. We distributed the sequence to 15 of our colleagues who were also uninformed about the topic of the research prior to watching the video. The results from this test group were compared to those obtained from the student population in the first test group. Note that the sequence number 57 was evaluated by 9 test subjects from the first group (students who watched the video with subtitles), and we accepted 12 questionnaires from the second test group (our colleagues who watched the video without subtitles).

When asked if they noticed any quality distortions during the screening (question A2), 44.44 % of the test subjects from the first group responded negatively, compared with a mere 6.67 % of our colleagues from the second group. Because they noticed fewer quality distortions, the average QoE rating from the subjects in the first test group was higher than the average rating from the second group (7.67 compared with 6.31, respectively). The MW *U* test returned a *p* value of 0.029, indicating a significant rating difference between the QoE ratings of these two test groups.

The differences between these two groups are also visible when comparing the annoyance levels caused by the perceived quality distortions (question A3.2). Test subjects from the second group experienced higher AALs as a function of packet loss intensity (2.55) compared with the results of the first group (1.11). The *p* value calculated using the MW *U* test is 0.025, i.e., the difference between these two ratings is significant.

These results confirmed our suspicions that the existence of video subtitles can impact the user experience. When reading the subtitles, almost half of the test subjects from the first test group failed to notice all 10 quality distortions that appeared on the screen, which affected their rating. We are motivated by these results to further investigate the role of video subtitles in other test sequences. However, we would have to conduct such a test on a different target group, because our colleagues are now aware of the purpose of our research.

4 Conclusions and outlook

Different challenges are faced when attempting to conduct real-life subjective evaluations of QoE for multimedia streaming. The service must be used during the everyday routine of the users, and the test subjects must be as uninformed as possible regarding the purpose of the test. Furthermore, network performances should be recorded during service usage for further analysis.

To overcome these challenges, the test sequences were generated in an emulated network environment using a 1-h documentary film. Second, the sequences were distributed to the test subjects on a DVD. This format enabled the subjects to consume the content in a real-life environment where they would typically watch similar content, e.g., a TV program. The sequences contained video artifacts that appeared during streaming in an emulated network environment; hence, the subjects experienced quality distortions of the types that sometimes occur during normal streaming sessions.

The results revealed that the test environment, the content properties and the video subtitles affected the users' experiences. The impact was found, and the dependencies between the following three objective parameters were interpreted: the packet loss rate, the number of packet loss occurrences in one streaming session and the duration of those occurrences. It was found that sequences with only one PLO are generally perceived as being of *Good* or *Excellent quality*, regardless of the PLR and the duration of a single PLO. However, it was observed that (a) users negatively perceive quality degradations when the PLR is $\geq$1 % if the degradations last at least 16 s; (b) if the video contains 7 or more PLOs, the duration of a single PLO becomes an important factor as the PLR increases; and (c) an increase in the number of PLOs significantly affects user QoE for PLRs of $\geq$0.5 %. Based on the obtained results, the objective parameters can be ranked by their order of importance in relation to their impact on user QoE as follows: (1) total duration of quality distortions in a video, (2) number of PLOs, (3) PLR, and (4) duration of a single PLO.

Furthermore, it can be argued that in a real-life context, the occasional decrease in network performance will not be adversely perceived by the service users. This implies that a certain level of flexibility exists when trying to match particular QoS demands of different services in IP networks. However, it should be stressed that we have used longer test sequences, which means that the impact of humans' short-term memory and recency effects must not be neglected. The evaluation of user QoE on different types of content (i.e., music videos, which are shorter than documentary films) may produce different results.

The content, if it is sufficiently entertaining to the viewers, can redeem the overall user experience despite the perceived quality distortions. More entertaining content causes users to be more forgiving of the occasional advent of various video artifacts. Furthermore, differences in user perception were observed if the content is consumed with company vs. without company. The subjects who had company exchanged their opinions about the video content and its quality during and after the screening. This increased the probability of memorizing the quality distortions; hence, the subjects who had company reported noticing more quality distortions compared with the subjects who watched the video alone. Finally, we found support for the hypothesis that the existence of subtitles can divert viewer focus from the image and have an impact on user experience.

Throughout this paper, various paths of future research were highlighted to include the following: the further investigation of quality distortions that can lead to the worst possible user QoE ratings, an analysis of the impact of video subtitles on a wider variety of test sequences and the subjective evaluation of QoE on different types of content. Furthermore, the correlations between different quality degradations and the level of user annoyance and QoE, disclosed in this study, will be used for the development of the inference system of the objective video quality assessment model for assessing the user QoE. In addition, we plan to conduct the analysis of user QoE for HTTP-based video streaming when different objective parameters come to the fore (for instance, network delay, video frame rate and bit rate, buffering time, and rebuffering frequency).

