

ITU-T

TELECOMMUNICATION
STANDARDIZATION SECTOR
OF ITU

P.917

(01/2019)

SERIES P: TELEPHONE TRANSMISSION QUALITY,
TELEPHONE INSTALLATIONS, LOCAL LINE
NETWORKS

Audiovisual quality in multimedia services

Subjective test methodology for assessing impact of initial loading delay on quality of experience

Recommendation ITU-T P.917

ITU-T P-SERIES RECOMMENDATIONS

TELEPHONE TRANSMISSION QUALITY, TELEPHONE INSTALLATIONS, LOCAL LINE NETWORKS

Vocabulary and effects of transmission parameters on customer opinion of transmission quality	Series	P.10
Voice terminal characteristics	Series	P.30
		P.300
Reference systems	Series	P.40
Objective measuring apparatus	Series	P.50
		P.500
Objective electro-acoustical measurements	Series	P.60
Measurements related to speech loudness	Series	P.70
Methods for objective and subjective assessment of speech quality	Series	P.80
Methods for objective and subjective assessment of speech and video quality	Series	P.800
Audiovisual quality in multimedia services	Series	P.900
Transmission performance and QoS aspects of IP end-points	Series	P.1000
Communications involving vehicles	Series	P.1100
Models and tools for quality assessment of streamed media	Series	P.1200
Telemeeting assessment	Series	P.1300
Statistical analysis, evaluation and reporting guidelines of quality measurements	Series	P.1400
Methods for objective and subjective assessment of quality of services other than speech and video	Series	P.1500

For further details, please refer to the list of ITU-T Recommendations.

Recommendation ITU-T P.917

Subjective test methodology for assessing impact of initial loading delay on quality of experience

Summary

Recommendation ITU-T P.917 defines a procedure for conducting behavioral studies targeted at investigating video streaming performance and its relation to users' quality of experience (QoE). The studies are to be conducted in controlled environments. Subjects are exposed to different initial loading delay conditions and other quality degradations typical for video streaming, and subjects may be asked to rate audiovisual quality or their experience. Additionally, their behaviour as a response to long loading times (i.e., aborting the video playback during its loading phase) may be investigated.

History

Edition	Recommendation	Approval	Study Group	Unique ID*
1.0	ITU-T P.917	2019-01-13	12	11.1002/1000/13843

Keywords

Initial loading delay, quality of experience, video streaming.

* To access the Recommendation, type the URL <http://handle.itu.int/> in the address field of your web browser, followed by the Recommendation's unique ID. For example, <http://handle.itu.int/11.1002/1000/11830-en>.

FOREWORD

The International Telecommunication Union (ITU) is the United Nations specialized agency in the field of telecommunications, information and communication technologies (ICTs). The ITU Telecommunication Standardization Sector (ITU-T) is a permanent organ of ITU. ITU-T is responsible for studying technical, operating and tariff questions and issuing Recommendations on them with a view to standardizing telecommunications on a worldwide basis.

The World Telecommunication Standardization Assembly (WTSA), which meets every four years, establishes the topics for study by the ITU-T study groups which, in turn, produce Recommendations on these topics.

The approval of ITU-T Recommendations is covered by the procedure laid down in WTSA Resolution 1.

In some areas of information technology which fall within ITU-T's purview, the necessary standards are prepared on a collaborative basis with ISO and IEC.

NOTE

In this Recommendation, the expression "Administration" is used for conciseness to indicate both a telecommunication administration and a recognized operating agency.

Compliance with this Recommendation is voluntary. However, the Recommendation may contain certain mandatory provisions (to ensure, e.g., interoperability or applicability) and compliance with the Recommendation is achieved when all of these mandatory provisions are met. The words "shall" or some other obligatory language such as "must" and the negative equivalents are used to express requirements. The use of such words does not suggest that compliance with the Recommendation is required of any party.

INTELLECTUAL PROPERTY RIGHTS

ITU draws attention to the possibility that the practice or implementation of this Recommendation may involve the use of a claimed Intellectual Property Right. ITU takes no position concerning the evidence, validity or applicability of claimed Intellectual Property Rights, whether asserted by ITU members or others outside of the Recommendation development process.

As of the date of approval of this Recommendation, ITU had not received notice of intellectual property, protected by patents, which may be required to implement this Recommendation. However, implementers are cautioned that this may not represent the latest information and are therefore strongly urged to consult the TSB patent database at <http://www.itu.int/ITU-T/ipr/>.

