A Quality-of-Experience Database for Adaptive Video Streaming

Zhengfang Duanmu, *Student Member, IEEE*, Abdul Rehman, and Zhou Wang, *Fellow, IEEE*

***Abstract*—The dynamic adaptive streaming over HTTP provides an inter-operable solution to overcome the volatile network conditions, but its complex characteristic brings new challenges to objective video quality-of-experience (QoE) measurement. To test the generalizability and to facilitate the wide usage of QoE measurement techniques in real-world applications, we establish a new database named Waterloo Streaming QoE Database III (SQoE-III). Unlike existing databases constructed with hand-crafted test sequences, the SQoE-III database, so far the largest and most realistic of its kind, consists of a total of 450 streaming videos created from diverse source content and diverse distortion patterns, with six adaptation algorithms of diverse characteristics under 13 representative network conditions. All streaming videos are assessed by 34 subjects, and a comprehensive evaluation is conducted on the performance of 15 objective QoE models from four categories with regards to their efficacy in predicting subjective QoE. Detailed correlation analysis and statistical hypothesis testing are carried out. The results of this paper shed light on the future development of adaptive bitrate streaming algorithm and video QoE monitoring system. The subjective database is available online at https://ece.uwaterloo.ca/∼zduanmu/tbc2018qoe/.**

***Index Terms*—Quality-of-experience, adaptive bitrate streaming, dynamic adaptive streaming over HTTP, subjective quality assessment.**

I. Introduction

IN THE past decade, there has been a tremendous growth in streaming media applications, thanks to the fast development of network services and the remarkable growth of smart mobile devices. Since the ratification of the Dynamic Adaptive Streaming over HTTP (DASH) standard in 2011 [1], video distribution service providers have invested significant effort in the transition from the conventional connection-oriented video transport protocols towards hypertext transfer protocol (HTTP) adaptive streaming protocols (HAS) due to its ability to traverse network address translations and firewall, reliability to deliver video packet, flexibility to react to volatile network conditions, and efficiency in reducing the server workload. DASH [2] achieves decoder-driven rate adaptation by providing video streams in a variety of bitrates and breaking them into small HTTP file segments. The media information of each segment is stored in a *manifest* file, which is created at server and transmitted to clients to provide the specification and location of each segment. Throughout the streaming process, the video player at the client adaptively switches among the available streams by selecting segments based on playback rate, buffer condition and instantaneous throughput [1]. Adaptive bitrate streaming (ABR) algorithms, that determine the bitrate of the next segment to download, are not defined within the standard but deliberately left open for optimization. The key is to define an optimization criterion that aims at maximizing viewer quality-of-experience (QoE).

Over the past decade, ABR has been a rapidly evolving research topic and has attracted an increasing amount of attention from both industry and academia [3]–[10]. However, thorough understand of realistic QoE impairment in common ABR scenarios is still lacking. Since the human visual system (HVS) is the ultimate receiver of streaming videos, subjective evaluation is the most straightforward and reliable approach to evaluate the QoE of streaming videos. The understanding of HVS would inspire development and validation of objective video QoE assessment methods. Furthermore, with many ABR algorithms at hand, it becomes pivotal to compare their performance, so as to find the best algorithm as well as directions for further improvement.

Even though subjective quality assessment studies provide reliable evaluations, they are inconvenient, time-consuming, and expensive. Many recent efforts have been made to develop objective video QoE models for ABR. However, most of them are designed and validated upon video databases that are limited in size, distortion patterns, or realistic settings, and are not publicly available. In addition, no QoE validation literature has reported comprehensive performance comparison of different objective QoE models. It is therefore important that objective QoE algorithms are tested on extensive subject-rated data. Furthermore, if such data, apart from being extensive in nature, is also publicly available, then other researchers can verify the results, and perform further development and comparative analysis.

In this paper, we aim to tackle the problems of subjective evaluation of objective QoE models and ABR algorithms. Our contributions are threefold. First, we build so far the largest database dedicated to subjective evaluation of HAS videos under realistic conditions. The database contains 20 source sequences of diverse content types and 450 streaming videos generated by 6 ABR algorithms under 13 wide-ranging and representative network conditions. Based on the video

Manuscript received September 26, 2017; revised January 2, 2018; accepted February 19, 2018. *(Corresponding author: Zhengfang Duanmu.)*

The authors are with the Department of Electrical and Computer Engineering, University of Waterloo, Waterloo, ON N2L 3G1, Canada (e-mail: zduanmu@uwaterloo.ca; abdul.rehman@uwaterloo.ca; zhou.wang@uwaterloo.ca).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TBC.2018.2822870

TABLE I
COMPARISON OF PUBLICLY AVAILABLE QOE DATABASES FOR HTTP-BASED ADAPTIVE VIDEO STREAMING

Database	Source Videos	Test Videos	Encoding Configurations	Test Case Formation	HAS-related Impairments	Resolution Adaptation
LIVEMVQA	10	200	H.264 at 4 levels	hand-crafted	switching or stalling	No
LIVEQHVS	3	15	**H.264 at 21 levels**	hand-crafted	switching	No
LIVEMSV	**24**	176	no compression	hand-crafted	stalling	No
Waterloo SQoE-I	20	180	H.264 at 3 levels	hand-crafted	initial buffering or stalling	No
Waterloo SQoE-II	12	**588**	H.264 at 7 levels	hand-crafted	switching	**Yes**
LIVE-Netflix Video QoE Database	14	112	H.264 at 6 levels	hand-crafted	**initial buffering & stalling & switching**	No
Waterloo SQoE-III	20	450	H.264 at 11 levels	**simulated**	**initial buffering & stalling & switching**	**Yes**

database, we carry out a subjective user study to evaluate and compare the QoE of the streaming videos. Second, we conduct a comprehensive evaluation on objective QoE models. 15 QoE algorithms from 4 categories including signal fidelity-based, network QoS-based, application QoS-based, and hybrid QoE models are assessed in terms of correlation with human perception. Statistical hypothesis tests are also performed to compare the QoE models in a statistically meaningful manner. Third, we evaluate 6 well-known ABR algorithms based on the subject-rated database. We find that no ABR algorithm produces the best QoE for all network conditions, and provide insights on the improvement of ABR algorithms. The results have significant implications on how video distributors can best use their resources to maximize user perceived QoE and how a practical real-time QoE monitoring system should be deployed.

II. RELATED WORK

Several well-known QoE databases have been widely used in the literature. The LIVE mobile video quality assessment database (LIVEMVQA) [11] consists of 10 reference and 200 distorted videos with 5 distortion types: H.264 compression, stalling, frame drop, rate adaptation, and wireless channel packet-loss. The single-stimulus continuous scale method [12] is adopted for testing, where both the instantaneous ratings as well as an overall rating at the end of each video is collected. It is the first publicly available subject-rated video database that contains practical distortions in the streaming process, though the distortion types are isolated and may not translate to combined degradations.

LIVE QoE database for HAS (LIVEQHVS) [13] contains three reference videos constructed by concatenating 8 high quality high definition video clips of different content. For each reference video, 5 bitrate-varying videos are constructed by adjusting the encoding bitrate of H.264 video encoder, resulting in a relatively small set of 15 quality-varying videos. Following a similar subjective experiment setup to LIVEMVQA, the authors collect both the instantaneous ratings and an overall rating at the end of each video. The importance of the hysteresis effect and nonlinear perception of the time-varying video quality is recognized.

Ghadiyaram *et al.* [14] performed a subjective study to understand the influence of dynamic network impairments such as stalling events on QoE of users watching videos on mobile devices. The constructed LIVEMSV database consists of 176 distorted videos generated from 24 reference videos with 26 hand-crafted stalling events. The authors adopted the single stimulus continuous quality evaluation procedure where the reference videos are also evaluated to obtain a difference mean opinion score (DMOS) for each distorted video sequence. The lack of video compression and quality switching reduces the relevance of the database to real-world HAS scenarios.

The Waterloo Streaming QoE Database I (SQoE-I) [15] focuses on the interaction between video presentation quality and playback stalling experiences. It contains 20 pristine high-quality 1920×1080 videos of diverse content. Each reference video is encoded into 3 bitrates with x264 encoder and then a 5-second stalling event is simulated at either the beginning or the middle point of the encoded sequences. In total, there are 200 video sequences including 20 source videos, 60 compressed videos, 60 initial buffering videos, and 60 mid-stalling videos. The most noteworthy finding of this study is that the video presentation quality of the freezing frame exhibits strong correlation with the dissatisfaction level of the stalling event with statistical analysis.

The Waterloo Streaming QoE Database II (SQoE-II) [16] involves 168 short and 588 long video clips with variations in compression level, spatial resolution, and frame-rate. The authors carry out an path-analytical experiments to address the confounding factors and better explore the space of quality adaptations. Albeit the interesting analysis, the database may not serve as a benchmark database due to the lack of stalling events.

The LIVE-Netflix Video QoE Database [17] is developed in parallel with our study in order to understand the influence of mixtures of dynamic network impairments such as rebuffing events and compression on QoE of users watching videos on mobile devices. The database consists of 112 distorted videos derived from 14 source content with 8 handcrafted playout patterns. In spite of the authors' effort in designing meaningful playout patterns, the test sequence are still hand-crafted, and thus may not reflect realistic scenarios. Only a small fraction of the source videos are made available to the public.

A summary of the aforementioned databases are given in Table I. Several other streaming video quality studies have been conducted in the past, mainly towards understanding the effects of network stream quality on QoE, validating the performance of ABR algorithms, and developing objective QoE models [4], [18]–[30]. Unfortunately, the results of these studies are not available to the public. Two excellent surveys on subjective QoE studies can be found in [31] and [32].

