IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received April 30, 2019, accepted May 8, 2019, date of publication May 10, 2019, date of current version May 24, 2019.

Digital Object Identifier 10.1109/ACCESS.2019.2916192

Initial Perceived Quality Evaluation for Video Streaming Services

JIARUN SONG[1], FUZHENG YANG[1,2], (Member, IEEE), AND WEI ZHANG[1], (Member, IEEE)
[1]State Key Laboratory of Integrated Services Networks, Xidian University, Xi'an 710071, China
[2]School of Electrical and Computer Engineering, Royal Melbourne Institute of Technology, Melbourne, VIC 3001, Australia

Corresponding author: Wei Zhang (wzhang@xidian.edu.cn)

This work was supported in part by the National Science Foundation of China under Grant 61601349, Grant 61601348, and Grant 61801364, and in part by the 111 Project under Grant B08038.

ABSTRACT Initial perceived quality (IPQ) reflects users' perception when waiting for videos to play, which directly determines the user's willingness to go on watching videos. However, IPQ has never been carefully analyzed and evaluated. In this paper, two types of subjective test methods were elaborately designed, where the IPQ scores were obtained. The accuracy of IPQ scores was checked according to the consistency of the rating results. Then, an objective evaluation model was proposed accordingly to evaluate IPQ. Meanwhile, the relationship between IPQ and users' unacceptability of initial loading delay was also investigated according to users' abandonment behavior during the waiting period. A probability model was proposed to evaluate the unacceptability of users using logistic regression analysis. The experimental results show that the proposed models can accurately evaluate IPQ and users' unacceptability of the initial loading delay. The proposed models can be used as guidelines for service providers to effectively improve the buffering strategy for video streaming services.

INDEX TERMS Initial perceived quality, video streaming, initial loading delay, user unacceptability.

I. INTRODUCTION

Recent years have witnessed the rapid growth of video streaming services and applications. According to Cisco's visual networking forecast, Internet video traffic would grow fourfold from 2016 to 2021, with a million minutes of video content crossing the network every second by 2021 [1]. Video streaming services have been widely used in various applications including video on demand, network video, and social media. The success of these video services largely depends on whether users' satisfaction can be guaranteed. Therefore, it is crucial for service providers to predict user's quality of experience (QoE) for service improvement and network optimizing [2]–[5].

Generally, in order to ensure the QoE of users for video streaming services, the service providers should provide high quality videos during the viewing period [6]–[15]. Meanwhile, it is also necessary to let users wait as short as possible when they click the play button for video display [16]–[20]. Particularly, with the popularity of sharing short user-generated video content, an increasing number of users have switched from the traditional video on demand services (e.g., films) to short video services (i.e., Facebook). They tend to spend less time on waiting for videos to play [21], [22]. In this circumstance, if the waiting time exceeds user's tolerance, they may give up watching the video immediately and choose alternative videos, no matter how good the video quality is. In this regard, compared with existing QoE studies which focused on the quality perception during viewing period, research towards the initial perception during the waiting time largely determines users' willingness to select the video services. Regrettably, it is still unclear how the initial loading delay affects users' initial perception. It is also unclear how much time can be spent on the initial loading of mobile videos before it becomes unacceptable for users. In order to guarantee the retention in video services, it is necessary for service providers to better understand the initial perceived quality (IPQ), namely, the perception corresponding to the duration from clicking the ''play'' button to the video display initiation.

Related to the influence of initial loading delay, a number of studies have been investigated in recent years [23]–[26]. Since the dynamic adaptive streaming over HTTP (DASH) videos has become a popular solution for most of the video

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

sharing websites and video streaming providers, such as YouTube and Netflix, the authors in [23] investigated user experience for DASH videos and developed a quantifiable measure of user experience. It was found that there was a linear relationship between the subjective impairment value and the initial loading delay. Rodriguez *et al.* investigated the initial loading delay on users QoE by a subjective test, which indicated that the relationship between QoE and the initial loading delay can be approximated to an exponential function [24]. The authors in [25] compared the influence of initial loading delay and interruption (stalling) during watching using the example of YouTube video streaming. Meanwhile, they also quantified the impact of initial delays on the user perceived QoE for different application scenarios by means of subjective laboratory and crowdsourcing studies.

However, it should be noted that all these above studies focused on the overall perceived quality in associate with other influential factors, where the impact of the initial loading delay is only considered as an influential factor of the overall perceived quality. How to accurately evaluate IPQ is still unclear. With this in mind, ITU-T Study Group 12 drafted a new standardized Recommendation named P.QUITS [27]. Following the mandates of P.QUITS, IPQ was investigated in this paper.

