Study of the Effects of Stalling Events on the Quality of Experience of Mobile Streaming Videos

Deepti Ghadiyaram*, Alan C. Bovik*, Hojatollah Yeganeh†, Roman Kordasiewicz† and Michael Gallant†
*The University of Texas at Austin
†Avvasi, Waterloo, Canada

***Abstract*—We have created a new mobile video database that models distortions caused by network impairments. In particular, we simulate stalling events and startup delays in over-the-top (OTT) mobile streaming videos. We describe the way we simulated diverse stalling events to create a corpus of distorted videos and the human study we conducted to obtain subjective scores. We also analyzed the ratings to understand the impact of several factors that influence the quality of experience (QoE). To the best of our knowledge, ours is the most comprehensive and diverse study on the effects of stalling events on QoE. We are making the database publicly available [1] in order to help advance state-of-the-art research on user-centric mobile network planning and management.**

***Index Terms*—Mobile video quality, rebuffering, quality of experience, subjective quality assessment.**

I. Introduction

Video streaming now dominates global mobile data traffic and it is estimated that it will account for over 66% of wireless traffic by the end of 2017 [2, 3]. Given limited wireless network capacity, the quality of experience (QoE), which describes an end user's holistic perception and satisfaction with a given communication network service, has become an essential measure of network performance. Understanding QoE of the video streaming services has gained immense attention during the last few years due to the dramatic shift towards over-the-top (OTT) video streaming on mobile networks. The paucity of available bandwidth causes volatile network conditions, which invariably result in rebuffering events and playback interruptions, commonly referred to as *stalling* or *rebuffering*[1]. Stalling impairs QoE and leads to reduced consumer satisfaction and user churn.

A subjective study to thoroughly understand the specific factors regarding video stream quality that effect viewers' QoE can help researchers better understand how increases in network video quality affect viewer behavior, leading to design choices that make more efficient use of network resources. These studies are critical for designing reliable models for objective evaluation of QoE that account for stalling events in a way that is consistent with subjective human evaluation, regardless of video content or the type and strength of rebuffering. Such models have the potential to motivate the design of solutions for streaming video networks that strike a balance between reducing network operational costs while delivering the highest possible quality video content to customers.

[1]Throughout this paper, stalling and rebuffering are used interchangeably.

Several valuable video quality studies have been conducted in the past towards understanding the effects of network stream quality on QoE [4, 5] and the development of objective QoE models [6-8]. The focus in these studies has been to investigate the influence of simple factors such as startup delays, total stall length, and the number of random video stalls on an end user's QoE. These factors have later been used to design objective QoE prediction models. Certain general conclusions such as that longer startup delays are more annoying than shorter ones have been reported in these studies [4, 5].

Our goal is to thoroughly study the influence of diverse, realistic stalling patterns by varying several QoE influencing parameters such as the position, frequency, and length of the stalls, and the type of video content on end users' QoE, to support the design of generalizable QoE models for mobile videos. The methods used in previous studies do not adequately advance our goal as they suffer from one or more of the following problems: (1) small, insignificant database size, (2) insufficient number of subjective judgments, (3) unknown video sources with limited variability in content, (4) lack of public availability of the database, (5) lack of fine and continuous scale ratings and (6) use of large display formats during the study which do not translate to the smaller resolutions of mobile devices.

Here, we summarize a new large-scale database that we designed to overcome all of the above limitations. The resulting **LIVE-Avvasi Mobile Video database** consists of 180 distorted videos generated from 24 reference videos with 26 unique stalling events and 4830 human opinions obtained from 54 subjects who viewed the videos on mobile devices. We summarize certain key aspects of the database construction such as the reference video selection and the generation of realistic rebuffering patterns, followed by the details on how the human study was conducted. We then provide a summary analysis of the influence of several critical factors on QoE such as position, frequency, length of the stalls, and varied video content.