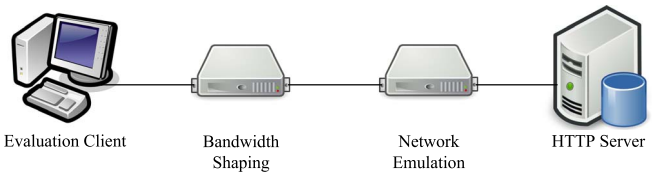


Fig. 1. Video streaming experimental setup.

All of the above studies suffer from the following problems: (1) the dataset is very limited in size; (2) hand-crafted stalling and quality switching patterns often do not reflect realistic scenarios (Specifically, Waterloo Streaming QoE-I database comprises videos with only one stalling event either at the beginning or the middle point; LIVEMSV database consists of videos with random stalling events; LIVEMVQA database contains videos with periodic stalling events; LIVE-Netflix Video QoE database includes 8 distortion profiles based on the authors' experience in streaming videos. Although such simplification makes the analysis of human QoE behavior easier, these hand-crafted distortion patterns can hardly represent distortions in the realistic adaptive streaming process that are dependent of the behavior of the ABR algorithms.); (3) the distortion types of video sequences are isolated; (4) spatial resolution adaptation commonly used in practice is not presented; and (5) the bitstream and network information, which are valuable to the development of ABR algorithms and objective QoE models, are not available. Realizing the need for an adequate and more relevant resource, we create a new database aiming for broader utility for modeling and analyzing contemporary HAS.

III. Adaptive Streaming Video Database and Subjective Quality Assessment

A. Video Database Construction

In order to generate meaningful and representative test videos, we conducted a set of DASH video streaming experiments, recorded the relevant streaming activities, and reconstructed the streaming session using video processing tools. We followed the recommendation in [33] and [34] to setup the testbed. The architecture of the testbed is depicted in Fig. 1 and consists of four modules: two computers (Ubuntu 14.04 LTS) with a 100Mbps direct network connection emulating a video client and server. DASH videos were pre-encoded and hosted on an Apache Web server. The main components of this architecture were the bandwidth shaping and the network emulation nodes which were both based on Ubuntu utilities. The bandwidth shaping node controlled the maximum achievable bandwidth for the client with the Linux traffic control system (tc) and the hierarchical token bucket (htb) which was a classful queuing discipline (qdisc). The available bandwidth for the client were adjusted every second according to bandwidth traces. The video client, where ABR algorithms were deployed, rendered videos at full screen while the video server was a simple HTTP server. After each video streaming session, a log file was generated on the client device, including selected bitrates, duration of initial buffering, start time, and end time of each stalling event. According to the recorded logs, we reconstructed each streaming session by concatenating streamed bitrate representations, appending blank frames to the test video to simulate initial buffering, and inserting identical frames at the buffering time instance to simulate stalling event. The loading indicator (for both initial buffering and stalling) was implemented as a spinning wheel. Detailed description of each module is given below.

TABLE II
Spatial Information (SI), Temporal Information (TI), Frame Rate (FPS), and Description of Reference Videos

Name	FPS	SI	TI	Description
BigBuckBunny	30	96	97	Animation, high motion
BirdOfPrey	30	44	68	Natural scene, smooth motion
Cheetah	25	64	37	Animal, camera motion
CostaRica	25	45	52	Natural scene, smooth motion
CSGO	60	70	52	Game, average motion
FCB	30	80	46	Sports, average motion
FrozenBanff	24	100	88	Natural scene, smooth motion
Mtv	25	112	114	Human, average motion
PuppiesBath	24	35	45	Animal, smooth motion
RoastDuck	30	60	84	Food, smooth motion
RushHour	30	52	20	Human, smooth motion
Ski	30	61	82	Sport, high motion
SlideEditing	25	160	86	Screen content, smooth motion
TallBuildings	30	81	13	Architecture, static
TearsOfSteel1	24	53	66	Movie, smooth motion
TearsOfSteel2	24	56	11	Movie, static
TrafficAndBuilding	30	66	15	Architecture, static
Transformer	24	72	56	Movie, average motion
Valentines	24	40	52	Human, smooth motion
ZapHighlight	25	97	89	Animation, high motion

Source Videos and Encoding Configuration: A video database of 20 pristine high-quality videos of size 1920×1080 were selected to cover diverse content, including humans, plants, natural scenes, architectures, screen content, and computer-synthesized sceneries. RushHour, TallBuildings, and TrafficAndBuilding were from the SJTU 4K video dataset [35]. All videos have the length of 10 seconds [36]. The detailed specifications of those videos are listed in Table II and a screenshot from each video is included in Fig. 2. Spatial information (SI) and temporal information (TI) [37] that roughly reflect the complexity of video content are also given in Table II. Apparently, the video sequences are of diverse spatio-temporal complexity and widely span the SI-TI space. Using aforementioned sequences as the source, each video was encoded with an x264 encoder into eleven representations using the encoding ladder shown in Table III to cover different quality levels. The choices of bitrate levels were based on the Netflix's recommendation [38] while representation eleven was appended to the original bitrate ladder to cover the high-quality representation suggested in the Apple's recommendation [39]. We followed Streamroot's encoding configuration recommendation [40] to remove scenecut and limit group-of-pictures (GoP) size. We segmented the test sequences with GPAC's MP4Box [41] with a segment length of 2 seconds for the following reasons. First, 2-second segments are widely used in the development of adaptation logics [29], [42]. In addition, it allows us to design test videos in an efficient way such that they cover a diverse adaptation patterns in a limited time.

Fig. 2. Snapshots of video sequences.

TABLE III
ENCODING LADDER OF VIDEO SEQUENCES

Representation index	Resolution	Bitrate (kbps)
1	320×240	235
2	384×288	375
3	512×384	560
4	512×384	750
5	640×480	1050
6	720×480	1750
7	1280×720	2350
8	1280×720	3000
9	1920×1080	4300
10	1920×1080	5800
11	1920×1080	7000

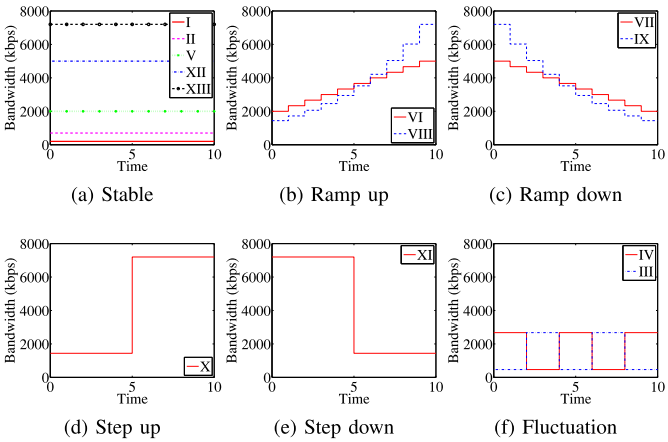


Fig. 3. Bandwidth profiles used in the experiment. The profiles are indexed from the lowest to the highest average bandwidth, where I and XIII represent the profiles with the lowest and highest average bandwidths, respectively.

Bandwidth shaping: The delay of network simulator was set to 80ms corresponding to what can be observed within long-distance fixed line connections or reasonable mobile networks, and thus is representative for a broad range of application scenarios as suggested in [33]. We used 13 network traces shown in Fig. 3 that are wide-ranging and representative including stationary as well as different scenarios indexed from the lowest to the highest average bandwidth. The average bandwidth of the network traces varies between 200Kbps and 7.2Mbps, covering all range of bitrates in the encoding bitrate ladder.

ABR algorithms: We acknowledge all the existing adaptation logics. In this study, we cover ABR algorithms of diverse characteristics, ranging from naïve de facto rate-based algorithm [2] to the state-of-the-art algorithms. We prototyped 6 ABR algorithms in an open source dynamic adaptive streaming player called dash.js [2] (version 2.2.0), which is the reference open-source implementation for the MPEG-DASH standard based on the HTML5 specification and is actively supported by leading industry participants. Appropriate modifications were made to each ABR algorithms as follows.

1. *Rate-based [2]:* The rate-based ABR algorithm, which is the default ABR controller in the DASH standard, picks the maximum available bitrate which is less than throughput prediction using the arithmetic mean of past 5 chunks. The original algorithm starts with a constant