Acknowledgments The authors wish to express special gratitude to all of our colleagues and students who participated in this research as well as to the unknown reviewers, who considerably contributed to improving the quality of the paper.

Appendix 1: The questionnaire used in the study

The questionnaire used in this study consisted of four pages. Pages 1 and 4 are not included in this appendix, because page 1 contained only the instructions on how to complete the questionnaire, whereas page 4 contained several general questions (regarding subject demographic information such as age group, etc.) and a blank space where the subjects were able to leave comments.

A) THE PERCEIVED VIDEO QUALITY

A1. Mark on the scale your opinion of the audiovisual quality of the video that you have just finished watching:

□□□□□□□□□□

0 1 2 3 4 5 6 7 8 9 10

Bad quality | Poor quality | Fair quality | Good quality | Excellent quality

A2. While watching the video, I noticed that the video quality was degraded on one or multiple occasions.

Is this statement true?
a) Yes. b) No.

A3. If you answered the previous question with **YES**, then please proceed to the next questions (**A3.1, A3.2, A3.3, A3.4 and A3.5**). If you answered the previous question with **NO**, then please skip to **Section B of the questionnaire**.

A3.1 What types of video quality degradation did you notice?

(mark your answer(s); multiple answers are possible)

a) The video appeared choppy (i.e., the reproduction was not "smooth").
b) The video was delayed in relation to the audio (synchronization issues).
c) The video image was incomplete (parts of the picture were not shown).
d) The video froze (the reproduction stopped).
e) Some parts of the video image appeared as if they were assembled from blocks.
f) The video image was split into several sections and it was clear that some sections were not a part of the current video image.
g) The video image appeared to be "broken" in some parts of the screen.
h) The video image contained colored blocks. It was clear that these blocks were not a part of the video image.
i) The audio was choppy.
j) The audio was delayed in relation to the video (synchronization issues).
k) The audio was incomplete (parts of the audio were not reproduced).
l) The audio reproduction stopped.
m) The reproduction of the entire content of the video stopped and then restarted after some amount of time.

If you experienced something that cannot be described by any of these answers, then please write what you experienced below:

A3.2 When you reflect back on the quality distortions that you experienced during the screening, you would say that they were:

□□□□□□□□□□

0 1 2 3 4 5 6 7 8 9 10

Imperceptible quality distortions | Perceptible quality distortion, but not annoying | Slightly annoying quality distortion | Annoying quality distortion | Very annoying quality distortion

A3.3 You noticed that distortions appeared in the video approximately _____ times *(write a number)*. You think that this was a:

□□□□□□□□□□

0 1 2 3 4 5 6 7 8 9 10

Negligible frequency | Low frequency, but not annoying | Medium frequency, slightly annoying | Annoyingly high frequency | Very annoying frequency

A3.4 If you were to quantify the total amount of time for which the quality distortions appeared on the screen, that time would be equal to _____ seconds *(write a number)*. You think that this was a:

□□□□□□□□□□

0 1 2 3 4 5 6 7 8 9 10

Negligible duration | Short duration, but not annoying | Medium duration, slightly annoying | Annoyingly long duration | Very annoying duration

A3.5 Considering the types of degradations that you noticed, their appearing frequency and total duration, how do you evaluate your experience of watching this video?

□□□□□□□□□□

0 1 2 3 4 5 6 7 8 9 10

Bad experience | Poor experience | Fair experience | Good experience | Excellent experience

B) ABOUT THE CONTENT, USER ENVIRONMENT, SOCIAL CONTEXT AND OTHER

B1. Mark on the scale how entertaining the video was to you.

0	1	2	3	4	5	6	7	8	9	10
	Least entertaining, boring		Partially entertaining		Entertaining		Mostly entertaining		Very entertaining	

B2. Did you watch the entire video without interruptions?

a) I watched the entire video without interruptions.

b) I watched the entire video, but with interruptions. No. of interruptions: _____ . *(write a number)*

c) I did not watch the entire video.

B3. How much noise was there in your surroundings while you were watching the video?

a) There was no noise whatsoever.

b) There was some noise, but not enough to distract me from watching.

c) It was a bit noisy, enough to disrupt my concentration for a short period of time.

d) There was a lot of noise, so much that I was unable to concentrate on the video for long time periods.