II. Details of the subjective study

A. Source sequences

We chose 24 High Definition (HD) creative commons licensed videos (that have audio included) from YouTube and Vimeo, 17 of which have a resolution of 1280×720 pixels and the rest 640×360 pixels. Any visual distortions due to aliasing were deemed minimal or invisible. In order to

Fig. 1. (Left) A sample stalled test sequence. (Right) Screenshot of the rating bar shown at the end of each presentation.

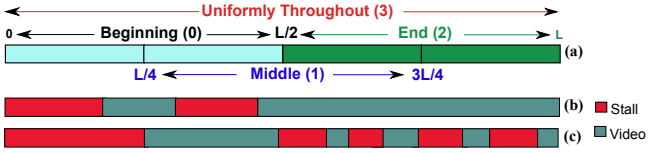

Fig. 2. Illustrating (a) different positions of stalls represented by a number between 0-3. (b) stall pattern **0_sfl** (short initial delay followed by few long stalls in the beginning.) (c) stall pattern **2_lmm** (long initial delay with many medium stalls towards the end.) as defined in Table II in any video sequence of length **L**.

focus exclusively on network impairments, we excluded any jittered or delayed videos, and thus all the 24 video sequences were selected to contain minimal spatial distortions and abrupt camera shakes. Moreover, the videos are semantically and temporally coherent and are long enough to be meaningful on their own. Their lengths (after adding rebuffering impairments) range between $29 - 134$ seconds. In order to understand the impact of video content type in the presence of network delays on QoE, we selected video sequences of varied categories that a typical video viewer is likely to encounter on the internet on a mobile device. All 24 videos were categorized into: sports (9), talk shows / documentaries (8), music (2), advertisements (3), and newscasts (2).

B. Distortion simulation

The studies in [6, 7] model the scenario where network congestion occurs at a constant rate, leading to periodic stalling patterns such that each stall is of a fixed duration. This, however is not realistic and thus, investigating the influence of arbitrarily occurring stall patterns on an end user's QoE is of significant interest. Consequently, the goal of our study was to develop a database of videos containing diverse, realistic stall patterns that will challenge automatic VQA algorithms that predict QoE scores across varying stall types and video content. Table I provides information about the parameters that we varied for each stalling pattern while Table II provides a brief overview of the diverse stall patterns we designed. Fig. 2 illustrates the different positions in a video sequence where the stalls were placed and a few sample stall patterns.

Each reference video was preprocessed to realistically introduce these impairments, thus generating 180 distorted videos. To be able to gather a reasonable number of subjective opinions on all the distorted videos without prohibitively increasing the study duration, we divided them into two groups A and B, each containing 90 video sequences. The 2 hour study session per group was divided to include as many videos as possible by

TABLE I
SUMMARY OF STALL FREQUENCY AND LENGTHS (IN SECONDS)

Number of stalls	Few (1-3)	Many (4-7)	
Stall length	Short (2-4)	Medium (5-9)	Long (10-15)
Initial delay	Short (0-7)	Long (8-20)	

TABLE II
SUMMARY OF DIFFERENT SIMULATED STALL PATTERNS. THE PREFIX **X** REFERS TO THE POSITION WHERE THE PATTERN IS INTRODUCED AND TAKES A VALUE 0-3 AS DEFINED IN FIG. 2(A).

Stalling patterns	**# of videos per pattern**
Only short initial delays (**shortInitial**)	5
Only long initial delays (**longInitial**)	4
Short initial + **f**ew **m**edium (**x_sfm**)	**B**(6), **M**(6), **E**(8)
Short initial + **f**ew **l**ong (**x_sfl**)	**B**(6), **M**(6), **E**(6)
Short initial + **m**any **m**edium (**x_smm**)	**B**(4), **E**(4), **U**(6)
Short initial + **m**any **s**hort (**x_sms**)	**B**(6), **E**(6), **U**(6)
Long initial + **f**ew **m**edium (**x_lfm**)	**B**(6), **M**(6), **E**(6)
Long initial + **f**ew **l**ong (**x_lfl**)	**B**(6), **M**(6), **E**(4)
Long initial + **m**any **m**edium (**x_lmm**)	**B**(4), **E**(4), **U**(5)
Long initial + **m**any **s**hort (**x_lms**)	**B**(6), **E**(6), **U**(4)

manually adjusting the distribution of the stall patterns (Table II) amongst the reference videos so that all of the stall events are reasonably represented. At least 2 and at most 4 sequences of any given distortion were included in each study group. Each group's playlist was further randomly divided into 6 sets each of 20 minute duration. To avoid contextual and memory effects in the judgment of a user's viewing experience, these sets were prepared to not contain any two adjacent sequences generated from the same reference video.

