International Telecommunication Union

ITU-T P.1203

TELECOMMUNICATION
STANDARDIZATION SECTOR
OF ITU

(10/2017)

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

Models and tools for quality assessment of streamed media

Parametric bitstream-based quality assessment of progressive download and adaptive audiovisual streaming services over reliable transport

Recommendation T P.1203

Recommendation ITU-T P.1203

Parametric bitstream-based quality assessment of progressive download and adaptive audiovisual streaming services over reliable transport

Summary

Recommendation ITU-T P.1203 provides the introductory document for a set of documents that describe model algorithms for monitoring the integral media session quality for TCP-type video streaming. The models comprise modules for short-term audio- and video-quality estimation. The per-one-second outputs of these short-term modules are integrated into estimates of audiovisual quality and, together with information about initial loading delay and media playout stalling events, further integrated into the final model output, the estimate of integral quality. The respective ITU-T work item has formerly been referred to as P.NATS (Parametric non-intrusive assessment of TCP-based multimedia streaming quality).

The structure of the set of recommendations reflects the different functionality of modules described in each document:

– ITU-T P.1203: Introductory document (this Recommendation)
– ITU-T P.1203.1: Video quality estimation module (short-term, providing per-one-second output information)
– ITU-T P.1203.2: Audio quality estimation module (short-term, providing per-one-second output information)
– ITU-T P.1203.3: Audiovisual integration and integration of final score, reflecting remembered quality for viewing sessions between 30 s and 5 min duration

The input used by the models consists of information obtained by prior stream inspection. Four different levels of inspection are included, resulting in models of different complexity both of the input information and the model algorithms, which are called "modes of operation" in the following:

– Mode 0: Information obtained from meta-information available during progressive download or adaptive streaming, for example from manifest files used in DASH, about codec and bitrate, and initial loading delay and stalling.
– Mode 1: All information from Mode 0, with additional video and audio frame information based on packet header inspection
– Mode 2: All information from Mode 1, and up to 2% (in Bytes) of the overall media stream information based on deep packet inspection and partial bitstream parsing
– Mode 3: All information from Mode 1, and complete media stream information based on bitstream parsing

The ITU-T P.1203-series of Recommendations addresses two application areas, which are respectively indicated in the module-related Recommendations [ITU-T P.1203.1], [ITU-T P.1203.2], [ITU-T P.1203.3]:

– Large-screen presentation as with fixed-network video streaming
– Mobile streaming on handheld devices such as smartphones

The ITU-T P.1203 module algorithms are no-reference, bitstream-based models.

History

Edition	Recommendation	Approval	Study group	Unique id[a]
1.0	ITU-T P.1203	2016-11-29	5	11.1002/1000/13158
1.1	ITU-T P.1203 (2016) Amd. 1	2017-01-19	12	11.1002/1000/13166
2.0	ITU-T P.1203	2017-10-29	12	11.1002/1000/13399

a) 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.

Keywords

Adaptive streaming, audio, audiovisual, IPTV, mean opinion score (MOS), mobile TV, mobile video, monitoring, multimedia, progressive download, QoE, TV, video.

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 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/.

Table of Contents

Page

Recommendation ITU-T P.1203

Parametric bitstream-based quality assessment of progressive download and adaptive audiovisual streaming services over reliable transport

1. Scope

This Recommendation describes a set of objective parametric quality assessment modules that together can be used to form a complete model to predict the impact of audio and video media encodings and observed IP network impairments on quality experienced by the end-user in multimedia streaming applications. The addressed streaming techniques comprise progressive download as well as adaptive streaming, for both mobile and fixed network streaming applications.

Four modes are defined to cover a range of use-cases, from monitoring streams where the video payload is fully encrypted through to unencrypted streams, and where there are no limitations on processing power so that deep packet inspection is possible.