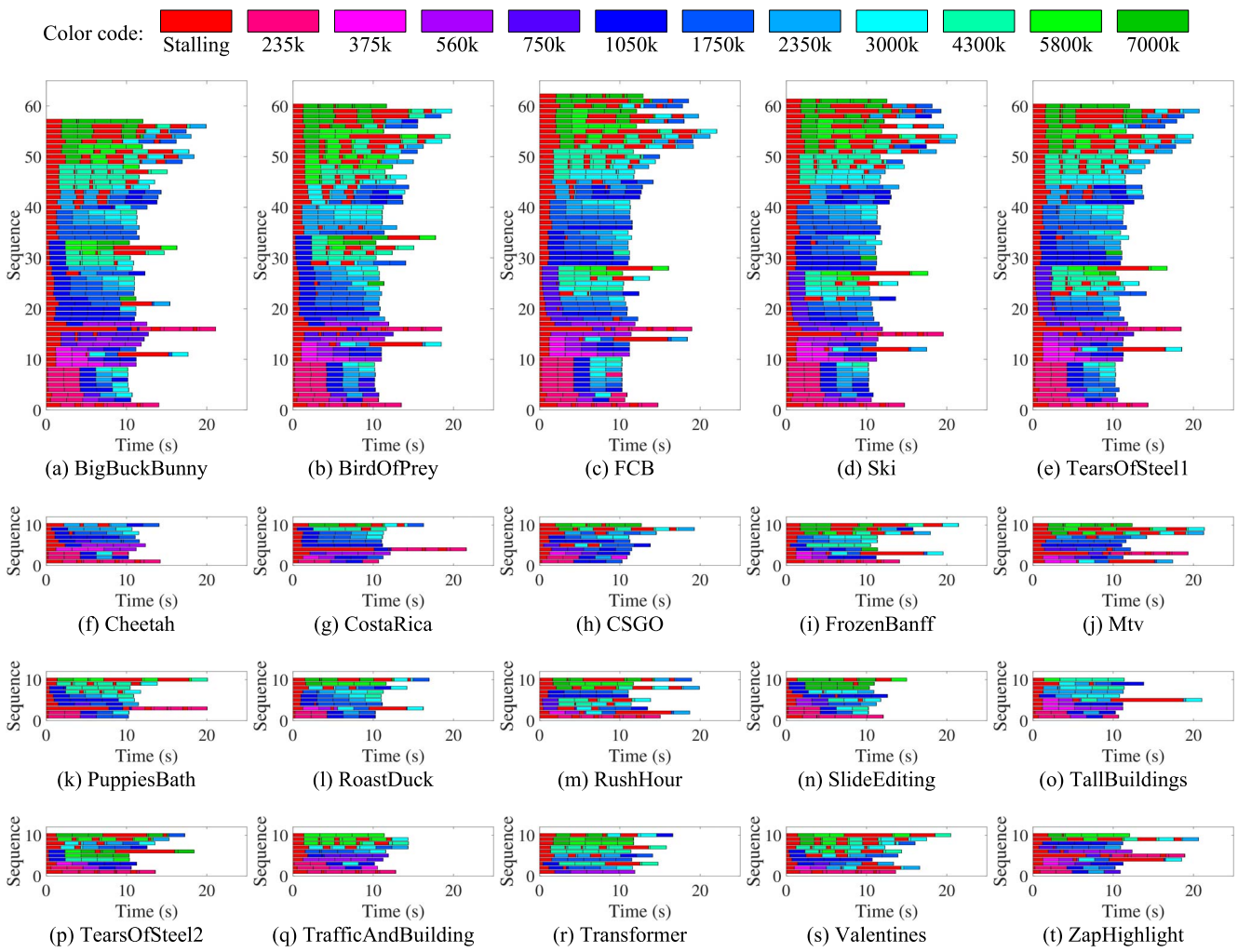


Fig. 4. Distortion profiles of the streaming video sequences in the subjective study. Each row in a subfigure represents a streaming video generated by one or multiple bitrate adaptation algorithms under one network profile.

bitrate if the viewing history is not available in the DOM storage. We set the initial bitrate to 1.2Mbps.

2. *BBA [3]:* We employed the function suggested by Huang *et al.* [3], where bitrate is chosen as a piecewise linear function of buffer occupancy. The algorithm always starts with the lowest bitrate till the buffer occupancy reaches a certain threshold called reservoir. Once reservoir is filled up, a higher bitrate is selected as the buffer occupancy increases till there is enough video segment in the buffer (upper reservoir) to absorb the variation caused by the varying capacity and by the finite chunk size, where the range from the lower to upper reservoir is defined as cushion. We set lower reservoir and cushion to be 2 and 5 seconds, respectively.
3. *AIMD [5]:* The algorithm picks the representation according to the bandwidth estimation using the previous downloaded chunk in an additive increase and multiplicative decrease manner. When the two thresholds for switching are not met, the algorithm keeps the selected bitrate.
4. *ELASTIC [7]:* This algorithm incorporates a PI-controller to maintain a constant duration of video in the buffer (5 seconds in the experiment). Since the bandwidth estimation module is not specified in the original implementation, we adopt the throughput prediction using harmonic mean of the past 5 chunks, because it is shown to be effective in previous studies [8].
5. *QDASH [4]:* QDASH picks an intermediate bitrate when there is a bandwidth drop to mitigate the negative impact of abrupt quality degradation. Without impacting the performance, we replace the proxy service for bandwidth estimation in the original implementation with the throughput prediction using harmonic mean of past 5 chunks for simplicity.
6. *FESTIVE [8]:* This rate-based algorithm balances both efficiency and stability, and incorporates fairness across players, which is not a concern of this paper. We assume there is no wait time between consecutive chunk downloads, and implement FESTIVE without the randomized chunk scheduling. Note that this does not negatively impact the player QoE. Specifically, FESTIVE calculates the efficiency score depending on the throughput prediction using harmonic mean of the past 5 chunks, as well as a stability score as a function of the bitrate

switches in the past 5 chunks. The bitrate is chosen to be the minimal stability score plus $\alpha = 12$ times efficiency score.

Since the selection of initial bitrate is not explicitly defined in AIMD, Elastic, QDASH, and FESTIVE, to provide a realistic simulation and to cover a diverse distortion pattern, we add random noise with standard deviation of 200Kbps to the initial bitrate in the actual trace as the selected initial bitrate.

In the end of the simulation, a total of 1,560 streaming sessions (20 source videos $\times$ 6 ABR algorithms $\times$ 13 bandwidth profiles) were recorded. Around 25% of the streaming videos were found to be duplications of each other by carefully examining the recorded streaming activity logs, and thus were discarded from the subjective experiment. The duplication was risen due to the intrinsic similarity among the ABR algorithms in certain bandwidth conditions. For example, most ABR algorithms stay at the lowest bitrate representations when the bandwidth condition is extremely poor. This results in 1,164 unique streaming videos. Due to the limited duration of the subjective experiment, we randomly selected 10 streaming sessions from the resulting streaming video pool for 15 contents and reconstructed all the streaming sessions of the other 5 contents. In summary, the Waterloo SQoE-III database consists of 20 reference videos and 450 simulated streaming videos, and of an average duration of 13 seconds. The number of samples originated by AIMD, BBA, Elastic, Festive QDash, Rate-based are 97, 97, 95, 101, 101, and 104, respectively. The detailed profile of the streaming videos is illustrated in Fig. 4.

B. Subjective Testing Methodology

The subjective testing adopts the single-stimulus methodology in which the reference videos are also evaluated in the same experimental session as the test streaming videos. The subjective experiment is setup as a normal indoor home settings with an ordinary illumination level, with no reflecting ceiling walls and floors. All videos are displayed at their actual pixel resolution on an LCD monitor at a resolution of 1920 $\times$ 1080 pixels with Truecolor (32bit) at 60Hz. The monitor is calibrated in accordance with the ITU-T BT.500 recommendations [12]. A customized graphical user interface is used to render the videos on the screen with random order and to record the individual subject ratings on the database. A total of 34 naïve subjects, including nineteen males and fifteen females aged between 18 and 35, participate in the subjective test. Visual acuity and color vision are confirmed from each subject before the subjective test. A training session is performed, during which, 4 videos that are different from the videos in the testing set are presented to the subjects. We used the same methods to generate the videos used in the training and testing sessions. Therefore, subjects knew what distortion types would be expected before the test session, and thus learning effects are kept minimal in the subjective experiment. Subjects were instructed with sample videos to judge the overall QoE considering all types of streaming activities in the session. For each subject, the whole study takes about 3 hours, which is divided into 6 sessions with five 7-minute breaks in-between. In order to minimize the influence of fatigue effect, the length of a session was limited to 25 minutes. The choice of a 100-point continuous scale as opposed to a discrete 5-point ITU-R Absolute Category Scale (ACR) has advantages: expanded range, finer distinctions between ratings, and demonstrated prior efficacy [43].

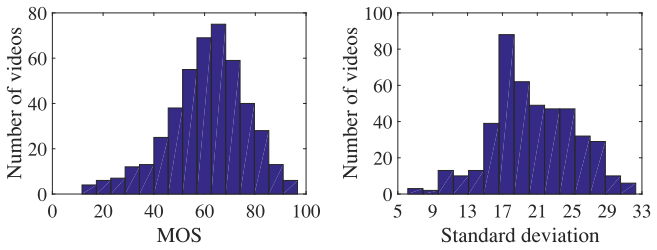


Fig. 5. MOS statistics of Waterloo SQoE-III database.

The raw subjective scores are converted to Z-scores per session to account for any differences in the use of the quality scale between sessions. Subsequently, 4 outliers are removed based on the outlier removal scheme suggested in [12], resulting in 30 valid subjects. After outlier removal, Z-scores are linearly rescaled to lie in the range of [0, 100]. The final quality score for each individual video is computed as the average of rescaled Z-scores, namely the mean opinion score (MOS), from all valid subjects. Fig. 5 plots the MOS scores across distorted videos for the subjective study, and shows the corresponding histograms for the MOS and the associated standard deviation in order to demonstrate that the distorted videos span most of the quality range. The average standard deviation and standard deviation of opinion scores parameter [44] in the MOS were 19 and 0.08, respectively. By comparison, however, subjects have a less degree of agreement in QoE for streaming videos with combined degradations than videos with isolated degradations in LIVEMVQA [11] and Waterloo SQoE-I [15].