C. Subjective Testing Methodology

We adopted a single stimulus continuous quality evaluation procedure (SSCQE) [9] to obtain quality ratings on video sequences where the subjects rated their viewing experience on a continuous scale ranging between **worst** (1) to **best** (5). 27 subjects were uniformly assigned to each group at random and the study was conducted in parallel for both groups on two different mobile devices. To minimize the effects of viewer fatigue, we divided the 2 hour study into three 40 minute sessions and a subject viewed and rated videos from 2 of the 6 sets in each session.The subject also viewed each of the reference videos once at an unknown and random point in the presentation to facilitate computation of **difference mean opinion scores (DMOS)** to facilitate a procedure known as hidden reference removal [10, 11].

All the subjects viewed the videos on an Apple iPhone 5 which ran a web application that communicated over WiFi with a web server that was hosted *in-house* to fetch and display the test sequences. The study setup was thoroughly tested to ensure that no external impairments were introduced by the WiFi while video streaming. The user interface of the application was designed to keep the video viewing experience close to that of the real YouTube mobile website, assuming the absence of any bitrate adaptations to focus exclusively on

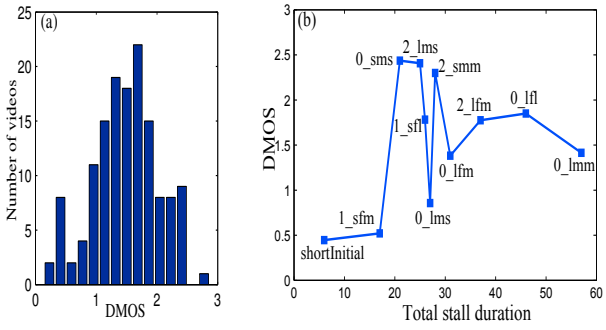

Fig. 3. (a) Histogram of DMOS for all video sequences from our database. (b) DMOS of stall patterns added to a specific reference video vs. total stall duration (i.e., sum of all the intermittent stalls). It is critical to note that DMOS value of 2_lms is more than 2_lfm's (also 2_smm vs. 0_lfm) despite their respective total stall lengths.

TABLE III
A PATTERN **P_INL** DEFINED IN FIGURES 3 (B), 4, AND 5. THE STALL POSITIONS ARE DEFINED WITH REGARDS TO THE TOTAL LENGTH OF EACH VIDEO.

p (stall position)	beginning (0), middle (1), end (2), uniformly throughout (3)
i (initial delay)	**s**hort (s), **l**arge (l)
n (number of stalls)	**f**ew (f), **m**any (m)
l (stall length)	**s**hort (s), **m**edium (m), **l**ong (l)

stalling.

The subjects were instructed to hold the phone in landscape mode at a comfortable viewing distance and angle. At the end of the presentation of each video, a rating screen was displayed (Fig. 1) on the screen with a continuous scale slider. The cursor was set at its center to avoid biasing the subject's perception of quality. The subject could control the slider using the touch screen and after a rating was entered, pressing the *Play Next Video* button on the rating screen presented the next test sequence. The web application automatically logs the submitted quality scores into a database.

The study was conducted at The University of Texas at Austin over a period of two weeks and the subject pool mostly consisted of undergraduate and graduate students from different disciplines. No vision tests were performed but the subjects were asked to wear corrective lenses during the study if they did so normally. Each subject was provided with instructions during all three sessions and had a three minute training phase during their first session, where they viewed two videos picked from the pool of test sequences that were used exclusively for subject training.

