Computer Networks 142 (2018) 194–207

Contents lists available at ScienceDirect

Computer Networks

journal homepage: www.elsevier.com/locate/comnet

A quantitative relationship between Application Performance Metrics and Quality of Experience for Over-The-Top video

Weiwei Li[a], Petros Spachos[b,*], Mark Chignell[a], Alberto Leon-Garcia[a], Leon Zucherman[c], Jie Jiang[c]

[a] Electrical and Computer Engineering, University of Toronto, Toronto, Ontario M5S 3G4, Canada
[b] School of Engineering, University of Guelph, Guelph, Ontario N1G 2W1, Canada
[c] TELUS Communications Company, Toronto, Ontario M1H 3J3, Canada

ARTICLE INFO

Article history:
Received 5 February 2018
Revised 19 April 2018
Accepted 22 May 2018
Available online 6 June 2018

Keywords:
QoE
QoS
OTT video streaming
Machine learning

ABSTRACT

Quality of Experience (QoE) is a measure of the overall level of customer satisfaction with a vendor. In telecommunications, consumer satisfaction is of great interest in the adoption of novel multimedia products and services. A number of factors can greatly influence the customer experience during a video session. Factors such as user perception, experience, and expectations are expressed by QoE while factors such as application and network performance are expressed by Quality of Service (QoS) parameters. This paper studies the relationship between QoS and QoE in a session-based mobile video streaming. Specific QoS Application Performance Metrics (APMs) are examined based on a QoE assessment database which is built for experimentation and contains 108 subjects. It is shown that these APMs are highly related to two QoE factors, Technical Quality (TQ) and Acceptability. Furthermore, Viewing Ration (VR) parameter and the corresponding Kendall correlation between VR and QoE factors proves that VR is a valuable metric for mapping QoS to QoE. We further generated the compacted decision tree to predict QoE factors through Rebuffering Ratio (RR), Non-interruption Content Viewing Ratio (VR_c), and Non-interruption Viewing Ratio (VR_s). Through extensive experimentation, a general relationship between APMs and QoE factors has been examined and a primary QoE model is proposed based on this relationship.

1. Introduction

One of the main goals of telecommunications is to provide services which satisfy consumers. However, the dramatic growth in data traffic is stressing network. Video service has occupied an important place in network services and video traffic has taken a huge amount of traffic on the Internet. A crucial requirement is to support video services to meet customer's expectation in terms of Quality of Experience (QoE). QoE is the degree of delight or annoyance of the user of an application or service. It results from the fulfillment of his or her expectations with respect to the utility and/ or enjoyment of the application or service in the light of the user's personality and current state [1].

QoE is a subjective measure of user's perception. People are still at the stop of determining the methodology of QoE assessment, exploring the connections between Quality of Service (QoS) and QoE, and deriving potential models for QoE estimation. Traditional QoS has focused on the video itself, while ignored that video is only a part of the whole service session for an Over-The-Top (OTT) video streaming. This is reasonable since OTT video becomes popular only recently. In order to propose a QoE model for next-generation networks, it is necessary to study QoE based on a life cycle of a video session [2].

In this paper, the relationship between Application Performance Metrics (APMs) and QoE factors collected by subjective experiments is studied. APMs are objective measurable metrics to represent the quality of a video. Two QoE factors are involved in this paper: Technical Quality (TQ) and Acceptability. TQ collects user's opinions from available options to understand QoE, while Acceptability is a binary measure to locate whether user accepts the video quality or not. The impact of failures which happened during a video display is also discussed. It is investigated whether the addition of failure requires new performance metrics, and how the failures impact user's assessment on QoE on mobile video streaming.

* Corresponding author.
E-mail address: petros@uoguelph.ca (P. Spachos).

https://doi.org/10.1016/j.comnet.2018.05.020
1389-1286/

The major contributions of this work are listed below:

- An extensive QoE experiment was designed and conducted. In the experiment, 108 subjects participated and successfully finished the experiment.
- A novel session-based QoE model is introduced for mobile OTT video streaming. The model proposes the use of the entire interaction during the life-cycle of the video section in order to determine the QoE.
- Following the introduced model and the International Telecommunication Union (ITU) standards, two types of failures are introduced: Accessibility and Retainability, in order to understand the QoE in the whole session.
- Along with the traditional Mean Opinion Score (MOS) scale, two new scales are used to examine whether the opinion scores remain stable after the introduction of the failures.
- Data correlation and four machine learning classification: Naive Bayes, Logistic Regression, k-NN Classification, and Decision Tree, were used to analyze the complex relationships among selected APMs and QoE factors and to compare the impact under various test conditions, the addition of failures and the change of scales.
- A primary QoE model is proposed which relies on a quantitative relationship between the OoS performance metrics and the QoE factors.

The remainder of the paper is structured as follows: In Section 2, the related work on QoE and APMs is reviewed. In Section 3 the session-based QoE and QoS are described followed by a description of the experimental design and methodology in Section 4. In Section 5 is the evaluation analysis to explain the relationship between APMs and QoE factors, and Section 6 proposed a primary QoE model of MOS based on this study. This work is concluded in Section 7.

2. Related work

2.1. QoE Overview

QoE of a service is determined by various factors, or categories [3–6]. Defining QoE categories is the basic problem for analyzing and researching the multi-faceted, user-oriented quality assessment problem. Two types of indicators are mentioned in QoE: Key Quality Indicator (KQI) and Key Performance Indicator (KPI). KPIs can be measured and calculated directly. The values of KPIs are derived from measurable network performance and non-network performance parameters. KQIs are used to capture the user's perception directly. KPIs are logically aggregated into KQIs, and one KPI can be a part of multiple KQIs at the same time [7]. Finally, KQIs are mapped into QoE factors.

Nokia proposed five main categories to characterize QoE: service availability, usability, integrity, reliability, and accessibility [4]. As mentioned in [6], there are many factors related to QoE, such as accessibility, server reliability/availability, usability, network quality, content effectiveness, and so on. QoE dimensions include technology performance, usability, subjective evaluation, expectations, and context [3].

This paper focuses on two KQIs: TQ and Acceptability. For KPIs, APMs related to failures are introduced. it provides a detailed description of APMs in a video session and how to connect APMs to QoE factors.

2.2. Classical QoE studies

Most QoE research aims to reveal the relationship between QoS and QoE. Khan et al. have studied the impact of QoS parameters on QoE and proposed a QoE adaptation scheme for video applications [8]. Imran et al. have utilized statistical techniques to evaluate QoE performance based on QoS parameters [9]. Alreshoodi and Woods have summarized recent studies on QoE/QoS correlation modes [10]. They have summarized three possible approaches for mapping the QoE/ QoS relationship: use QoS to map QoE, use QoE to map QoS and use some QoS and QoE to estimate other QoS and QoE. They conclude that the problem is which approach is efficient enough.

The ITU-T Study Group is active in developing standards for video QoE evaluation [11–14]. QUALINET developed systematic methodologies for QoE on multimedia applications [15,16]. Joskowicz et al. summarized parametric models for video quality estimation in recent years [17], and proposed their own parametric model for perceptual video quality [18]. Dalal et al. implemented a video QoE assessment framework for real-time video streaming [19]. Based on real-time models, a packet scheduling algorithm was utilized to minimize a defined cost function [20,21].

2.3. Tendencies in QoE assessment

Recently, more and more researchers discussed QoE measurement by proposing new approaches. Oyman et al. considered how to develop performance metrics to accurately evaluate QoE for adaptive streaming services [22]. They found that rebuffering is the single most influential impairment relating to QoE and they used rebuffering percentage to estimate the user satisfaction.

Machine learning is being introduced into the QoE/ QoE model study. Balachandran et al. proposed a data-driven machine learning approach to tackle the complex relationship between the quality metrics and QoE measurement [23]. Mushtaq et al. have discussed different machine learning approaches to assess the correlations between QoS and QoE [24]. Six classifiers are tested based on their data to investigate the correlation between QoS and QoE in video streaming service. Chen et al. have discussed QoS parameters impacting users' satisfaction and proposed a video rate adaptation scheme to improve viewer QoE [25,26].

Another tendency in QoE assessment is the deployment of Acceptability. T. Hoßfeld et al. have pointed that QoE studies should not be limited to the study of MOS. They have classified a set of objective and subjective QoE metrics and indicated that acceptance is a key QoE metric [27]. Menkovskis et al. have implemented a QoE model to predict whether the quality of network service is acceptable ('Yes') or unacceptable ('No') [28]. Their model is based on a decision tree, and they declared that the accuracy is over 90%. Other work on acceptability QoE model is proposed by Song and Tjondronegoro [29] They have generated a logistic regression model to map QoS parameters to acceptability.

Meanwhile, there has been considerable recent research in QoE evaluation, especially in developing QoE models for OTT video applications [30–33]. An early indication of the need to assess QoE for an entire session was discussed in [34]. Human factors that influence QoE, such as context, human memory, and attention effects, were investigated in [31,34]. Moorthy et al. implemented studies including subjective testing, subjective opinion evaluation and objective algorithm performance analysis [33]. Their QoE evaluation emphasized the impact of rate adaptation.

Mok et al. [32] and De Pessemier et al. [30] have studied the impacts of impairments on QoE directly instead of correlating network performance to QoE. Dorian et al. have proposed the impact of video quality on QoE factor, however, they did not discuss the possible failures although they mention the concept of a video session life, which included "stopped/exit" [31].

Comparing to above research, this paper focuses on finding the impact of impairment and failures, which has a close explanation about the QoS side compared to previous work.

3. Session-based QoE and QoS

In this section, the proposed session-based QoE and QoS model is described, followed by the main QoE factors that are examined and the QoS performance metrics which affect it.

3.1. Session QoE with impairments and failures

Traditionally, QoE assessment focused on Integrity impairments [14]. Integrity indicates the degree to which a session unfolds without excessive impairments. However, this approach does not include whether a video can successfully start and/ or end normally and without any problems.