To accurately evaluate IPQ, two main aspects should be covered, i.e., how to obtain the subjective IPQ and how to objectively model it. Targeting the first aspect, we elaborately designed two types of subjective test methods where the IPQ scores were rated at different stages of viewing period in the two methods and the accuracy of the acquired IPQ was checked according to the consistency of the rating results. Then, an objective evaluation model was proposed to predict IPQ. Meanwhile, the relationship between IPQ and user unacceptability of initial loading delay was also analyzed according to the abandonment behavior during waiting period. A probability estimation model was then proposed using the logistic regression analysis. The contributions of this work include: (1) a subjective implementation to accurately obtain IPQ; (2) a method for service providers to accurately evaluate IPQ, aiming at effectively improving the buffering strategy; (3) a guideline to clarify user unacceptability of initial loading delay for video streaming services.

The remainder of this paper is organized as follows. Section II describes the experiment design, including test platform establishing, experimental settings and procedures. In Section III, IPQ is analyzed and evaluated according to the subjective test results and the relationship between IPQ and user unacceptability of initial loading delay is analyzed. The performance of the proposed models and the conclusion are given in Section IV and V, respectively.

II. DESIGN OF EXPERIMENT

The main purposes of this experiment are to investigate user perception on the video quality during the initial loading period and clarify user (un)acceptability of the initial loading delay. In order to study IPQ for video streaming services, we elaborately designed two methods to obtain the subjective opinion of users, where the quality was rated at different stages of viewing process. The IPQ scores derived from the more reasonable method were employed for IPQ analysis. For user (un)acceptability of the initial loading delay, it refers to a binary measure to locate the threshold of minimum acceptable quality that fulfills user quality expectations and needs for a certain application or system [28]. The psychological research indicates that human's internal states can be evaluated using human behaviors on the basis of drive and incentive theory [29]. Therefore, we recorded user abandonment behavior during the initial waiting period to characterize user acceptability of the initial loading delay. The details of experiments are described as follows:

A. EXPERIMENTAL MATERIALS

In the subjective test, we employed 126 audiovisual sequences that are downloaded from YouTube, including news, sports, entertainment, advertisement, comedy, and landscapes. For each type of content, there were 21 audiovisual clips. The video resolution included five levels, namely the 320P, 480P, 720P, 1080P, and 1440P. These sequences were divided into Group 1 for IPQ analysis and model training as well as Group 2 for performance validation. More specifically, 84 audiovisual sequences are in Group 1. The durations of these sequences are 10s, 60s, and 180s. The other 42 audiovisual sequences are in Group 2, where the sequence durations are 20s, 40s, and 120s, respectively.

For each group, the initial loading delay has 14 different lengths, namely, 0.1s, 0.2s, 0.3s, 0.5s, 0.7s, 1s, 2s, 4s, 6s, 8s, 10s, 15s, 20s, and 30s. The video resolution, video duration, and initial loading delay for each sequence were randomly assigned. Detailed parameter settings for the Group 1 and Group 2 are listed in Table 1 and Table 2, respectively. Therefore, there were 84 audiovisual sequences in Group 1 and each parameter setting corresponds to two audiovisual sequences. In Group 2, there were 42 audiovisual sequences and each parameter setting corresponds to one sequence.

B. TEST PLATFORM

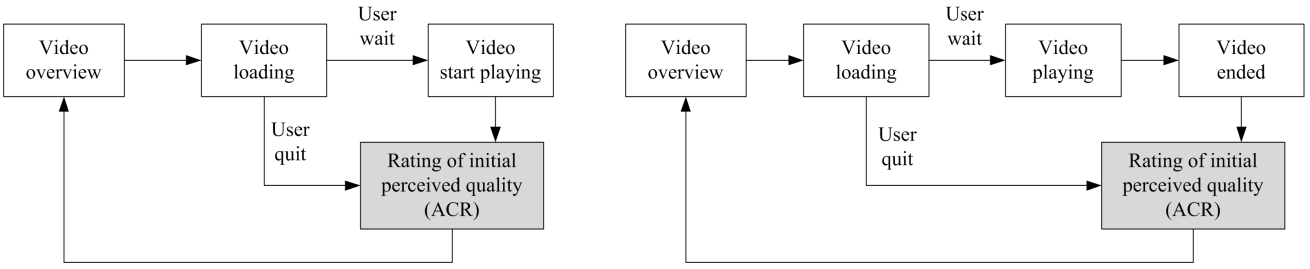
In order to obtain IPQ, a subjective test platform based on the android system was developed in this work. The test platform simulates the YouTube application on the mobile terminal to provide users the real-world viewing experience. The main interfaces are shown in Fig. 1(a) and Fig. 1(b), respectively. The test platform was able to accurately control the initial loading buffer length to provide the video services with different initial loading delays. Users can rate the IPQ score on a pop-up window which is illustrated in Fig. 1(c). Meanwhile, the platform also records the abandonment behavior if users close the video sequence before the video starting to play.

C. TEST PROCEDURE

Two subjective rating methods were designed to obtain the IPQ scores, as shown in Fig. 2. Considering user's IPQ was

(a) (b) (c)

FIGURE 1. Screenshots of test platform. (a) Main interface (b) Play interface (c) Rating interface.