IV. Performance of Objective QoE Models

A. Video Quality Assessment Models

Modern video quality assessment (VQA) algorithms tackle the QoE problem by measuring the signal fidelity of a test video with respect to its pristine version. However, most VQA models do not consider the impact of playback interruption. Since VQA models serve as the major tools to measure the QoE of offline videos, it is imperative to understand whether they can be applied to streaming videos. In this regard, we evaluate a wide variety of VQA algorithms including PSNR, SSIM [45], MS-SSIM [46], STRRED [49], VQM [48], VMAF [50], SSIMplus [47], and VIIDEO [51] against human subjective scores on two datasets to test their generalizability on streaming videos, where dataset D_A includes the videos without stalling and dataset D_B contains all 450 streaming videos. The implementations of the VQA models are obtained from the original authors. Spearman's rank-order correlation coefficient (SRCC) [52] is employed for performance evaluation by comparing MOS and objective QoE scores. Since none of the full-reference VQA algorithms supports cross-resolution video quality evaluation except for SSIMplus, we up-sampled

TABLE IV
PERFORMANCE COMPARISON OF VQA MODELS ON WATERLOO SQOE-III DATABASE

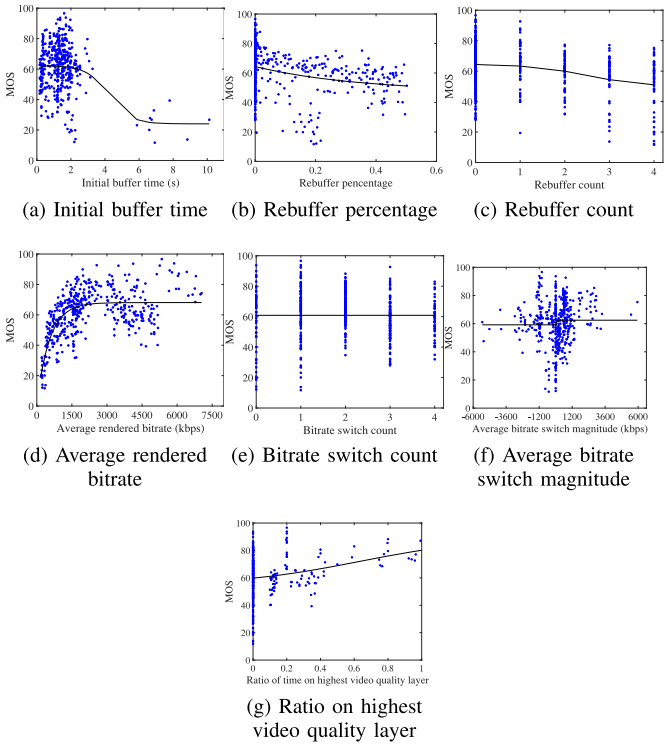
VQA model	SRCC		Computation time (normalized based on PSNR)
	D_A	D_B	
PSNR	0.6676	0.4606	**1**
SSIM [45]	0.7448	0.5240	8.38
MS-SSIM [46]	0.7438	0.5217	13.65
SSIMplus [47]	**0.8298**	**0.5617**	**1.28**
VQM [48]	**0.8192**	**0.5650**	31.90
STRRED [49]	0.6760	0.4706	160.29
VMAF [50]	0.7977	0.5613	23.41
VIIDEO [51]	0.4388	0.3506	140.46

all representation to 1920×1080 and then apply the VQA on the up-sampled videos because it is the size of display in the subjective experiment. Table IV summarizes the evaluation results, where the top 2 VQA models for each evaluation criterion are highlighted in bold. It can be observed that SSIMplus and VQM are the best-performing VQA models, while VIIDEO, the only no-reference model in the test is the weakest, suggesting that there remains significant room for improvement of no-reference VQA algorithms. It is also worth noting that there are significant performance drops from D_A to D_B, suggesting that QoE assessment for streaming video is a complex problem that requires more sophisticated modeling than what has been covered in traditional VQA models.

B. Industrial Standard QoE Features

DASH industry forum proposed a set of standard client-side QoE media metrics [53]. We evaluate five industry-standard QoE metrics [53], along with the average magnitude of switches that is also recognized as a major influencing factor of QoE [31]. We summarize the metrics as follows.

1. *Initial buffer time (T_i):* Measured in seconds, this metric represents the duration from the time that the player initiates a connection to a video server till the time that sufficient player video buffer has filled up and the player starts rendering video frames.
2. *Rebuffer percentage (P_r):* This metric is the fraction of the total session time (i.e., playing plus rebuffer time) spent in buffering. This is an aggregate metric that can capture periods of long video "freeze" observed by a user. It is computed as $\frac{\sum_i \text{ duration of rebuffer event } i}{\text{session duration}}$.
3. *Rebuffer count (C_r):* Rebuffer percentage does not capture the frequency of induced interruptions observed by a user. For example, a video session that experiences "video stuttering" where each interruption is small but the total number of interruptions is high, may not have a high buffering ratio, but may be just as annoying to a user.
4. *Average rendered bitrate ($\overline{B}$):* Measured in kilobytes per second, this metric is the most widely used video presentation quality measure in streaming applications. It is the average of the bitrates played weighted by the duration each bitrate is played.
5. *Bitrate switch count (C_s):* A single video session can have multiple bitrates played in HAS. Number of switches is usually used to quantify the flicker effects introduced by the quality variation.
6. *Average bitrate switch magnitude ($\overline{B}_s$):* Measured in kilobytes per switch, this metric was also identified as an influencing factor of flicker effect. Conventional wisdom dictates that people prefer multiple switches with smaller bitrate differences to abrupt quality variation. It is computed as $\frac{\sum_{i=2}^{n} |\text{bitrate}_i - \text{bitrate}_{i-1}|}{\#\text{ of switches}}$, where n is the number of segments.
7. *Ratio on the highest video quality layer (P_h):* Previous studies have argued that the effect of bitrate switch count is negligible compared to the percent of time on the highest quality layer [17], [30], [31], [54]–[56]. A few adaptation logics also employ the feature as a QoE measure [57]. Therefore, while it has not been included as an industrial standard QoE measure, we examine the influence of ratio on the highest video quality layer on QoE.



(a) Initial buffer time (b) Rebuffer percentage (c) Rebuffer count

(d) Average rendered bitrate (e) Bitrate switch count (f) Average bitrate switch magnitude

(g) Ratio on highest video quality layer

Fig. 6. Standard quality metrics versus MOS.

TABLE V
SRCC BETWEEN STANDARD CLIENT-SIDE QOE METRICS AND MOS

Quality metric	SRCC
Initial buffer time	-0.0303
Rebuffer percentage	-0.2733
Rebuffer count	-0.2505
Average rendered bitrate	**0.5118**
Bitrate switch count	0.1334
Average bitrate switch magnitude	0.1583
Ratio on highest video quality layer	0.1172

Fig. 6 shows the scatter plots of the 7 aforementioned quality metrics versus MOS. We evaluate the performance of each metric using SRCC and summarize the result in Table V. Fig. 6 illustrates that only average rendered bitrate has a roughly

monotonic relationship with MOS in general. However, the correlation is not high and it exhibits a strong nonlinear relationship with respect to MOS. In particular, bitrates in the range of 2,500 kbps to 7,200 kbps yield a very similar QoE. Furthermore, the moderate correlation between bitrate and quality is expected to drop further when video sequences encoded from various codecs and implementations are mixed together. Thus, existing video delivery optimization frameworks that strive for higher bitrate in all ranges not only result in inefficient use of network, but also do not necessarily provide a better QoE. On the other hand, the two second-order statistics of bitrate - bitrate switch count and average bitrate switch magnitude - have relatively little impact on MOS. Our experimental results also contradict the conclusion from previous studies that the ratio on the highest video quality layer exhibited considerably high correlation with QoE. One possible explanation is that the number of quality layers in previous studies is relatively small. Since the criterion only considers information about the highest quality layer, it effectively "throws away" information about the remaining quality layers distribution. Such simplification may not generalize well, especially when the quality difference between consecutive layers is not significantly large as presented in our study. In addition, despite the general trend of MOS with respect to initial buffer time, rebuffer percentage, rebuffer count, bitrate switch count, and average bitrate switch magnitude, none of them is sufficient to predict QoE accurately. Therefore, it is difficult to compare the performance of ABR algorithms and optimization frameworks with the statistics of isolated metrics, which unfortunately remains as the major validation approach in practice. Moreover, we augment the correlation analysis with ANalysis Of Variance (ANOVA) on the MOS data to reveal the statistical significance of each metric on MOS, where the significance level p-value is set to 0.05. We choose bin sizes that are appropriate for each quality metric of interest: 1-second bin, 5% bin, unit bin, 360 kbps-sized bin, unit bin, 600 kbps-sized bin, and 5% bin for initial buffering time, rebuffer percentage, rebuffer count, average rendered bitrate, bitrate switch count, average bitrate switch magnitude, and ratio on the highest video quality layers, respectively. The results of ANOVA suggest that initial buffer time is the only factor that is statistically insignificant to MOS.

Given the poor performance of isolated quality metrics, a natural question is: Does combination of metrics provide more insights? We plot the cross metric correlation in Fig. 7, where most metric pairs are quite independently to each other, which indicates that metrics may supplement each other and there is a potential for a combination of metrics to provide a better performance. Thus, we randomly divide the video data into disjoint 80% training and 20% testing subsets, and apply linear regression on the training subset and then test on the testing subset. To mitigate any bias due to the division of data, the process is repeated 1000 times. SRCC between the predicted and the ground truth quality scores are computed at the end of each iteration. The median correlation and its corresponding regression model are reported in Table VI. For clarity, rather than showing all combinations, we include 2, 3, and 4 variant regression models with the highest correlations.

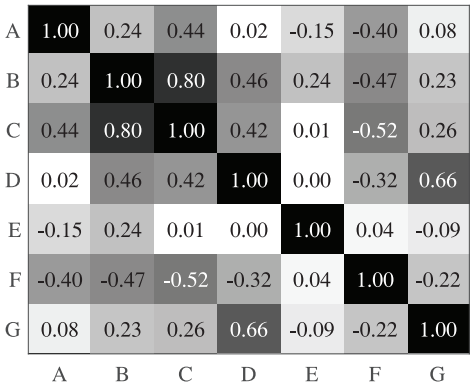


Fig. 7. Metric correlation matrix. A: Initial buffer time; B: rebuffer percentage; C: rebuffer count; D: average rendered bitrate; E: bitrate switch count; F: average bitrate switch magnitude; G: ratio on highest video quality layer.