With the popularity of OTT video streaming, research on QoE proposed that user perception of a video service should be studied during the life-cycle of a video session [2,31]. Experimental results have shown that the customer experience in a service is significantly impacted by the entirety of interactions during the session of a customer with a service. Consequently, as a type of service, the QoE of video streaming should be determined by the entire interaction during the life-cycle of the video session.

When the user requests a video, it follows the steps below:

(i) wait for the video to start,
(ii) watch the video along with some possible impairments, and
(iii) quit the video service normally or abnormally due to unexpected problems.

Integrity impairments cover only the second step of the previous process. In this work, two more components, Accessibility and Retainability failures are proposed, to understand QoE in the whole session. Concerning the QoE measurement of an entire session, a session-based QoE evaluation which includes three components has been proposed in our previous work [35–40].

Accessibility and Retainability are called failures since either represents an abnormal termination of video service. Accessibility failures will cover the first step of the video session process and Retainability failures will cover the third step of the video session process.

The ITU standard sets six primary components to the quality of telecommunication services [41,42]. These are: Support, Operability, Accessibility, Retainability, Integrity, and Security. A service session typically contains Accessibility, Retainability, and Integrity. Hence, only these three components in a video session are discussed.

Accessibility. It refers to the successful start of the session. When the subject attempts to initiate the session, the session may or may not start successfully. If the session fails to start, an Accessibility failure has occurred. Service accessibility may be related to availability, security (authentication), access, coverage etc.

Retainability. It is the capability to continue the session until its completion, or until it is interrupted by the subject. If the session is terminated permanently due to a failure, this is a Retainability failure. In general, *Retainability* characterizes connection losses.

Integrity. It indicates the degree to which a session unfolds without excessive impairments. Even when a session does not experience any of the previous two failures, there are a number of service-specific impairments that may impact the QoE of the service. For instance, throughput, delay, delay variation (or jitter) can impact the perceived quality of the service.

The definitions of these three components follow the International Telecommunication Union (ITU) standard [41]. Our investigation of the session-based QoE is based on these definitions and focuses on mobile video service. This is because mobile video service occupies a huge data traffic and encounters more Integrity, Accessibility, Retainability issues comparing to non-mobile service.

3.2. QoE factors

In this work, two main QoE factors are examined to represent user's perception:

Technical Quality (TQ): All technical features that the user can perceive during the whole video session. These features include but are not limited to the video blockiness, video freezing, video blurriness, and audio sync issues. This definition is proposed in a study focusing on quantifying the influence of rebuffering on QoE of mobile video [30], and the terminology of TQ is widely used for QoE subjective assessment [15,30]. In this work, the technical features are the impairments and failures that we designed and discuss in Section 4.3.

For rating the TQ, three different rating scales were employed:

Scale A: Excellent - Good - Fair - Poor - Bad. Scale A strictly follows the ITU standards and it is a 5-point rating scale.

Scale B: Excellent - Good - Fair - Poor - Bad - Terrible. Scale B extends the rating scale of Scale A on the negative side by adding one more choice at the bottom (Terrible). This is proposed in order to decide whether the user's evaluation of impairments and failures tends to go to the negative side when more failures are shown.

Scale C: Excellent - Good - Fair - Poor - Bad -Terrible - Worst Possible. Scale C also extends on the basis of Scale A with two more negative choices.

The design for Scale B and Scale C is to decide whether the opinion scores are stable when even worse opinion scores are provided with the appearance of failures, which is the first time introduced in the QoE assessment.

Acceptability: Acceptability refers to the subject's decision to either accept or reject a product or a service by utilizing 2-point likert scale (answering *Yes* or *No*). In QoE research, acceptability is treated as a whole offer – including price, cost, and system – and relies on directly querying subjects regarding the acceptability of the quality level experienced. In [43], the authors have defined acceptability in the context of mobile video QoE as *"a binary measure to locate the threshold of minimum acceptable quality that fulfills subject quality expectations and needs for a certain application or system"*.

At the same time, consumer acceptance is of great interest in the adoption of novel multimedia products and services. The ITU definition of QoE is based on the notion of subject acceptability of a service [12], while most QoE systems follow MOS to measure a subject's acceptability. Study on acceptability can help explore the possibility and effectiveness of new QoE models based on binary QoE assessment.

For acceptability, the traditional 2-point likert scale was employed.

3.3. QoS performance metrics

QoE and QoS are complementary but distinct measures of service quality. QoS of OTT video streaming is focused on network performance and its measurement involves network-centric metrics for service assessment. Objective QoS metrics are important in assessing network performance. However, network performance cannot directly represent the user's perception of quality regarding

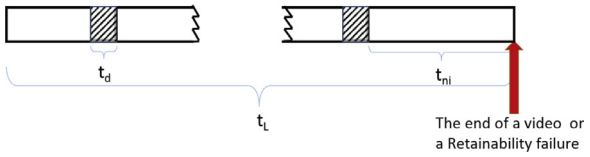


Fig. 1. An example of impairment/ failure timing.

services. Many other factors, such as psychological aspects, end-to-end service processes, and context, should be considered in QoE evaluation.

An important component of QoE study is how to evaluate the impact of QoS parameters on QoE assessment. For Integrity impairments, rebuffering event is considered. Our work proposed a session-based QoE evaluation, hence it needs to examine the impact of Retainability/ Accessibility failures on video QoE as well. As a result, the QoS parameters used in our work are associated with both Integrity impairment and Retainability/ Accessibility failures, which were seldom discussed before.

The following QoS parameters are examined:

- Initial buffering duration: The duration of an initial buffering.
- Rebuffering Number (RN): The number of rebuffering events happened during a video session.
- Rebuffering time: The time point at which a rebuffering event happened.
- Rebuffering duration (t_d): The duration of a rebuffering event.
- Video length (t_v): The original length of a video. This is the total length of a pristine video, without any Integrity impairment, Retainability failure or Accessibility failure.
- The total display time(t_L): The total display time of a video and rebuffering duration.
- Content viewing time: The content viewing time of a video from a specific user. Content viewing time is a performance metric introduced due to the addition of failure types since Integrity impairment assumes that user can watch the whole video despite encountering impairments.

A graphical description of the different timing is shown in Fig. 1. t_{ni} represents the content display time after the last impairment. In our experiment, all rebuffering have the same duration.This is not a realistic scenario in daily applications. However, deploying variable lengths of rebuffering duration means that we need to consider the impact of rebuffering time lengths on QoE evaluation. As we described above, we are more interested in exploring the QoE evaluation with the addition of Retainability and Accessibility failures. Considering the time and money consumption of recruiting participants, we simplify some properties on Integrity impairments. More details are discussed in paper [44].

Note that $t_v \neq t_L - RN \cdot t_d$ when Retainability failure happens. This is because when a Retainability failure happened, the user cannot see any more content of this video. Content viewing time is decided by a video session. If there is a Retainability failure, the content viewing time is smaller than t_v. Meanwhile, the value of t_v is a fixed value when the video is selected.

Based on these objective metrics, four performance metrics to quantify QoS for OTT video streaming are proposed. All metrics are Application Performance Metrics (APMs), hence, they are performance metrics on the application level during the video playback.

(i) Rebuffering Ratio (RR): The ratio of total rebuffering time versus the total display time of a video (rebuffering time + content viewing time). It can be defined as:

$$RR = \frac{t_d \cdot RN}{t_L} \tag{1}$$

Table 1
Questionnaire for each video.

No.	Rating criterion/ Question	Rating scale
1	Is the technical quality of this video acceptable?	Yes/No
2	Your evaluation of the technical quality in the video is:	Scale A/ Scale B/ Scale C
3	The content of the video is:	Scale A
4	Your overall viewing experience (Content + Technical quality) during the video play back is:	Scale A/ Scale B/ Scale C

(ii) Non-interruption Content Viewing Ratio (VR_c): The content display time after the last impairment versus the total content viewing time. It is equal to:

$$VR_c = \frac{t_{ni}}{t_L - RN \cdot t_d} \tag{2}$$

(iii) Non-interruption Viewing Ratio (VR_s): The content display time after the last impairment versus the total content viewing time. It can be defined as:

$$VR_s = \frac{t_{ni}}{t_v} \tag{3}$$

(iv) Content Viewing Ratio (VR): The content viewing time versus the total length of a video. It reflects the level of video completion. It is equal to:

$$VR = \frac{t_L - RN \cdot t_d}{t_v} \tag{4}$$

4. Experiment design and methodology

This section describes the subjective assessment methodology that was followed. A laboratory-controllable subjective experiment was designed, with specific impairments and failures [44].

4.1. Experimental setup

The experiment took place in a controlled environment at the University of Toronto. The conditions of participation were to have normal or corrected to normal vision and to not have participated in a video quality assessment experiment in the six months prior to the date of the experiment. All subjects were aged over 18 years.

108 subjects participated and finished the experiment. The majority of participants are engineering students from the University of Toronto. They are familiar with OTT video services, such as YouTube and Netflix. The age of above 90% participants are between 18 and 30. The ratio of males to females is 2: 1 (some participants chose to not expose their gender).

All the subjects used the same computer with the same configuration. Each subject evaluated 31 video sessions in total, which lasted around 90 min. The average length of each video is 96.7 s. The videos are displayed in random order to control possible effects. The resolution of video clips is 864×482 and a frame-rate of 30 frames-per-second (fps), which is a comparative number to the resolution of mobile displays in the real world. The complete videos were between 73 and 123 s in length with an average of 94.1 s. The video sessions consisted of 23 short movie trailers (teaser-trailers) and 8 short movies.

The Absolute Category Rating (ACR) method, which was recommended from ITU [14,45], was employed for the experiments. Every participant answered four questions related to the video quality immediately after the video. The questions are listed in Table 1. Q1 and Q2 are related to TQ and Acceptability. The other two questions are related to CQ (Content Quality) and OX (Overall eXperience).

