IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received July 22, 2021, accepted August 21, 2021, date of publication August 24, 2021, date of current version August 31, 2021.

Digital Object Identifier 10.1109/ACCESS.2021.3107619

INTENSE: In-Depth Studies on Stall Events and Quality Switches and Their Impact on the Quality of Experience in HTTP Adaptive Streaming

BABAK TARAGHI, (Member, IEEE), MINH NGUYEN, (Member, IEEE), HADI AMIRPOUR, (Member, IEEE), AND CHRISTIAN TIMMERER, (Senior Member, IEEE)
Christian Doppler Laboratory ATHENA, Alpen-Adria-Universität Klagenfurt, 9020 Klagenfurt am Wörthersee, Austria

Corresponding author: Babak Taraghi (babak.taraghi@aau.at)

This work was supported in part by the Austrian Federal Ministry for Digital and Economic Affairs, in part by the National Foundation for Research, Technology, and Development, and in part by Christian Doppler Research Association.

This work involved human subjects or animals in its research. The authors confirm that all human/animal subject research procedures and protocols are exempt from review board approval.

ABSTRACT With the recent growth of multimedia traffic over the Internet and emerging multimedia streaming service providers, improving Quality of Experience (QoE) for HTTP Adaptive Streaming (HAS) becomes more important. Alongside other factors, such as the media quality, HAS relies on the performance of the media player's Adaptive Bitrate (ABR) algorithm to optimize QoE in multimedia streaming sessions. QoE in HAS suffers from weak or unstable internet connections and suboptimal ABR decisions. As a result of imperfect adaptiveness to the characteristics and conditions of the internet connection, stall events and quality level switches could occur and with different durations that negatively affect the QoE. In this paper, we address various identified open issues related to the QoE for HAS, notably *(i)* the minimum noticeable duration for stall events in HAS; *(ii)* the correlation between the media quality and the impact of stall events on QoE; *(iii)* the end-user preference regarding multiple shorter stall events versus a single longer stall event; and *(iv)* the end-user preference of media quality switches over stall events. Therefore, we have studied these open issues from both objective and subjective evaluation perspectives and presented the correlation between the two types of evaluations. The findings documented in this paper can be used as a baseline for improving ABR algorithms and policies in HAS.

INDEX TERMS Crowdsourcing, HTTP adaptive streaming, quality of experience, quality switches, stall events, subjective evaluation, objective evaluation.

I. INTRODUCTION

In the last decade, multimedia traffic has been increased significantly. According to a study by Cisco [1], video data accounted for 75% of the global internet traffic in 2017, and this figure is estimated to reach 82% by 2022. This trend exposes challenges in providing high Quality of Experience (QoE) for end-users due to network limitations and increased quality and latency demands [2]. The advent of the HTTP Adaptive Streaming (HAS) technique was a major technical milestone for multimedia delivery over the Internet. In HAS, a multimedia file is encoded at multiple bitrates, video resolutions, audio sample rates, and other factors, *i.e.*, *representations*. Each representation will then be split into temporal segments and stored on a simple HTTP server. The media player, on the client-side, runs an adaptive bitrate (ABR) algorithm to select segments of the most suitable representation to be downloaded. Due to network fluctuations, especially in a mobile network [3], an ABR algorithm might request segments from various representations with different quality levels during the streaming session, leading to quality switches and stall events [4], [5]. A quality switch is determined when the qualities of two contiguous segments are different. Quality switches can be divided into two groups: *(a) upward quality switch*, where the quality of a segment is higher than that of the previous one, *(b) downward quality switch*, in which segment's quality is lower than the previous segment's quality. Please note

The associate editor coordinating the review of this manuscript and approving it for publication was Diego Oliva.

that subjects are more critical to downward quality switches than upward quality switches [6]. An analysis of a trace containing 5 million media streaming sessions in [7] shows that more than 36% of the streaming sessions have at least one downward quality switch. A stall event occurs when the media playback stops playing as there is no downloaded segment available in the client's playback buffer. To inform the end-user about such an event, the media player will show an indicator, often in the form of a spinning wheel. We identify the stall events by two attributes *stall event frequency* and *stall event duration* in a streaming session. These attributes are often taken into account to calculate the predicted Mean Opinion Score (MOS) by QoE models such as in [8]–[10]. QoE is defined as the measure of the delight or annoyance of a customer's experiences with a service [11]. MOS often presents QoE in HAS. Many objective QoE models [8], [12], [13] use metrics such as media bitrate, quality switches, and stall events to predict the MOS for a streaming session algorithmically, and some also include the startup delay [14]. To calculate the perceived or actual MOS, it is recommended to conduct subjective evaluations and seek out the subjects' opinions on a specific streaming session. Subjective evaluation reveals actual or perceived MOS, but it is considered time-consuming and costly [15]. Although stall events and quality switches are considered disturbing and sometimes annoying for end-users [16], [17], their impact on the QoE and the end-user preference over the two phenomenons have not been fully understood.

This study offers insights into the impact of the aforementioned multimedia streaming session defects on the end-users QoE. By conducting objective and subjective evaluations, we define the contributions of this paper as five-fold:

- **Minimum Noticeable Stall event Duration (MNSD) Evaluation.** We have investigated the minimum threshold of a stall event duration that is noticeable by end-users. Therefore, stall events with a smaller duration than the determined threshold are considered not detectable and, thus, do not affect the perceived QoE.
- **Stall event vs. Quality level switch (SvQ) Evaluation.** From a high-level approach to ABR algorithms, when the network condition is not favorable, there are two main possible scenarios. First, the ABR scheme continues the media playback with the same representation but suffers from a stall event. The second scenario is to decrease the quality without introducing a stall event. We assessed the end-user preference regarding these two scenarios.
- **Short stall events vs. a Longer stall event (SvL) Evaluation.** Frequent changes in media representation selection (*i.e.*, high quality variation) may result in multiple short stall events. In contrast, the client may alternatively reduce the quality variation and have one longer stall event. We studied the impact of multiple short stall events in contrast with a single longer stall event on the QoE from both predicted and perceived MOS perspectives.
- **Relation of Stall event impact on the QoE with Video Quality level (RSVQ) Evaluation.** Zeng *et al.* [18] conclude that stall events have a higher negative impact on the QoE when the video is at a high-quality level. However, Yamagishi and Hayashi [19] disagree with this finding based on their subjective experiments. In this work, we examine the possible relation between the stall events' effect on the QoE and the media quality.
- **Objective QoE Models Comparison.** We have compared the state-of-the-art QoE objective evaluation models with the MOS retrieved from our subjective tests and studied their correlations.

The remainder of the paper is organized as follows. Section II reviews related work, followed by the evaluation setup in Section III. In Section IV, the experimental results, statistical analysis, and key findings are described in detail. Finally, Section V concludes the paper and outlines future work.

II. RELATED WORK

Hossfeld *et al.* [20] conducted subjective tests to study the impacts of initial delays and stall events on the QoE and compared their effects. The subjective test results demonstrate that QoE for a given initial delay depends on the application. Besides, it is shown that initial delays are less harmful to QoE compared to stall events. Another subjective study evaluated the trade-off between the startup *delay* before a video starts playing due to the buffer filling with *interrupts* in the middle or stall events to determine the buffer size that maximizes the QoE [21]. Their analysis reveals that stall events have a two times more significant impact on the QoE than the buffering for a given amount of time. The influence of stall events is investigated in [22]. The subjects mostly preferred the media playback fluidity to a media playback with higher quality but with stall events. The acceptability of the quality of a video session was also measured as a function of the waiting time during video playback. Users accept video sessions with a high probability (more than 75%) if their overall waiting time is less than 20s. On the other hand, video sessions with more than 60s waiting times are generally (more than 75%) not accepted by users. Effects of stall events on the QoE of mobile streaming videos were studied in [23]. Based on the subjective evaluation results, the following conclusions were made: *(i)* (stall event position) video sessions with a stall event toward the end of the video have a higher Differential Mean Opinion Score (DMOS) than the same stall events' pattern at the beginning of the video; *(ii)* (stall frequency) the higher the frequency of stall events, the higher the DMOS value; *(iii)* (stall duration) video sessions with longer stall events have higher DMOS than those with shorter stall events; *(iv)* (sequence length) the same stall events' pattern has a higher (negative) influence on the shorter sequences than on longer sequences; *(v)* (total stall duration) the total duration of stall events has less impact than their position, frequency, and length. Meanwhile, the work in [24] provided

opposite findings. Subjective test results showed that the negative effect of stall events is gradually decreased when their location goes toward the end of the video session. Also, shorter stall event intervals have a higher negative impact than longer stall event intervals.