The subject rejection procedure in [9] was used to reject 1 subject from group B while the remaining scores were averaged and a DMOS for each video was computed [11]. DMOS is representative of the *perceived viewing experience* of each video whose values ranged between $[0, 2.43]$ and $[0, 2.87]$ with a standard error of the sample mean equal to 0.07 and 0.08 for playlists A and B respectively. Fig. 3 (a) shows the distribution of the DMOS values from both groups. A higher DMOS value indicates bad QoE.

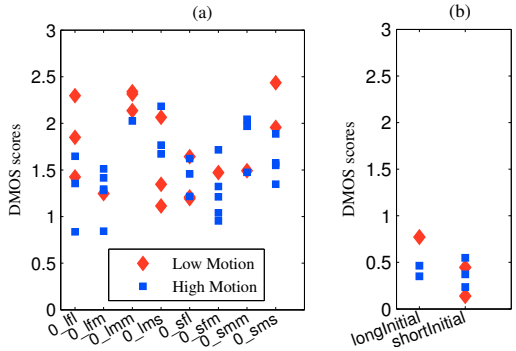

Fig. 4. (a) Few stall patterns that occur in the first half of the test sequences having varied spatio-temporal activity. (b) DMOS of test sequences containing only initial delays.

III. EVALUATION OF SUBJECTIVE OPINION

We now describe our method of categorizing the videos and our analysis of the scores obtained.

Global spatio-temporal motion categories: Motion information plays a significant role in the perception of video quality [12] and QoE. Thus, we characterized each reference video's spatial-temporal content at each pixel (i, j) in the form of horizontal and vertical motion vectors computed between consecutive frames using Horn-Schunck's optical flow model [13]. The mode (m_k) of the magnitudes of these motion vectors at each pixel (i, j) was computed for every pair of consecutive frames. The mean $(\bar{\mu})$ of these mode values computed across the entire video sequence was found to be a good indicator of the magnitude of global motion in [12]. We followed the same approach and observed that the mean magnitudes of global motion $(\bar{\mu})$ of the reference videos could be clustered into two groups - one with lower mean values and the other with higher. If μ_{low} is the mean of all the lower mean values and μ_{high} that of all the higher ones, we computed $\mu_{threshold}$ to be the midpoint between μ_{low} and μ_{high}. We categorized 13 videos into the **high motion** category since their mean values $\bar{\mu}$ were greater than $\mu_{threshold}$ and the remaining 11 videos into the **low motion** category.

We now present our analysis of the combined scores from both groups to understand the impact of several QoE influencing factors such as varied severities of stall events, their frequency and positions, and their interplay with the spatial-temporal motion in a given video sequence. Fig. 4(a) shows the impact of the stalls we simulated to occur at the beginning of few video sequences (prefix **0**_), along with the underlying test sequences' motion categories.

Motion category: From Fig. 4(a), it is evident that most stalling patterns have had an impact on the user experience to a degree that appears to be independent of a sequence's motion category. This was observed to hold true for all of the other stall patterns as well.

Position of the stalls: A closer look at the DMOS of test sequences with specific stall patterns simulated at different positions is shown in Fig. 5(a). This plot indicates that the subjects exhibited a hysteresis (recency) "after-effect" [14], thus a previous unpleasant viewing experience caused by stall events in the first half (prefix **0** as defined in Fig. 2) or the