4.2. Experimental procedure

The experiment procedure for each participant has the following phases:

(i) Registration phase. The participant first signed the consent form and answered some general profile questions. During this phase demographics, video viewing habits, and video quality preferences were collected.
(ii) Training phase. After the registration phase, each participant took a short training on QoE evaluation along with the definition of the different terms/ words in the questionnaire. The participant watched 5 videos and answered the questionnaire used for video evaluation. These 5 videos included either Integrity impairments and Retainability failures or Accessibility failure in a predetermined order. The responses to the questionnaire were not used in any analysis. The only purpose of this phase is to familiarize the participant with the procedure.
(iii) Evaluation phase. After the training, the experiment started. The participant watched the videos based on the group she/ he belongs. Between the videos, there is a short break to answer the questions. During the video playback, there were two breaks after every 10 video clips.

4.3. Impairment and failure types

Each video has one type of impairment/ failure. The type of impairment/ failure is the test condition. Each participant encountered the same amount of impairments/ failures, in a randomized order.

Three categories of impairment/ failure issues were designed for the experiment, which follows the life-cycle of a video session:

Test sequences containing Integrity impairments during playback (I). In our experiments, only rebuffering events were used to present Integrity impairments. One video sequence may have more than one rebuffering event during playback.

Test sequences containing Retainability failures during playback (R). In our experiments, a Retainability failure may happen with/without Integrity impairments. As long as the Retainability failure happened, the video session ended.

Test sequences containing Accessibility Failures during playback (AF). An Accessibility failure is a failure which occurs before any content of the video sequence is displayed.

In this experiment, we have videos without any impairment/ failure (I0), videos encountered Accessibility Failure (AF), three types of Integrity impairments (I1-I3), and three types of Retainability failures (R0-R2). The description of each impairment/ failures is listed as follows [44]:

- *I*0: There is no impairments and failure. The video is pristine.
- *I*1: The video has a single temporary interruption of 10 s duration happening at 15 s.
- *I*2: The video has two 10 s temporary interruptions happening at 15 s and 30 s of the content display time.
- *I*3: The video has three 10 s temporary interruptions happening at 15 s, 30 s, and 45 s of the content display time.
- *R*0: A permanent interruption happening at 70 s of the content display time.
- *R*1: One 10 s temporary interruptions happening at 15 s; and a permanent interruption happening at 30 s of the content display time.
- *R*2: Two 10 s temporary interruptions happening at 15 s and 30 s; and a permanent interruption happening at 50 s of the content display time.

Table 2
Performance metrics for integrity impairments and failures.

Type	RN	RR	VR_c	VR_s	VR
I0	0	0	1	1	1
I1	1	$\frac{t_d}{t_d+t_v}$	$\frac{t_v-t_{f1}}{t_v}$	$\frac{t_v-t_{f1}}{t_v}$	1
I2	2	$\frac{2t_d}{2t_d+t_v}$	$\frac{t_v-t_{f2}}{t_v}$	$\frac{t_v-t_{f2}}{t_v}$	1
I3	3	$\frac{3t_d}{3t_d+t_v}$	$\frac{t_v-t_{f3}}{t_v}$	$\frac{t_v-t_{f3}}{t_v}$	1
R0	0	0	1	$\frac{t_{R0}}{t_v}$	$\frac{t_{R0}}{t_v}$
R1	1	$\frac{t_d}{t_d+t_{R1}}$	$\frac{t_{R1}-t_{f1}}{t_{R1}}$	$\frac{t_{R1}-t_{f1}}{t_v}$	$\frac{t_{R1}}{t_v}$
R2	2	$\frac{2t_d}{2t_d+t_{R2}}$	$\frac{t_{R2}-t_{f2}}{t_{R2}}$	$\frac{t_{R2}-t_{f2}}{t_v}$	$\frac{t_{R2}}{t_v}$
A	1	1	0	0	0

*t_d: The duration of each rebuffering.
*t_v: The content time of each video.
*t_{Ri}: The content viewing time for Ri, i = 0, 1, 2
*t_{fi}: The time point of the i-th rebuffering happened at the content time, i=1, 2, 3.

Table 3
Groups arrangement of the participants, along with the rating scales and impairment/ failures types.

Group	Subjects	Scale	Type	Scenario
G1	36	A	I0-I3	*A_I*
G2	24	A	I0-I3, R0-R2, AF	*A_IF*
G3	24	B	I0-I3, R0-R2, AF	*B_IF*
G4	24	C	I0-I3, R0-R2, AF	*C_IF*

- *AF*: The video never starts to play. The video player display "failure-to-play" message immediately.

Performance metrics of corresponding types are listed in Table 2. VR_c, VR_s, and VR are new performance metrics proposed due to the addition of failure types since the definition of Integrity assumes that user can watch the whole video despite encountering impairments. We can see that VR is always equal to one for Integrity impairments, and the difference between VR_c and VR_s is on the calculation of Retainability failure types. The reason we propose these new metrics is because we found that the length of content viewing time impacts the evaluation of Retainability failures [36]. We will discuss whether these new performance metrics should be used to sketch the impact of failures in the following sections.

4.4. Group formation

We divided the 108 subjects into four groups and each group finished their experiment with different rating scales. Table 3 shows the details about the rating scale assignment and group arrangement.

We can see that G1 followed the ITU standard model. The rating scale is a 5-point scale (Scale A), and they only evaluated Integrity impairments during the experiment, as shown in the fourth column of Table 3. G1 is a special group which is used as a comparison. The participant in G1 evaluated all video clips without the appearance of any failure.

The participants in G2, G3, and G4 evaluated both impairments and failures. To further explore the impact of Retainability and Accessibility failures, we also employed various rating scales on G2, G3, and G4. The purpose of various scales is to extend negative choices for rating since our experiment introduces more negative scenarios (failure types) than usual.

In the following analysis, we will use the Scenarios shown on the fifth column of Table 3 to represents each group. We will use 5 to −1 to represent these ratings, i.e. Excellent = 5, Good = 4, Fair = 3, Poor = 2, Bad = 1, Terrible = 0, and Worst Possible = −1.

5. Evaluation analysis

In this section, we will analyze the relationship between the performance metric and the QoE factors.

5.1. Analysis overview

In our experimental analysis, we used two methods: data correlation and machine learning classifier approaches.

5.1.1. Correlation

Correlation is a statistical measure of association between two variables. A correlation coefficient is a direct approach to reflect relationships between a pair of variables. However, the disadvantage is that correlation cannot reveal the interactions if there are more than two variables.

Correlation can help us to decide whether a specific APM is related to QoE factors or not. Kendall correlation was used to measure the relevance between one APM and one QoE factor. Follow the suggestion in [31], the Kendall correlation is a rank correlation which does not have any assumption on the distribution or the joint distribution of variables; while Pearson correlation assumes a linear correlation between variables. Considering that VR_c, VR_s, and *VR* are related to failures which have rarely been discussed, it is important to examine whether they should be used as effective performance metrics for QoE assessment.

5.1.2. Machine learning classifier approaches

We have used various machine learning classifiers to analyze the complex relationships among selected APMs and QoE factors and to compare the impact under various test conditions (the addition of failures and the change of scales). Machine learning classifier is a black box approach to analyzing the association between APMs and QoE factors. The advantage is that it provides a clear output (QoE factors) by the input (APMs). Although at the same time, it hides details from data. This is why the accuracy of prediction decreases when the requirement of granularity becomes higher, as stated in [23].

In this work, our goal is to examine the primary association between APMs and QoE factors. We first go through four simple and widely-used classifiers discussed in [10,23,24]: Naive Bayes, Logistic Regression, k-NN Classification, and Decision Tree. Then, we select the classifier which is most stable accompanying high accuracy across all cases.

5.2. Impairment/ failure types vs QoE factors

MOS and 95CI (Confidence Interval) of impairment/ failure types under specific scenarios(*A_I*, *A_IF*, *B_IF*, and *C_IF*) is shown in Fig. 2.

As it can be seen, the MOS values of these types are different. This implies that the users' perception of one type of impairment/ failure is varied from others, no matter which scenario is examined. Generally speaking, the MOS of impairments decreases with the increase of rebuffering times ($I0 > I1 > I2 > I3$). Meanwhile, the MOS values of Retainability failures are lower than those of impairments.

Fig. 2 indicates that there is a connection between impairment/ failure types and QoE factors. However, how impairment/ failure types impact QoE factors needs further examination. Fig. 3 shows the corresponding values of RR, VR_s, and VR of impairment/ failure types based on all samples of *A_IF*. On one hand, it shows that *RR* distinguish the level of rebuffering issues since impairments with same RNs are clustered together along the RR values. On the other hand, it shows that VR and VR_s represents the characteristics of failures.

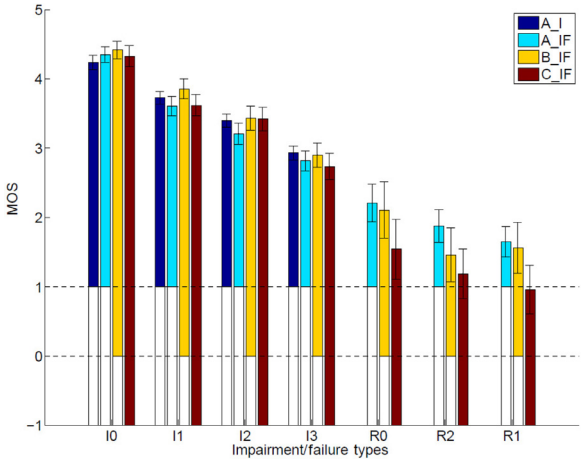


Fig. 2. MOS under four scenarios.