© ITU 2019

All rights reserved. No part of this publication may be reproduced, by any means whatsoever, without the prior written permission of ITU.

Table of Contents

		Page
1	Scope.....	1
2	References.....	1
3	Definitions	2
	3.1 Terms defined elsewhere	2
	3.2 Terms defined in this Recommendation.....	2
4	Abbreviations and acronyms	2
5	Conventions	2
6	Background.....	2
7	Factors influencing user behaviour and ratings	3
	7.1 Test platform	4
	7.2 Source content	4
	7.3 Test conditions.....	5
	7.4 Assignment of SRCs to HRCs.....	6
	7.5 Overall test procedure.....	6
	7.6 Subjective test environment and set-up	10
	7.7 Subjects and subjective test control.....	12
	Appendix I – Example instructions to subjects.....	14
	Appendix II – Test questionnaire.....	16
	Appendix III – Pilot test on the impact of initial loading delays	18
	Bibliography.....	25

Introduction

Initial loading delay during video streaming is known to impact the quality of experience (QoE) of end-users. However, the specific relationship between initial loading delay and a user's likelihood of aborting a playback – or the perceived overall quality of the initial loading delay and/or entire media session – varies for different services. Furthermore, it depends on a user's expectations. It is important for Internet service providers (ISPs) to be able to assess the impact of initial loading delay on users' QoE, as numbers quantifying these relationships are not readily available.

In this Recommendation, a procedure for conducting studies with human test subjects is described. These studies are called behavioral, since interaction with a (simulated) video streaming service is the main focus. Results from a study conducted following this Recommendation may be used for several purposes, including but not limited to developing prediction models that estimate:

- the likelihood of users abandoning a video playback during its initial loading phase; and/or
- the perceived overall quality of a video playback impacted by initial loading, stalling, and audiovisual quality changes; and/or
- the perceived quality of the initial loading experience.

The test procedures specified in this Recommendation aim to bridge the gap between traditional video quality assessment methods such as those specified in ITU-T P.910 and large-scale data collections that over-the-top (OTT) video providers may conduct.

First, background and related work on tests aimed at assessing the impact of initial loading delay are described. A general overview of factors that influence user behaviour and subsequent ratings is given. Then, this Recommendation, ITU-T P.917, describes a test procedure, which is to be conducted in a controlled environment.

Recommendation ITU-T P.917

Subjective test methodology for assessing impact of initial loading delay on quality of experience

1 Scope

This Recommendation describes a test methodology for use in controlled environments with human subjects. The scope of this Recommendation is currently limited to the following factors:

- test environment: controlled environment (see [ITU-T P.913], clause 8.1);
- video length: short videos (< 5 mins);
- simulated platform: video on demand (VOD);
- test devices: PCs or mobile phones.

The technical scope is limited to progressive download streaming and adaptive streaming type services (using reliable transport), which includes:

- over-the-top (OTT) services, as well as operator managed video services (over TCP);
- video over both mobile and fixed connections;
- the protocols HTTP/TCP/IP, RTMP/TCP/IP, HLS/HTTP/TCP/IP, and DASH/HTTP/TCP/IP. This Recommendation is agnostic to the specific network delivery method (HTTP or DASH or other), with one exception that it assumes reliable delivery (TCP/IP) that results in apparent loading times (indicated with a "buffering" symbol to the user);
- video services typically using container formats such as Flash (FLV), MP4, WebM, 3GP, and MPEG2-TS. This test method is agnostic to the type of container format.

The methods described in this Recommendation may be used in other technical contexts where users browse content and have to wait for media playback (e.g., music streaming), but it has not been validated in such contexts.

2 References

The following ITU-T Recommendations and other references contain provisions which, through reference in this text, constitute provisions of this Recommendation. At the time of publication, the editions indicated were valid. All Recommendations and other references are subject to revision; users of this Recommendation are therefore encouraged to investigate the possibility of applying the most recent edition of the Recommendations and other references listed below. A list of the currently valid ITU-T Recommendations is regularly published. The reference to a document within this Recommendation does not give it, as a stand-alone document, the status of a Recommendation.