B4. Mark the statements that are applicable to you. If there aren't any or you are unsure, proceed to the next question.

(mark your answer(s); multiple answers are possible)

a) When I watch DVDs as I usually do, their quality is often degraded.

b) There is a possibility that my DVD player that I used to watch this video may be broken or malfunctioning.

c) I watched the video on a screen with a 4:3 aspect ratio.

d) I watched the video on a screen with a 16:9 aspect ratio.

e) My screen supports the HD format (Full HD or HD ready).

f) I watched the video on a CRT screen.

B5. What was the social context in which you watched the video?

a) I watched the video alone.

b) I watched the video in the company of ____ persons. (*write a number*)

B6. If you answered the previous question with b), did that person(s) suggest to you in any way that the quality of the video was degraded?

a) No, I noticed on my own that the quality was degraded.

b) Yes, without the person(s) in my company, I would not have noticed the quality degradations in the video.

c) No one noticed any quality degradations.

B7. What do you think about the video subtitles?

(mark your answer(s); multiple answers are possible)

a) Without the subtitles, I would not have understood the content.

b) They were useful, but I would be able to watch the video without them.

c) The subtitles were only distracting me.

d) The quality of the subtitles was good.

e) The quality of the subtitles was poor.

f) Instead of the subtitles, I would prefer a Croatian narrator.

B8. Do you see and hear well?

a) Yes, I see and hear well (either with or without visual and hearing aids).

b) I have impaired hearing.

c) I have impaired sight.

d) I have impaired sight and hearing.

B9. Where you tired while watching the video?

a) Yes, I was tired.

b) No, I was rested.

B10. Did you complete this questionnaire immediately after watching the video?

a) Yes.

b) No.

B11. Were you familiar with the topic of this research prior to watching the video?

a) Yes.

b) No.

Table 8 The number of rejected questionnaires for each specific criterion

Rejection criterion	No. of rejected questionnaires
Responses to question A3.1 not relating to the actual quality degradations in a test sequence	32
Responses to questions A1 and A3.5 differing by more than 8 points	36
Abnormal rating on question A3.3 (noticed number of quality distortions ≤2 but rated ≥6 on the annoyance scale)	39
Abnormal rating on question A3.4 (total duration of all video quality distortions ≤10 s but rated ≥6 on the annoyance scale)	40
Response "c" provided to question B2	12
Response "d" provided to question B3	16
Response "a" and/or "b" provided to question B4	23
Response "b" provided to question B6	19
Response "c" provided to question B6 when response "a" was provided to question A2	10
Response "b", "c" or "d" provided to question B8	6
Response "b" provided to question B10	19
Response "a" provided to question B11	15
The questionnaire was not fully completed	24

Appendix 2: Rejected questionnaires

The number of rejected questionnaires for each specific criterion is shown in Table 8. Note that some questionnaires were rejected for several criteria simultaneously; thus, the sum of the numbers of rejected questionnaires in the second column of the table exceeds 228.