Abrupt and smooth quality switches, high and low-frequency switching, as well as corresponding stall events, were studied in [25]. It was observed that smooth switching is not significantly better than abrupt switching when adapting towards lower quality. It was also observed that frequent quality adaptation is not perceived considerably worse than videos with less frequent quality adaptation. Stall events also showed a similar influence on the QoE as video adaptation.

Staelens *et al.* [26] subjectively evaluated the trade-off between video stall event duration and initial video quality in the case of camera switching during adaptive streaming of sports content. The subjective results indicate that short stall events do not significantly influence the overall quality ratings. However, the quality perception is strongly influenced by the video quality at the moment of camera switching. Besides, substantial-quality fluctuations should be avoided.

Tavakoli *et al.* [27] subjectively assessed the quality of an adaptation model in HAS. To design the experiment, the following factors have been considered: *(i)* adaptation strategy (gradually or rapidly change in the quality), *(ii)* adaptive streams (4 streams (bitrate) with 600, 1000, 3000, and 5000 kbps), *(iii)* content type (seven types), and *(iv)* segment size (2 and 10 seconds). In general, 3 Mbps and 5 Mbps (in constant quality) compressed videos were not perceived differently, while both were preferred to the increasing scenarios, both rapidly and gradually increasing scenarios. For 1 Mbps, both the consistent quality and the increasing quality scenarios show similar preferences. All the adaptation strategies showed better QoE compared to the 600 kbps encoded video. The overall results show similar performance for increasing quality scenarios, *i.e.*, gradually increasing quality strategies. Regarding the scenarios to decrease the quality, a strong preference was found for the gradual bit rate changes for segments with a duration of 10 seconds compared to the others. It was also found that content type and Spatio-temporal information significantly impact adaptation strategies and video playback.

Garcia *et al.* [28] studied the impact of initial delay, stall events, and video bitrate on High Definition (HD) audiovisual sequences. Two short (30s) and long (60s) sequences were considered. The authors provided four main findings. First, the subjective evaluation results showed that startup delay has a meager impact on perceived quality. Second, the time interval between stall events does not influence the impact of stall events, at least for 30s video sequences. Third, different stall event durations and frequencies do not have a significant impact on the QoE. Finally, it was shown that the impact of stall events is independent of the video content at a high bitrate on the perceived video quality. However, the quality of each tested video is unchanged. Thus, the relation between stall events and media quality switching is not investigated.

Duanmo *et al.* [29] investigated the human responses to the combined effect of video compression, initial buffering, and stall events. Their subjective evaluation found that subjects tend to give a higher penalty to the video with higher quality at the freezing frame with stall events at the same temporal instance and of the same duration. Yamagishi and Hayashi [19] developed a model to predict the quality of adaptive-bitrate-streaming services as a function of video resolution, the audio and video bitrate, bitrate adaptation, stall events, and the segment duration. Their work found that stall events on high media qualities result in minor impairment than ones that occur when the media is of lower quality. Meanwhile, Zeng *et al.* [18] presents a contradictory statement. In this paper, we will examine these opposite conclusions based on our subjective results.

Bampis *et al.* [30] conducted subjective tests on the LIVE-Netflix Video QoE Database to evaluate the temporal effects on the QoE. The authors found that subjects prefer transient quality drops to stall events only on low complexity video content. Although the database simulates a video streaming application, it contains mixtures of quality changes and stall events. Therefore, the minimum noticeable stall event duration cannot be inferred from the experimental results.

Rodríguez *et al.* [31] found that *switching frequency*, *switching type* (*i.e.*, spatial and temporal resolutions), and *switching temporal location* are three critical factors of media quality switches that impact the QoE. Tran *et al.* [17] conducted subjective tests to formulate a multi-factor QoE model. They found that stall events with more than 2 seconds length result in more severe degradation of the user's QoE than any quality switches.

Unlike these related works, in this paper, we do not only investigate different aspects of the impact of stall events (*i.e.*, stall duration and frequency) on the QoE but also the relationship between stall events and quality level switching on QoE impairment is evaluated. We have conducted extensive crowdsourcing subjective evaluations to provide reliable results of user's perceptions. Also, an in-depth analysis of the obtained subjective results and comparison among recent state-of-the-art QoE models have been made.

III. EVALUATION SETUP

To study the effect of stall events on QoE, we have conducted both objective and subjective evaluations. We will describe the characteristics of video sequences, our testbed, objective QoE evaluation models, and the setup to conduct the subjective evaluations in this section. To mimic different stall events occurrence and quality level switches in HAS, we have designed multiple patterns described in the following subsections.

A. TEST SEQUENCES

The following open source movies are used for the evaluations as proposed in [32]. To encode and package the DASH

TABLE 1. Bitrate ladder of the test sequences.

Index	Resolution	Video Bitrate	Audio Bitrate
1	320x240	235Kbps	128Kbps
2	512x384	560Kbps	128Kbps
3	640x480	1050Kbps	128Kbps
4	1280x720	2350Kbps	128Kbps
5	1920x1080	7000Kbps	128Kbps

content, we have used FFmpeg[1] with the following configurations and parameters. The bitrate ladder that we have used for packaging the DASH assets is shown in Table 1.

1) Sintel, the Durian Open Movie Project[2]
2) Valkaama[3]
3) Big Buck Bunny[4]
4) Tears of Steel[5]

The above movies are encoded and packaged with the following parameters: AAC for audio coding, AVC/H.264 (x264) for video coding, segment duration of two seconds, and group-of-pictures (GoP) length of 24 frames, frame per second (fps) of 24, MP4 segment type, and DASH format. We used two different parts with two minutes duration from each selected multimedia sequence to extend the data set varieties. Each video was encoded into five representations using the encoding bitrate ladder highlighted in Table 1. The choices of bitrate levels and encoding configurations were based on Netflix's recommendation [33] and Apple's recommendation [34], which is presented in [35]. As proposed in [36], we followed Streamroot's encoding configuration recommendation [37] to remove scene cuts and limit the GoP size. Segments have two seconds duration as proposed in [32], [38].

B. STALL EVENTS' PATTERNS

To investigate the stated multimedia streaming session defects, namely, stall events and quality switches in HAS, we have designed eleven stall events' patterns for MNSD, SvQ, SvL, and RSVQ evaluations that will be described in this subsection.

1) MINIMUM NOTICEABLE STALL EVENT DURATION EVALUATION

For determining the minimum noticeable stall event duration threshold, we have studied the stall events with a step of one millisecond and starting from one millisecond to the maximum one-second duration. To cancel the effect of media content on stall event noticeability, we have repeated each stall event four times and in four different points of different test sequences. We have randomly distributed the 4000 stall events in 32 test sequences which were cut out with a length of one minute from the original test sequences introduced in Subsection III-A. As a result, we have produced 800 test sequences with five stall events in each. Each stall event has been populated with a minimum gap time of three seconds from the previous and the next stall event in the same test sequence.

2) STALL EVENT VS. QUALITY LEVEL SWITCH EVALUATION

We have designed two sets of stall events' patterns to understand the subjects' preferences between stall events and quality switches. These stall events' patterns are depicted in Figure 1. In Set A, two cases are being covered. First, the streaming session starts with representation index one from Table 1 (*i.e.*, lowest bitrate), and it continues for twenty seconds with the same representation. We have added a stall event with a duration of six seconds, which mimics an average of a real-life setting's long stall event duration [17], and then we introduce an upward quality switch to the third representation index and continue the playback for another twenty seconds. As a second case compared with the first case, we continue the playback for forty seconds with representation index one without any stall event. In Set B, first case, the streaming session starts from representation index three from Table 1, and it continues for twenty seconds with the same representation. We have added a stall event with a duration of six seconds, and then the playback continues with the same representation for another twenty seconds. As a second case to be compared with the first case in Set B, the playback starts with representation index three and continues for twenty seconds. Then a downward quality switch occurs, and the playback continues with representation index one for another twenty seconds. The goal of designing these stall event patterns is to allow the subjects to examine each streaming session and express their opinion over streaming sessions with stall events and quality switches.