- | | |
|----------------|---|
| [ITU-T P.10] | Recommendation ITU-T P.10/G.100 (2017), <i>Vocabulary for performance, quality of service and quality of experience.</i> |
| [ITU-T P.910] | Recommendation ITU-T P.910 (2008), <i>Subjective video quality assessment methods for multimedia applications.</i> |
| [ITU-T P.913] | Recommendation ITU-T P.913 (2016), <i>Methods for the subjective assessment of video quality, audio quality and audiovisual quality of Internet video and distribution quality television in any environment.</i> |
| [ITU-T P.1203] | Recommendation ITU-T P.1203 (2017), <i>Parametric bitstream-based quality assessment of progressive download and adaptive audiovisual streaming services over reliable transport.</i> |

[ITU-T P.1203.1] Recommendation ITU-T P.1203.1 (2019), *Parametric bitstream-based quality assessment of progressive download and adaptive audiovisual streaming services over reliable transport – Video quality estimation module*.

3 Definitions

3.1 Terms defined elsewhere

This Recommendation uses the following terms defined elsewhere:

3.1.1 initial loading delay [ITU-T P.1203].

3.1.2 stalling [ITU-T P.1203].

3.1.3 integral quality [ITU-T P.1203].

3.1.4 quality of experience [ITU-T P.10].

3.2 Terms defined in this Recommendation

This Recommendation defines the following term:

3.2.1 overall quality: Synonym for integral quality as defined in clause 3.1.3.

3.2.2 quality of initial loading experience: The degree to which a user's expectations of the initial loading delay are fulfilled.

4 Abbreviations and acronyms

This Recommendation uses the following abbreviations and acronyms:

ACR	Absolute Category Rating
FHD	Full High-Definition
HD	High Definition
HRC	Hypothetical Reference Circuit
ILQ	Initial Loading Quality
MOS	Mean Opinion Score
OTT	Over-The-Top
PVS	Processed Video Sequence
QHD	Quad HD (2560 × 1440)
QoE	Quality of Experience
SRC	Source
VOD	Video on Demand

5 Conventions

None.

6 Background

In the literature, it has been shown that user experience with regard to initial loading delay is impacted by several factors, including human, system, and contextual factors (for factors influencing quality of experience (QoE), see [b-Reiter]). These have been analyzed in various studies, such as [b-Dobrian]

and [b-Krishnan] for the case of large-scale data from real streaming services. The studies have shown that previous experiences of users (leading to expectations about their Internet performance) heavily influence their engagement patterns. The problem with these studies is that the underlying data are not available and hence do not allow building a model. Also, mean opinion scores (MOSs) are not captured for these types of data.

In a laboratory context, there have been attempts at presenting users with patterns of video delivery degradations and studying their behavioral responses or quality ratings, such as [b-Mok] and [b-Robitza]. [b-Mok] captured quality rating responses plus user behaviour (such as seeking or pausing the video) from participants, but the paper is not detailed enough regarding the test paradigm description in order to allow reproducing the experiment.

[b-Robitza] found that a third of test participants – when they do not know that initial loading or stalling will happen – will react differently than in normal life, thus not showing the expected behaviour. The users had not been informed that problems will occur when attempting to load a video. This test series, showed that it is possible to elicit realistic responses from video loading degradations, such as getting people to reload the browser window or to select another video, but a systematic test of different loading patterns will be impossible with this paradigm, without testing a large number of people.

To develop a model that predicts a subject's initial loading experience for a video session, a MOS methodology may be used. This is in line with other models that use the MOS as a simplification for quantifying the user's QoE, without giving the user any other means to indicate a bad experience. However, in practice, the experience of the user may translate to a certain corrective action (e.g., cancelling the video playback). Hence, the option of quitting the video loading process will be given to the test subjects.

7 Factors influencing user behaviour and ratings

When gathering MOS or assessing user abandonment, it is expected that users will be influenced by several factors during the test. These factors may bias the ratings or user behaviour, making ratings either more or less critical when compared to a rating that would be taken in real-life, or leading to users not behaving as they would normally do, for example at home. Table 1 lists factors influencing a user's rating.