(a) (b)

FIGURE 2. Procedures of two test methods. (a) Rating IPQ immediately when the video start playing (b) Rating at the end of the video.

generated at the beginning of the video service, the first test method (Method 1) was designed to rate the IPQ score immediately when the video started playing, as shown in Fig. 2(a). Users could choose to quit the video in advance during the loading period if they were tired of waiting for a long time to play. They were also allowed to go on watching the video after rating the IPQ score if they were interested in the video content. The second method (Method 2) was designed with consideration of user's rating habit, where users evaluated the IPQ score at the end of each video, as illustrated in Fig. 2(b). Users were also allowed to quit the video during the whole watching process. Both the rating methods were designed based on a single stimulus procedure according to absolute category rating (ACR) [30], [31], and a 5-point rating scale was used to obtain the IPQ score [32].

Twenty non-expert subjects were invited to perform the two subjective rating tests using the sequences of Group 1. All the subjects were university students in different grades and they were screened for visual acuity and color blindness. The subjective tests were carried out following the guide-lines specified by VQEG [31], including the selection of test method, presentation of the test material and determination of grading scales, etc. The test environment was set following the instructions defined in [30], as illustrated in Fig. 3(a). All the videos played on the mobile phones, where the resolution of the screen was 2560 × 1440 and the screen size was 5.1 inch. The mobile device was illustrated in Fig. 3(b). Before the formal test, the subjects were asked to watch five examples to get familiar with the operation of test platform and the rating procedures. The distance

(a) (b)

FIGURE 3. Test environment and devices (a) test environment (b) test devices.

between the subjects and screen was set to be 3H (H is the height of the screen). To avoid fatigue, each test did not exceed 30 minutes and there was a 20 minutes break between two tests. Finally, we obtained 1680 IPQ rating samples and abandonment behavior records, respectively. The IPQ value was measured in terms of Mean Opinion Score (MOS) and 84 MOS values were obtained for IPQ analysis and model training. These MOS values constitute the training dataset TR.

According to the results analysis of the Group 1, we then chose one of the more reasonable methods to conduct the corresponding test using the sequences of Group 2. Twenty non-expert subjects invited to perform the subjective test. The procedure was the same with that in the training test. Finally, a total of 840 rating samples were obtained and 42 MOS values were acquired for validation, which constituted the validation dataset VL.

TABLE 1. Distribution of parameter settings for group 1.

ID	Initial delay(s)	Resolution	Video duration(s)	ID	Initial delay(s)	Resolution	Video duration(s)	ID	Initial delay(s)	Resolution	Video duration(s)
1	0.1	320P	10	15	0.1	320P	60	29	0.1	1080P	180
2	0.2	480P	10	16	0.2	480P	60	30	0.2	1440P	180
3	0.3	720P	10	17	0.3	720P	60	31	0.3	320P	180
4	0.5	1080P	10	18	0.5	1080P	60	32	0.5	480P	180
5	0.7	1440P	10	19	0.7	1440P	60	33	0.7	720P	180
6	1	320P	10	20	1	320P	60	34	1	1080P	180
7	2	480P	10	21	2	480P	60	35	2	1440P	180
8	4	720P	10	22	4	720P	60	36	4	320P	180
9	6	1080P	10	23	6	1080P	60	37	6	480P	180
10	8	1440P	10	24	8	1440P	60	38	8	720P	180
11	10	320P	10	25	10	320P	60	39	10	1080P	180
12	15	480P	10	26	15	480P	60	40	15	1440P	180
13	20	720P	10	27	20	720P	60	41	20	320P	180
14	30	1080P	10	28	30	1080P	60	42	30	480P	180

TABLE 2. Distribution of parameter settings for group 2.

ID	Initial delay(s)	Resolution	Video duration(s)	ID	Initial delay(s)	Resolution	Video duration(s)	ID	Initial delay(s)	Resolution	Video duration(s)
1	0.1	1440P	20	15	0.1	720P	40	29	0.1	1080P	120
2	0.2	320P	20	16	0.2	1080P	40	30	0.2	1440P	120
3	0.3	480P	20	17	0.3	1440P	40	31	0.3	320P	120
4	0.5	720P	20	18	0.5	320P	40	32	0.5	480P	120
5	0.7	1080P	20	19	0.7	480P	40	33	0.7	720P	120
6	1	1440P	20	20	1	720P	40	34	1	1080P	120
7	2	320P	20	21	2	1080P	40	35	2	1440P	120
8	4	480P	20	22	4	1440P	40	36	4	320P	120
9	6	720P	20	23	6	320P	40	37	6	480P	120
10	8	1080P	20	24	8	480P	40	38	8	720P	120
11	10	1440P	20	25	10	7200P	40	39	10	1080P	120
12	15	320P	20	26	15	1080P	40	40	15	1440P	120
13	20	480P	20	27	20	1440P	40	41	20	320P	120

III. ANALYSIS OF TEST RESULTS

In this section, the accuracy of IPQ scores obtained by these two test methods was compared using the TR dataset. The scores corresponding to the more reasonable method were chosen as the IPQ groundtruth. To better understand the IPQ characteristics, the distribution of user rating scores was further analyzed. Finally an objective evaluation model for IPQ was proposed. Moreover, in order to clarify how much time could be spent on the initial loading of video before it becomes unacceptable, we further analyzed the relationship between IPQ and user unacceptability of initial loading delay.