The model described is restricted to information provided to it by an appropriate bit stream analysis module or set of modules. The model described here is applicable to progressive download and adaptive streaming, where the quality experienced by the end user is affected by audio-coding and/or video degradations due to coding, spatial re-scaling or variations in video frame rates, as well as delivery degradations due to initial loading delay, stalling (which are both caused by rebuffering at the client), and media adaptations. Here, a "media adaptation" refers to events where the player switches video playback between a known set of media quality levels while adapting to network conditions. Each of the quality levels typically differs in a significant video and/or audio (and thus audiovisual) quality change. These quality changes are most readily observed by changes in bitrate, resolution, frame rate, and similar attributes.

The model predicts a mean opinion score (MOS) on a 5-point absolute category rating (ACR) scale (see [ITU-T P.910]) as a global multi-media MOS score (as defined in [ITU-T P.911], for instance). In addition, the well-defined modules provide several diagnostic outputs.

The primary applications for this model are monitoring of transmission quality for operations and maintenance purposes. The ITU-T P.1203 model for adaptive- and progressive-download-type media streaming may be deployed both in end-point locations and at mid-network monitoring points. Note, however, that the present Recommendation only describes the model that maps input parameters obtained from a probe located at a specific point in the network or in the client to the global multimedia MOS-scores and related quality diagnostic indicators, as described above.

The model associated with this Recommendation cannot provide a comprehensive evaluation of transmission quality as perceived by an *individual end-user* because its scores reflect the perceived impairments due to coded audiovisual media data being transmitted over an IP connection with certain performance, and does not include specific terminal devices or user-specific. The scores predicted by a general quality model necessarily reflect *average perceptual impairments*.

Effects such as those due to audio levels, signal noise and effects due to source generation such as video shake or certain colour properties (and other similar video or audio factors) and other impairments related to the payload are not reflected in the scores computed by this model. Therefore, it is possible to have high scores with this model, yet have a poor overall stream quality.

As a consequence, this Recommendation can be used for applications such as:

– In-service quality monitoring for specific IP-based audiovisual services, as specified in more detail below.

– Benchmarking of different service implementations. However, it cannot be used for direct benchmarking of different encoder implementations.

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 G.1022]*Buffer models for media streams on TCP transport*, 1st edition.

[ITU-T H.264]*Advanced video coding for generic audiovisual services*, 14th edition.

[ITU-T P.1201.1]*Parametric non-intrusive assessment of audiovisual media streaming quality – Lower resolution application area*, 1st edition.

[ITU-T P.1201.2]*Parametric non-intrusive assessment of audiovisual media streaming quality – Higher resolution application area*, 1st edition.

[ITU-T P.1202]*Parametric non-intrusive bitstream assessment of video media streaming quality*, 1st edition.

[ITU-T P.1202.1]*Parametric non-intrusive bitstream assessment of video media streaming quality – Lower resolution application area*, 1st edition.

[ITU-T P.1203.1]*Parametric bitstream-based quality assessment of progressive download and adaptive audiovisual streaming services over reliable transport – Video quality estimation module*, 3rd edition.

[ITU-T P.1203.2]*Parametric bitstream-based quality assessment of progressive download and adaptive audiovisual streaming services over reliable transport – Audio quality estimation module*, 2nd edition.

[ITU-T P.1203.3]*Parametric bitstream-based quality assessment of progressive download and adaptive audiovisual streaming services over reliable transport – Quality integration module*, 3rd edition.

[ITU-T P.1401]*Methods, metrics and procedures for statistical evaluation, qualification and comparison of objective quality prediction models*, 2nd edition.

[ITU-T P.800.1]*Mean opinion score (MOS) terminology*, 4th edition.

[ITU-T P.910]*Subjective video quality assessment methods for multimedia applications*, 5th edition.

[ITU-T P.911]*Subjective audiovisual quality assessment methods for multimedia applications*, 1st edition.

3. Definitions

3.1. Terms defined elsewhere

This Recommendation uses the following terms defined elsewhere:

3.1.1. mean opinion score (MOS): [ITU-T P.800.1]

3.1.2. rebuffering: [ITU-T G.1022]

3.1.3. stalling (or stall): [ITU-T G.1022]

3.2. Terms defined in this recommendation

This Recommendation defines the following terms:

3.2.1. model, model algorithm: An algorithm with the purpose of estimating the subjective (perceived) quality of a media sequence.