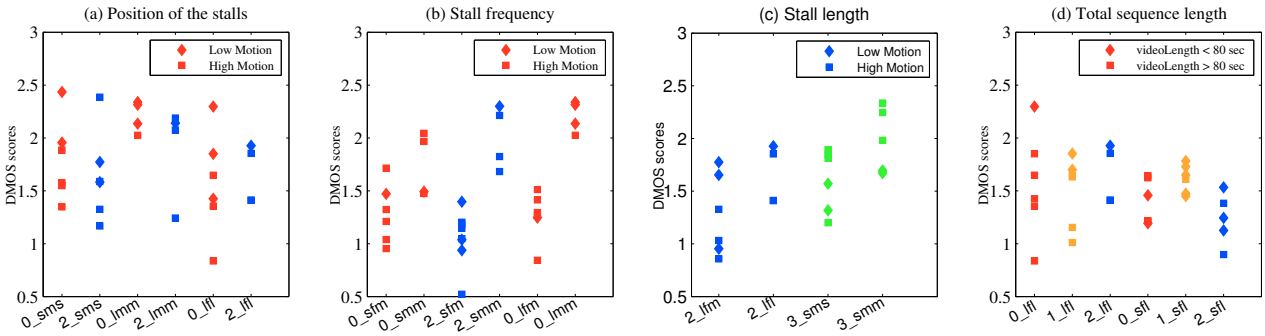

Fig. 5. DMOS of a few stall patterns illustrating the impact of (a) position of the stalls, (b) stall frequency, (c) stall length, and (d) total duration of a video sequence on QoE. Note that frequency and stall length have a significant impact on QoE.

middle of the video penalized their overall perceived quality. However, looking closely at Fig. 3(b), a general conclusion can be made that the *same stall pattern* towards the end of the video lends to higher DMOS than at the beginning of the video (**0**_lms vs **2**_lms).

Stall Frequency: Comparing the DMOS of videos with the *same stall patterns* but with varied frequency of occurrence (eg: 2_s**f**m vs. 2_s**m**m i.e., few medium vs. many medium length stalls), videos with many stall events had much higher DMOS than those with fewer stalls (Fig. 5(b)).

Stall length: When the *stall pattern* and *frequency* were maintained and only the stall length was varied, videos with shorter stalls had lower DMOS than those with longer stalls, irrespective of their position of occurrence (eg: 3_sm**m** vs. 3_sm**s** i.e., many medium vs. many short stalls in Fig. 5(c)).

Length of initial delay: To understand the impact of the length of the starting delay on QoE, we compared the DMOS of videos with just short and long initial delays (delay lengths are defined in Table I). From Fig. 4(b) (**long**Initial vs. **short**Initial), it can be observed that subjects were not too frustrated with the inital delay and thus the opinion scores did not seem to vary greatly around this factor. However, a more focussed study with more videos having initial delays of several different (preferably longer) lengths will help us better understand the impact of this factor on QoE.

Total sequence length: To understand the impact of the total video length on QoE, we compared the DMOS of sequences that varied in their total length but were distorted with the *same stall events* (Fig. 5(d)). It was observed that shorter videos had a slightly more serious influence on end users' QoE.

Total stall length: In Fig. 3 (b), we summed the individual intermittent stall lengths of different patterns added to a specific reference video and observed the corresponding DMOS values. It is significant to observe that more than the total stall duration, the interplay of stall position, frequency, and stall length influences QoE the most. Thus, **2_smm** has a higher DMOS than **0_lfm**, despite having a lesser total stall length.

To summarize, the frequency of occurrence of stalls, their position, and their lengths have a more serious impact on an end user's QoE when compared with factors such as motion information and total video length. Longer, more frequent stalls have a significant impact on a user's QoE, irrespective of their position of occurrence. Stalls towards the end of the video usually lead to poor QoE. That more frequent stalls lead to poor QoE has already been reported earlier [8] but the finding was based on a subjective study conducted using only a single video. Hopfeld *et. al* [6] also observed the same effect in a study they conducted by varying the number and length of the stalls. Our large-scale study reveals that this influence persists even in the presence of diverse video content containing realistic, arbitrarily occuring stall patterns. However, a more focussed study to understand the influence of the position of the stalls on QoE can be conducted where longer test sequences are used such that the *hysteresis* effect is reduced, and stalls occurring in the first half have a lesser impact on the overall QoE. Other factors such as the effect when a user likes (or does not like) the content in a video with stalls, or the effect of coincided spatial artifacts in a video along with stalls are possible topics for future work.