A. IPQ GROUNDTRUTH DETERMINATION

Obtaining accurate quality groundtruth is the basis for studying IPQ. As IPQ is rarely studied in the literature, how to obtain the groundtruth is still an open issue. Here, two types of subjective test methods were elaborately designed to obtain the IPQ, where the IPQ scores were rated either

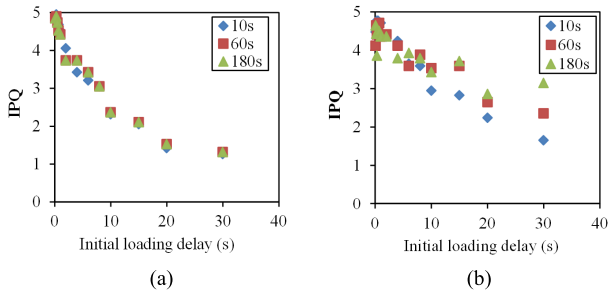


FIGURE 4. Relationship between IPQ and initial loading delay for different subjective test methods (a) Method 1, (b) Method 2.

at the beginning or the end of video playback, respectively. The accuracy of two types of IPQ results was then evaluated according to the consistency of the rating results. We chose the more reasonable results between the two methods as the groundtruth for further analysis.

Fig. 4 illustrates the relationships between the initial loading delay and IPQ obtained by the two subjective rating methods under different video durations. It can be found that IPQ values are gradually decreasing with the increase of initial loading delay for both methods. However, the trends between initial loading delay and IPQ obtained by Method 1 are almost the same under different video durations, as illustrated in Fig. 4(a). The IPQ scores obtained by this method seems only related to the initial loading delay, where the Pearson Correlation Coefficients (PCC) and the Spearman Rank Order Correlation Coefficients (SROCC) between IPQ and initial loading delay are -0.985 and -0.983 respectively. In contrast, the tendencies between the initial loading delay and IPQ obtained by Method 2 are variable with different video durations, as illustrated in Fig. 4(b). The IPQ scores obtained in this way are related to multiple influential factors, such as the initial loading delay and video duration. The underlying reason is that for Method 1, the subjects rate the IPQ scores immediately when videos start playing, and therefore the recorded IPQ is closer to the user's true initial perception. For Method 2, the subjects rate the IPQ scores at the end of video. In this case, IPQ is no longer the user's true initial feelings, which is affected by the viewing experience. Here, we choose the results obtained by Method 1 as the IPQ groundtruth for further analysis.

B. DISTRIBUTION OF IPQ RATING SCORES

In order to better clarify the characteristics of IPQ, we further analyzed the distribution of IPQ rating scores under different initial loading delays, as listed in Table 3. It can be found that when the initial loading delay is not larger than 4 seconds, more than 90% of rating scores are 4 and 5 point, and only 0.21% of rating scores are 2. Most users can obtain a favorable IPQ in such a case. When the initial loading delay is ranged from 6 seconds to 8 seconds, 11.67%, 55.83%, and 32.50% of scores are rated as 2, 3, and 4, respectively. Since over 80% of scores are rated as 3 and 4, the initial loading delay seems still acceptable. When the initial loading delay is 10s, nearly 58.33% of scores are rated as 2 and 41.67 % are rated as 3. As the initial loading delay grows to 15 seconds, most of scores are rated as 2 and only 15% of scores are rated as 3. IPQ obviously becomes worse than those with lower initial loading delay. When the initial loading delay is as high as 20 seconds, all the rating scores are rated as 1 and 2. The percentage of the scores that rated as 1 is 61.67%, which continues to increase when the initial loading delay reaches up to 30s.

TABLE 3. Distribution of IPQ rating scores.

Initial loading delay	Percentage of user rating scores				
	1	2	3	4	5
0-4s	0	0.21%	9.58%	28.13%	62.08%
6-10s	0	11.67%	55.83%	32.50%	0
10s	0	58.33%	41.67%	0	0
15s	5.00%	80.00%	15.00%	0	0
20s	61.67%	38.33%	0	0	0
30s	71.67%	28.33%	0	0	0

Fig. 5 visualizes the distribution of the percentage of rating scores under different initial loading delays. Each initial loading delay corresponds to a bar. For each bar, there are several sub-bars with different colors that denote different rating scores. The height of each sub-bar indicates the percentage of the rating scores at a certain value. It clearly shows that higher initial loading delay corresponds to lower initial perceived quality. When the initial loading delay is larger than 10 seconds, IPQ becomes considerably poor. Therefore, in order to provide an excellent video service, it is suggested that the initial loading delay should not exceed 4 seconds. However, if the conditions are not guaranteed, we can try to control the initial loading delay within 10s, where most of users can still get an acceptable quality.