TABLE VI
MEDIAN SRCC ACROSS 1000 TRAIN-TEST COMBINATIONS OF REGRESSION MODELS

Regression model	SRCC
-64.4P_r+0.0073$\overline{B}$+50.8	0.7264
-63.1P_r+0.0079$\overline{B}$+0.0010$\overline{B}_s$+49.7	0.7743
-2.3T_i-56.5P_r+0.0070$\overline{B}$+0.0007$\overline{B}_s$+54.0	0.7800

For all metrics, the combination with the average rendered bitrate provides the highest correlation while the combination of average rendered bitrate and rebuffer percentage achieves the highest correlation to MOS amongst bi-variant regression models. What is also worth mentioning is that although bitrate switch count and average bitrate switch magnitude are weakly correlated with MOS, the performance of linear regression model can be notably improved by taking these video quality variation indicators into consideration. The results encourage exploration of advanced models that predict human perception of time-varying video quality.

C. Evaluation of Objective QoE Models

Using the above database, we test the performance of 13 state-of-the-art QoE models from three categories: network QoS-based, [60], application QoS-based [9], [58], [59], [61]–[63], and hybrid models of application QoS and signal fidelity [10], [28], [54]–[56], [64]. A description of the 13 QoE models is shown in Table VII. While various of state-of-the-art temporal pooling strategies are shown to perform well on videos of time-varying quality [27], [65], we do not validate the algorithms for the following reasons. First, as we have shown in Section IV-B, the impact of switching is relatively small compared to stalling. Second, how to apply these pooling strategies on videos with stalling events is an open question. It is also worth noting that we do not include the enhanced version of VsQM [66] in the performance comparison because the algorithm only applies to video sequences with three segments. For fairness, all models are tested using their default parameter settings. For Xue *et al.* 's [63], we set $c = 0.05$ such that the model achieves its optimal performance on the current database. For the models that do not explicitly account for the duration of

TABLE VII
COMPARISON OF EXISTING VIDEO QOE MODELS

QoE model	Stalling & initial buffering		Presentation quality		Switching
	Regression function	Influencing factors	Regression function	Influencing factors	Regression function
Liu's [58]	linear	stalling length	linear	bitrate	—
Yin's [9]	linear	stalling length	linear	bitrate	linear
FTW [59]	exponential	stalling length, # of stalling	—	—	—
Bentaleb's [10]	linear	# of stalling, stalling length	linear	SSIMplus	linear
NARX-QoE [55]	NARX	# of stalling, stalling length,	NARX	STRRED	NARX
ATLAS [56]	support vector regression	# of stalling, stalling length,	support vector regression	STRRED	support vector regression
Kim's [60]	—	—	exponential	packet loss packet jitter bandwidth efficiency	—
Mok's [61]	linear	stalling length, stalling frequency, initial buffering length	—	—	—
VsQM [62]	exponential	average stalling length per segment, # of stalling per segment, period per segment	—	—	—
Xue's [63]	logarithmic	stalling length, # of stalling, bit count of the stalling segment	linear	QP	—
Liu's [28]	polynomial	# of stalling, stalling length, magnitude of motion vector	exponential	VQM	quadratic
SQI [15]	combination of exponentials	# of stalling, stalling length, video quality of stalling segment	linear	SSIMplus	—
P.NATS [54]	polynomial + random forest	# of stalling, stalling length, average stalling interval	random forest	O.21, O.22	random forest

TABLE VIII
PERFORMANCE COMPARISON OF QOE MODELS ON WATERLOO SQOE-III DATABASE. A: SIGNAL FIDELITY-BASED; B: APPLICATION QOS-BASED; C: NETWORK QOS-BASED; AND D: HYBRID MODELS

QoE model	Type	SRCC performance	Computation complexity	
			server (s)	client (s)
SSIMplus [47]	A	0.5617	0.98	0
VQM [48]	A	0.5650	24.40	0
Liu's [58]	B	0.5145	0	0.01
Yin's [9]	B	0.7143	0	0.01
FTW [59]	B	0.2745	0	0.01
Bentaleb's [10]	D	0.6322	0.98	0.01
NARX-QoE [55]	D	0.4236	161.57	1.83
ATLAS [56]	D	0.1941	161.57	0.85
Kim's [60]	C	0.0196	0	0.01
Mok's [61]	B	0.1702	0	0.02
VsQM [62]	B	0.2010	0	0.01
Xue's [63]	B	0.3840	0	0.14
Liu's [28]	D	0.8039	24.40	0.05
SQI [15]	D	0.7707	0.98	0.01
P.NATS [54]	D	**0.8454**	0.02	0.07

initial buffering, we follow the recommendation in [54] by considering it as a special stalling event with a discounting factor of $\frac{1}{3}$. For NARX-QoE [55], we follow the original authors' recommendation by initializing the model with SQI [15]. The criteria described in Section IV are employed for performance evaluation by comparing MOS and objective QoE. Table VIII summarizes the evaluation results of the QoE models from three categories along with the two top VQA algorithms in terms of prediction accuracy and computational complexity. The computational complexity is measured as the average computation time required to assess per second of video (using a computer with Intel Core i7-4790 processor at 3.60 GHz). Scatter plots of objective scores vs. MOS for all the algorithms on the entire Waterloo SQoE-III database, along with the best fitting logistic functions, are shown in Fig. 8. These test results provide some useful insights regarding the general approaches used in QoE models. First of all, the stalling-centric QoE models [59], [61], [62] do not perform well. The major reason is that these models (i.e., FTW [59] and Mok *et al.* 's [61]) do not take the presentation quality of the videos into consideration, which is shown to be a major influencing factor of QoE. Indeed, this is quite apparent from our test results, where even PSNR, a very crude presentation quality measure that does not take into account any initial buffering or stalling at all, performs significantly better than stalling-centric methods. Second, the network QoS-based QoE model Kim *et al.* 's [60] also performs poorly because it ignores the characteristics of source video and ABR algorithms. As we have shown in the Section V that either utilizing different ABR algorithms at the same network condition or using the same bitrate to encode different video content could lead to drastically different QoE. Third, although Xue *et al.* 's [63] performs reasonably well in previous study [15], it does not well predict subjective QoE on the database with more diverse distortion patterns. A plausible explanation may be that quantization parameter (QP) in video codec setting is not a good indicator of perceptual quality, especially when there are multiple spatial resolutions in the bitrate ladder. Fourth, it is clear that all QoE models except for SQI [15] and P.NATS [54] fail to provide an adequate alignment to the clusters with and without stalling, suggesting that it is important to capture the interactions between video presentation quality and the impact

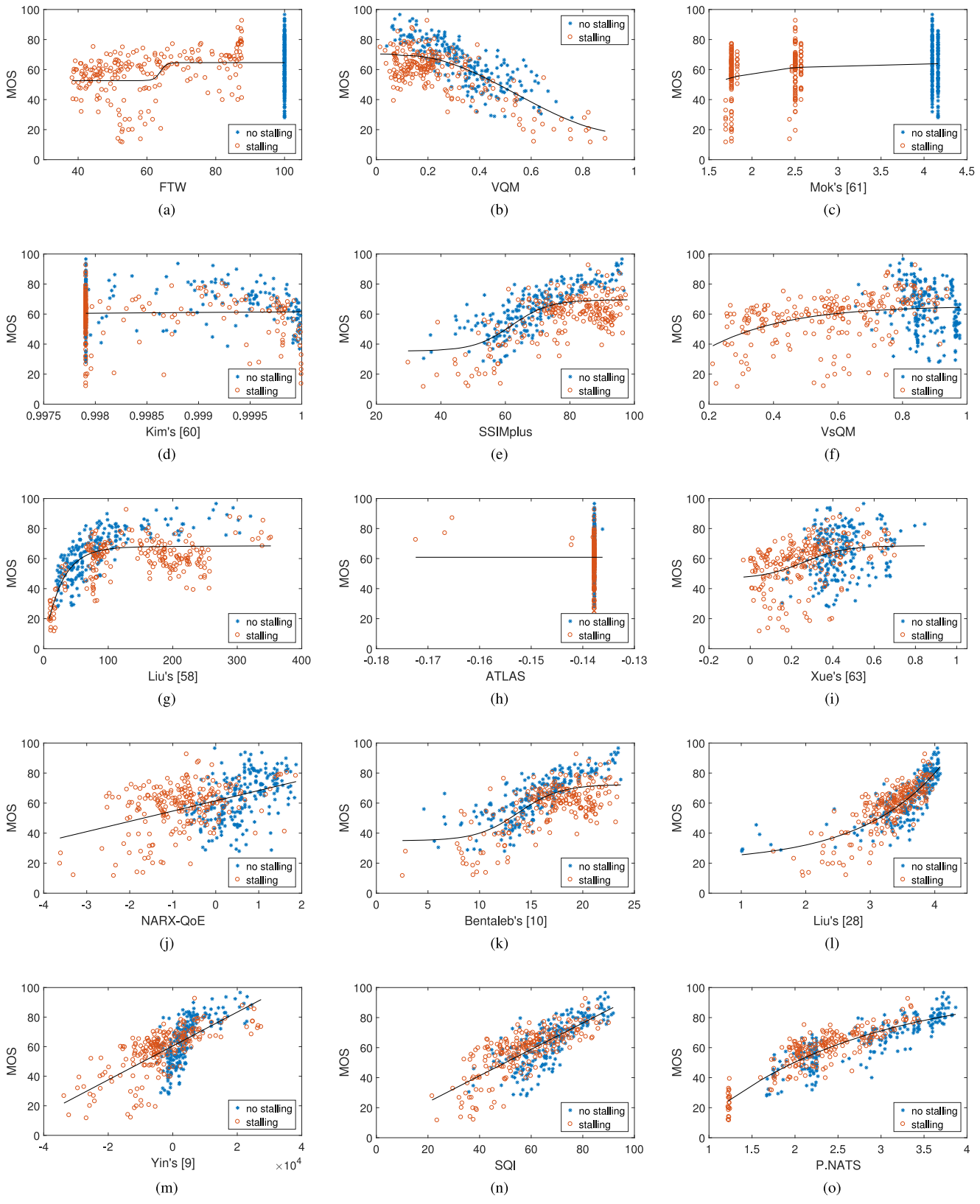


Fig. 8. Scatter plots of model predicted QoE vs. MOS. Also shown are the best fitting logistic functions as solid curves.

of stalling. Fifth, P.NATS [54], an objective QoE model based on random forest regression of 14 features, achieves the highest prediction accuracy with a clear margin over all other models. The inference to be drawn from this is that there exist interactions among video presentation quality, quality adaptation, and the experience of stalling. Simply computing

the experience of each dimension followed by weighted averaging would lead to suboptimal performance. What is also worth mentioning is that the hybrid models NARX-QoE and ATLAS perform relatively poor, suggesting that the ratio on the highest video quality layer is not a robust feature in QoE prediction. At last, all models overestimate the QoE of live video sequence FCB at low bitrates, suggesting that a content type-aware QoE model may further improve the performance of existing QoE models.