3) SHORT STALL EVENTS VS. A LONGER STALL EVENT EVALUATION

The idea here is to understand if the end-users prefer to see multiple short stall events or a longer stall event in streaming sessions. To address the stated question, we have designed six stall events' patterns shown in Figure 2. To establish a baseline for all the evaluations related to the SvL question, we have also introduced a *(0-0)* stall events' pattern. *(0-0)* pattern will be interpreted as zero stall events with a duration of zero seconds for each, whereas in the second stall events' pattern, we have a *(1-4)* stall events' pattern, which means one stall event with a duration of four seconds would occur in the streaming session. In contrast, we also have a *(4-1)* stall events' pattern. To expand the statistical population, we have also introduced stall events with longer durations. *(1-8)*, which stands for one stall event with a duration of eight seconds, *(4-2)* represents four stall events with durations of two seconds, and *(8-1)* is explained as eight stall events with a duration of one second for each. By studying the results from subjective evaluations, we would be able to conclude which one, *e.g.*, a *(1-4)* stall events' pattern that represents one longer stall event or a *(4-1)* stall events' pattern that represents multiple short stall events is preferred by subjects.

[1]https://ffmpeg.org, accessed Apr. 12, 2021.
[2]https://durian.blender.org, accessed Apr. 12, 2021.
[3]http://www.valkaama.com, accessed Apr. 12, 2021.
[4]https://peach.blender.org, accessed Apr. 12, 2021.
[5]https://mango.blender.org, accessed Apr. 12, 2021.

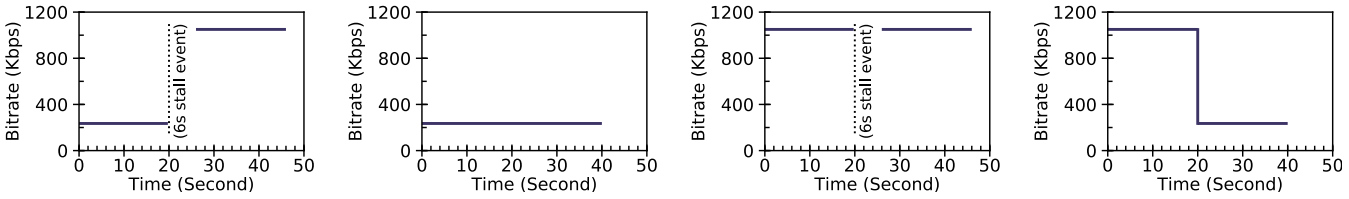


FIGURE 1. Stall events' patterns for SvQ evaluation. (a) A pattern with 6s stall event but upward quality switch, (b) A pattern without a stall event and continuous low quality streaming, (c) A pattern with high video quality streaming but with a 6s stall event, (d) A pattern with a downward quality switch without stall event.

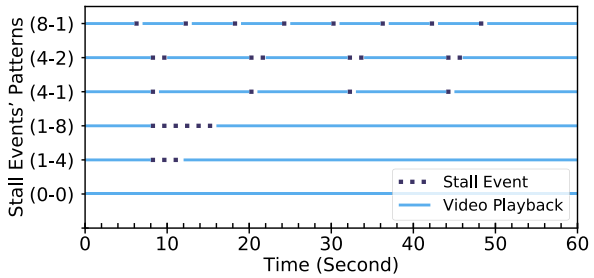


FIGURE 2. Stall events' patterns for SvL evaluation. In stall event's pattern *(a-b)*, *a* represents the number of stalls and *b* represents the duration of each stall.

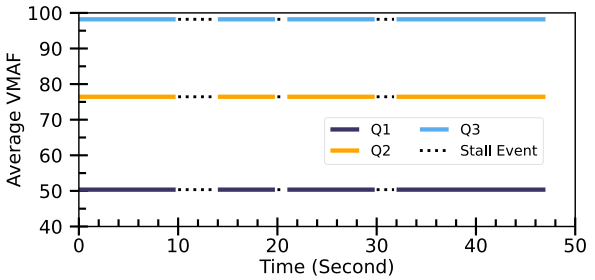


FIGURE 3. Stall events' patterns for RSVQ evaluation. For all video quality levels, *i.e.*, Q1, Q1, and Q3, three stall events with durations of 4s, 1s, and 2s are added.

4) RELATION OF STALL EVENT IMPACT ON THE QoE WITH VIDEO QUALITY LEVEL EVALUATION

To better understand the impact of stall events on the QoE at different video quality levels, we have conducted evaluations with the following streaming session properties. We have calculated the perceptual video quality of the test sequences with forty seconds length using the VMAF[6] objective metric [39]. VMAF uses a machine-learning algorithm (Support Vector Machine (SVM) regressor) to predict the quality of the video. We have used three representations from our bitrate ladder shown in Table 1 with 1, 2, and 5 indices. The selected representations cover a reasonable range of qualities as follows. The average VMAF score for test sequences encoded and packaged into the representation index one is 50.35, 76.44 for the second representation, and 98.20 for the fifth representation. We have used four test sequences with the mentioned video qualities and produced 12 experiments without stall events. We have added three stall events with one, two, and four seconds of durations in the identical test sequences and created another 12 experiments to assess subjects' opinions on streaming sessions with and without stall events concerning video quality. The total length of each experiment with stall events is 47 seconds. Figure 3 depicts this stall events' pattern.

C. SUBJECTIVE EVALUATION SETUP

We have used *Serverless Architecture*[7] and *AWS Lambda* [40] to develop a HAS subjective evaluation portal. The implementation follows the best practices proposed in [41] and the defined standards in ITU-T P.910 [42]. By leveraging the *HTML5 media element* [43] features, we have developed a media player to be used for playing the prepared experimental test sequences. We will describe the architecture, procedures, and measurements we designed to conduct subjective evaluations in this subsection.

To conduct extensive subjective evaluation and have many participants, we have used *Amazon Mechanical Turk*[8] (MTurk). MTurk is a crowdsourcing website to hire remotely located *crowd-workers* to perform discrete on-demand tasks. We have created multiple campaigns on the MTurk website, namely, MNSD, SvQ, SvL, and RSVQ campaigns which will be described in Subsections III-C1, III-C2, III-C3, and III-C4 respectively, after introducing the baselines and common design among all campaigns. When participants click a link that we have shared through the campaigns, an AWS lambda function will return the required libraries, HTML, and JavaScript files. After reading the instruction, the user enters an identity number (MTurk worker id), creating a database record. Next, the server will prepare a manifest file specific to the user. A list of test sequences selected by a prioritization algorithm will be populated into the manifest file and marked as locked in the database for the evaluation period. The prioritization algorithm works based on total votes and the number of not expired requested locks for that specific test sequence. The custom media player parses the manifest file and starts downloading the test sequences from the *AWS S3* [44] bucket. When the first test sequence is fully downloaded to the client browser, the playback starts. To store the test sequences within the subject's device (*i.e.*, web browser), we use *IndexedDB*.[9] While the first test sequence

[6]https://github.com/Netflix/vmaf, accessed Apr. 14, 2021.
[7]https://www.serverless.com, accessed Apr. 15, 2021.
[8]https://www.mturk.com, accessed Apr. 16, 2021.
[9]https://www.w3.org/TR/IndexedDB-2, accessed Apr. 16, 2021.

is being played, the other test sequences will still be downloaded in the background without interrupting the current playback. Via another AWS Lambda function, the subjects' votes and opinions will be captured and stored in the database. Once a subject casts their vote for a test sequence, that experiment id will be removed from the locked array and stored in the votes array mapped to the voted score alongside the answer to the reliability question, if applicable. At the end of the evaluation session, we generate a completion code. Only those crowd-workers who can provide the completion code will be compensated. Each campaign has its time constraints, and if the participant takes more than that time to cast their votes, they cannot be compensated, and also, their provided score will not be counted in the final results.

In SvQ, SvL, and RSVQ campaigns, each time the subject fully observes a test sequence, a popup window will be shown and allow the participant to rate their experience on a scale of 1 to 5, where 1 represents the worst experience and 5 stands for an excellent experience. There will also be a reliability question asked for each test sequence. Once all test sequences have been evaluated, the votes and answers to the reliability question will be stored in the database.

1) MINIMUM NOTICEABLE STALL EVENT DURATION CAMPAIGN

Out of 800 test sequences produced in the objective evaluation phase, a set of 16 randomized and prioritized test sequences will be selected and locked for 30 minutes in the database and then populated in the manifest file to be consumed by the media player. Participants will go through a trial test sequence after reading the instructions. The votes for the trial test sequence do not count for the final results. The voting procedure here is that subjects can click a ''Capture Stall Event'' button each time a stall event occurs. When each test sequence is finished, a reliability question related to the media content will be asked. Once all test sequences have been evaluated, the votes and answers to the reliability question will be stored in the database.