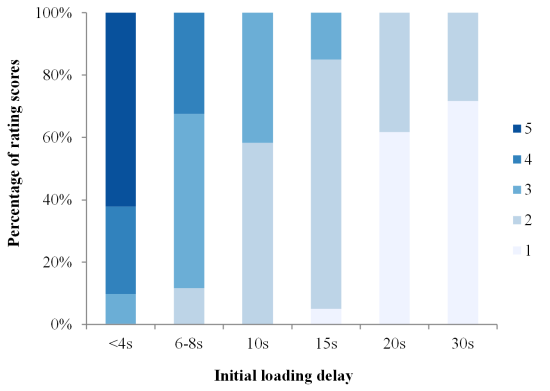


FIGURE 5. Distribution of the percentage of rating scores under different initial loading delays.

C. IPQ MODELLING

Considering IPQ determines user's willingness to watch videos, we then established an objective evaluation model for IPQ. The model can be used as a reference for service providers to optimize the initial buffering strategy. According to the data in Fig. 4, it is obvious that IPQs are gradually decreasing with the increase of initial loading delay under different video durations. However, the downward trend of the curve gradually slows down, especially when the delay is more than 20s. The relationship between IPQ and initial delay can be fitted by a negative exponential function as:

$$Q_{IP} = a \cdot \exp(-b \cdot V_{IDL}) + c \quad (1)$$

where Q_{IP} is the initial perceived quality, V_{IDL} is the initial loading delay. According to the data in Fig. 4, parameters a, b, c can be obtained by the least square fitting method using MATLAB R2015b [33]. The values of a, b, c under different video durations are listed in Table 4. It can be found that the values of a, b, c under different video durations are also very similar, respectively. Therefore, the values of a, b, c can be calculated using the average values under different video durations, namely 3.956, 0.093, and 1.025, respectively.

TABLE 4. Values of a, b, c under different video durations.

Video duration	a	b	c
10s	3.959	0.093	1.013
60s	3.981	0.091	1.034
180s	3.927	0.094	1.027
Average	3.956	0.093	1.025

D. DISTRIBUTION OF ACCEPTABILITY FOR INITIAL LOADING DELAY

Generally, video service providers and content providers expect less viewer abandonment, more viewer engagement, and higher audience retention. Besides understanding user's IPQ during waiting period, clarifying user acceptability of initial loading delay is also desired. Since users are not informed about the video source quality and content type during the initial waiting period, here we employ user abandonment behavior during the initial loading period to represent user unacceptability of the initial loading delay. If a user quit the video during waiting period, this video session is evaluated as ''unacceptable''. Conversely, the video session is evaluated as ''acceptability'' if users do not quit. Table 5 lists the percentage of user unacceptability under different initial loading delays. It can be found that the user unacceptability rate gradually increases along with the rise of the initial loading delay. When the initial loading delay is smaller than 4 seconds, the IPQ rating scores are ranged from 3 to 5 (MOS $>$ 3.75) and all these video sessions are evaluated as ''acceptability''. However, when the initial loading delay is as high as 30s, the IPQ rating scores are ranged from 1 to 2

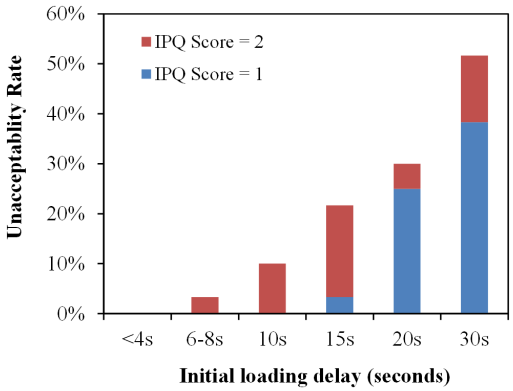


FIGURE 6. Distribution of the unacceptability rate under different initial loading delays.

TABLE 5. Distribution of user unacceptability.

Initial loading delay	Total number of ratings	IPQ rating scores	IPQ (MOS)	Rate of unacceptability
<4s	480	3, 4, 5	>3.75	0%
6-8s	120	2, 3, 4	2.89-3.28	3.33%
10s	60	2, 3	2.57	10.00%
15s	60	1, 2, 3	1.98	21.67%
20s	60	1, 2	1.62	30.00%
30s	60	1, 2	1.24	51.67%

(MOS $=$ 1.24) and over half of these video sessions are evaluated as ''unacceptability''.

Fig. 6 visualizes user unacceptability rate under different initial loading delays. It is found that all these ''unacceptability'' correspond to the rated IPQ scores of 1 and 2. More specifically, when the initial loading delay shorter than 10 seconds, users has less abandonment behavior (13.33%) and the IPQ scores for the abandonment are 2. As the initial loading delay increases to 15 seconds, most IPQ scores (18.33%) were still rated at 2. When the initial loading delay reaches 20 and 30 seconds, the user abandonment rate is as high as 30% and 51.67%, respectively. In such cases, most IPQ scores for the abandonment are 1. Accordingly, in order to maintain a high retention rate, it is suggested that the initial loading delay should be controlled within 10 seconds, which is consistent with our previous conclusions.