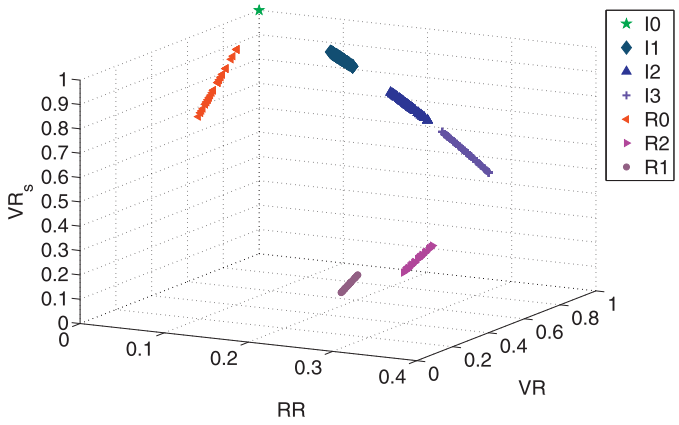


Fig. 3. 3-D view: APMs of impairment/failure types, *A_IF*.

Table 4
One-way ANOVA: quantification of impairment/ failure type.

Scenario	*df*	*F*-value	*p*-value
A_I	3	115.27	$< .0001$
A_IF	6	114.25	$< .0001$
B_IF	6	97.22	$< .0001$
C_IF	6	114.95	$< .0001$

For example, *R*1 has one rebuffering issue and then encountered a failure, while *R*2 has two rebuffering issues and a relatively longer content viewing time. Hence *R*1 shows generally smaller RR values while R2 has larger VR values. These impairment/failure types deployed in our experiment explains the characteristics of impairments and failures well, which is important for a subjective experiment.

One-way Analysis of variance (ANOVA) was conducted for each scenario to quantify the main effect of impairment/ failure types. Table 4 shows that the impact of impairment/ failure types is significant across all scenarios.

To verify the impact of impairment/ failure types on acceptability, the acceptability rate of each impairment/ failure type is shown in Fig. 4. This is the rate in which customers agree that TQ is acceptable.

It is clear that there is a huge drop between impairments and failures, which is not reflected by Fig. 2. It can be inferred that acceptability reflects users' perception from a different aspect and can capture assessment which is not direct in MOS.

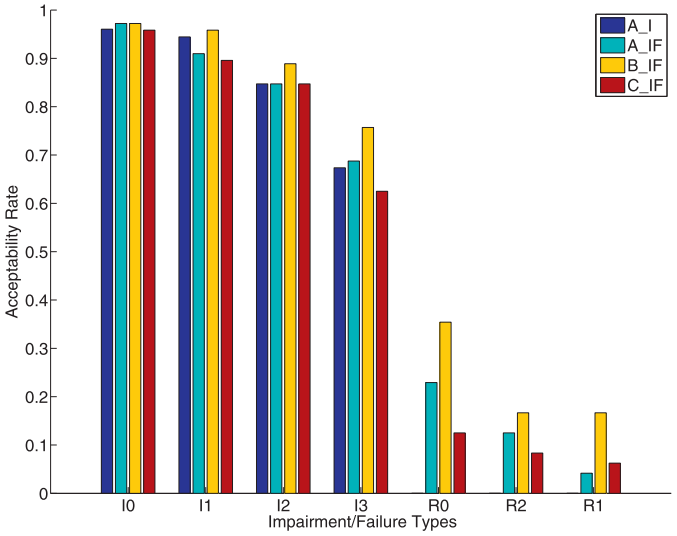


Fig. 4. Acceptability rate under four scenarios.

Table 5
Correlation coefficients between APMs and TQ.

	A_I	*A_IF*	*B_IF*	*C_IF*
	Kendall Correlation Coefficients			
RN	−0.4259	−0.2761	−0.2596	−0.2306
RR	−0.3872	−0.4166	−0.3833	−0.3912
VR_c	0.3872	0.4166	0.3833	0.3912
VR_s	0.3872	0.4944	0.4574	0.4762
VR	–	0.4581	0.4336	0.4794
	Pearson Correlation Coefficients			
RN	−0.4854	−0.3110	−0.2555	−0.52065
RR	−0.4852	−0.4954	−0.4425	−0.4219
VR_c	0.4851	0.5086	0.4617	0.4447
VR_s	0.4851	0.6382	0.6059	0.6184
VR	–	0.5318	0.5460	0.5914

5.3. APMs versus QoE factors

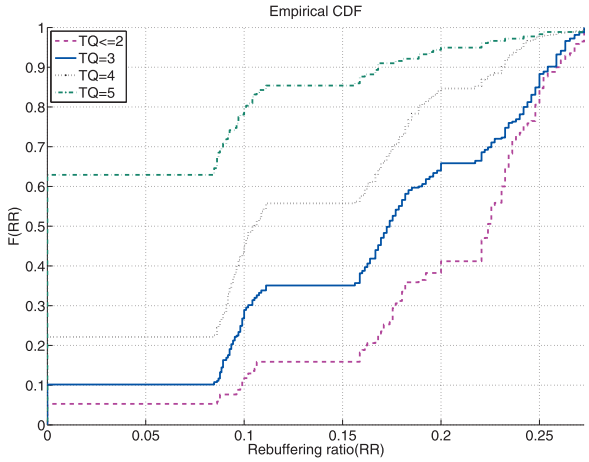
The APMs are compared for the two QoE factors: TQ and Acceptability.

5.3.1. APMs under TQ levels

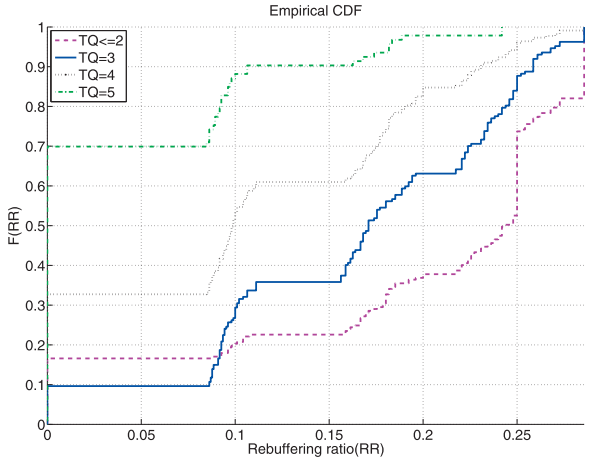
Table 5 summarizes the Kendall and Pearson correlation coefficients between APMs and TQ under the four scenarios (*A_I*, *A_IF*, *B_IF*, and *C_IF*). As it can be inferred, the correlation between RN and TQ drops a lot when failures are introduced. This means RN cannot reflect the impact caused by rebuffering with the appearance of failures. On the other side, RR measures the impact of rebuffering in a stable manner even with failures.

At the same time, the absolute value of Kendall correlation of RR is the same as the absolute value of that of VR_c under the same scenarios. This indicates that RR and VR_c represent the same property from different aspects. Note that the Pearson correlations of (RR, TQ) and that of (VR_c, TQ) are slightly different.

The correlation between VR and TQ under *A_I* is not available because of VR = 1 for all impairment types. For other cases, the Kendall correlation coefficients between VR_s/VR and TQ are above 0.4. The value 0.4 is close to the Kendell correlation coefficients between RR and TQ in *A_I*, i.e. the traditional QoE assessment with Integrity impairments only. It indicates that the associations between VR_s/VR and TQ in the session-based QoE evaluation is at the same level of associations between RR and TQ in the traditional QoE assessment. Note that RR is a common metric used for traditional QoE assessment. It means that VR_s, and VR should be considered as RR for our further analysis on the session-based QoE.



(a) *A_I* case.



(b) *A_IF* case.

Fig. 5. Empirical cdf for *RR* vs. TQ.

The empirical Cumulative Distribution Function (CDF) of RR based on TQ levels is shown in Fig. 5. With the decrease of TQ, the shapes of CDFs are different in both *A_I* and *A_IF* while the general tendencies in both cases are similar. It seems about 80% of highest ratings (TQ = 5) fall below RR < 0.1. At the same time, 80% of ratings of TQ ≤ 2 fall below RR < 0.25. This indicates that RR represents the characteristics of TQ levels regardless of whether failures are included in the evaluation or not.

The empirical CDF of VR_s is shown in Fig. 6. Similar to Fig. 5, it is clear that under different levels of TQ, the CDFs of VR_s is distinguishing.

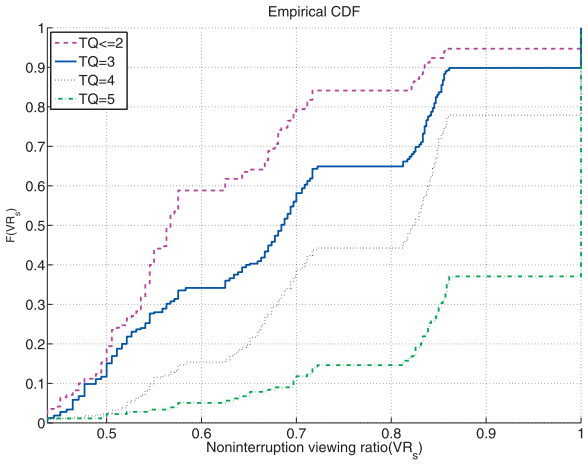
The empirical CDF of VR in the *A_IF* case, since VR is a constant in the *A_I* case, is shown in Fig. 7. When TQ>3, around 90% of VR is equal to 1, i.e., the whole content of a video has been shown.

5.3.2. APMs vs. Acceptability

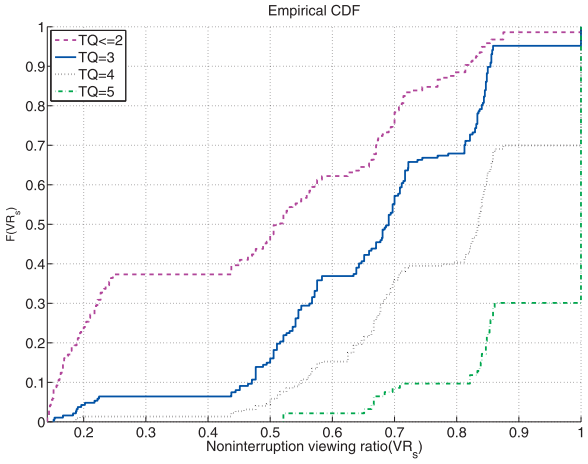
Table 6 shows the correlations between APMs and Acceptability when Acceptability is viewed as a binary scale ('Yes = 1', and 'No = 0').