We carry out a statistical significance analysis by following the approach introduced in [67]. First, a nonlinear regression function is applied to map the objective quality scores to predict the subjective scores. We observe that the prediction residuals all have zero-mean, and thus the model with lower variance is generally considered better than the one with higher variance. We conduct a hypothesis testing using F-statistics. Since the number of samples exceeds 50, the Gaussian assumption of the residuals approximately hold based on the central limit theorem [68]. The test statistic is the ratio of variances. The null hypothesis is that the prediction residuals from one quality model come from the same distribution and are statistically indistinguishable (with 95% confidence) from the residuals from another model. The results are summarized in Table IX, where a symbol '1' means the row model performs significantly better than the column model, a symbol '0' means the opposite, and a symbol '-' indicates that the row and column models are statistically indistinguishable. The performance of QoE models can be roughly clustered into three levels, wherein P.NATS [54], Liu *et al.*'s [28], and SQI [15] are statistically superior to all other QoE models. It is worth noting that even though the SQI model [15] does not take into account quality adaptation/switching completely, it still achieves highly competitive performance. While the two top performers of application QoS-based models Liu *et al.*'s [58] and Yin *et al.*'s [9], the two top performers of signal fidelity-based models SSIMplus [47] and VQM [48], and the worst hybrid model Bentaleb *et al.*'s [10] are statistically inferior to the tier-1 models, they outperform the last group which mainly consists of QoS-based models. It is quite apparent that hybrid QoE models exhibit clear advantages.

There is inherent variability amongst subjects in the quality judgment of a streaming video. It is important not to penalize an algorithm if the differences between the algorithm scores and MOS can be explained by the inter-subject variability. Therefore, we follow the recommendation in [67] to compare the objective QoE models with the theoretical null model. Specifically, we compute the ratio between the variances of residuals between the individual ratings of all streaming videos and the corresponding MOS and the residual between individual ratings and the algorithm prediction of QoE (after non-linear regression). The ratio of two variances forms the F-statistic under central limit theorem. The null hypothesis is that the variance of the model residual is statistically indistinguishable (with 95% confidence) to the variance of the null residual. A threshold F-ratio can be determined based on the degrees of freedom in the numerator and denominator, along with the confidence level, where the numerator and denominator degrees of freedom in the F-test is obtained by subtracting one from the number of samples. Values of the F-ratio larger than the threshold cause us to reject the null hypothesis, and vice verse. The variance of the residuals from the null model and each of the 15 objective QoE models are shown in Fig. 9, wherein none of the QoE models is equivalent to the theoretical null model, suggesting that there remains considerable opportunity to improve the performance of QoE models.

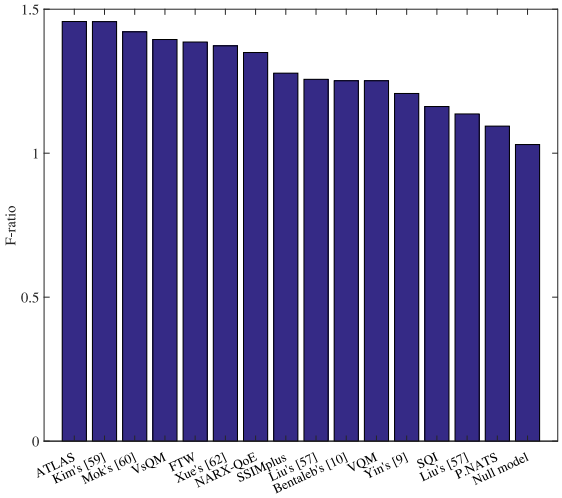


Fig. 9. F-ratios comparison of objective QoE models and theoretical null model.

V. Evaluation of ABR Algorithms

We use MOS of the 6 ABR algorithms described in the previous section to evaluate and compare their performance. The mean of MOS values across different content under 13 network profiles for the ABR algorithms are summarized in Table X. It is worth mentioning that this only provides a rough comparison of the relative performance of the ABR algorithms in the "startup phase". Besides, computational complexity is not a factor under consideration.

From the subjective test results, we have several observations. First, the video quality at which the content is streamed has a significantly higher impact on sports content, *e.g.,* FCB, than on other content. In particular, none of the video sequences of average bitrate lower than 800 kbps received a rating higher than 60. This is consistent with previous study [25]. Second, BBA [3], which spends 60% of the time at bitrates lower than 1,000 kbps even under the best network condition in the experiment, provides the lowest QoE under most network conditions. Similarly, due to the conservative switching strategy that the player only switches to the next level and uses a lower rate of upward switches at higher representation levels, FESTIVE [8] (the algorithm increases the bitrate at bitrate level k only after k chunks) performs poorly under the ramp up network condition VIII. This suggests that a consistently low video presentation quality is not tolerated by subjects. Third, FESTIVE [8] achieves the best performance under the ramp down network condition VII although it consumes the lowest bitrate among bandwidth-aware algorithms due to its multiplicative (0.85) factor on the estimated bandwidth. This conservative strategy

TABLE IX
STATISTICAL SIGNIFICANCE MATRIX BASED ON F-STATISTICS ON THE WATERLOO SQOE-III DATABASE. A SYMBOL "1" MEANS THAT THE PERFORMANCE OF THE ROW MODEL IS STATISTICALLY BETTER THAN THAT OF THE COLUMN MODEL, A SYMBOL "0" MEANS THAT THE ROW MODEL IS STATISTICALLY WORSE, AND A SYMBOL "-" MEANS THAT THE ROW AND COLUMN MODELS ARE STATISTICALLY INDISTINGUISHABLE

	SSIMplus [47]	VQM [48]	Liu's [58]	Yin's [9]	FTW [59]	Bentaleb's [10]	Kim's [60]	Xue's [63]	Mok's [61]	VsQM [62]	NARX-QoE [55]	ATLAS [56]	Liu's [28]	SQI [15]	P.NATS [54]
SSIMplus [47]	-	-	-	-	1	-	1	1	1	1	1	1	0	0	0
VQM [48]	-	-	-	-	1	-	1	1	1	1	1	1	0	0	0
Liu's [58]	-	-	-	-	1	-	1	1	1	1	1	1	0	0	0
Yin's [9]	-	-	-	-	1	-	1	1	1	1	1	1	0	0	0
FTW's [59]	0	0	0	0	-	0	-	-	-	-	-	-	0	0	0
Bentaleb's [10]	-	-	-	-	1	-	1	1	1	1	1	1	0	0	0
Kim's [60]	0	0	0	0	-	0	-	-	-	0	-	-	0	0	0
Xue's [63]	0	0	0	0	-	0	-	-	-	-	-	-	0	0	0
Mok's [61]	0	0	0	0	-	0	-	-	-	-	-	-	0	0	0
VsQM [62]	0	0	0	0	-	0	-	-	-	-	-	-	0	0	0
NARX-QoE [55]	0	0	0	0	-	0	1	-	-	-	-	1	0	0	0
ATLAS [56]	0	0	0	0	-	0	-	-	-	0	-	-	0	0	0
Liu's [28]	1	1	1	1	1	1	1	1	1	1	1	1	-	-	0
SQI [15]	1	1	1	1	1	1	1	1	1	1	1	1	-	-	0
P.NATS [54]	1	1	1	1	1	1	1	1	1	1	1	1	1	1	-

TABLE X
PERFORMANCE OF ABR ALGORITHMS UNDER DIFFERENT NETWORK PROFILES

	I	II	III	IV	V	VI	VII	VIII	IX	X	XI	XII	XIII	Average
AIMD	22.33	45.46	59.40	62.09	63.98	77.54	62.44	71.61	63.31	65.19	64.58	69.36	73.40	61.28
BBA	**25.97**	50.24	47.42	46.33	50.08	55.49	55.92	55.69	55.11	53.43	55.11	55.11	55.11	50.25
ELASTIC	22.33	49.08	56.42	63.41	**67.77**	**77.74**	63.79	70.51	56.52	65.18	**67.54**	64.48	**79.91**	62.48
FESTIVE	22.33	50.01	45.95	**67.31**	59.21	74.70	**75.59**	56.27	61.75	**67.48**	64.08	72.92	76.86	**62.68**
QDASH	22.33	45.25	54.59	64.52	64.82	77.36	66.33	**76.32**	**67.29**	63.85	65.22	65.45	75.91	61.67
Rate-based	23.32	**50.59**	**60.80**	59.92	66.72	72.27	66.85	71.00	54.92	61.91	59.41	**76.79**	77.45	61.59

helps tolerate the buffer fluctuation caused by variability in chunk size and reduces the likelihood of stalling, especially at high bitrates because a sudden bandwidth drop may result in longer stalling time at higher bitrates. Based on the two observations, we conclude that a QoE-driven ABR algorithm should adopt a stateful bitrate selection that performs aggressively at low bitrates and conservatively at high bitrates. While FESTIVE [8] takes the stateful approach, bitrate level is not a proper indicator of state because it does not generalize well to different size of bitrate ladder. Interestingly, previous studies [8] proved that such stateful design converges to a fair share of bandwidth if there are multiple competitors. Fourth, the rate-based algorithm [2] performs at least as good as other bandwidth-aware algorithms under network conditions I, II, and III but poorly otherwise. This may be explained by the startup strategy. Since the rate-based algorithm [2] starts with a constant bitrate regardless of the network condition while the other bandwidth-aware algorithms start with bitrates around the initial bandwidth, the initial bitrates of the rate-based algorithm [2] is the highest among the bandwidth-aware algorithms under the first three network conditions and the lowest under other network conditions. This suggests that the fast startup strategy that begins with low bitrates is not appreciated by the subjects. This phenomenon is also orally confirmed by the participants. Fifth, QDASH [4] that temporarily trades the buffer occupancy for high bitrates during bandwidth drop outperforms all other algorithms under the ramp down network condition XI, which confirms that smooth quality degradations are preferred over abrupt transitions [4]. From the algorithm design space point of view, both rate-based and buffer-based algorithms discard useful information, and thus result in suboptimal solution. Sixth, not a single algorithm provides the best perceptual quality under all network profiles. On average, the performance of 5 out of the 6 models under testing is fairly close to each other. This suggests that there is still room for future improvement. In particular, proper combination of the ideas used in different ABR algorithms has the potential to further improve the performance.