E. MEASURING UNACCEPTABILITY OF INITIAL LOADING DELAY

Considering the binary nature of the acceptability of the initial loading delay evaluation, here we performed a logistic regression analysis to model the acceptability of the initial loading delay. Since significant correlation between the unacceptability and the initial loading delay is observed, we selected the initial loading delay (denoted as V_{IDL}) as a predict variable and the unacceptability of initial loading

delay is chosen as the dependent variable. The result of this logistic regression analysis is a model for the probability that the user will not accept the initial loading delay of the video (denoted as p), which can be expressed as follows:

$$p = \frac{\exp^{-\lambda_1+\lambda_2 \cdot V_{IDL}}}{1+\exp^{-\lambda_1+\lambda_2 \cdot V_{IDL}}} \tag{2}$$

where $\lambda_1 = 3.332$ and $\lambda_2 = 0.117$ are the logit coefficients. The logistic regression results show that the predictor variable V_{IDL} has a significant and positive contribution to the prediction of the outcomes, indicating that the probability of unacceptable increases as the initial loading delay increasing. The pseuso R^2 (Nagelkerke) of the model is 0.420 and the Chi-Square (χ^2) is 239.594. This statistical test confirms that the data is distributed for the proposed logistic regression model [34]. According to the prediction model, a critical point is reached if the initial loading delay becomes more than 29 seconds, since the probability of unacceptable result is then higher than 50%.

IV. PERFORMANCE EVALUATION

To validate the efficiency of the proposed model, it is desired to compare its performance with existing models using the validation dataset (i.e., VL dataset). To our best knowledge, however, there were no objective models specifically designed for user's initial perceived quality evaluation. As the model of ITU-T P.1203 can output a video quality when only considering the influence of the initial loading delay [26], we thus compared the proposed IPQ evaluation model with the P.1203.

Three metrics suggested by VQEG [30] were employed in the performance validation, namely, the PCC for linearity, the SROCC for monotonicity, and the root-mean-squared error (RMSE) for accuracy. Generally, a smaller value of the RMSE and a larger value of the PCC and SROCC indicated a superior performance. Table 6 lists the summary of the evaluation performance of the proposed model and P.1203 model. It can be found that the PCC and SROCC between the predicted IPQ by the proposed models and the subjective IPQ are 0.987 and 0.987, respectively. The RMSE for the proposed model is 0.216. For the P.1203 model, the PCC and SROCC between the predicted IPQ and subjective IPQ are 0.797 and 0.900, respectively. The RMSE for the P.1203 model is 0.923. This result reflects that the proposed model can accurately predict IPQ.

To check whether the proposed model is significantly better than P.1203, statistical analysis was further conducted. Following the test method recommended in [35], the test was based on the residual between the subjective IPQ and the IPQ predicted by either of the proposed model and P.1203. Before being able to run the test, the Kolmogorov-Smirnov test (K-S test) was employed to check the Gaussianity of the difference between the subjective IPQ and the predicted IPQ for different models. Experimental results showed that the p-values of the proposed model ($p = 0.099$) and P.1203 ($p = 0.266$) are both larger than 0.05, indicating that the scores follow the normal distribution. The paired t-test was further performed based on the residuals between the subjective IPQ and the predicted IPQ by two methods. The test results showed that the p-value was 0.002, indicating that there is a significant difference between the prediction performance of the proposed model and P.1203.

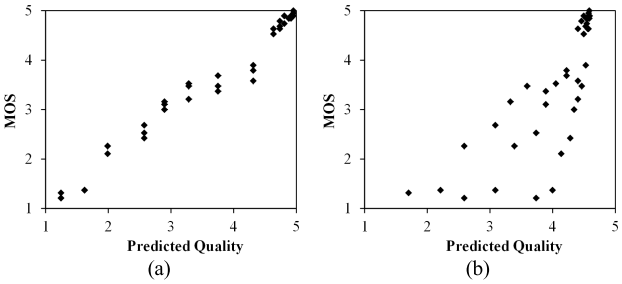


FIGURE 7. Scatter plots of the predicted quality and MOS. (a) Proposed model (b) P.1203 model.

TABLE 6. Performance comparison.

	PCC	SROCC	RMSE
Proposed model	0.987	0.987	0.216
P.1203 model	0.797	0.900	0.923

Fig. 7 further visualizes the performance of the models using the scatter plots of the predicted IPQ and subjective IPQ (in terms of MOS). There shows a strong linear relationship between the predicted video quality by the proposed model and MOS, which indicates that the quality predicted by the proposed model is very close to the user actual perception. In contrast, the difference between the initial perceived quality predicted by the P.1203 model and MOS is considerably large.