These coefficients agree with the conclusions from Table 5. It also indicates that acceptability and TQ are similar to each other, although TQ and acceptability explain the user's perception of different granularities and consideration.

The empirical CDFs of RR and VR_s under Acceptability are shown in Fig. 8 and Fig. 9, respectively. The cumulative tendency



(a) *A_I* case.



(b) *A_IF* case.

Fig. 6. Empirical cdf for VR_s vs. TQ.

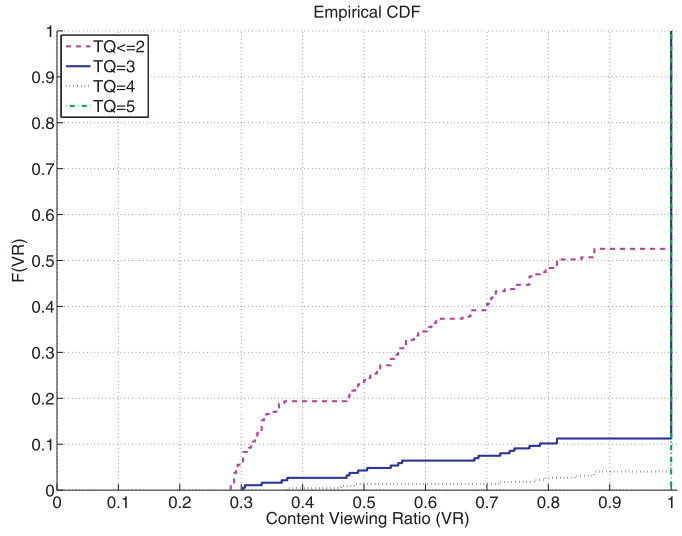
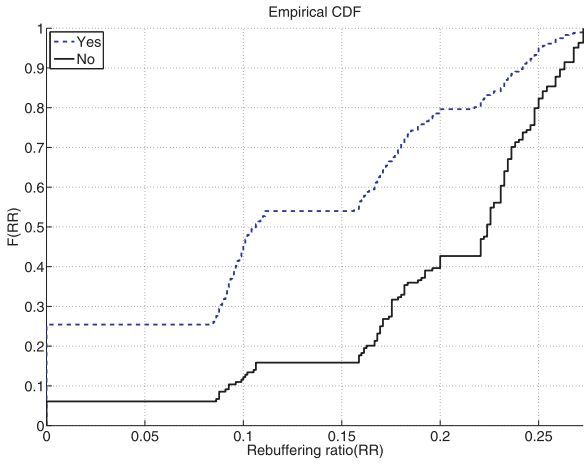


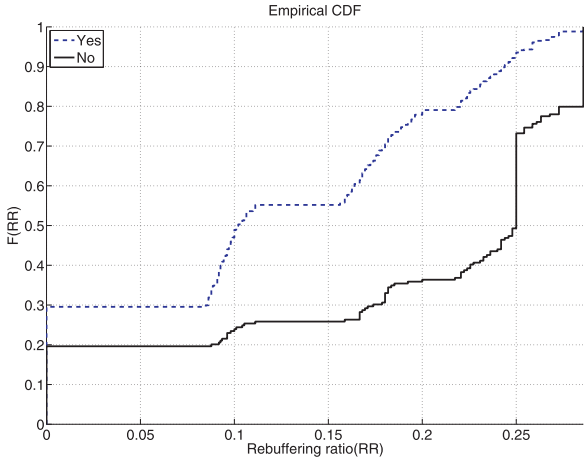
Fig. 7. Empirical cdf for VR vs. TQ, *A_IF*.

Table 6
Correlation coefficients between APMs and Acceptability.

	A_I	*A_IF*	*B_IF*	*C_IF*
RN	−0.2782	−0.0775	−0.0629	−0.0880
RR	−0.2528	−0.3042	−0.2763	−0.2961
VR_c	0.2528	0.3042	0.2763	0.2961
VR_s	0.2528	0.4057	0.3728	0.4130
VR	–	0.6101	0.5958	0.6072



(a) *A_I* case.



(b) *A_IF* case.

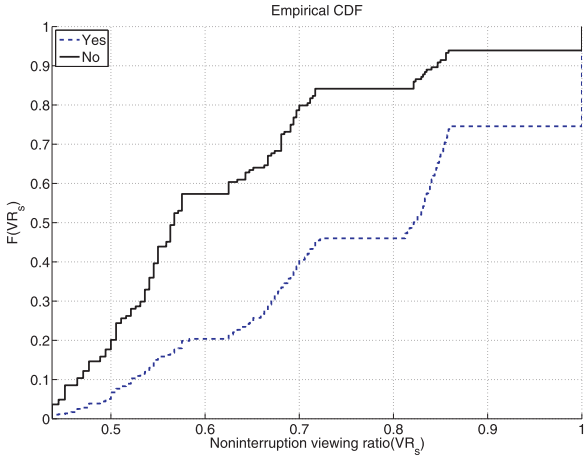
Fig. 8. Empirical cdf for RR vs. Acceptability.

of RR is similar in *A_I* and *A_IF* cases, while the CDF of VR_s shows a large difference when $VR_s \leq 0.4$ in the two cases. It indicates our previous judgment: VR_s is useful to present the presence of failures.

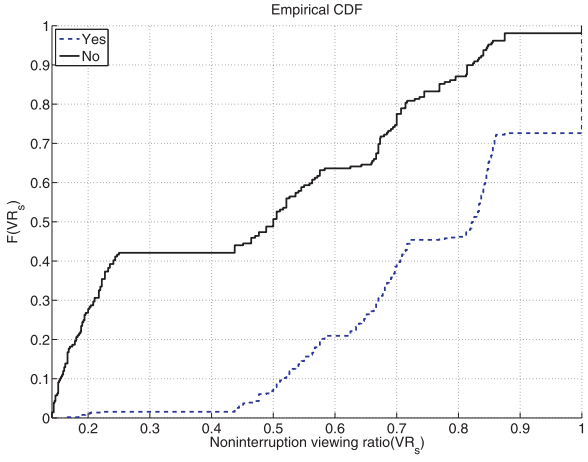
The CDF of VR under the *A_IF* scenario(VR=1 for all samples in *A_I*) is shown in Fig. 10. It is obvious that the majority of 'Yes' needs VR=1.

5.4. Machine learning classifiers comparison

Considering that TQ and acceptability in our experiments are categorical variables, we used machine learning classification methods to model the relationships among selected APMs and QoE factors. These candidate methods were Naive Bayes, Logistic Regression, k-NN Classification, and Decision Tree. K-fold-cross-



(a) *A_I* case.



(b) *A_IF* case.

Fig. 9. Empirical cdf for VR_s vs. Acceptability.

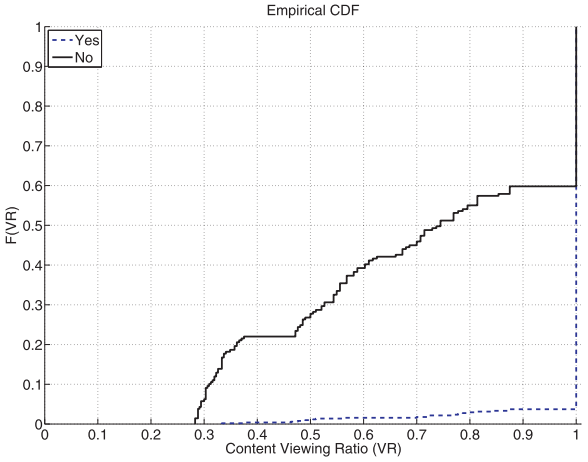
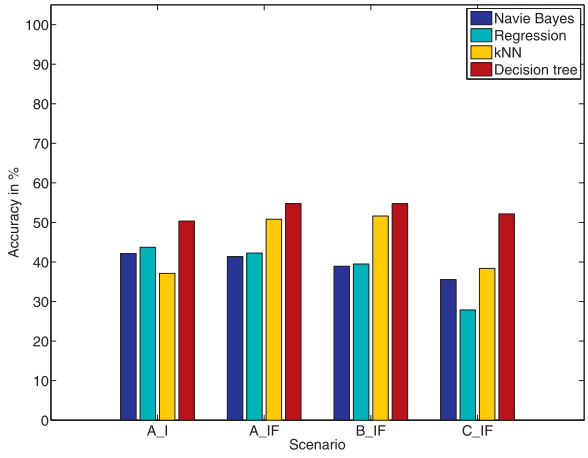


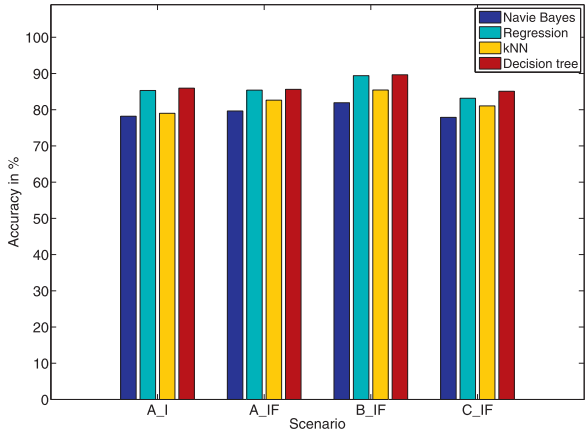
Fig. 10. Empirical cdf for VR vs. Acceptability, *A_IF* case.

validation was employed to find the one providing the highest mean accuracy.