3.2.2. sequence: An audiovisual stream composed of multiple non-overlapping segments.

3.2.3. bitstream: The part of an IP-based transmission where the actual audiovisual, video, or audio content is available in encoded and packetized form.

3.2.4. media adaptation: Events where the player switches video playback between a known set of media quality levels while adapting to network conditions, by downloading and decoding individual segments in sequence.

3.2.5. initial loading delay: Refers to the time in seconds between the initiation of video playback by the user and the actual start of the playback. In the scope of this Recommendation, initial loading delay and stalling during playback are distinguished.

3.2.6. output sampling interval: A 1-second duration of parsed video or audio (stalling is not considered part of this time), where 1 s output shall correspond to rating of 10 s sequence that has the same characteristics as the 1 s under consideration. The output sampling interval of Pa and Pv must match what the Pq module expects as input.

3.2.7. adaptation set: Refers to a set of distinct media quality levels to be used for HTTP adaptive streaming, between which the player can perform media adaptation.

3.2.8. media quality level: A particular encoding setting applied to a video or audio stream.

3.2.9. segment: An audiovisual file belonging to one particular media quality level.

3.2.10. stalling: Stalling is caused by rebuffering events at the client side, which could be a result of video data arriving late. Usually, rebuffering events are indicated to the viewer, e.g., in the form of a spinning wheel, and result in stalling of the media playout.

3.2.11. integral quality: The quality as perceived by a subject in a subjective test, which corresponds to the scope of this Recommendation. Artefacts presented in the subjective tests typically include a combination of audio compression, video compression, and stalling effects.

4. Abbreviations and acronyms

This Recommendation uses the following abbreviations and acronyms:

AAC	Advanced Audio Coding
AAC-LC	Advanced Audio Coding – Low Complexity
AC3	Audio Coding 3
ACR	Absolute Category Rating
AMR-NB	Adaptive Multi-Rate – Narrowband
AMR-WB	Adaptive Multi-Rate – Wideband
ARQ	Automatic Repeat Request
DASH	Dynamic Adaptive Streaming over HTTP
GOP	Group of Pictures

HAS	HTTP-based adaptive streaming
HD	High Definition
HE-AAC	High-Efficiency Advanced Audio Coding
HTTP	Hypertext Transfer Protocol
I-	Inline-(frame)
MOS	Mean Opinion Score
MPEG	Motion Pictures Expert Group
PCAP	Packet Capture Format
PCC	Pearson Correlation Coefficient
PVS	Processed Video Sequence
QoE	Quality of Experience
RMSE	Root Mean Square Error
RTP	Real-time Transport Protocol
RTSP	Real Time Streaming Protocol
SD	Standard Definition
TCP	Transmission Control Protocol
TS	Transport Stream

5. Conventions

This Recommendation uses the following conventions:
– Pa designates the audio quality estimation module [ITU-T P.1203.2].
– Pv designates the video quality estimation module (see [ITU-T P.1203.1]).
– Pq designates the quality integration module (see [ITU-T P.1203.3]).

6. Areas of application

The application areas for ITU-T P.1203 are:
– Progressive download streaming and adaptive streaming (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. Note that the model is agnostic to the specific network delivery method (HTTP or DASH or other), with one exception that it assumes reliable delivery (TCP/IP).
  • Video services typically using container formats such as Flash (FLV), MP4, WebM, 3GP, and MPEG2-TS. Note that the model is agnostic to the type of container format.

6.1. Application range for the models

Table 1 below shows the application range of the model based on what the model has actually been developed for.