VI. CONCLUSION AND FUTURE WORK

We introduced the Waterloo SQoE-III database containing 450 streaming videos that were derived from diverse source videos using session reconstruction. The dataset is diverse in terms of video content, and is both realistic and diverse in distortion types. We assessed 15 QoE models with statistical analysis and shed light on the development of both ABR and QoE measurement algorithms. The database is made publicly available to facilitate future QoE research. We hope that the Waterloo SQoE-III database will provide fertile ground for years of future research. Given the sheer quantity of data, we believe that our foregoing analysis is the tip of the ice-berg of discovery. We invite further analysis of the data towards understanding and producing better models of human behavior.

It is important to mention some of the limitations of the Waterloo SQoE-III database. First, while conventional wisdom [11], [12], [36], [37] provides support to the short videos as experiment materials, many recent studies suggest that longer videos of up to 30 seconds may be required to be able to test the impact of switching patterns [13], [31]. The impact of video length in adaptive streaming is still an open question; thus, we consider the conclusions from the experiment at this point provisional. Second, although we have tried our

best to construct a database that comprise as many content type, segment length, rendering device, and audience profile as possible, the experiment is by no means exhaustive. In the design of the database and experiment, we find that the biggest challenge arises from the limited capacity of subjective testing (which is caused by multiple factors including the cost, the time involved, and the potential fatigue effect of the subjects). With such a strong limitation on capacity, it is extremely difficult to accommodate exhaustive test sequences. As a result, certain assumptions and simplifications have to be made to reduce the combinations.

Future research may be carried out in many directions. First, other existing and future QoE models may be tested and compared by making use of the database. Second, objective QoE models that incorporate spatio-temporal aspects of videos and that predict human reactions to spatial adaptation and temporal adaptation could ultimately help video streaming engines allocate resources in a smarter way. Third, optimization of the existing video streaming frameworks based on QoE models is another challenging problem that is worth further investigations.

REFERENCES

[1] T. Stockhammer, "Dynamic adaptive streaming over HTTP: Standards and design principles," in *Proc. ACM Conf. Multimedia Syst.*, San Jose, CA, USA, 2011, pp. 133–144.

[2] DASH Industry Forum. *For Promotion of MPEG-DASH 2013*. Accessed: Dec. 12, 2017. [Online]. Available: http://dashif.org

[3] T.-Y. Huang, R. Johari, N. McKeown, M. Trunnell, and M. Watson, "A buffer-based approach to rate adaptation: Evidence from a large video streaming service," *ACM SIGCOMM Comput. Commun. Rev.*, vol. 44, no. 4, pp. 187–198, Feb. 2015.

[4] R. K. P. Mok, X. Luo, E. W. W. Chan, and R. K. C. Chang, "QDASH: A QoE-aware DASH system," in *Proc. ACM Conf. Multimedia Syst.*, Chapel Hill, NC, USA, 2012, pp. 11–22.

[5] C. Liu, I. Bouazizi, and M. Gabbouj, "Rate adaptation for adaptive HTTP streaming," in *Proc. ACM Conf. Multimedia Syst.*, San Jose, CA, USA, 2011, pp. 169–174.

[6] Z. Wang, K. Zeng, A. Rehman, H. Yeganeh, and S. Wang, "Objective video presentation QoE predictor for smart adaptive video streaming," in *Proc. SPIE*, vol. 9599. San Diego, CA, USA, 2015, pp. 1–13.

[7] L. De Cicco, V. Caldaralo, V. Palmisano, and S. Mascolo, "Elastic: A client-side controller for dynamic adaptive streaming over http (DASH)," in *Proc. IEEE Int. Packet Video Workshop*, San Jose, CA, USA, 2013, pp. 1–8.

[8] J. Jiang, V. Sekar, and H. Zhang, "Improving fairness, efficiency, and stability in http-based adaptive video streaming with festive," in *Proc. ACM Int. Conf. Emerg. Netw. Exp. Technol.*, Nice, France, 2012, pp. 97–108.

[9] X. Yin, A. Jindal, V. Sekar, and B. Sinopoli, "A control-theoretic approach for dynamic adaptive video streaming over HTTP," *ACM SIGCOMM Comput. Commun. Rev.*, vol. 45, no. 4, pp. 325–338, Oct. 2015.

[10] A. Bentaleb, A. C. Begen, and R. Zimmermann, "SDNDASH: Improving QoE of HTTP adaptive streaming using software defined networking," in *Proc. ACM Int. Conf. Multimedia*, Amsterdam, The Netherlands, 2016, pp. 1296–1305.

[11] A. K. Moorthy, L. K. Choi, A. C. Bovik, and G. De Veciana, "Video quality assessment on mobile devices: Subjective, behavioral and objective studies," *IEEE J. Sel. Topics Signal Process.*, vol. 6, no. 6, pp. 652–671, Oct. 2012.

[12] "Methodology for the subjective assessment of the quality of television pictures," ITU, Geneva, Switzerland, ITU-Recommendation BT.500-12, Nov. 1993.

[13] C. Chen *et al.*, "Modeling the time—Varying subjective quality of HTTP video streams with rate adaptations," *IEEE Trans. Image Process.*, vol. 23, no. 5, pp. 2206–2221, May 2014.

[14] 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 Glob. Conf. Signal Inf. Process.*, Atlanta, GA, USA, 2014, pp. 989–993.

[15] Z. Duanmu, K. Zeng, K. Ma, A. Rehman, and Z. Wang, "A quality-of-experience index for streaming video," *IEEE J. Sel. Topics Signal Process.*, vol. 11, no. 1, pp. 154–166, Feb. 2017.

[16] Z. Duanmu, K. Ma, and Z. Wang, "Quality-of-experience of adaptive video streaming: Exploring the space of adaptations," in *Proc. ACM Int. Conf. Multimedia*, Mountain View, CA, USA, 2017, pp. 1752–1760.

[17] C. G. Bampis *et al.*, "Study of temporal effects on subjective video quality of experience," *IEEE Trans. Image Process.*, vol. 26, no. 11, pp. 5217–5231, Nov. 2017.

[18] Y. Qi and M. Dai, "The effect of frame freezing and frame skipping on video quality," in *Proc. IEEE Int. Conf. Intell. Inf. Hiding Multimedia Signal Process.*, Pasadena, CA, USA, 2006, pp. 423–426.

[19] A. Floris, L. Atzori, G. Ginesu, and D. D. Giusto, "QoE assessment of multimedia video consumption on tablet devices," in *Proc. IEEE Globecom Workshops*, Anaheim, CA, USA, Dec. 2012, pp. 1329–1334.

[20] L. Atzori, A. Floris, G. Ginesu, and D. D. Giusto, "Quality perception when streaming video on tablet devices," *J. Vis. Commun. Image Represent.*, vol. 25, no. 3, pp. 586–595, Apr. 2014.

[21] N. Staelens *et al.*, "Assessing quality of experience of IPTV and video on demand services in real-life environments," *IEEE Trans. Broadcast.*, vol. 56, no. 4, pp. 458–466, Dec. 2010.

[22] T. Hossfeld *et al.*, "Initial delay vs. interruptions: Between the devil and the deep blue sea," in *Proc. IEEE Int. Conf. Qual. Multimedia Exp.*, 2012, pp. 1–6.

[23] A. Sackl, S. Egger, and R. Schatz, "Where's the music? Comparing the QoE impact of temporal impairments between music and video streaming," in *Proc. IEEE Int. Conf. Qual. Multimedia Exp.*, 2013, pp. 64–69.

[24] P. Ni, R. Eg, A. Eichhorn, C. Griwodz, and P. Halvorsen, "Flicker effects in adaptive video streaming to handheld devices," in *Proc. ACM Int. Conf. Multimedia*, Scottsdale, AZ, USA, 2011, pp. 463–472.

[25] F. Dobrian *et al.*, "Understanding the impact of video quality on user engagement," *ACM SIGCOMM Comput. Commun. Rev.*, vol. 41, no. 4, pp. 362–373, Aug. 2011.

[26] Y. Liu *et al.*, "A study on quality of experience for adaptive streaming service," in *Proc. IEEE Int. Conf. Commun. Workshop*, Budapest, Hungary, 2013, pp. 682–686.

[27] A. Rehman and Z. Wang, "Perceptual experience of time-varying video quality," in *Proc. IEEE Int. Conf. Qual. Multimedia Exp.*, Dec. 2013, pp. 218–223.

[28] Y. Liu, S. Dey, F. Ulupinar, M. Luby, and Y. Mao, "Deriving and validating user experience model for DASH video streaming," *IEEE Trans. Broadcast.*, vol. 61, no. 4, pp. 651–665, Dec. 2015.