The mean accuracy using the above four machine learning classifiers is shown in Fig. 11. Decision Tree provides the highest accuracy in almost all scenarios, which agrees with the conclusion in [23]. It is also the most stable classifier across all cases. Decision Tree is employed in the following analysis. Decision Tree



(a) TQ = f(PMs).



(b) Accept = f(PMs).

Fig. 11. Machine learning classifiers.

Table 7
Four levels of RR, VR, and VR_s based on data.

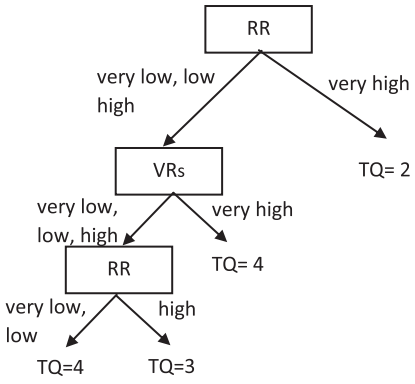
Level	RR	VR	VR_s
Very low	<0.05	<0.45	<0.3
Low	$0.05-0.125$	$0.45-0.65$	$0.3-0.6$
High	$0.125-0.2$	$0.65-0.9$	$0.6-0.9$
Very high	>0.2	>0.9	>0.9

and other classifiers we used are implemented by the scikit-learn tool [46] and MATLAB. Both the scikit-learn tool and MATLAB use Classification And Regression Tree (CART) [47] to build a tree. In general, the time complexity of CART algorithm is $O(mn\log n)$, where m is the number of the input parameters, n is the total number of data.

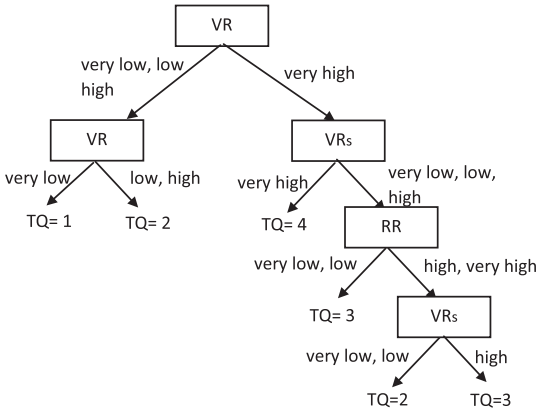
5.5. The impact of failures appearance

To clearly reveal the effect of APMs under various scenarios by classification tree, we use the compacted decision tree as Balachandran et al. [23] did. We classify RR, VR_s, and VR into four levels: (very low, low, high and very high), as shown in Table 7. The reason we divided APMs into these four levels is that these levels of APMs can classify the designed impairment/ failure types, as shown in Fig. 3.

Decision Tree is biased towards classes which occupy the majority number of samples. However, we have more positive sam-

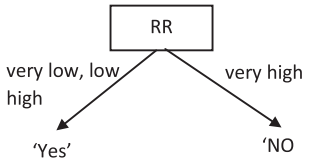


(a) TQ=f(RR, VR_s, VR), *A_I*.

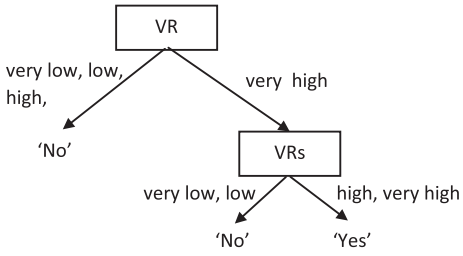


(b) TQ=f(RR, VR_s, VR), *A_IF*.

Fig. 12. Compacted decision tree for *A_I* and *A_IF*.



(a) Accept=f(RR, VR_s, VR), *A_I*.



(b) Accept=f(RR, VR_s, VR), *A_IF*.

Fig. 13. Compacted decision tree for *A_I* and *A_IF*.

Table 8
Mean accuracy (%) of decision trees, *A_I* and *A_IF*.

	A_I	*A_IF*
TQ = f(RR, VR_s)	42.64%	48.19%
TQ = f(RR, VR, VR_s)	39.68%	39.86%
Accept = f(RR, VR_s)	80.11%	81.11%
Accept = f(RR, VR, VR_s)	75.90%	78.19%

ples (TQ>3 or acceptability is 'Yes') than negative ones. In fact, we value the information hidden for the negative cases more, thus we assign a higher weight to negative cases. Note that we only plan to obtain a general understanding of the impact caused by failures, especially when the types of APMs are limited (only three metrics are considered). For developing a predictive QoE model, this strategy might not be useful.

The structure of compacted decision tree to interpret TQ = f(RR, VR, VR_s) for *A_I* and *A_IF*. is shown in Fig. 12. Note that VR=1 in all *A_I* samples, however, it is interesting to investigate whether VR will become a valuable predictor if failures appear.

According to the experimental results, if Integrity impairments appear, RR is the main predictor to decide TQ levels in the compacted tree. On the other hand, when failures are taken into account, VR plays a more important role to determine TQ levels. This proves that the impact of failures should be considered in QoE of OTT video streaming. The changed structure indicates that a predictive model for both failures and impairments should consider VR first, and the impact of RR and VR_s should be discussed separately under different VR levels.

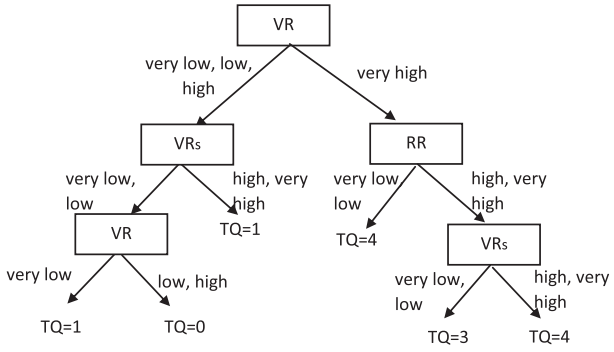
It is also important to notice that the compacted decision tree does not include all possible TQ levels. For example, TQ=1 is missed in the decision tree based on *A_I*, while TQ = 5 is missed in the case of *A_IF*.

Two are the main reasons for this phenomenon: first, the basis of Decision Tree is information gain, which leads the tree being biased to samples with larger sizes under the same conditions; and second, we compacted the level of APMs instead of using exact values, which provides a general structure for APMs with the price of losing precision of the tree. However, the compacted levels are enough to explore the general relationship between APMs and between APMs and QoE factors which is discussed in this work.

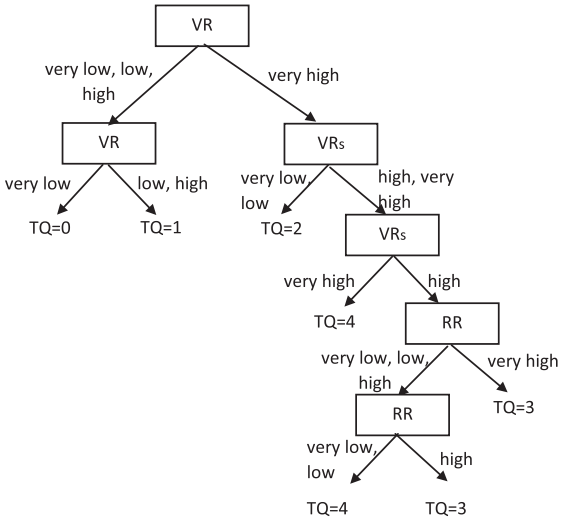
The Decision Tree of Accept = f(VR, RR_s) of *A_I* and *A_IF* is shown in Fig. 13. It is clear that the relationship between APMs of both trees is similar to the corresponding structures shown in Fig. 12.

Table 8 shows the mean accuracy of compacted decision trees generated based on (RR, VR_s) and (RR, VR, VR_s). The accuracy of trees based on (RR, VR_s) is from our previous work [48]. The 'Yes/No' choice of acceptability leads to higher accuracy.

It can be inferred that the selection of the granularity is important in deriving a predictive model, which is stated in [23]. Considering that the available types of APMs might be limited in an OTT video, it is invaluable to discuss whether to select TQ or Acceptability as the indicator for QoE. Another thing is that the accuracy of f(RR, VR, VR_s) is slightly lower than f(RR, VR_s) among all cases, even when one more predictor, VR, is added in the latter. As was explained before, the purpose of the compacted decision tree is to examine t the generic relationship between APMs. However, the decision tree is based on information entropy, which means the relationship between APMs will impact its accuracy. If a primary

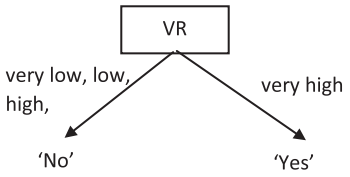


(a) TQ=f(RR, VR, VR$_s$), B_IF.

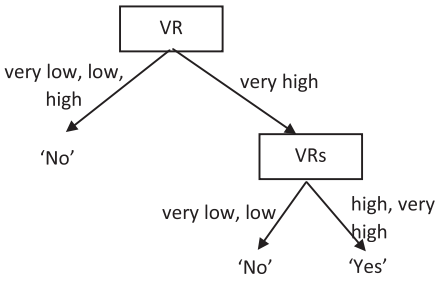


(b) TQ=f(RR, VR, VR$_s$), C_IF.

Fig. 14. Compacted decision tree for TQ = f(RR, VR$_s$), B_IF and C_IF.



(a) Accept=f(RR, VR, VR$_s$), B_IF.



(b) Accept=f(RR, VR, VR$_s$), C_IF.

Fig. 15. Compacted decision tree for Accept = f(RR, VR$_s$), B_IF and C_IF.

Table 9
Mean accuracy (%) of decision trees, B_IF and C_IF.

	B_IF	C_IF
TQ = f(RR, VR$_s$)	43.05%	43.75%
TQ = f(RR, VR, VR$_s$)	39.72%	35.69%
Accept = f(RR, VR$_s$)	82.92%	82.50%
Accept = f(RR, VR, VR$_s$)	86.94%	79.72%

Table 10
The numerical number of levels of *RR*, *VR*, and *VR$_s$*.