Table 1 — Application areas, test factors, and coding technologies for which ITU-T P.1203 has been verified and is known to produce reliable results

Applications for which the model is intended
In-service mid-point or client-side monitoring of encrypted HTTP/TCP based VoD/Live streaming services (mode 0, mode 1). This assumes that the required input for mode 0 or mode 1 is made available for the model, despite the stream being encrypted. See Table 4 for details.
In-service mid-point or client-side monitoring of non-encrypted HTTP/TCP based VoD/Live streaming services (mode 0, mode 1, mode 2 and mode 3).
Test factors for which the model has been validated
Video compression degradations: ITU-T H.264/AVC High profile, 75 kbit/s – 12.5 Mbit/sFor details regarding codec parameters see the Pv module recommendation [ITU-T P.1203.1]
Audio compression degradations tested during standard development: AAC-LC, 32-196 kbit/sFor details regarding codec parameters see the audio module Pa [ITU-T P.1203.2] NOTE: The audio quality module Pa is assumed to be valid also for other codecs, since it is identical to the audio coding component in [ITU-T P.1201.2] and [ITU-T_P.1201], which has been tested for a larger number of audio codecs. Further audio codecs validated as part of the development of [ITU-T_P.1201] are, with the bitrate range from 24-196 kbit/s: AAC-LC, HE-AACv2, AC3, MPEG-LII. See [ITU-T P.1203.2] for details.
Video content: Video contents of different spatio-temporal complexityFor details regarding tested video content see the Pv module [ITU-T P.1203.1]
Initial loading delay and stalling degradations: For details regarding specifics of initial loading delay and stalling see the Pq module [ITU-T P.1203.3]
Display Resolutions: Full HD (1920x1080)
Display device: PC/TV monitors, mobile phone (Samsung Galaxy S5)
Media adaptation: Video quality variations caused by switching between different quality levels. For details regarding quality layer properties see [ITU-T P.1203.1]
Frame Rates: 8-30 frames per second

Table 2 — Application areas, test factors, and coding technologies for which further investigation of ITU-T P.1203 is needed

Test factors for which the model has not been validated
Broad variations in audio quality; models were not validated for poor audio quality together with high video quality. Audio bitrate was varied but audio quality hardly seems to change/affect the overall audiovisual quality score.

Table 3 — Application areas, test factors, and coding technologies for which ITU-T P.1203 is not intended to be used

Applications for which the model is not intended
In-service monitoring of video UDP-based streaming, where packet loss introduces visible quality degradations
Direct comparison/benchmarking of encoder implementations, and thus of services that employ different encoder implementations
Evaluation of visual quality including display/device properties
Test factors for which the model should not be applied
Audio/video sync distortions
Packet loss distortions
Video codecs for which the model is not validated (MPEG2, ITU-T H.265, VP9, etc.)
Transcoding solutions
The effects of noise, delay, colour correctness

6.2. Modes of operation

The modes of operation are defined in Table 4, which also provides more information on input. Additional details are available in [ITU-T P.1203.1]. Meta-data is defined here as being header information and information on the I.GEN interface as defined in clause 7.1.

Table 4 — ITU-T P.1203 modes of operation

Mode	Encryption	Input	Complexity
0	Encrypted media payload and media frame headers	Meta-data	Low
1	Encrypted media payload	Meta-data and frame size/type information	Low
2	No encryption	Meta-data and up-to 2% of the media stream	Medium
3	No encryption	Meta-data and any information from the video stream	Unlimited

7. Building blocks

The module layout of the ITU-T P.1203 model is depicted in Figure 1.