IV. CONCLUSIONS AND FUTURE WORK

We presented preliminary results and analysis from a recently conducted large-scale subjective study to evaluate the effects of network impairments on end users' viewing experience in OTT mobile video streaming. The **LIVE-Avvasi Mobile VQA database** includes 180 videos derived from 24 reference videos by incorporating a wide variety of stall events, along with the associated opinion scores from 54 subjects on their viewing experience. Furthermore, we presented a simple statistical analysis of the ratings to understand the impact of several factors on QoE. We intend to make use of this data to understand the influence of dynamic network impairments such as stalling events on QoE. This will help us design perceptually-aware video assessment algorithms that will be able to better predict QoE, and which will in turn be useful in designing solutions for resource allocation and rate adaptation of OTT video streaming. We believe our publicly available database is the first of its kind to include such diverse stalling events and a large number of subjective scores and videos [1]. We hope that our database will serve as a valuable resource to enable the efforts of the research community to build better QoE predictive models and to benchmark their performance to advance the state-of-the-art in user-centric mobile network planning and management.

REFERENCES

[1] (2014) LIVE-Avvasi Mobile Video Quality Database. [Online]. Available: http://live.ece.utexas.edu /research/quality/live-avvasi-mobilevideo.html

[2] CISCO Corp, Cisco Visual Networking Index: Global Mobile Data Traffic Forecast Update, 2012-2017. [Online] Available: http://www.cisco.com/en/US/solutions/collateral/ns341/ns525/ns537/ns705/ns827/white_paper_c11-520862.html

[3] A.C. Bovik, "Automatic prediction of perceptual image and video quality," *IEEE Proc.*, vol. 101, no. 9, pp. 2008-2024, Sept. 2013.

[4] F. Dobrian, V. Sekar, A. Awan, I. Stoica, D. Joseph, A. Ganjam, and H. Zhang, "Understanding the impact of video quality on user engagement," *ACM SIGCOMM Comp. Commun. Review*, vol. 41, no. 4, pp 362-373, 2011.

[5] S.S. Krishnan and R.K. Sitaraman,"Video stream quality impacts viewer behavior: inferring causality using quasi-experimental designs," *Proc. of ACM Conf. on Internet measurement*, pp. 211-224, 2012.

[6] T. Hossfeld, M. Seufert, M. Hirth, T. Zinner, P. Tran-Gia and R. Schatz, "Quantification of YouTube QoE via Crowdsourcing," *IEEE Int. Sym. on Multimedia (ISM)*, pp. 494-499, 2011.

[7] 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," *Fourth Int. Workshop on Quality of Multimedia Exp. (QoMEX)*, pp. 1-6, 2012.

[8] R.K.P. Mok, E.W.W. Chan, and R.K.C. Chang, "Measuring the quality of experience of HTTP video streaming," *IFIP/IEEE Int. Sym. on Integrated Network Management (IM)*, pp. 485-492, 2011.

[9] (2000) ITU-R Recommendation BT.500-11, Methodology for the subjective assessment of the quality of television pictures, International Telecommunications Union, Tech. Rep.

[10] M.H. Pinson and S. Wolf, "A new standardized method for objectively measuring video quality," *IEEE Trans. on Broadcasting*, vol. 50, no. 3, pp. 312-322, 2004.

[11] K. Seshadrinathan, R. Soundararajan, A.C. Bovik, and L.K. Cormack, "Study of Subjective and Objective Quality Assessment of Video," *IEEE Trans. Image Process.*, vol.19, no. 6, pp. 1427-1441, 2010.

[12] M. Saad, A.C. Bovik, and C. Charrier, "Blind Prediction of Natural Video Quality," *IEEE Trans. on Image Proc.*, vol. PP no. 99, 2014.

[13] B.K. Horn and B.G. Schunck, "Determining optical flow," *Tech. Symp. East. Int. Society for Optics and Photonics*, pp. 319-331, 1981.

[14] K. Seshadrinathan and A.C. Bovik, "Temporal hysteresis model of time varying subjective video quality," *IEEE Int. Conf. on Acoustics, Speech and Signal Process. (ICASSP)*, pp 1153-1156, 2011.