2) STALL EVENT VS. QUALITY LEVEL SWITCH CAMPAIGN

In the SvQ campaign, we have produced 16 test sequences. Each participant has to vote for all the test sequences; therefore, no locking or prioritization algorithm is implemented. The participants have 20 minutes to read the provided instructions, evaluate all the test sequences, and cast their votes.

3) SHORT STALL EVENTS VS. A LONGER STALL EVENT CAMPAIGN

We have produced 24 test sequences for the SvL campaign, and each participant will be asked to cast their votes for eight test sequences. The test sequences will be selected randomly and prioritized by the algorithm, which is explained in Subsection III-C. Each evaluation session for this campaign is designed to last 20 minutes, and the locking duration is 12 minutes for the selected test sequences.

4) RELATION OF STALL EVENT IMPACT ON THE QoE WITH VIDEO QUALITY LEVEL CAMPAIGN

In the RSVQ campaign, similar to the SvL campaign, we have 24 test sequences produced, and each participant is asked to cast their votes for eight test sequences. Lock duration, prioritization algorithm, and session duration are also similar to the SvL campaign explained in III-C3.

5) STATISTICAL ANALYSIS

Finally, we have performed ANOVA, Post Hoc, and Homogeneity analysis to present richer results and conduct in-depth studies over the obtained results from the subjective evaluations. We used an alpha level of 0.05 for all statistical tests. We have assessed the variances in the results from the SvQ and RSVQ evaluations by performing One-Way ANOVA and presented the outcomes in Table 4. For SvL evaluation, we have executed Post Hoc (Games-Howell) and Homogeneity (Tukey B) analysis to exhibit in-depth studies over the findings, and the results are presented in Table 2 and Table 3. The discussion over the findings from statistical analysis for each evaluation is given in subsections of Section IV-A respectively.

D. OBJECTIVE EVALUATION SETUP

This subsection introduces our testbed, evaluation procedures, QoE models, and analysis methods used to execute the objective evaluations and study the results.

1) TESTBED AND STALL EVENTS GENERATING

To conduct the objective evaluations, we have used an open-source software called CAdViSE[10] [45] as our testbed. This testbed provides a cloud-based platform to evaluate HAS sessions under various network conditions. By storing the client requests and media player events as session logs, we can reproduce the same streaming session, including occurred stall events. To evaluate designed stall events' patterns described in Subsection III-B, we have created virtual streaming session logs and then used an open-source application[11] that utilizes the session logs and stitches the segments from packaged assets together.

We have used FFmpeg to concatenate the video and audio segments. Using the same software, we can cut the exact required duration of a fabricated stall video and inject it between the actual video segments. To mimic stall events in a real-life setting, we also use the last frame of the previous segment as a background for the fabricated stall events. When the audiovisual files are combined, we use ITU-T P.1203 *Standalone Implementation*[12] [46], [47] (P.1203 for short) to extract a JSON file as a feed to the QoE model. The JSON file will be passed to the model to retrieve the MOS. The current implementation allows us to consider all the stall events with a duration of equal to or more than one

[10] https://github.com/cd-athena/CAdViSE, accessed Apr. 15, 2021.
[11] https://github.com/cd-athena/HASClipStitcher, accessed Apr. 15, 2021.
[12] https://github.com/itu-p1203/itu-p1203, accessed Apr. 15, 2021.

TABLE 2. Analysis of Variance (Post Hoc, Games-Howell) for the obtained results from SvL subjective evaluation.

I [1]	J [1]	Mean Difference (I-J)	Std. Error	P [2]
(4-2)	(8-1)	0.125	0.180	0.982
(4-1)	(4-2)	0.087	0.161	0.994
	(8-1)	0.212	0.180	0.844
(1-8)	(4-1)	0.388	0.151	0.110
	(4-2)	0.475	0.151	0.024
	(8-1)	0.600	0.170	0.007
(1-4)	(1-8)	0.287	0.142	0.334
	(4-1)	0.675	0.153	0.000
	(4-2)	0.762	0.153	0.000
	(8-1)	0.887	0.172	0.000
(0-0)	(1-4)	0.425	0.135	0.024
	(1-8)	0.712	0.133	0.000
	(4-1)	1.100	0.144	0.000
	(4-2)	1.187	0.145	0.000
	(8-1)	0.312	0.165	0.000

[1] I and J, Stall Events' Pattern
[2] P, Significant at the $P < 0.05$ level.

TABLE 3. Analysis of Variance (Homogeneous Subsets, Tukey B[a]), means for groups in homogeneous subsets for the results obtained from SvL subjective evaluation.

Stall Events' Pattern	N	1 [1]	2 [1]	3 [1]
(8-1)	80	3.23		
(4-2)	80	3.35		
(4-1)	80	3.44		
(1-8)	80		3.83	
(1-4)	80		4.11	
(0-0)	80			4.54

[a] Uses harmonic mean sample size = 80.
[1] 1, 2, and 3, are subsets for alpha = 0.05.

TABLE 4. Analysis of variance (One-Way ANOVA, between groups) for the obtained results from SvQ and RSVQ subjective evaluations.

Evaluation	df [1]	Mean Square	F [2]	P [3]
SvQ, Set A Case I & II	1	6.475	5.944	0.015
SvQ, Set B Case I & II	1	11.752	12.240	0.000
RSVQ, Q1 & Q1 + Stall	1	8.411	7.188	0.008
RSVQ, Q2 & Q2 + Stall	1	56.860	59.807	0.000
RSVQ, Q3 & Q3 + Stall	1	55.153	62.770	0.000

[1] df, The total number of values minus 1.
[2] F statistic, the division of two mean squares.
[3] P, Significant at the $P < 0.05$ level.

millisecond. Another step in this application is to concatenate the audiovisual files (audio and video files) using FFmpeg to produce a final mp4 file ready for the subjective evaluations phase, as explained in Subsection III-C.

2) QoE MODELS

To predict the quality of the streaming sessions, we have used three state-of-the-art objective QoE models, ITU-T P.1203 [14], BiQPS [48], and FINEAS [8]. The first two models belong to the bitstream classification, whereas FINEAS is one of the parametric models, which only utilize the parameters extracted from packet headers (*e.g.*, bitrate, framerate) and objective metrics such as rebuffering duration and rebuffering frequency to estimate the QoE [49].

ITU-T P.1203 is selected as it has been utilized widely to evaluate audiovisual of HAS in the context of live streaming and video-on-demand [50]–[52]. The P.1203 model proposed within ITU-T Study Group 12 integrates predictions based on a large set of training and validation databases. There are three modules in ITU-T P.1203: *(i) short-term video-quality module, (ii) short-term audio-quality model, (iii) quality integration module.* The first module comprises four modes: 0, 1, 2, and 3, in which mode 0 takes the least metadata as the input, whereas mode 3 needs full access to the bitstream. To tradeoff between the complexity of predictions and the accuracy, we use mode 1, which includes audio/video codec, video resolution, framerate, audio/video bitrate, frame type, and frame size.

A recently introduced open software BiQPS [48] is a QoE model predicting the QoE of video streaming sessions based on a Long-Short Term Memory (LSTM) network. This model considers five parameters, including stall event duration, quantization parameter, bitrate, video resolution, and framerate as inputs. These inputs are fed to an LSTM network with bidirectional and attention mechanisms, and the output is the predicted QoE score with a range of 1 to 5.

The FINEAS QoE model computes the QoE score based on a mathematical formula that considers both quality variations and stall events as presented in Equation 1.

$$QoE = 5.67 \times \frac{\bar{q}}{q_{max}} - 6.72 \times \frac{\hat{q}}{q_{max}} + 0.17 - 4.95 \times F, \quad (1)$$

where $\bar{q}$ and $\hat{q}$ are the average quality levels and their standard deviations, respectively. F models the impact of stall frequency ϕ and average stall duration ψ as shown in Equation 2.

$$F = \frac{7}{8} \times \max\left(\frac{\ln(\phi)}{6} + 1, 0\right) + \frac{1}{8} \times \left(\frac{\min(\psi, 15)}{15}\right). \quad (2)$$

3) STATISTICAL ANALYSIS

We evaluate the selected objective QoE evaluation models against subjective results using three statistical metrics, namely, *(i)* Pearson Correlation Coefficient (PCC), *(ii)* Spearman's Rank-order Correlation Coefficient (SRCC), and *(iii)* Root-Mean-Squared Error (RMSE). One should note that a higher PCC and SRCC and a smaller RMSE are declaring a better performance of the examined QoE model.