Table 1 – Factors influencing ratings

Factor	Comment / Explanation	Possible influence on ratings	Possible alleviation
Intrinsic motivation to watch video	Do users want to watch the upcoming video because they are interested in it or the continuation of the test?	Might yield less critical ratings if users do not want to watch the video and do not care about its content	Show more interesting content; do not repeat content
Extrinsic motivation to watch video	Are users being motivated to watch the video?	Might yield less critical ratings if users have no reason to watch the video	Give users a realistic task that relates to video content; do not repeat content
Test hypothesis communicated to users	What do subjects think this test is about? (What do the researchers want to know?) Subjects may want to "please" experimenter	Might skew ratings depending on whether subjects want to fulfill this hypothesis	Give written instructions to subjects; be clear about hypothesis

Table 1 – Factors influencing ratings

Factor	Comment / Explanation	Possible influence on ratings	Possible alleviation
Realism of the test environment	How realistic / ecologically valid is the simulated test environment? (e.g., presenting within a real browser vs. just video playback software)	Might skew ratings	Increase realism of test environment platform (software); be clear about simulated physical context (e.g., home)
Assumption of usage context	What are subjects told they should <i>imagine</i> they are doing?	Might skew ratings	Be clear about usage context (e.g., VOD vs. live vs. duration of content)

The following clause describes a method to gather user ratings and observe user behaviour (e.g., quitting during video loading) in a laboratory context. Where indicated, modifications to the method are allowed. If deviations from the method are made, they must be properly documented.

7.1 Test platform

The test must be conducted on a platform that simulates a real video platform with which users may already be familiar. It should therefore use similar navigation patterns and have similar performance. This is expected to yield more realistic ratings and behaviour.

The test platform may be a PC-based or mobile phone-based platform. It must be stated that different browsing and interaction patterns are commonly used for such platforms.

The user interface of both platforms should have two key components: a navigation bar and a video grid. Through the navigation bar, subjects can browse different video categories, and the related videos will be shown on the video grid as a list of thumbnails.

When subjects select a specific video by clicking its thumbnail, the video will start to play after an initial loading period. The subject may be allowed to abort the video playback at any time during the initial loading period, through either a dedicated button in the interface (e.g., an X button), or a hardware button on the device (e.g., a "back" button).

7.2 Source content

7.2.1 Genres

The audiovisual material sources (SRCs) must be representative of typical video clips on major VoD websites. The SRCs may, for example, belong to any of the following genres:

- movies, movie trailers;
- TV shows;
- sports;
- music videos;
- animation;
- news;
- documentaries;
- comedies.

The test platform may expose these genres as navigation categories to enable subjects to browse through the content based on their interests.

7.2.2 Characteristics and length

The SRCs must provide well-synchronized audio and video content. Lip-synchronicity must be guaranteed for non-dubbed video material and must be good enough for dubbed video material. SRCs should also provide a natural combination of the presented audio and video signal content. Dubbed video material is allowed in countries where users typically watch dubbed video.

It is recommended to use interesting and/or engaging SRCs, such that subjects will be presented videos that they will be motivated to watch. This is expected to increase the validity of the captured behaviour and ratings.

For the scope of the presented test methodology, the duration of each video sample should be less than three mins.

7.3 Test conditions

During the test procedure, subjects may be shown different audiovisual quality and different combinations of initial loading delays and stalling with varying length and frequency of the latter. A particular instantiation of initial loading delay, stalling, and audiovisual quality profile is called a hypothetical reference circuit (HRC).

7.3.1 Types of degradations

The audiovisual quality may be intentionally degraded using different video codecs, and/or by reducing video resolution, bitrate, and/or framerate, as typically done for HTTP adaptive streaming services.

Within an HRC, audiovisual quality may change (quality switching) to simulate network bandwidth fluctuations. Typical quality switching patterns may be included in the HRC design, such as ramp-ups (switching from low to high quality after initial loading).

Initial loading delays may vary between short (0.1 seconds) and long (60 seconds).

Stalling events may vary between short (0.1 seconds) and long (60 seconds). There is no limit to the number of stalling events, but they should be kept reasonable.

7.3.2 Resolutions, codecs and codec settings

The encoding schemes of video streaming providers differ depending on the type of streamed video and the video codecs being used. The particular choice of video codec and encoding settings will therefore depend on the application under study.

Examples of common video coding resolutions and video bitrate ranges can be found in [ITU-T P.1203.1].

The lowest and highest video resolutions used in a test may differ between mobile and PC platforms, respectively. For example, a mobile phone may not be able to display 2160p video.

7.3.3 Reference conditions

Table 2 lists possible reference initial loading and video coding conditions (HRCs) to be shown in the test. These should be shown to every subject in order to produce anchors for high, medium and low quality ratings, respectively.