References

1. Kaikkonen, A., Kekäläinen, A., Cankar, M., Kallio, T., Kankainen, A.: Usability testing of mobile applications: a comparison between laboratory and field testing. J. Usability Stud. (2005) 1, 1:4–17. http://uxpajournal.org/usability-testing-of-mobile-applications-a-comparison-between-laboratory-and-field-testing/. Accessed 21 May 2014
2. Sun, X., May, A.: A comparison of field-based and lab-based experiments to evaluate user experience of personalised mobile devices. Adv. Hum. Comput. Interact. **2013**, 1–10 (2013). doi:10.1155/2013/619767
3. Matulin, M., Mrvelj, Š.: State-of-the-practice in evaluation of quality of experience in real-life environments. Promet **25**(3), 255–263 (2013). doi:10.7307/ptt.v25i3.1195
4. Reichl, P., Fröhlich, P., Baillie, L., Schatz, R., Dantcheva, A.: The liliput prototype: a wearable lab environment for user tests of mobile telecommunication applications. In: Proceedings of the 2007 Conference on Human Factors in Computing Systems, pp. 1833–1838. San Jose, California (2007). doi:10.1145/1240866.1240907
5. Jumisko-Pyykkö, S., Hannuksela, MM.: Does context matter in quality evaluation of mobile television? In: Proceedings of the 10th Conference on Human-Computer Interaction with Mobile Devices and Services, pp. 63–72, Amsterdam, Netherland, (2008). doi:10.1145/1409240.1409248
6. Staelens, N., Moens, S., Van den Broeck, W., Mariën, I., Vermeulen, B., Lambert, P., Van de Walle, R., Demeester, P.: Assessing the perceptual influence of H.264/SVC signal-to-noise ratio and temporal scalability on full length movies. In: Proceedings of the First International Workshop on Quality of Multimedia Experience, pp. 29–34, San Diego, California, USA, San Diego, California, (2009). doi:10.1109/QOMEX.2009.5246982
7. Staelens, N., Moens, S., Van den Broeck, W., Mariën, I., Vermeulen, B., Lambert, P., Van de Walle, R., Demeester, P.: Assessing quality of experience of IPTV and video on demand services in real-life environments. IEEE Trans. Broadcast. **56**(4), 458–466 (2010). doi:10.1109/TBC.2010.2067710
8. Ickin, S., Wac, K., Fiedler, M., Janowski, L., Hong, J.H., Dey, A.K.: Factors influencing quality of experience of commonly used mobile applications. IEEE Commun. Mag. **50**(4), 48–56 (2012). doi:10.1109/MCOM.2012.6178833
9. Van den Broeck, W., Jacobs, A., Staelens, N.: Integrating the everyday-life context in subjective video quality experiments. In: Proceedings of the 4th International Workshop on Quality of Multimedia Experience (QoMEX), pp. 19–24, Yarra Valley, Australia, (2012). doi:10.1109/QoMEX.2012.6263848
10. Nidhi Aggarwal, N.: A review on video quality assessment. In: Proceedings of the Recent Advances in Engineering and Computational Sciences (RAECS), pp. 1–6, Chandigarh, India, (2014). doi:10.1109/RAECS.2014.6799645
11. Rodríguez, D.Z., Rosa, R.L., Costa, E.A., Abrahão, J., Bressan, G.: Video quality assessment in video streaming services considering user preference for video content. IEEE Trans. Consum. Electron. **60**(3), 436–444 (2014). doi:10.1109/TCE.2014.6937328
12. Xu, Q., Huang, Q., Yao, Y.: Online crowdsourcing subjective image quality assessment. In: Proceedings of the 20th ACM International Conference on Multimedia, pp. 359–368, Nara, Japan (2012). doi:10.1145/2393347.2393400
13. Gardlo, B., Ries, M., Hossfeld, T., Schatz, R.: Microworkers vs. facebook: the impact of crowdsourcing platform choice on