IV. RESULTS, ANALYSIS AND FINDINGS

This section first provides the subjective evaluation results for the defined scenarios following the objective evaluation results. We then present the correlations between subjective and objective scores to understand the results better, and finally, we summarize our findings.

A. SUBJECTIVE EVALUATION RESULTS

We had 713 participants in our subjective evaluations through held campaigns on the MTurk platform. Four hundred fifty-two participants managed to complete the assigned evaluations and cast their votes on time. There were 176, 108,

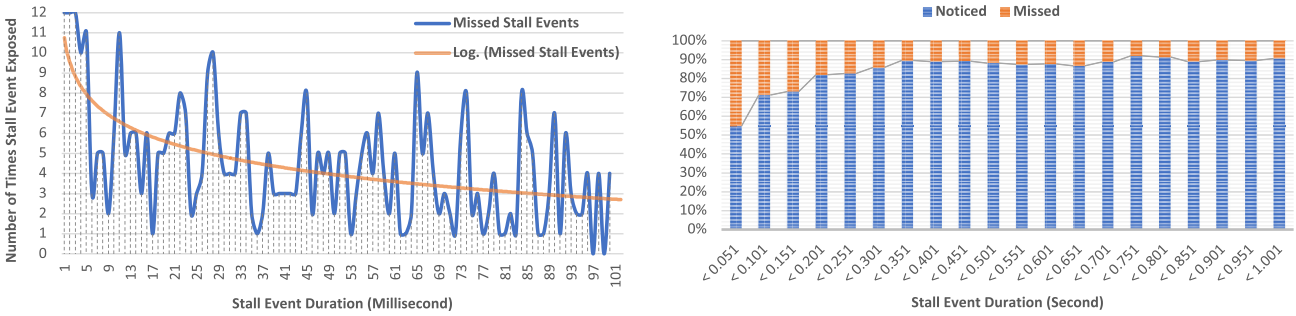
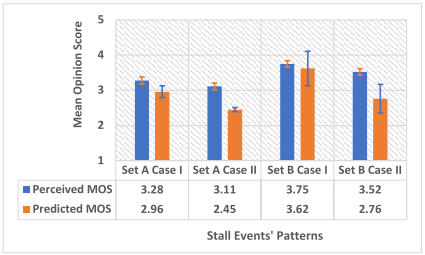
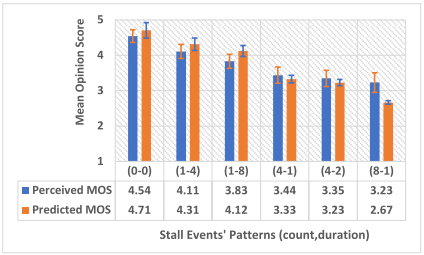


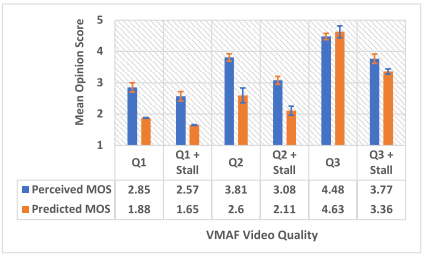
FIGURE 4. Subjective minimum noticeable stall event duration (MNSD) evaluation results. (left) The number of missed stall events: For stall events with less than 4ms duration, all subjects miss the stall event, but with the increasing stall duration this number is reduced. (right) The percentage of missed stall events vs. the percentage of noticed stall events for different stall event durations.



(a) Stall event vs. Quality level switch (SvQ) Evaluation.



(b) Short stall events vs. a Longer stall event (SvL) Evaluation.



(c) Relation of Stall event impact on the QoE with Video Quality level (RSVQ) Evaluation.

FIGURE 5. Perceived and predicted MOS for (a) SvQ, (b) SvL, and (c) RSVQ Evaluations.

61, and 107 participants in MNSD, SvQ, SvL, and RSVQ evaluation campaigns, respectively.

1) MINIMUM NOTICEABLE STALL EVENT DURATION EVALUATION

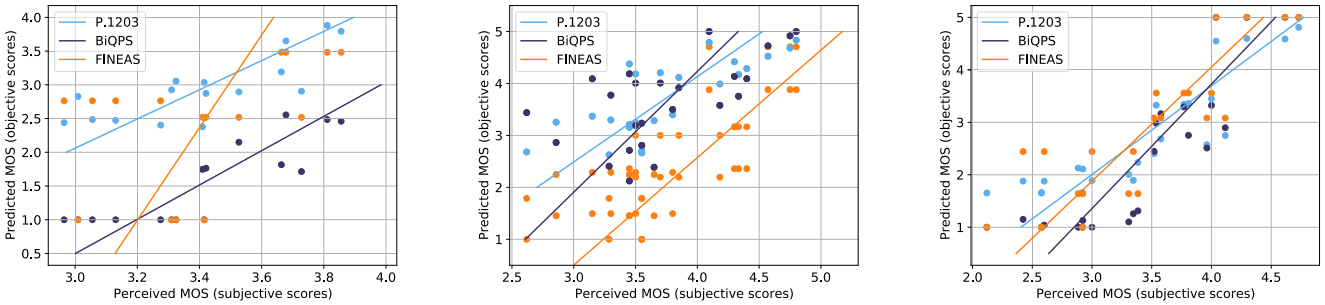
The results from subjective evaluation for the MNSD evaluation campaign are shown in Figure 4. As observed, the decrease of noticed stall events starts from stall events with a duration of less than 0.301 seconds. More than 45% of the subjects could not notice the stall events with a duration of less than 0.051 seconds. We have determined that any stall event with a duration of less than 0.004 seconds was not noticeable for the participants in the MNSD evaluation.

2) STALL EVENT VS. QUALITY LEVEL SWITCH EVALUATION

The subjective evaluation results for the two sets A and B are illustrated in Figure 5a. By performing statistical analysis over the results which are presented in Table 4, it is shown that there is a preference for Case I in Set A and also Case I in Set B compared to Case II in both sets, respectively. The difference in subjects opinions was significant, *i.e.*, $F(1, 890) = 5.944, p = 0.015$ for Set A and $F(1, 866) = 12.240, p > 0.001$ for Set B. This preference means that subjects tend to watch a higher quality version even if it is obtained by adding a stall event with a duration of six seconds. The results contrast with [22], where the subjects mostly preferred the media playback fluidity to a media playback with higher quality but with stall events.

3) SHORT STALL EVENTS VS. A LONGER STALL EVENT EVALUATION

The results for six stall events' patterns used in the SvL evaluations are given in Figure 5b, and statistical analysis results are provided in Table 2. The analysis results demonstrate a preference for a longer stall event over stall events with high frequency but with the same total duration as the longer stall event. For example, the stall events' pattern *(1-4)*, *i.e.*, one stall event with a duration of four seconds, obtained a 4.11 MOS, while the stall events' pattern *(4-1)*, *i.e.*, four stall evens with a duration of one second for each, obtained a 3.44 MOS. The difference in subjects opinions between *(1-4)* and *(4-1)* stall events' patterns was significant, *i.e.*, $F(5, 474) = 21.126, p > 0.001$. Similarly, the stall events' pattern *(1-8)*, *i.e.*, one stall event with a duration of eight seconds, obtained a 3.83 MOS, while the stall events' pattern *(8-1)*, *i.e.*, eight stall evens with a duration of one second, obtained a 3.23 MOS. The difference in subjects opinions between *(1-8)* and *(8-1)* stall events' patterns, was significant, *i.e.*, $F(5, 474) = 21.126, p = 0.007$. We have also studied the homogeneity of the stall events' patterns for SvL subjective evaluations, and the results are given in Table 3. As it can be seen, stall events' patterns *(8-1)*, *(4-2)*, and *(4-1) i.e.*, stall events with higher frequency and short durations, are considered the first homogeneous subset and stall events' patterns *(1-8)* and *(1-4) i.e.*, stall events with longer durations and less frequency, are considered the second homogeneous subset. The distribution of



(a) Correlations of the results obtained from the SvQ objective and subjective evaluations.

(b) Correlations of the results obtained from the SvL objective and subjective evaluations.

(c) Correlations of the results obtained from the RSVQ objective and subjective evaluations.