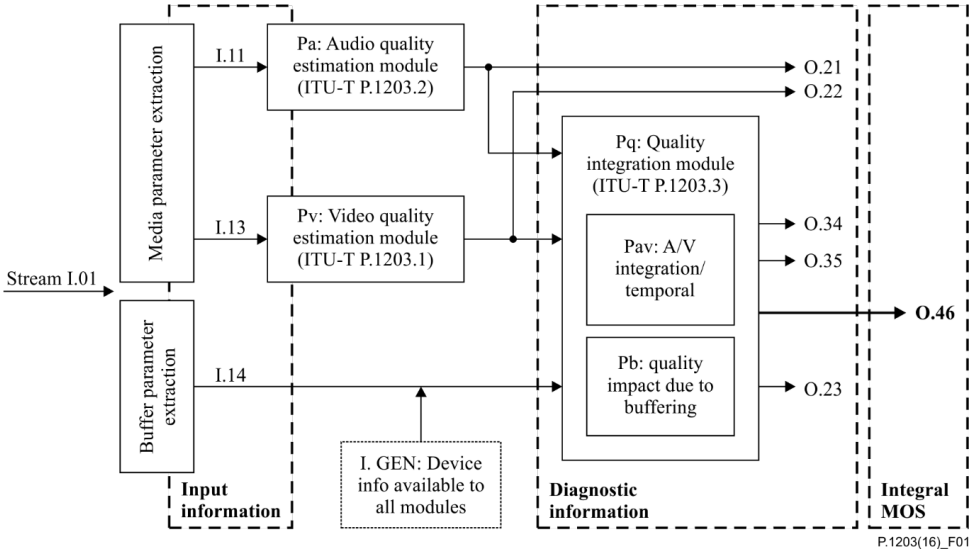

Figure 1 — Building blocks of the ITU-T P.1203 model

7.1. Model input interfaces

The ITU-T P.1203 model will receive media information and prior knowledge about the media stream or streams. In various modes of operation, the following inputs may be extracted or estimated in different ways, which is outside the scope of this Recommendation but may be added in future annexes. The model receives the following input signals regardless of the mode of operation:

I. GEN Display resolution and device type. The device type is defined as follows:
- PC/TV: screen size 24 inches or larger and smaller than or equal to 100 inches.
- Mobile: screen size 10 inches or smaller.

I.11 Audio coding information, as specified in Table 5, entries "I.11".

I.13 Video coding information, as specified in Table 5, entries "I.13".

I.14 Initial loading delay and stalling event information as described in Table 5, entries "I. 14".

7.2. Specification of Inputs I.11, I.13 and I.14

Table 5 — I.11, I.13 and I.14 inputs description

Id	Description	Values	Frequency	Available to modes
I.GEN				
0	The resolution of the image displayed to the user	Number of pixels (WxH) in displayed video	Per media session	All
1	The device type on which the media is played	"PC" or "mobile"	Per media session	All
I.11				
2	Target Audio bit-rate	Bitrate in kbit/s	Per media segment	All
3	Segment duration	Duration in seconds	Per media segment	All
4	Audio frame number	Integer, starting with 1	Per media segment	1, 2, 3
5	Audio frame size	Size of the frame in bytes	Per audio frame	1, 2, 3
6	Audio frame duration	Duration in seconds	Per audio frame	1, 2, 3
7	Audio codec	One of: AAC-LC, AAC-HEv1, AAC-HEv2,AC3	Per media segment	All
8	Audio sampling frequency	In Hz	Per media segment	All
9	Number of audio channels	2	Per media segment	All
10	Audio bit-stream	Encoded audio bytes for the frame	Per audio frame	2, 3
I.13				
11	Target Video bit-rate	Bit-rate in kbit/s	Per media segment	All
12	Video frame-rate	Frame rate in frames per second.	Per media segment	All
13	Segment duration	Duration in seconds	Per media segment	All
14	Video encoding resolution	Number of pixels (WxH) in transmitted video	Per media segment	All
15	Video codec and profile	H264-high	Per media segment	All
16	Video frame number	Integer, starting at 1, denoting the frame sequence number in encoding order	Per video frame	1, 2, 3
17	Video frame duration	Duration of the frame in seconds	Per video frame	1, 2, 3
18	Frame presentation timestamp	The frame presentation timestamp	Per video frame	1, 2, 3
19	Frame decoding timestamp	The frame decoding timestamp	Per video frame	1, 2, 3
20	Video frame size	The size of the encoded video frame in bytes	Per video frame	1, 2, 3
21	Type of each picture	"I" or "Non-I" for mode 1"I"/"P"/"B" for modes 2, 3	Per video frame	1, 2, 3
22	Video bit-stream	Encoded video bytes for the frame	Per video frame	2, 3