[29] S. Lederer, C. Müller, and C. Timmerer, "Dynamic adaptive streaming over HTTP dataset," in *Proc. ACM Conf. Multimedia Syst.*, Chapel Hill, NC, USA, 2012, pp. 89–94.

[30] T. Hoßfeld, M. Seufert, C. Sieber, and T. Zinner, "Assessing effect sizes of influence factors towards a QoE model for HTTP adaptive streaming," in *Proc. IEEE Int. Conf. Qual. Multimedia Exp.*, Singapore, 2014, pp. 111–116.

[31] M. Seufert *et al.*, "A survey on quality of experience of HTTP adaptive streaming," *IEEE Commun. Surveys Tuts.*, vol. 17, no. 1, pp. 469–492, 1st Quart., 2014.

[32] M.-N. Garcia *et al.*, "Quality of experience and HTTP adaptive streaming: A review of subjective studies," in *Proc. IEEE Int. Conf. Qual. Multimedia Exp.*, Singapore, 2014, pp. 141–146.

[33] C. Timmerer, M. Maiero, and B. Rainer, "Which adaptation logic? An objective and subjective performance evaluation of HTTP-based adaptive media streaming systems," *CoRR*, vol. abs/1606.00341, p. 11, Jun. 2016. [Online]. Available: http://arxiv.org/abs/1606.00341

[34] C. Müller, S. Lederer, and C. Timmerer, "An evaluation of dynamic adaptive streaming over HTTP in vehicular environments," in *Proc. ACM Workshop Mobile Video*, Chapel Hill, NC, USA, 2012, pp. 37–42.

[35] L. Song, X. Tang, W. Zhang, X. Yang, and P. Xia, "The SJTU 4K video sequence dataset," in *Proc. IEEE Int. Conf. Qual. Multimedia Exp.*, 2013, pp. 34–35.

[36] P. Fröhlich *et al.*, "QoE in 10 seconds: Are short video clip lengths sufficient for quality of experience assessment?" in *Proc. IEEE Int. Conf. Qual. Multimedia Exp.*, 2012, pp. 242–247.

[37] "Subjective video quality assessment methods for multimedia applications," ITU, Geneva, Switzerland, ITU-Recommendation BT.910, Sep. 1999.

[38] *Per-Title Encode Optimization*, Netflix Inc., Scotts Valley, CA, USA, 2015. [Online]. Available: http://techblog.netflix.com/2015/12/per-title-encode-optimization.html
[39] "Best practices for creating and deploying HTTP live streaming media for apple devices," Apple Inc., Cupertino, CA, USA, Tech. Note TN2224, 2016. [Online]. Available: https://developer.apple.com/library/content/technotes/tn2224/_index.html
[40] 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/
[41] J. Le Feuvre, C. Concolato, and J.-C. Moissinac, "GPAC: Open source multimedia framework," in *Proc. ACM Int. Conf. Multimedia*, Augsburg, Germany, 2007, pp. 1009–1012.
[42] A. Zambelli. *Smooth Streaming Technical Overview*. Accessed: Dec. 13, 2017. [Online]. Available: http://www.iis.net/learn/media/on-demand-smooth-streaming/smoothstreamingtechnical-overview
[43] K. Ma *et al.*, "Group MAD competition—A new methodology to compare objective image quality models," in *Proc. IEEE Int. Conf. Comput. Vis. Pattern Recognit.*, Las Vegas, NV, USA, 2016, pp. 1664–1673.
[44] T. Hoßfeld, P. E. Heegaard, M. Varela, and S. Möller, "QoE beyond the MOS: An in-depth look at QoE via better metrics and their relation to MOS," *Qual. User Exp.*, vol. 1, no. 1, p. 2, Dec. 2016.
[45] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: From error visibility to structural similarity," *IEEE Trans. Image Process.*, vol. 13, no. 4, pp. 600–612, Apr. 2004.
[46] Z. Wang, E. P. Simoncelli, and A. C. Bovik, "Multiscale structural similarity for image quality assessment," in *Proc. IEEE Asilomar Conf. Signals Syst. Comput.*, vol. 2. Pacific Grove, CA, USA, 2003, pp. 1398–1402.
[47] A. Rehman, K. Zeng, and Z. Wang, "Display device-adapted video quality-of-experience assessment," in *Proc. SPIE*, vol. 9394. San Francisco, CA, USA, Feb. 2015, pp. 1–11.
[48] M. H. Pinson and S. Wolf, "A new standardized method for objectively measuring video quality," *IEEE Trans. Broadcast.*, vol. 50, no. 3, pp. 312–322, Sep. 2004.
[49] R. Soundararajan and A. C. Bovik, "Video quality assessment by reduced reference spatio-temporal entropic differencing," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 23, no. 4, pp. 684–694, Apr. 2013.
[50] Z. Li, A. Aaron, L. Katsavounidis, A. Moorthy, and M. Manohara. (Jun. 2016). *Toward a Practical Perceptual Video Quality Metric*. [Online]. Available: http://techblog.netflix.com/2016/06/toward-practical-perceptual-video.html
[51] A. Mittal, M. A. Saad, and A. C. Bovik, "A completely blind video integrity oracle," *IEEE Trans. Image Process.*, vol. 25, no. 1, pp. 289–300, Jan. 2016.
[52] VQEG. (Apr. 2000). *Final Report From the Video Quality Experts Group on the Validation of Objective Models of Video Quality Assessment*. [Online]. Available: http://www.vqeg.org/
[53] DASH Industry Forum. *DASH-IF Position Paper: Proposed QoE Media Metrics Standardization for Segmented Media Playback*. Accessed: Nov. 14, 2017. [Online]. Available: http://dashif.org/wp-content/uploads/2016/10/ProposedMediaMetricsforSegmentedMediaDelivery-r12.pdf
[54] W. Robitza, M.-N. Garcia, and A. Raake, "A modular HTTP adaptive streaming QoE model—Candidate for ITU-T P.1203 ('P. NATS')," in *Proc. IEEE Int. Conf. Qual. Multimedia Exp.*, Erfurt, Germany, 2017, pp. 1–6.
[55] C. G. Bampis, Z. Li, and A. C. Bovik, "Continuous prediction of streaming video QoE using dynamic networks," *IEEE Signal Process. Lett.*, vol. 24, no. 7, pp. 1083–1087, Jul. 2017.
[56] C. G. Bampis and A. C. Bovik, "Learning to predict streaming video QoE: Distortions, rebuffering and memory," *CoRR*, vol. abs/1703.00633, 2017. [Online]. Available: http://arxiv.org/abs/1703.00633
[57] H. Mao, R. Netravali, and M. Alizadeh, "Neural adaptive video streaming with pensieve," in *Proc. ACM SIGCOMM*, Aug. 2017, pp. 197–210.
[58] X. Liu *et al.*, "A case for a coordinated Internet video control plane," in *Proc. ACM SIGCOMM*, Helsinki, Finland, 2012, pp. 359–370.
[59] T. Hoßfeld, R. Schatz, E. Biersack, and L. Plissonneau, "Internet video delivery in YouTube: From traffic measurements to quality of experience," in *Data Traffic Monitoring and Analysis*. Heidelberg, Germany: Springer, Jan. 2013, pp. 264–301.
[60] H. J. Kim, D. G. Yun, H.-S. Kim, K. S. Cho, and S. G. Choi, "QoE assessment model for video streaming service using QoS parameters in wired-wireless network," in *Proc. IEEE Int. Conf. Adv. Commun. Technol.*, 2012, pp. 459–464.
[61] R. K. P. Mok, E. W. W. Chan, and R. K. C. Chang, "Measuring the quality of experience of HTTP video streaming," in *Proc. IFIP/IEEE Int. Symp. Integr. Netw. Manag.*, Dublin, Ireland, 2011, pp. 485–492.
[62] D. Z. Rodriguez, J. Abrahao, D. C. Begazo, R. L. Rosa, and G. Bressan, "Quality metric to assess video streaming service over TCP considering temporal location of pauses," *IEEE Trans. Consum. Electron.*, vol. 58, no. 3, pp. 985–992, Aug. 2012.
[63] J. Xue, D.-Q. Zhang, H. Yu, and C. W. Chen, "Assessing quality of experience for adaptive HTTP video streaming," in *Proc. IEEE Int. Conf. Multimedia Expo*, Chengdu, China, 2014, pp. 1–6.
[64] Z. Duanmu, A. Rehman, K. Zeng, and Z. Wang, "Quality-of-experience prediction for streaming video," in *Proc. IEEE Int. Conf. Multimedia Expo*, Seattle, WA, USA, Jul. 2016, pp. 1–6.
[65] J. Park, K. Seshadrinathan, S. Lee, and A. C. Bovik, "Video quality pooling adaptive to perceptual distortion severity," *IEEE Trans. Image Process.*, vol. 22, no. 2, pp. 610–620, Feb. 2013.
[66] D. Z. Rodríguez, R. L. Rosa, E. C. Alfaia, J. I. Abrahão, and G. Bressan, "Video quality metric for streaming service using DASH standard," *IEEE Trans. Broadcast.*, vol. 62, no. 3, pp. 628–639, Sep. 2016.
[67] H. R. Sheikh, M. F. Sabir, and A. C. Bovik, "A statistical evaluation of recent full reference image quality assessment algorithms," *IEEE Trans. Image Process.*, vol. 15, no. 11, pp. 3440–3451, Nov. 2006.
[68] D. C. Montgomery, *Applied Statistics and Probability for Engineers*, 6th ed. New York, NY, USA: Wiley, 2013.

Zhengfang Duanmu, photograph and biography not available at the time of publication.

Abdul Rehman, photograph and biography not available at the time of publication.

Zhou Wang, photograph and biography not available at the time of publication.