experimental results. In: Proceedings of the 4th International Workshop on Quality of Multimedia Experience (QoMEX), pp. 35–36, Yarra Valley, Australia, (2012). doi:10.1109/QoMEX.2012.6263885
14. Hoßfeld, T., Keimel, C., Hirth, M., Gardlo, B., Habigt, J., Diepold, K., Tran-Gia, P.: Best practices for QoE crowdtesting: QoE assessment with crowdsourcing. IEEE Trans. Multimed. **16**(2), 541–558 (2014). doi:10.1109/TMM.2013.2291663
15. Sladojevic, S., Culibrk, D., Mirkovic, M., Coll, DR., Borba, GR.: Logging real packet reception patterns for end-to-end quality of experience assessment in wireless multimedia transmission. In: Proceedings of the IEEE International Conference on Multimedia and Expo Workshops (ICMEW), pp. 1–6, San Jose, California (2013). doi:10.1109/ICMEW.2013.6618453
16. Staelens, N., De Meulenaere, J., Claeys, M., Van Wallendael, G., Van den Broeck, W., De Cock, J., Van de Walle, R., Demeester, P., De Turck, F.: Subjective quality assessment of longer duration video sequences delivered over HTTP adaptive streaming to tablet devices. IEEE Trans. Broadcast. **60**(4), 707–714 (2014). doi:10.1109/TBC.2014.2359255
17. Ickin, S., Fiedler, M., Wac, K., Arlos, P., Temiz, C., Mkocha, K.: VLQoE: video QoE instrumentation on the smartphone. Multimed. Tools Appl. **74**(2), 381–411 (2015). doi:10.1007/s11042-014-1919-0
18. Fröhlich, P., Egger, S., Schatz, R., Mühlegger, M., Masuch, K., Gardlo, B.: QoE in 10 seconds: Are short video clip lengths sufficient for quality of experience assessment? In: Proceedings of the 4th International Workshop on Quality of Multimedia Experience (QoMEX), pp. 242–247, Yarra Valley, Australia (2012). doi:10.1109/QoMEX.2012.6263851
19. Li, W., Rehman, HU., Kaya, D., Chignell, M., Leon-Garcia, A., Zucherman, L., Jiang, J.: Video quality of experience in the presence of accessibility and retainability failures. In: Proceedings of the 10th International Conference on Heterogeneous Networking for Quality, Reliability, Security and Robustness (QShine), pp. 1–7, Rhodes, Greece (2014). doi:10.1109/QSHINE.2014.6928651
20. Hands, D.S., Avons, S.E.: Recency and duration neglect in subjective assessment of television picture quality. Appl. Cogn. Psychol. **15**(6), 639–657 (2001). doi:10.1002/acp.731
21. Datta, P., Izdebski, L., Kumar, N., Suh, K.: "It came to me in a stream…" The upward arc of online video, driven by consumers. Cisco white paper. https://www.cisco.com/web/about/ac79/docs/sp/Online-Video-Consumption_Consumers.pdf (2012). Accessed 17 October 2014
22. Farrokhi, F., Mahmoudi-Hamidabad, A.: Rethinking convenience sampling: defining quality criteria. Theory Pract. Lang. Stud. **2**(4), 784–792 (2012). doi:10.4304/tpls.2.4.784-792
23. International Telecommunication Union: Subjective video quality assessment methods for multimedia applications. International Telecommunication Union (ITU-T Rec. P.910). http://www.itu.int/rec/T-REC-P.910-200804-I (2008). Accessed 15 April 2014
24. Borowiak, A., Reiter, U.: Long duration audiovisual content: impact of content type and impairment appearance on user quality expectations over time. In: Proceedings of the 5th International Workshop on Quality of Multimedia Experience (QoMEX), pp. 200–205, Klagenfurt, Austria, (2013). doi:10.1109/QoMEX.2013.6603237
25. Menkovski, V., Exarchakos, G., Liotta, A., Cuadra Sánchez, A.: Estimations and remedies for quality of experience in multimedia streaming. In: Proceedings of the 3rd International Conference on Advances in Human-Oriented and Personalized Mechanisms, Technologies and Services, pp. 11–15, Nice, Italy, (2010). doi:10.1109/CENTRIC.2010.14
26. Reichl, P., Egger, S., Schatz, R., D'Alconzo, A.: The logarithmic nature of QoE and the role of the Weber-Fechner Law in QoE assessment. In: Proceedings of the IEEE International Conference on Communications Workshops (ICC), pp. 1–5, Cape Town, South Africa (2010). doi:10.1109/ICC.2010.5501894
27. Fiedler, M., Hoßfeld, T.: Quality of experience-related differential equations and provisioning-delivery hysteresis. In: Proceedings of the 21st ITC Specialist Seminar on Multimedia Applications-Traffic, Performance and QoE, Miyazaki, Japan, https://www.diva-portal.org/smash/get/diva2:835271/FULLTEXT01.pdf (2010). Accessed 5 July 2016
28. Fiedler, M., Hoßfeld, T., Tran-Gia, P.: A generic quantitative relationship between quality of experience and quality of service. IEEE Netw. **24**(2), 36–41 (2010). doi:10.1109/MNET.2010.5430142
29. Jelassi, S., Rubino, G., Melvin, H., Youssef, H., Pujolle, G.: Quality of experience of VoIP service: a survey of assessment approaches and open issues. IEEE Commun. Surv. Tutor. **14**(2), 491–513 (2012). doi:10.1109/SURV.2011.120811.00063
30. Shen, Y., Liu, Y., Liu, Q., Yang, D.: A method of QoE evaluation for adaptive streaming based on bit rate distribution. In: Proceedings of the IEEE International Conference on Communications Workshops (ICC), pp. 551–556, Sydney, Australia (2014). doi:10.1109/ICCW.2014.6881256
31. Pinson, MH., Boyd, KS., Hooker, J., Muntean, K.: How to choose video sequences for video quality assessment. In: Proceedings of the Seventh International Workshop on Video Processing and Quality Metrics for Consumer Electronics (VPQM-2013), pp. 79–85, Scottsdale, Arizona (2013)
32. Matulin, M., Mrvelj, Š.: Subjective evaluation of quality of experience for video streaming service. In: Proceedings of the 2nd Research Conference In Technical Disciplines, pp. 60–64, Žilina, Poland, http://bib.irb.hr/datoteka/739619.RCITD-2014.pdf (2014). Accessed 12 January 2015