Id	Description	Values	Frequency	Available to modes
I.14				
23	Stalling/initial loading event start	The start time of the stalling event in seconds relative to the start of the original video clip, expressed in media time (not wall clock time) NOTE: This is 0 for initial loading delay.	Per stalling event	All
24	Event duration	The duration of the stalling event in seconds	Per stalling event	All

7.3. Stalling

Only the following state transitions are considered in ITU-T P.1203:

a) Initial stalling to Playing
b) Playing to Stalling
c) Playing to End
d) Stalling to Playing.

Note that user-initiated state transitions are outside of the scope of this work item. More specifically pausing, seeking, user initiated quality change, user initiated play or user initiated end are all not considered.

7.4. Measurement window specification

The Pv [ITU-T P.1203.1] and Pa [ITU-T P.1203.2] modules provide one video or audio quality score per output sampling interval, respectively, and do not perform any kind of long-term temporal integration of input features or output scores. This integration is handled in the integration module Pq specified in [ITU-T P.1203.3].

Pv and Pa modules must therefore apply a sliding **measurement window** for the input data acquisition and output score calculation. The measurement window is defined as:

> *audio or video information of one or more segments belonging to a specific media quality level, used as input to the Pv or Pa module at output time ts.*

At any output time *ts*, the Pv ([ITU-T P.1203.1]) and Pa ([ITU-T P.1203.2]) modules can use information from the measurement window $[ts - T/2, ts + T/2]$, with T = 20 s, to generate the output sample according to the output sampling interval (see clause 7.5).

None of the following information must be used from outside the measurement window:

– bitstream data;
– previously calculated scores;
– extracted bitstream features, meta information, or any kind of indicators.

If the measurement window contains segments of multiple media quality levels, only contiguous adjacent segments of the same media quality level as the segment to which *ts* corresponds must be used as input to the Pa and Pv modules.

The timing of the measurement window input specification is visualized in Figure 2.

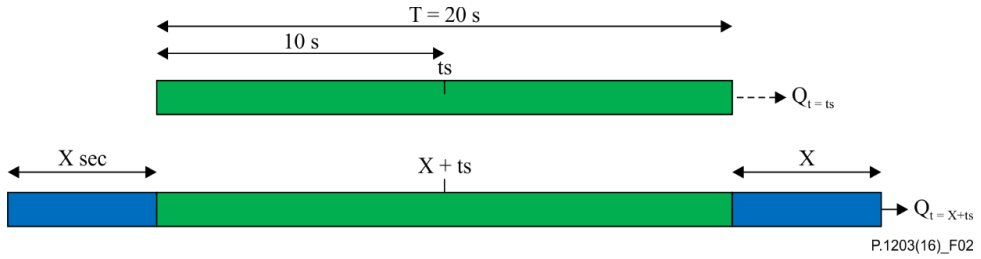

Figure 2 — Measurement window

7.4.1. Implementation of measurement window

The measurement window must be implemented as described in clause 7.4.1.1 to clause 7.4.1.3.

7.4.1.1. Frame extraction

From the input segments that form the sequence, each audio sample or video frame (depending on whether Pv or Pa is used) must be extracted (from now on simply called "frame"). Each frame must carry the information as described in the rows of Table 5, I.13 that are "per-frame", depending on the mode in which the module operates.

If frames cannot be extracted from physical bitstreams or video frame metadata (i.e., if the module is operating in mode 0), artificial frames must be generated by producing $S \cdot R$ frames for a segment of length S with a frame rate of R. Each frame must have a duration and decoding timestamp (DTS), with the DTS strictly monotonically increasing.

For example, for a segment of 5 s length with a frame rate of 25 Hz, 125 frames must be generated, with each frame having a duration of 0.04 s, and the frames having DTSs of [0.04, 0.08, 0.12, ...].