FIGURE 6. **Perceived and predicted MOS correlations for (a) SvQ, (b) SvL, and (c) RSVQ evaluations.**

MOS values in the two subsets *i.e.*, 3.23, 3.35, and 3.44 for the first subset and 3.83 and 4.11 for the second subset indicate an overall preference of subjects for longer stall events with less frequency. Our results confirm the conclusions of [19], [23] for stall event duration and frequency, *i.e.*, the longer the duration of the stall event or the higher the frequency of stall events, the lower the MOS would be. However, our results contrast with [24] where shorter stall events showed a higher negative impact on the QoE.

4) RELATION OF STALL EVENT IMPACT ON THE QoE WITH VIDEO QUALITY LEVEL EVALUATION

In the RSVQ evaluation, the impact of stall event occurrence on the QoE is evaluated at different video quality levels. The average subjective test scores are illustrated in Figure 5c, and the statistical analysis of the results is presented in Table 4. It can be seen that stall events have a minor penalty on the QoE when the quality of videos is low (Q1). However, for the middle and high-quality videos, the stall event occurrence has a higher penalty on the perceived QoE than the same stall event at a low-quality video. The statistical analysis shows that the difference in subjects opinions was significant between all video qualities, *i.e.*, Q1, Q2, and Q3 with and without stall events, $F(1, 426) = 7.188, p = 0.008$ for Q1, $F(1, 426) = 59.807, p > 0.001$ for Q2, and $F(1, 428) = 62.770, p > 0.001$ for Q3. Our results are consistent with [18], [29], where subjects tend to give a higher penalty to the stall events that occur at the higher quality video, but in contrast to [19] where authors found that stall events on high qualities result in minor impairment than ones that occur when the video is of a low quality.

B. OBJECTIVE EVALUATION RESULTS

We calculate PCC, SRCC, and RMSE for three stall events' patterns, *i.e.*, SvQ, SvL, and RSVQ. We illustrate the scatter plots of the predicted quality for all objective models with the first-order regression line for the SvQ stall events' pattern in Figure 6a. We also show the scatter plots for SvL and RSVQ stall event patterns in Figure 6b and Figure 6c, respectively. We finally collected all objective and subjective scores from these three subjective tests and showed the scatter plot for the overall scores in Figure 7. Table 5 summarizes the statistical metrics for all of these scatter plots.

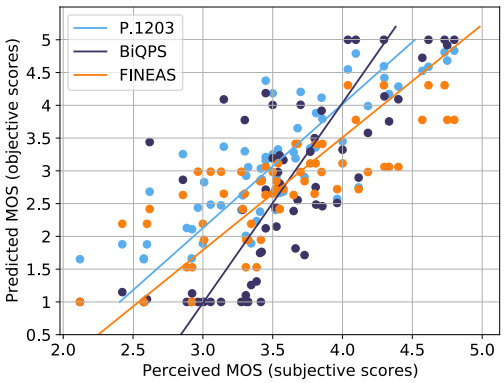


FIGURE 7. **Correlations of perceived and predicted MOS for SvQ, SvL, and RSVQ evaluations.**

In the SvQ test, where both stall events and quality switches occur in some test sequences, it can be seen that the BiQPS model achieves the best performance with the highest PCC, SRCC, and smallest RMSE. In particular, its PCC, SRCC and RMSE are 0.851, 0.850, and 0.144, respectively. On the contrary, the FINEAS model provides the worst QoE prediction with PCC and SRCC less than 0.5 and 0.241 RMSE. It can be attributed to the severe punishment for the standard deviation's video quality and stall events. When the quality is increased significantly, as shown in Figure-1(a) so that its standard deviation is large, the predicted QoE score of Set A - Case I is decreased remarkably and smaller than that of Set A - Case II in Figure 1. Meanwhile, the subjective results show an opposite trend (see Figure 5a).

Regarding the SvL evaluation, the FINEAS model predicts MOS the most precisely. Its RMSE is 0.33, whereas this figure for other models is more than 0.35. Also, with a PCC of 0.807, FINEAS indicates a strong relationship with the perceived MOS at the end-user. The reason may be that the coefficients related to stall events in Eq (2) were tuned carefully in [9] and as this evaluation only considers stall events, Eq (2) supports the FINEAS model to provide good results.

TABLE 5. Correlations of the perceived MOS with predicted MOS in each evaluation and the overall scores.

Evaluation		P.1203	BiQPS	FINEAS
SvQ	PCC	0.787	**0.851**	0.479
	SRCC	0.744	**0.850**	0.400
	RMSE	0.170	**0.144**	0.241
SvL	PCC	0.774	0.624	**0.807**
	SRCC	0.731	0.579	**0.757**
	RMSE	0.354	0.437	**0.330**
RSVQ	PCC	0.878	**0.888**	0.867
	SRCC	**0.941**	0.908	0.881
	RMSE	0.331	**0.319**	0.345
Overall	PCC	**0.831**	0.762	0.781
	SRCC	**0.814**	0.757	0.774
	RMSE	**0.326**	0.379	0.366

Meanwhile, the QoE predicted by the BiQPS model and MOS are moderately correlated with PCC and SRCC lower than 0.7, and BiQPS has the lowest accuracy with RMSE of 0.437. This can be attributed to the dataset [53] used in BiQPS's training phase. The training data contains a limited number of patterns, including stalling events and no quality switches. Also, there are only up to 6 stalling events in the dataset.

Similar to SvQ evaluation, when predicting the QoE for RSVQ videos, the BiQPS model achieves the best result. Its RMSE is 0.319, which is lower than other models by 3.6% to 7.5%. An interesting finding here is that SRCC is higher than PCC in every considered model, which means the relationship between each model and MOS is monotonic but not linear.

From the above analysis, we can see that BiQPS and FINEAS do not provide consistently high performance among different evaluations. Their predictions are the best for some tests but may become the worst in other cases. Therefore, overall their PCCs and SRCC are less than 0.8, and SRCCs are more than 0.36. Meanwhile, the P.1203 model shows the best performance for all evaluations with the highest PCC and SRCC (more than 0.8) and the most minor RMSE 0.326.

V. CONCLUSION AND FUTURE WORK

In this paper, we have conducted extensive objective and subjective evaluations to assess the impact of stall events and quality switches on the perceived and predicted QoE in HTTP Adaptive Streaming. Various findings have been presented in this study. *(i)* Stall events with a duration of less than four milliseconds are not noticeable by the end-users; therefore, it is negligible in MOS prediction by QoE models. *(ii)* End-users would like a high-quality video despite a long stall event, rather than a smooth low or decreased-quality video. Thus, ABR algorithms may consider keeping the video in high quality even though the network condition temporally drops so that a stall event occurs. *(iii)* A longer stall event is preferred over multiple short stall events with the same total duration. *(iv)* The impact of stall events on the QoE is decreased when they occur at a low-quality level. *(v)* The P.1203 QoE evaluation model provides the best performance with a relatively high correlation and small error compared to other state-of-the-art objective QoE evaluation models. The FINEAS QoE model shows a good prediction when the video quality varies slightly. The BiQPS QoE model could benefit from training with a more extensive dataset to improve its performance.

Future work may utilize the evaluation setup and findings presented in this paper to conduct further objective and subjective evaluations taking into account *(a)* additional QoE models or/and *(b)* different content/context configurations (*e.g.*, test sequences, encoding parameters, stall events' patterns).