Symbol	Very Low	Low	High	Very high
L_{RR}	1	2	3	4
L_{VR}	1	2	3	4
L_{VR_s}	1	2	3	4

Table 11
Linear regression for *A_I*.

$MOS = f(PMs)$	MSE
$MOS = 4.6317 - 0.4217L_{RR}$	0.0020
$MOS = 3.7290 - 0.4217L_{RR} + 0.2006L_{VR_s}$	< 0.0001

model based on decision tree needs to be developed, the relationship among APMs should be explored in more details.

The structure of decision trees for TQ = f(RR, VR, VR$_s$) in Scale B and Scale C is shown in Fig. 14. The structure of the two decision trees are close to the structure shown in Fig. 12(b): VR is the primary determinant of TQ levels.

In these scenarios, the traditional ITU MOS scale is changed to the introduced extended scales. The main reason for this design is to reveal more details about the impact of Retainability and Accessibility failures. TQ tends to lower levels compared to the tree in the *A_IF* case. It is obvious that the structures of the two trees are close to TQ = f(RR, VR, VR$_s$) in *A_IF*, thus the extended scales will not impact the relationship among APMs. However, the extended point, Terrible (TQ = 0), appears in the tree indicating that users tend to evaluate failure types worse than Bad (TQ = 1), but avoid Worst Possible (TQ = −1).

The decision tree of Accept = f(RR, VR$_s$, VR) under B_IF and C_IF is shown in Fig. 15. It indicates that the impact caused by failures is stable across all cases if evaluated by Acceptability. A comparison of the accuracy of Accept = f(RR, VR$_s$, VR) and TQ = f(RR, VR$_s$, VR) is shown in Table 9 and the accuracy of Accept = f(RR, VR$_s$) and TQ = f(RR, VR$_s$) in [48]. It shows that acceptability provides higher accuracy compared to TQ.

6. Primary model discussion

It has been proved that VR$_s$ and VR should be considered when failure appears in the QoE assessment. In this part, a simple QoE model is proposed by running regression analysis to compare the usage of these new predictors.

In the following models, we used levels of APMs, and we have:

- *MOS* denotes the MOS of TQ.
- L_{RR} denotes the level of RR.
- L_{VR} denotes the level of VR.
- L_{VR_s} denotes the levels of VR$_s$.

The mapping between levels of RR, VR and VR$_s$ and numerical level for regression models are shown in Table 10.

$MOS = f(L_{RR})$ can achieve a high accuracy in linear regression analysis, as shown in Table 11. At the same time, predicting *MOS* based on L_{RR} and L_{VR_s} in the *A_IF* scenario has a high goodness of fit if we only estimate on impairment types, as shown in Table 12. However, if we add L_{VR} as a predictor and estimate im-

Table 12
Linear regression for *A_IF*, for impairment types.

$MOS = f(PMs)$	MSE
$MOS = 4.7396 - 0.4990L_{RR}$	0.0095
$MOS = 4.0499 - 0.4071L_{RR} + 0.1533L_{VR_s}$	0.0083

Table 13
Linear regression for *A_IF*, impairment and failure types.

$MOS = f(PMs)$	MSE
$MOS = 1.4568 - 0.1749L_{RR} + 0.5829L_{VR}$	0.2021
$MOS = 0.6450 + 0.3364L_{VR} + 0.4577L_{VR_s}$	0.1561
$MOS = -1.3059 + 0.4292L_{RR} + +1.0438L_{VR}$ $+1.1600L_{VR_s}$	8.8439

pairment and failure types together, the MSE (Mean Squared Error) will highly increase, especially in the $MOS = f(L_{RR}, L_{VR}, L_{VR_s})$ case, shown in Table 13. Therefore, we propose to use the QoE models in the following Eq. (5):

$$MOS = g_1(L_{VR}) \cdot f_1(L_{RR}, L_{VR_s}) + (1 - g_1(L_{VR})) \cdot f_2(L_{RR}, L_{VR_s}) \quad (5)$$

where

$$g_1(L_{VR}) = \begin{cases} 1, & \text{if } L_{VR} = 4 \\ 0 & \text{if } L_{VR} \neq 4 \end{cases} \quad (6)$$

Based on the experimental data:

$$MOS = g_1(L_{VR}) \cdot (4.0499 - 0.4071L_{RR} + 0.1533L_{VR_s}) + (1 - g_1(L_{VR})) \cdot 0.2794L_{RR} + 0.6430L_{VR_s} \quad (7)$$

and $MSE = 0.0085$.

This QoE model was tested on *B_IF* and *C_IF* and the MSE is 0.0007 and 0.0038, respectively. Because the limited types of impairments and failures, this model is a coarse-grained model for QoE estimation. However, it indicates the importance of VR and VR_s when failure appears.

7. Conclusion

In this paper, the relationship between performance metrics and QoE factors through a data-driven machine learning approach was examined. A session-based QoE model was used and two new QoE metrics were examined to evaluate the performance of the new model. Further the impairments, two failures were introduced in the experiments. Through extensive experimentation, it was found that the feature of failures requires new performance metrics to be introduced in the QoE evaluation.

Furthermore, the traditional multi-point scale was compared to the binary likert scale. According to experimental results, multiple levels are not necessary for all OTT video services. Depending on the requirement of accuracy and the purpose of QoE assessment, acceptability might be a valuable indicator of the user's perception. An extended scale was also examined and if it is necessary for the addition of failures. The extended scales can help users distinguish different TQ levels.

Finally, a primary QoE model is proposed based on the experimental results and analysis. The introduced model follows the changes in QoE due to the addition of failures without missing the impairments.

References

[1] K. Brunnström, S.A. Beker, K. De Moor, A. Dooms, S. Egger, M.-N. Garcia, T. Hossfeld, S. Jumisko-Pyykkö, C. Keimel, M.-C. Larabi, B. Lawlor, P. Le Callet, S. Möller, F. Pereira, M. Pereira, A. Perkis, J. Pibernik, A. Pinheiro, A. Raake, P. Reichl, U. Reiter, R. Schatz, P. Schelkens, L. Skorin-Kapov, D. Strohmeier, C. Timmerer, M. Varela, I. Wechsung, J. You, A. Zgank, Qualinet white paper on definitions of quality of experience, Qualinet White Paper on Definitions of Quality of Experience Output from the Fifth Qualinet Meeting, Novi Sad, March 12, 2013, 2013.

[2] A. Leon-Garcia, L. Zucherman, Generalizing MOS to assess technical quality for end-to-end telecom session, in: Globecom Workshops (GC Wkshps), 2014, 2014, pp. 681–687, doi:10.1109/GLOCOMW.2014.7063511.

[3] S. Barakovic, J. Barakovic, H. Bajric, QoE dimensions and QoE measurement of NGN services, in: Proceedings of the 18th Telecommunications Forum (TELFOR), 2010.

[4] N. Corporation, Nokia: Quality of Experience (QoE) of Mobile Services: Can It be Measured and Improved? White Paper, Nokia, 2005.

[5] ITU-T, QoE Factors in Web-Browsing, Telecommunication Standardization Sector of ITU Recommendation G.1031.

[6] L. Skorin-Kapov, Survey and challenges of QoE management issues in wireless networks, J. Comput. Netw. Commun. 2013 (2013) 28pages.

[7] M. Kunapareddy, G. Godbole, Delivering Network Centric Customer Experience, Technical Report, Tech Mahindra Limited, 2011.

[8] A. Khan, L. Sun, E. Jammeh, E. Ifeachor, Quality of experience-driven adaptation scheme for video applications over wireless networks, IET Commun. 4 (11) (2010) 1337–1347, doi:10.1049/iet-com.2009.0422.

[9] R. Imran, M. Odeh, N. Zorba, C. Verikoukis, Spatial opportunistic transmission for quality of experience satisfaction, J. Vis. Commun. Image Represent. 25 (3) (2014) 578–585, doi:10.1016/j.jvcir.2013.08.014.

[10] M. Alreshoodi, J. Woods, Survey on QoE\QoS correlation models for multimedia services, CoRR (2013) arXiv:1306.0221.

[11] ITU-R, Methodology for the Subjective Assessment of the Quality of Television Pictures, Telecommunication Standardization Sector of ITU Recommendation BT. 500-13.

[12] ITU-T, Telephone Transmission Quality, Telephone Installations, Local Line Networks, Recommendation Series P, Telecommunication standardization sector of ITU (Jan. 2007) P.10/G.100 (2006) /Amd.1.

[13] ITU-T, Quality of Experience Requirements for IPTV Services, Telecommunication Standardization Sector of ITU Recommendation G.1080.

[14] ITU-T, Subjective Video Quality Assessment Methods for Multimedia Applications, Recommendation P.910, Telecommunication standardization sector of ITU.

[15] A. Raake, S. Egger, Quality and quality of experience, in: S. Moeller, A. Raake (Eds.), Quality of Experience: Advanced Concepts, Applications and Methods, Springer, 2014.

[16] I. Wechsung, K.D. Moor, Quality of Experience versus User Experience, in: S. Moeller, A. Raake (Eds.), Quality of Experience: Advanced Concepts, Applications and Methods, Springer, 2014.

[17] J. Joskowicz, R. Sotelo, J. Lopez Arado, Comparison of parametric models for video quality estimation: towards a general model, in: Proceedings of the 2012 IEEE International Symposium on Broadband Multimedia Systems and Broadcasting (BMSB), 2012, pp. 1–7, doi:10.1109/BMSB.2012.6264245.

[18] J. Joskowicz, J.C. Ardao, A parametric model for perceptual video quality estimation, Telecommun. Syst. 49 (1) (2012) 49–62, doi:10.1007/s11235-010-9352-9.

[19] A. Dalal, A. Bouchard, S. Cantor, Y. Guo, A. Johnson, Assessing QoE of on-demand TCP video streams in real time, in: Proceedings of the 2012 IEEE International Conference on Communications (ICC), 2012, pp. 1165–1170, doi:10.1109/ICC.2012.6364073.