Note that artificial frames do not carry additional payload information.

7.4.1.2. Determination of score calculation

An empty list must be initialized, which will hold frames. The last output time and the accumulated sequence duration must be set to 0. Then, for every frame extracted as described in clause 7.4.1.1, that frame is added to the list.

If the accumulated frame duration of the list is greater than 20 s, the first frame in the list is removed.

If the last output time is 0 and the accumulated sequence duration is smaller than 11, no score shall be output. Otherwise, if the accumulated sequence duration minus 10 is greater or equal to the last output time plus 1, the frame list and the corresponding output sample timestamp is forwarded to the Pa/Pv module to calculate the score, and the last output time is increased by 1.

The following pseudocode shows the procedure described above, which is called for every frame extracted:

```
if acc_frame_dur + frame.duration > 20:
    removed_frame = remove first item from frames
    acc_frame_dur -= removed_frame.duration

add frame to frames
acc_frame_dur += frame.duration
acc_sequence_dur += frame.duration

if last_score_output_at == 0 and acc_sequence_dur < 10 + 1:
    do nothing

if acc_sequence_dur - 10 >= last_score_output_at + 1:
    last_score_output_at += 1
```

```
    forward frames and output_sample_timestamp to module
```

When the last extracted frame has been added to the list and the above procedure has been run, the measurement window must be flushed according to the following procedure.

A final output sample timestamp is calculated by rounding down the total sequence duration (see clause 7.5). Then, repeatedly increasing the output sample timestamp by 1, frames that should not be considered are removed from the list, that is, if their DTS is lower than the current output sample timestamp minus 10. The remaining frames are then forwarded to the module for score calculation.

The following pseudocode explains the above procedure:

```
final_sample_timestamp = floor(acc_sequence_dur)
output_sample_timestamp = last_score_output_at + 1

while output_sample_timestamp <= final_sample_timestamp:
    while frames[0].dts < output_sample_timestamp - 10:
        remove first frame in frames

    forward frames and output_sample_timestamp to module

    output_sample_timestamp += 1
```

7.4.1.3. Determination of contiguous adjacent segments

When the list of frames from the measurement window and the output sample timestamp are forwarded to the Pa or Pv module, the module has to determine which frames shall be considered for calculating the score at the output sample timestamp. It does so with the following procedure, where *frames* refers to the frames forwarded from the procedure in clause 7.4.1.2:

```
for index, frame in frames:
    if frame.dts < output_sample_timestamp:
        last_index = index
output_sample_relative = last_index

output_sample_frame = frames[output_sample_relative]
target_media_quality_level = output_sample_frame.media_quality_level
window = [output_sample_relative]

if frames[output_sample_relative-1].media_quality_level == target_media_quality_
level:
    i = output_sample_relative - 1
    window = [i] + window
    if i-1 != -1:
        i -= 1
        while frames[i].media_quality_level == frames[i+1].media_quality_level:
            window = [i] + window
            if i-1==-1:
                break
            else:
                i-=1

if output_sample_relative + 1 != length(frames):
    if frames[output_sample_relative+1].media_quality_level ==
target_media_quality_level:
    i = output_sample_relative + 1
    window.append(i)
    if i+1 != len(frames):
        i += 1
        while frames[i].media_quality_level == frames[i-1].media_quality_level:
            window.append(i)
```

```
if i+1 == len(frames):
    break
else:
    i+=1
```

At the end of the procedure, "window" identifies the indices of frames (0-based) from the list of forwarded frames that must be used for calculating the quality score (see the respective procedures in [ITU-T P.1203.1] and [ITU-T P.1203.2].

7.5. Model output information

The Pa and Pv modules provide one score per output sampling interval, thus one score every 1 s (see clause 3.2.6).

The output sampling interval of the Pa and Pv modules has no relation to a media segment, or the media segments used in the ITU-T P.1203 context, since the length of the media segments is not necessarily in complete seconds.

There should not be any output score for frames at the end of a sequence, when those frames do not add up to a complete second. The quality score is calculated at the closest frame boundary at or after each integer second from the start of the stream.