Moreover, we also validate the performance of the prediction model for unacceptability of the initial loading delay using VL dataset. Fig. 8 visualizes the predicted results by plotting the probability of the unacceptability of a video as a function of initial loading delay. The logistic curve of the predicted probability is compared with the subjective evaluations of the unacceptability of initial loading delay (denoted

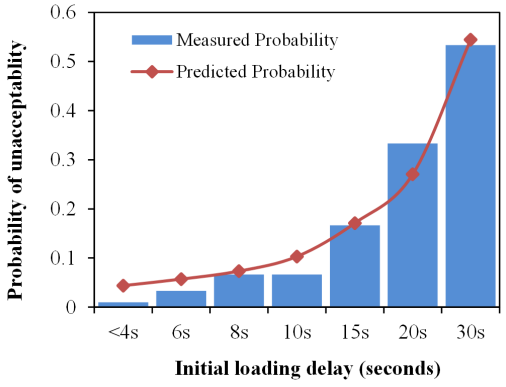


FIGURE 8. Probability of unacceptability of the initial loading delay.

as "measured probability"). The RMSE between the predicted probability and the measured probability is 0.035, which indicates that the predicted probability obtained using logistic regression is a good fit of the measured probability.

V. CONCLUSION

The IPQ directly determines whether a user will continue to watch videos. However, IPQ has never been carefully studied related to the probability of unacceptability. In this paper, we designed two types of subjective test methods to obtain IPQ. Experimental results showed that the regularity of IPQ scores rated immediately when the video starts playing is better than that rated at the end of the video. After the characteristic analysis of experimental results, it is found that the initial loading delay should not exceed 4 seconds for an excellent video service. If the above conditions are not guaranteed, the initial loading delay should be controlled within 10 seconds, where most of users are still in their tolerance for waiting. Moreover, considering the influence of the initial loading delay, an objective evaluation model was proposed to evaluate IPQ and a logistic model was proposed to predict user unacceptability of the video, respectively. Experimental results showed that the proposed models can accurately evaluate IPQ and user acceptability. The proposed models can be used as guidelines for service providers to effectively improve the buffering strategy for video streaming services according to user initial perceptions.