[20] A. Reis, J. Chakareski, A. Kassler, S. Sargento, Quality of experience optimized scheduling in multi-service wireless mesh networks, in: Proceedings of the 17th IEEE International Conference on Image Processing (ICIP), 2010, pp. 3233–3236, doi:10.1109/ICIP.2010.5651785.

[21] S. Tao, J. Apostolopoulos, R. Guerin, Real-time monitoring of video quality in IP networks, IEEE/ACM Trans. Netw. 16 (5) (2008) 1052–1065, doi:10.1109/TNET.2007.910617.

[22] O. Oyman, S. Singh, Quality of experience for HTTP adaptive streaming services, IEEE Commun. Mag. 50 (4) (2012) 20–27, doi:10.1109/MCOM.2012.6178830.

[23] A. Balachandran, V. Sekar, A. Akella, S. Seshan, I. Stoica, H. Zhang, Developing a predictive model of quality of experience for internet video, SIGCOMM Comput. Commun. Rev. 43 (4) (2013) 339–350, doi:10.1145/2534169.2486025.

[24] M. Mushtaq, B. Augustin, A. Mellouk, Empirical study based on machine learning approach to assess the QoS/QoE correlation, in: Networks and Optical Communications (NOC), 2012 17th European Conference on, 2012, pp. 1–7, doi:10.1109/NOC.2012.6249939.

[25] Y. Chen, Q. Chen, F. Zhang, Q. Zhang, K. Wu, R. Huang, L. Zhou, Understanding viewer engagement of video service in Wi-Fi network, Comput. Netw. 91 (C) (2015) 101–116, doi:10.1016/j.comnet.2015.08.006.

[26] Y. Chen, F. Zhang, F. Zhang, K. Wu, Z. Q., QoE-aware dynamic video rate adaptation, in: IEEE Global Communications Conference (GLOBECOM) 2015, 2015.

[27] T. Hoßfeld, P.E. Heegaard, M. Varela, S. Möller, QoE beyond the MOS: an in-depth look at QoE via better metrics and their relation to MOS, Qual. User Exp. 1 (1) (2016) 1–23, doi:10.1007/s41233-016-0002-1.

[28] V. Menkovski, G. Exarchakos, A. Liotta, Machine learning approach for quality of experience aware networks, in: Intelligent Networking and Collaborative Systems (INCOS), 2010 2nd International Conference on, 2010, pp. 461–466, doi:10.1109/INCOS.2010.86.

[29] W. Song, D. Tjondronegoro, Acceptability-based QoE models for mobile video, Multim. IEEE Trans. 16 (3) (2014) 738–750, doi:10.1109/TMM.2014.2298217.

[30] T. De Pessemier, K. De Moor, W. Joseph, L. De Marez, L. Martens, Quantifying the influence of rebuffering interruptions on the user's quality of experience during mobile video watching, Broadcast. IEEE Trans. 59 (1) (2013) 47–61, doi:10.1109/TBC.2012.2220231.
[31] F. Dobrian, V. Sekar, A. Awan, I. Stoica, D. Joseph, A. Ganjam, J. Zhan, H. Zhang, Understanding the impact of video quality on user engagement, in: Proceedings of the ACM SIGCOMM 2011 Conference, in: SIGCOMM '11, ACM, New York, NY, USA, 2011, pp. 362–373, doi:10.1145/2018436.2018478.
[32] R. Mok, E. Chan, R. Chang, Measuring the quality of experience of HTTP video streaming, in: Integrated Network Management (IM), 2011 IFIP/IEEE International Symposium on, 2011, pp. 485–492, doi:10.1109/INM.2011.5990550.
[33] A. Moorthy, L.K. Choi, A. Bovik, G. De Veciana, Video quality assessment on mobile devices: subjective, behavioral and objective studies, IEEE J. Sel. Top. Signal Process. 6 (6) (2012) 652–671, doi:10.1109/JSTSP.2012.2212417.
[34] M. Söderlund, Behind the satisfaction facade: an exploration of customer frustration, in: Proceedings of the 32nd European Marketing Academy Conference, 2003.
[35] W. Li, H.-U. Rehman, D. Kaya, M. Chignell, A. Leon-Garcia, L. Zucherman, J. Jiang, Video quality of experience in the presence of accessibility and retainability failures, in: Heterogeneous Networking for Quality, Reliability, Security and Robustness (QShine), 2014 10th International Conference on, 2014, pp. 1–7, doi:10.1109/QSHINE.2014.6928651.
[36] W. Li, H.-U. Rehman, M. Chignell, A. Leon-Garcia, L. Zucherman, J. Jiang, Impact of retainability failures on video quality of experience, in: Signal-Image Technology and Internet-Based Systems (SITIS), 2014 Tenth International Conference on, 2014, pp. 524–531, doi:10.1109/SITIS.2014.106.
[37] W. Li, P. Spachos, M. Chignell, A. Leon-Garcia, L. Zucherman, J. Jiang, Impact of technical and content quality on overall experience of OTT video, in: 2016 13th IEEE Annual Consumer Communications Networking Conference (CCNC), 2016, pp. 930–935, doi:10.1109/CCNC.2016.7444912.
[38] P. Spachos, W. Li, M. Chignell, A. Leon-Garcia, L. Zucherman, J. Jiang, Acceptability and quality of experience in over the top video, in: Communication Workshop (ICCW), 2015 IEEE International Conference on, 2015, pp. 1693–1698, doi:10.1109/ICCW.2015.7247424.
[39] W. Li, P. Spachos, M. Chignell, A. Leon-Garcia, J. Jiang, L. Zucherman, Capturing user behavior in subjective quality assessment of OTT video service, in: 2016 IEEE Global Communications Conference (GLOBECOM), 2016, pp. 1–6, doi:10.1109/GLOCOM.2016.7841652.
[40] J. Jiang, P. Spachos, M. Chignell, L. Zucherman, Assessing unreliability in OTT video QoE subjective evaluations using clustering with idealized data, in: 2016 Digital Media Industry Academic Forum (DMIAF), 2016, pp. 235–239, doi:10.1109/DMIAF.2016.7574940.
[41] CCITT, Concepts, Models, Objectives, Dependability Planning - Terms and Definition Related to the Quality of Telecommunication Services, ReSeries E.800, Telecommunication Standardization Sector of ITU.
[42] ITU-T, IMT-2000 References to Release 10 of GSM-Evolved UMTS Core Network, Q.1741.8, Telecommunication Standardization Sector of ITU.
[43] S. Jumisko-Pyykkö, V.K. Malamal Vadakital, M.M. Hannuksela, Acceptance threshold: A Bidimensional research method for user-Oriented quality evaluation studies, Int. J. Digit. Multim.Broadcast. 2008 (2008) 1–21, doi:10.1155/2008/712380.
[44] P. Spachos, T. Lin, W. Li, M. Chignell, A. Leon-Garcia, J. Jiang, L. Zucherman, Subjective QoE assessment on video service: laboratory controllable approach, in: 2017 IEEE 18th International Symposium on a World of Wireless, Mobile and Multimedia Networks (WoWMoM), 2017, pp. 1–9, doi:10.1109/WoWMoM.2017.7974323.
[45] ITU-T, Subjective Audiovisual Quality Assessment Methods for Multimedia Applications, Recommendation P.911, Telecommunication standardization sector of ITU.
[46] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, E. Duchesnay, Scikit-learn: machine learning in Python, J. Mach. Learn. Res. 12 (2011) 2825–2830.
[47] L. Breiman, Classification and Regression Trees, New York: Routledge, 1984.
[48] W. Li, P. Spachos, M. Chignell, A. Leon-Garcia, L. Zucherman, J. Jiang, Understanding the relationships between performance metrics and QoE for over-the-top video, in: 2016 IEEE International Conference on Communications (ICC), 2016, pp. 1–6, doi:10.1109/ICC.2016.7511100.

Weiwei Li received her B. Eng. degree from College of Electrical Engineering and Information Technology, Sichuan University, China, in 2008 and her MASc. and Ph.D. degrees from the Department of Electrical and Computer Engineering, University of Toronto, in 2011 and 2017. Her research interests include Quality of Experience for Over-The-Top Video Streaming, pattern recognition algorithm.

Petros Spachos is an Assistant Professor at School of Engineering, University of Guelph, Canada. He received the Diploma degree in Electronic and Computer Engineering from the Technical University of Crete, Greece, and the M.A.Sc. and the Ph.D. degree both in Electrical and Computer Engineering from the University of Toronto, Canada. His research interests include experimental wireless networking and mobile computing with a focus on wireless sensor networks, smart cities and the Internet of Things.

Mark Chignell received the Ph.D. degree in psychology from the University of Canterbury, Christchurch, New Zealand and the Masters degree in industrial and systems engineering from Ohio State University, Columbus, OH, USA. He is a Professor of mechanical and industrial engineering and the Director of the Knowledge Media Design Institute, University of Toronto, Toronto, Canada. His research interests include making people smarter and more effective through better design of user interfaces and applications.

Alberto Leon-Garcia is Distinguished Professor in Electrical and Computer Engineering at the University of Toronto. He is a Fellow of the Institute of Electronics an Electrical Engineering "For contributions to multiplexing and switching of integrated services traffic". Professor Leon-Garcia is author of the textbooks: Probability and Random Processes for Electrical Engineering, and Communication Networks: Fundamental Concepts and Key Architecture. His research is on application platforms for smart applications including smart cities.

Leon Zucherman is currently a Research Associate at the University of Toronto, Canada. He holds a Ph.D. and a M.A.Sc. degree from University of Toronto. His research interests include Customer Experience, Quality of Experience and Quality of Service, mainly in the telecom sector.

Jie Jiang has over 20 years' experience in OSS, network planning and engineering. Jie also specializes in business and system analysis, he has been working in Customer Experience area for over 10 years.