REFERENCES

[1] Cisco. *Global—2022 Forecast Highlights*. Accessed: Apr. 10, 2021. [Online]. Available: https://www.cisco.com/c/dam/m/e_us/solutions/service-provider/vni-forecast-highlights/pdf/Global_2022_Forecast_Highlights.pdf
[2] O. Oyman and S. Singh, "Quality of experience for HTTP adaptive streaming services," *IEEE Commun. Mag.*, vol. 50, no. 4, pp. 20–27, Apr. 2012.
[3] C. Müller, S. Lederer, and C. Timmerer, "An evaluation of dynamic adaptive streaming over HTTP in vehicular environments," in *Proc. 4th Workshop Mobile Video (MoVid)*, 2012, pp. 37–42.
[4] T. C. Thang, H. T. Le, A. T. Pham, and Y. M. Ro, "An evaluation of bitrate adaptation methods for HTTP live streaming," *IEEE J. Sel. Areas Commun.*, vol. 32, no. 4, pp. 693–705, Apr. 2014.
[5] S. Akhshabi, A. C. Begen, and C. Dovrolis, "An experimental evaluation of rate-adaptation algorithms in adaptive streaming over HTTP," in *Proc. 2nd Annu. ACM Conf. Multimedia Syst.*, Feb. 2011, pp. 157–168.
[6] N. Cranley, P. Perry, and L. Murphy, "User perception of adapting video quality," *Int. J. Hum.-Comput. Stud.*, vol. 64, no. 8, pp. 637–647, 2006.
[7] D. Bhat, R. Deshmukh, and M. Zink, "Improving QoE of ABR streaming sessions through QUIC retransmissions," in *Proc. 26th ACM Int. Conf. Multimedia*, Oct. 2018, pp. 1616–1624.
[8] S. Petrangeli, J. Famaey, M. Claeys, S. Latré, and F. De Turck, "QoE-driven rate adaptation heuristic for fair adaptive video streaming," *ACM Trans. Multimedia Comput., Commun., Appl.*, vol. 12, no. 2, pp. 1–24, Mar. 2016.
[9] M. Claeys, S. Latré, J. Famaey, T. Wu, W. Van Leekwijck, and F. De Turck, "Design and optimisation of a (FA) *Q*-learning-based HTTP adaptive streaming client," *Connection Sci.*, vol. 26, no. 1, pp. 25–43, Jan. 2014.
[10] W. Shi, Y. Sun, and J. Pan, "Continuous prediction for quality of experience in wireless video streaming," *IEEE Access*, vol. 7, pp. 70343–70354, 2019.
[11] K. Brunnström *et al.*, "Qualinet white paper on definitions of quality of experience," Tech. Rep. hal-00977812, 2013. [Online]. Available: https://hal.archives-ouvertes.fr/hal-00977812
[12] C. M. Lentisco, L. Bellido, J. C. Q. Cuellar, E. Pastor, and J. L. H. Arciniegas, "QoE-based analysis of DASH streaming parameters over mobile broadcast networks," *IEEE Access*, vol. 5, pp. 20684–20694, 2017.
[13] X. Yin, A. Jindal, V. Sekar, and B. Sinopoli, "A control-theoretic approach for dynamic adaptive video streaming over HTTP," in *Proc. ACM Conf. Special Interest Group Data Commun.*, Aug. 2015, pp. 325–338.
[14] *Parametric Bitstream-Based Quality Assessment of Progressive Download and Adaptive Audiovisual Streaming Services Over Reliable Transport—Video Quality Estimation Module*, document ITU-R P1203-01. Accessed: Jul. 10, 2021. [Online]. Available: http://handle.itu.int/11.1002/ps/P1203-01
[15] S. Chikkerur, V. Sundaram, M. Reisslein, and L. J. Karam, "Objective video quality assessment methods: A classification, review, and performance comparison," *IEEE Trans. Broadcast.*, vol. 57, no. 2, pp. 165–182, Jun. 2011.
[16] S. Tavakoli, S. Egger, M. Seufert, R. Schatz, K. Brunnström, and N. Garcia, "Perceptual quality of HTTP adaptive streaming strategies: Cross-experimental analysis of multi-laboratory and crowdsourced subjective studies," *IEEE J. Sel. Areas Commun.*, vol. 34, no. 8, pp. 2141–2153, Aug. 2016.
[17] H. T. T. Tran, N. P. Ngoc, A. T. Pham, and T. C. Thang, "A multi-factor QoE model for adaptive streaming over mobile networks," in *Proc. IEEE Globecom Workshops (GC Wkshps)*, Dec. 2016, pp. 1–6.

[18] K. Zeng, H. Yeganeh, and Z. Wang, "Quality-of-experience of streaming video: Interactions between presentation quality and playback stalling," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Sep. 2016, pp. 2405–2409.

[19] K. Yamagishi and T. Hayashi, "Parametric quality-estimation model for adaptive-bitrate-streaming services," *IEEE Trans. Multimedia*, vol. 19, no. 7, pp. 1545–1557, Jul. 2017.

[20] T. Hossfeld, S. Egger, R. Schatz, M. Fiedler, K. Masuch, and C. Lorentzen, "Initial delay vs. interruptions: Between the devil and the deep blue sea," in *Proc. 4th Int. Workshop Quality Multimedia Exp.*, Jul. 2012, pp. 1–6.

[21] J. Allard, A. Roskuski, and M. Claypool, "Measuring and modeling the impact of buffering and interrupts on streaming video quality of experience," in *Proc. 18th Int. Conf. Adv. Mobile Comput. Multimedia*. New York, NY, USA: Association for Computing Machinery, Nov. 2020, pp. 153–160, doi: 10.1145/3428690.3429173.

[22] T. De Pessemier, K. De Moor, W. Joseph, L. De Marez, and L. Martens, "Quantifying the influence of rebuffering interruptions on the user's quality of experience during mobile video watching," *IEEE Trans. Broadcast.*, vol. 59, no. 1, pp. 47–61, Mar. 2013.

[23] D. Ghadiyaram, A. C. Bovik, H. Yeganeh, R. Kordasiewicz, and M. Gallant, "Study of the effects of stalling events on the quality of experience of mobile streaming videos," in *Proc. IEEE Global Conf. Signal Inf. Process. (GlobalSIP)*, Dec. 2014, pp. 989–993.

[24] R. Wang, Y. Geng, Y. Ding, Y. Yang, and W. Li, "Assessing the quality of experience of HTTP video streaming considering the effects of pause position," in *Proc. 16th Asia–Pacific Netw. Oper. Manage. Symp.*, Sep. 2014, pp. 1–4.

[25] S. Egger, B. Gardlo, M. Seufert, and R. Schatz, "The impact of adaptation strategies on perceived quality of HTTP adaptive streaming," in *Proc. Workshop Design, Quality Deployment Adapt. Video Streaming*. New York, NY, USA: Association for Computing Machinery, 2014, pp. 31–36, doi: 10.1145/2676652.2676658.

[26] N. Staelens, P. Coppens, N. Van Kets, G. Van Wallendaef, W. Van den Broeck, J. De Cock, and F. De Turek, "On the impact of video stalling and video quality in the case of camera switching during adaptive streaming of sports content," in *Proc. 7th Int. Workshop Quality Multimedia Exp. (QoMEX)*, May 2015, pp. 1–6.

[27] S. Tavakoli, K. Brunnström, K. Wang, B. Andrén, M. Shahid, and N. Garcia, "Subjective quality assessment of an adaptive video streaming model," *Proc. SPIE*, vol. 9016, pp. 197–208, Feb. 2014, doi: 10.1117/12.2040131.

[28] M. N. Garcia, D. Dytko, and A. Raake, "Quality impact due to initial loading, stalling, and video bitrate in progressive download video services," in *Proc. 6th Int. Workshop Quality Multimedia Exp. (QoMEX)*, Sep. 2014, pp. 129–134.

[29] Z. Duanmu, A. Rehman, K. Zeng, and Z. Wang, "Quality-of-experience prediction for streaming video," in *Proc. IEEE Int. Conf. Multimedia Expo (ICME)*, Jul. 2016, pp. 1–6.

[30] C. G. Bampis, Z. Li, A. K. Moorthy, I. Katsavounidis, A. Aaron, and A. C. Bovik, "Study of temporal effects on subjective video quality of experience," *IEEE Trans. Image Process.*, vol. 26, no. 11, pp. 5217–5231, Nov. 2017.

[31] D. Z. Rodríguez, Z. Wang, R. L. Rosa, and G. Bressan, "The impact of video-quality-level switching on user quality of experience in dynamic adaptive streaming over HTTP," *EURASIP J. Wireless Commun. Netw.*, vol. 2014, no. 1, pp. 1–15, Dec. 2014.

[32] S. Lederer, C. Müller, and C. Timmerer, "Dynamic adaptive streaming over HTTP dataset," in *Proc. 3rd Multimedia Syst. Conf.*, 2012, pp. 89–94.

[33] Netflix. (2015). *Per-Title Encode Optimization*. [Online]. Available: http://techblog.netflix.com/2015/12/per-title-encode-optimization.html

[34] Apple. (2016). *Best Practices for Creating and Deploying HTTP Live Streaming Media for Apple Devices*. [Online]. Available: https://developer.apple.com/library/content/technotes/tn2224/_index.html

[35] B. Taraghi, A. Bentaleb, C. Timmerer, R. Zimmermann, and H. Hellwagner, "Understanding quality of experience of heuristic-based HTTP adaptive bitrate algorithms," in *Proc. Workshop Netw. Oper. Syst. Support Digit. Audio Video (NOSSDAV)*, Jul. 2021, pp. 82–89, doi: 10.1145/3458306.3458875.