REFERENCES

[1] "Cisco visual networking index: Forecast and methodology 2016–2021," Cisco, San Jose, CA, USA, White Paper 2018, 2017, vol. 1.
[2] S. Tang, X. Qin, and G. Wei, "Network-based video quality assessment for encrypted HTTP adaptive streaming," *IEEE Access*, vol. 6, pp. 56246–56257, 2018.
[3] K. Miller, D. Bethanabhotla, G. Caire, and A. Wolisz, "A control-theoretic approach to adaptive video streaming in dense wireless networks," *IEEE Trans. Multimedia*, vol. 17, no. 8, pp. 1309–1322, Aug. 2015.
[4] *Quality of Experience Requirements for IPTV Services*, document ITU-T Rec. G.1080, Dec. 2008.
[5] B. Jiang, J. Yang, Q. Meng, B. Li, and W. Lu, "A deep evaluator for image retargeting quality by geometrical and contextual interaction," *IEEE Trans. Cybern.*, to be published.
[6] F. Yang and S. Wan, "Bitstream-based quality assessment for networked video: A review," *IEEE Commun. Mag.*, vol. 50, no. 11, pp. 203–209, Nov. 2012.
[7] P. Juluri, V. Tamarapalli, and D. Medhi, "Measurement of quality of experience of video-on-demand services: A survey," *IEEE Commun. Surveys Tuts.*, vol. 18, no. 1, pp. 401–418, 1st Quart., 2016.
[8] Y. Chen, K. Wu, and Q. Zhang, "From QoS to QoE: A tutorial on video quality assessment," *IEEE Commun. Surveys Tuts.*, vol. 17, no. 2, pp. 1126–1165, 2nd Quart., 2015.
[9] Z. Wang, L. Lu, and A. C. Bovik, "Video quality assessment based on structural distortion measurement," *Signal Process., Image Commun.*, vol. 19, no. 2, pp. 121–132, 2004.
[10] J. Yang, C. Ji, B. Jiang, W. Lu, and Q. Meng, "No reference quality assessment of stereo video based on saliency and sparsity," *IEEE Trans. Broadcast.*, vol. 64, no. 2, pp. 341–353, Jun. 2018.
[11] S. Chikkerur, V. Sundaram, M. Reisslein, and L. J. Karam, "Objective video quality assessment methods: A classification, review, and performance comparison," *IEEE Trans. Broadcast.*, vol. 57, no. 2, pp. 165–182, Jun. 2011.
[12] C. G. Bampis, Z. Li, I. Katsavounidis, and A. C. Bovik, "Recurrent and dynamic models for predicting streaming video quality of experience," *IEEE Trans. Image Process.*, vol. 27, no. 7, pp. 3316–3331, Jul. 2018.
[13] Z. Chen, W. Zhou, and W. Li, "Blind stereoscopic video quality assessment: From depth perception to overall experience," *IEEE Trans. Image Process.*, vol. 27, no. 2, pp. 721–734, Feb. 2018.
[14] Q. Wu, H. Li, F. Meng, and K. N. Ngan, "Toward a blind quality metric for temporally distorted streaming video," *IEEE Trans. Broadcast.*, vol. 64, no. 2, pp. 367–378, Jun. 2018.
[15] M. T. Vega, C. Perra, F. De Turck, and A. Liotta, "A review of predictive quality of experience management in video streaming services," *IEEE Trans. Broadcast.*, vol. 64, no. 2, pp. 432–445, Jun. 2018.
[16] R. K. P. Mok, E. W. W. Chan, and R. K. C. Chang, "Measuring the quality of experience of HTTP video streaming," in *Integrated Network Management*. Piscataway, NY, USA: IEEE, 2011, pp. 485–492.
[17] T. De Pessemier, K. De Moor, W. Joseph, L. De Marez, and L. Martens, "Quantifying the influence of rebuffering interruptions on the user's quality of experience during mobile video watching," *IEEE Trans. Broadcast.*, vol. 59, no. 1, pp. 47–61, Mar. 2013.
[18] M.-N. Garcia, D. Dytko, and A. Raake, "Quality impact due to initial loading, stalling, and video bitrate in progressive download video services," in *Proc. IEEE 6th Int. Workshop QoMEX*, Sep. 2014, pp. 129–134.
[19] M. Seufert, S. Egger, M. Slanina, T. Zinner, T. Hoßfeld, and P. Tran-Gia, "A survey on quality of experience of HTTP adaptive streaming," *IEEE Commun. Surveys Tuts.*, vol. 17, no. 1, pp. 469–492, 1st Quart., 2015.
[20] M. Choi, J. Kim, and J. Moon, "Wireless video caching and dynamic streaming under differentiated quality requirements," *IEEE J. Sel. Areas Commun.*, vol. 36, no. 6, pp. 1245–1257, Jun. 2018.
[21] S. S. Krishnan and R. K. Sitaraman, "Video stream quality impacts viewer behavior: Inferring causality using quasi-experimental designs," *IEEE/ACM Trans. Netw.*, vol. 21, no. 6, pp. 2001–2014, Dec. 2013.
[22] J. Song, R. Wang, F. Yang, Z. Ma, and Q. Zhao, "Initial perceived quality analysis for DASH video streaming," in *Proc. IEEE Visual Commun. Image Process. (VCIP)*, Apr. 2019, pp. 1–4.
[23] 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.
[24] 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.
[25] T. Hoßfeld, S. Egger, R. Schatz, M. Fiedler, K. Masuch, and C. Lorentzen, "Initial delay vs. interruptions: Between the devil and the deep blue sea," in *Proc. Int. Workshop Qual. Multimedia Exper. (QoMEX)*, Jul. 2012, pp. 1–6.
[26] *Parametric Bitstream-Based Quality Assessment of Progressive Download and Adaptive Audiovisual Streaming Services Over Reliable Transport*, document ITU-T Rec. P.1203, Nov. 2016.
[27] *Subjective Test Plan for Assessing User Experience of Initial Loading of Streaming Video (P.QUITS)*, document ITU-T SG12-C123, Feb. 2018.
[28] S. Jumisko-Pyykkö, V. K. MalamalVadakital, and M. M. Hannuksela, "Acceptance threshold: A bidimensional research method for user-oriented quality evaluation studies," *Int. J. Digit. Multimedia Broadcast.*, vol. 2008, Jul. 2008, Art. no. 712380.
[29] J. S. Nevid, *Psychology: Concepts and Applications*, 3rd ed. Boston, MA, USA: Houghton Mifflin, 2009.
[30] *Methods for the Subjective Assessment of Video Quality, Audio Quality and Audiovisual Quality of Internet Video and Distribution Quality Television in any Environment*, document ITU-T Rec. P.913, Jan. 2014.
[31] *Methodology for the Subjective Assessment of the Quality of Television Pictures*, document ITU-R Rec. BT.500-13, Jan. 2012.
[32] Q. Huynh-Thu, M.-N. Garcia, F. Speranza, P. Corriveau, and A. Raake, "Study of rating scales for subjective quality assessment of high-definition video," *IEEE Trans. Broadcast.*, vol. 57, no. 1, pp. 1–14, Mar. 2011.
[33] (2019). *MATLAB*. [Online]. Available: https://www.mathworks.com/
[34] M. H. Kutner, C. J. Nachtsheim, J. Neter, and W. Li, *Applied Linear Statistical Models*, 5th ed. New York, NY, USA: McGraw-Hill, 2005.
[35] "Final report from the video quality experts group on the validation of objective models of video quality assessment, phase II (FR-TV2)," Video Quality Experts Group, Geneva, Switzerland, Tech. Rep. COM9-C-60-E, 2003.