For all outputs, the 1-5 quality scale is used, where "1" means "bad" quality and "5" means "excellent" quality, as specified in [ITU-T P.910].

The ITU-T P.1203 model outputs are as follows:

- O.21: Audio coding quality per output sampling interval
  - Per-one-second scores provided per session and on a 1-5 quality scale.
- O.22: Video coding quality per output sampling interval
  - Per-one-second scores provided per session and on a 1-5 quality scale.
- O.23: Perceptual stalling indication
  - Single score on a 1-5 quality scale for the session.
- O.34: Audiovisual segment coding quality per output sampling interval
  - Multiple segment scores provided per session.
  - Window-size same as for/synced with O.21, O.22.
- O.35: Final audiovisual coding quality score
  - Single score for the session, on a 1-5 quality scale.
  - Includes aspects of temporal integration.
- O.46: Final media session quality score
  - Single score for the session, on a 1-5 quality scale.

Includes initial loading delay and stalling aspects.

8. Overview of databases used for model development

For model development and validation, a total of 30 databases were created. Each database consists of a set of processed video sequences (PVSs). Within one database, each PVS was derived from a unique source video. The source videos of each database were of fixed duration in-between 1-5 minutes. The number of PVSs in each database was chosen such that the total video duration presented to subjects is around 60 minutes. In more detail, 60 PVSs were used for databases of 1-minute duration, with fewer PVSs for the longer durations, and with a minimum of 14 PVSs for the source videos of 5-minute duration. In total, 1064 PVSs were used.

The source video sequences were processed by rescaling, encoding, and segmenting to form a set of quality representations for each video content segment. A processed video sequence was created by

selecting one representation for each video content segment, with possibly introducing initial loading delay or stalling between the segments.

Video was encoded with [ITU-T H.264] using the libx264 codec with high10 profile and two-pass encoding, using different target bitrates. Scene cut detection was switched off and the maximum number of consecutive B-frames set to 3. The GOP duration was fixed for each video, but was variable in some of the databases.

Audio was encoded with AAC using the libfdk_aac codec.

For each database, an ACR-type subjective test was performed to collect ratings on the 5-point scale. Out of the 30 subjective tests, 19 were performed using a full-HD PC monitor for playback, and 11 were performed using a mobile phone with a 5-inch display.

Out of the 30 databases, 17 were initially shared for model development. The remaining 13 databases were used for model selection.

Overall performance p is determined by a weighted average of the per-database mean squared error (MSE). In more detail, the mean squared error MSE_k of database k is weighted by a weight w, summed over all databases and normalized,

$$\rho = \frac{1}{N}\sum_{k=1}^{M} w_k \times MSE_k \tag{8-1}$$

where the weight $w\_k = 0.25$ if the database k is part of the initially shared databases, and $w = 0.75$ otherwise. The total number of databases M is $M = 30$, and the normalisation constant N is given by $N = \sum_{k=1}^{M} w_k$.

9. Description of the ITU-T P.1203 model algorithms

Detailed descriptions of the individual modules can be found in the respective Recommendations, and their annexes – [ITU-T P.1203.1] for the video quality estimation modules, [ITU-T P.1203.2] for the audio quality estimation module and [ITU-T P.1203.3] for the quality integration module.

Appendix I

Performance figures

(This appendix does not form an integral part of this Recommendation.)

The performance of the overall ITU-T P.1203 model on the databases described in the body of this Recommendation and for the different modes is summarized in the table below:

Performance measure	Mode 0	Mode 1	Mode 2	Mode 3
RMSE	0.465	0.415	0.381	0.333
Pearson correlation	0.814	0.842	0.868	0.892

Note that the calculation of the performance figures above was performed after final training of the model on all available subjective test databases. That means that the figures are slightly optimistic compared to if they would have been calculated based on completely unknown databases.

To compensate for between-test bias effects, the test scores have been mapped to the model output values with a linear mapping applied per each database.