[36] Z. Duanmu, A. Rehman, and Z. Wang, "A quality-of-experience database for adaptive video streaming," *IEEE Trans. Broadcast.*, vol. 64, no. 2, pp. 474–487, Jun. 2018.

[37] E. Beavers. (2014). *How to Encode Multi-Bitrate Videos in MPEG-Dash for MSE Based Media Players*. [Online]. Available: https://blog.streamroot.io/encode-multi-bitrate-videos-mpeg-dash-mse-based-media-players/

[38] A. Zambelli. (2009). *Smooth Streaming Technical Overview*. [Online]. Available: https://docs.microsoft.com/en-us/iis/media/on-demand-smooth-streaming/smooth-streaming-technical-overview

[39] Z. Li, A. Aaron, I. Katsavounidis, A. Moorthy, and M. Manohara, "Toward a practical perceptual video quality metric," Netflix TechBlog, Los Gatos, CA, USA, Tech. Rep., 2016, p. 2, vol. 6. [Online]. Available: http://techblog.netflix.com/2016/06/toward-practical-perceptual-video.html

[40] Amazon Web Services. *AWS Lambda Documentation*. Accessed: Jul. 13, 2021. [Online]. Available: https://docs.aws.amazon.com/lambda/

[41] T. Hossfeld, C. Keimel, M. Hirth, B. Gardlo, J. Habigt, K. Diepold, and P. Tran-Gia, "Best practices for QoE crowdtesting: QoE assessment with crowdsourcing," *IEEE Trans. Multimedia*, vol. 16, no. 2, pp. 541–558, Feb. 2014.

[42] *Subjective Video Quality Assessment Methods for Multimedia Applications*, document ITU-T P.910, International Telecommunication Union, 1999.

[43] W3C. *HTML5*. Accessed: Jul. 13, 2021. [Online]. Available: https://www.w3.org/TR/2011/WD-html5-20110113/video.html

[44] Amazon Web Services. *Amazon Simple Storage Service Documentation*. Accessed: Jul. 13, 2021. [Online]. Available: https://docs.aws.amazon.com/s3/

[45] B. Taraghi, A. Zabrovskiy, C. Timmerer, and H. Hellwagner, "CAdViSE: Cloud-based adaptive video streaming evaluation framework for the automated testing of media players," in *Proc. 11th ACM Multimedia Syst. Conf.*, May 2020, pp. 349–352, doi: 10.1145/3339825.3393581.

[46] A. Raake, M.-N. Garcia, W. Robitza, P. List, S. Göring, and B. Feiten, "A bitstream-based, scalable video-quality model for HTTP adaptive streaming: ITU-T P.1203.1," in *Proc. 9th Int. Conf. Quality Multimedia Exp. (QoMEX)*, May 2017, pp. 1–6. [Online]. Available: http://ieeexplore.ieee.org/document/7965631/

[47] W. Robitza, S. Göring, A. Raake, D. Lindegren, G. Heikkilä, J. Gustafsson, P. List, B. Feiten, U. Wüstenhagen, M.-N. Garcia, K. Yamagishi, and S. Broom, "HTTP adaptive streaming QoE estimation with ITU-T rec. P.1203: Open databases and software," in *Proc. 9th ACM Multimedia Syst. Conf.*, Jun. 2018, pp. 466–471.

[48] H. T. T. Tran, D. Nguyen, and T. C. Thang, "An open software for bitstream-based quality prediction in adaptive video streaming," in *Proc. 11th ACM Multimedia Syst. Conf.*, May 2020, pp. 225–230.

[49] N. Barman and M. G. Martini, "QoE modeling for HTTP adaptive video streaming—A survey and open challenges," *IEEE Access*, vol. 7, pp. 30831–30859, 2019.

[50] M. Nguyen, C. Timmerer, and H. Hellwagner, "H2BR: An HTTP/2-based retransmission technique to improve the QoE of adaptive video streaming," in *Proc. 25th ACM Workshop Packet Video*. New York, NY, USA: Association for Computing Machinery, Jun. 2020, pp. 1–7, doi: 10.1145/3386292.3397117.

[51] H.-F. Bermudez, J.-M. Martinez-Caro, R. Sanchez-Iborra, J. L. Arciniegas, and M.-D. Cano, "Live video-streaming evaluation using the ITU-T P.1203 QoE model in LTE networks," *Comput. Netw.*, vol. 165, Dec. 2019, Art. no. 106967.

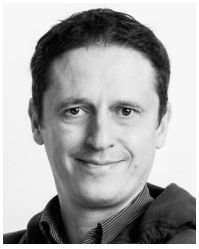
[52] A. Bentaleb, C. Timmerer, A. C. Begen, and R. Zimmermann, "Bandwidth prediction in low-latency chunked streaming," in *Proc. 29th ACM Workshop Netw. Oper. Syst. Support Digit. Audio Video*, 2019, pp. 7–13.

[53] H. T. T. Tran, D. V. Nguyen, D. D. Nguyen, N. P. Ngoc, and T. C. Thang, "An LSTM-based approach for overall quality prediction in HTTP adaptive streaming," in *Proc. IEEE Conf. Comput. Commun. Workshops (INFOCOM WKSHPS)*, Apr. 2019, pp. 702–707.

BABAK TARAGHI (Member, IEEE) received the bachelor's degree in information technology, in 2015 and the master's degree in software engineering from University Technology Malaysia, in 2017. He is currently pursuing the Ph.D. degree in ATHENA project with the Institute of Information Technology (ITEC), Alpen-Adria-Universität Klagenfurt (AAU), with a focus on adaptive video streaming. He possesses a strong background in software development engineering with more than 15 years of professional experience in software solution design and software construction. His research interests include multimedia communication, streaming, adaptation, and quality of experience. Further information is available at https://tiny.one/tbabak.

MINH NGUYEN (Member, IEEE) received the Engineering degree in electronics and telecommunications from Hanoi University of Science and Technology, Vietnam, in 2018. He is currently pursuing the Ph.D. degree in ATHENA project with the Institute of Information Technology (ITEC), Alpen-Adria-Universität Klagenfurt (AAU). His research interests include adaptive video streaming, multimedia networking, computer vision, and QoS/QoE evaluation.

HADI AMIRPOUR (Member, IEEE) received the dual B.Sc. degree in electrical and biomedical engineering and the M.Sc. degree in electrical engineering. He is currently pursuing the Ph.D. degree with the Institute of Information Technology (ITEC), Alpen-Adria-Universität Klagenfurt (AAU). He was involved in the project EmergIMG, a Portuguese consortium on emerging imaging technologies, funded by the Portuguese funding agency and H2020. He is currently working at ATHENA, and his research interests include video streaming, image and video compression, quality of experience, emerging 3D imaging technology, and medical image analysis. Further information at https://hadiamirpour.github.io.



CHRISTIAN TIMMERER (Senior Member, IEEE) received the M.Sc. (Dipl.-Ing.) and Ph.D. (Dr.techn.) degrees (for research on the adaptation of scalable multimedia content in streaming and constraint environments) from Alpen-Adria-Universität Klagenfurt (AAU), in January 2003 and in June 2006, respectively. He is currently an Associate Professor with the Institute of Information Technology (ITEC) and the Director of the Christian Doppler Laboratory ATHENA (https://athena.itec.aau.at/). His research interests include immersive multimedia communication, streaming, adaptation, quality of experience, and sensory experience. He was the General Chair of WIAMIS 2008, QoMEX 2013, MMSys 2016, and PV 2018. He has participated in several EC-funded projects, notably DANAE, ENTHRONE, P2P-Next, ALICANTE, SocialSensor, COST IC1003 QUALINET, and ICoSOLE. He also participated in ISO/MPEG work for several years, notably in the area of MPEG-21, MPEG-M, MPEG-V, and MPEG-DASH, where he also served as a Standard Editor. In 2013, he co-founded Bitmovin (http://www.bitmovin.com/) to provide professional services around MPEG-DASH, where he holds the position of the Chief Innovation Officer (CIO)—the Head of Research and Standardization. He is a member of ACM, specifically IEEE Computer Society, IEEE Communications Society, and ACM SIGMM. He was a Guest Editor for the three Special Issues on the IEEE JOURNAL ON SELECTED AREAS IN COMMUNICATIONS (JSAC) and also served as an Associate Editor for IEEE TRANSACTIONS ON MULTIMEDIA. Further information available at http://blog.timmerer.com.