ITUPublications
Recommendations

International Telecommunication Union
Standardization Sector

Recommendation

ITU-T P.1204 (10/2023)

SERIES P: Telephone transmission quality, telephone installations, local line networks

Models and tools for quality assessment of streamed media

Video quality assessment of streaming services over reliable transport for resolutions up to 4K

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	P.10-P.19
Voice terminal characteristics	P.30-P.39
Reference systems	P.40-P.49
Objective measuring apparatus	P.50-P.59
Objective electro-acoustical measurements	P.60-P.69
Measurements related to speech loudness	P.70-P.79
Methods for objective and subjective assessment of speech quality	P.80-P.89
Voice terminal characteristics	P.300-P.399
Objective measuring apparatus	P.500-P.599
Measurements related to speech loudness	P.700-P.709
Methods for objective and subjective assessment of speech and video quality	P.800-P.899
Audiovisual quality in multimedia services	P.900-P.999
Transmission performance and QoS aspects of IP end-points	P.1000-P.1099
Communications involving vehicles	P.1100-P.1199
Models and tools for quality assessment of streamed media	**P.1200-P.1299**
Telemeeting assessment	P.1300-P.1399
Statistical analysis, evaluation and reporting guidelines of quality measurements	P.1400-P.1499
Methods for objective and subjective assessment of quality of services other than speech and video	P.1500-P.1599

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

Recommendation ITU-T P.1204

Video quality assessment of streaming services over reliable transport for resolutions up to 4K

Summary

Recommendation ITU-T P.1204 is the introductory Recommendation for a set of Recommendation that describe model algorithms for monitoring the video quality for streaming using reliable transport (e.g., adaptive streaming based on the hypertext transfer protocol (HTTP) over the transmission control protocol (TCP), quick user datagram protocol (UDP) Internet connections (QUIC)).

The ITU-T P.1204 series of Recommendations comprises different variants of models for sequence-related (between 5 and 10 s) and per-1-second video-quality estimation. The variants differ in the type of input information they use: bitstream based, pixel-based, and hybrid (using both bitstream and pixel information).

In principle, the per-1-second outputs of these video-quality models can be used together with an audio-quality model for integration into audiovisual quality and, together with information about initial loading delay and media playout stalling events, further into a final per-session model output, an estimate of integral per-session quality (see e.g., ITU-T P.1203, ITU-T P.1203.2, ITU-T P.1203.3).

Recommendation ITU-T P.1204 was developed in collaboration with the Video Quality Experts Group (VQEG).

The structure of the set of Recommendations reflects the different functionalities of modules described in each Recommendation:

– ITU-T P.1204 (2023), *Video quality assessment of streaming services over reliable transport for resolutions up to 4K*.

– ITU-T P.1204.3 (2020), *Video quality assessment of streaming services over reliable transport for resolutions up to 4K with access to full bitstream information*.

– ITU-T P.1204.4 (2022), *Video quality assessment of streaming services over reliable transport for resolutions up to 4K with access to full and reduced reference pixel information*.

– ITU-T P.1204.5 (2023), *Video quality assessment of streaming services over reliable transport for resolutions up to 4K with access to transport and received pixel information*.

The ITU-T P.1204.x-series of Recommendations addresses three application areas, which are respectively indicated in the module-related ITU-T P.1204.3, ITU-T P.1204.4 and ITU-T P.1204.5:

– large-screen presentation as with fixed-network video streaming;

– mobile streaming on handheld devices such as smartphones;

– presentation on tablet-type devices.

History *

Edition	Recommendation	Approval	Study Group	Unique ID
1.0	ITU-T P.1204	2020-01-13	12	11.1002/1000/14155
2.0	ITU-T P.1204	2023-10-29	12	11.1002/1000/15698

Keywords

Adaptive streaming, mean opinion score (MOS), mobile (MO), monitoring, multimedia, OTT, 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/software copyrights, 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 appropriate ITU-T databases available via the ITU-T website at http://www.itu.int/ITU-T/ipr/.

Table of Contents

Recommendation ITU-T P.1204

Video quality assessment of streaming services over reliable transport for resolutions up to 4K

1 Scope

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

Five model types are defined to cover a range of use-cases, from monitoring bitstreams where the video payload is fully encrypted, unencrypted bitstreams, and where deep packet inspection is possible, or where the bitstream is available at the encoding premises up to measurement using the pixel information available e.g., from the client side. The models thus have a wide range of applications, from encoding optimization over client-side quality of experience (QoE) assessment to network or service optimization, or benchmarking purposes. The models in the [ITU-T P.1204] series of Recommendations are bitstream based, pixel-based and hybrid based.

The models described here are applicable to progressive download and adaptive streaming or other streaming applications with reliable transport, where the quality experienced by the end user is affected by video degradations due to coding, spatial re-scaling, or variations in video frame rates. Quality assessment of adaptive streaming includes aspects of media adaptation, which may be handled in integration modules such as [ITU-T P.1203.3], and not in the video modules. This Recommendation is able to handle various video codecs (i.e., [ITU-T H.264], [ITU-T H.265] high-efficiency video coding (HEVC), video payload type 9 (VP9), AOMedia Video 1 (AV1)[1], resolutions up to 4K or ultra-high definition-1 (UHD-1) and frame rates of up to 60 frames/s. The video-quality module Pv of [b-ITU-T P.1203], i.e., [ITU-T P.1203.1], only addresses [ITU-T H.264] and full high definition (HD) with up to 30 frames/s.

The models predict a mean opinion score (MOS) on a five-point absolute category rating (ACR) scale (see [ITU-T P.910]) as an overall video quality MOS (5 to 10 s). In addition to the overall quality score, the video quality models produce a per-one-second quality score, suitable for diagnostics or integration into an integral quality score for longer sessions (*cf.* e.g., [ITU-T P.1203.3] for 1 to 5 minute duration sessions).

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

Effects due to source generations such as signal noise, video shake, certain colour properties (and other similar video factors), as well as other impairments related to the payload are not reflected in the scores computed by this model.

As a consequence, this Recommendation can be used for applications such as the following.

– In-service quality monitoring for specific IP-based audiovisual services, as specified in more detail in clause 6.1.

[1] AV1 codec is currently supported by [ITU-T P.1204.4] and [ITU-T P.1204.5].

– Performance and quality assessment of live networks (including video encoding) considering the effect due to the encoding bitrate, encoding resolution, and encoding frame rate.
– Laboratory testing of video systems.
– Benchmarking of different service implementations.
– Benchmarking of different encoder implementations. Note that only the full/reduced reference pixel-based model type can be used for direct benchmarking of this type.
– Evaluation of transcoding solutions.

In particular, targeted applications are progressive download streaming and adaptive streaming (using reliable transport), which includes the following.

– Over-the-top (OTT) services, as well as operator-managed video services (over the transmission control protocol).
– Video over both mobile (MO) and fixed connections.
– The streaming protocols HTTP live streaming (HLS) or dynamic adaptive streaming over HTTP (DASH) used with the hypertext transfer protocol (HTTP) or HTTP2 over TCP/IP or quick user datagram protocol Internet connections (QUIC), or real-time messaging protocol (RTMP) over TCP/IP). Note that the model is agnostic to the specific application or transport layer protocol, with the exception that it assumes reliable delivery of video packets.
– Video services typically use container formats based on the ISO/IEC base media file format such as moving picture experts group-4 (MPEG-4) part 14 (MP4), or other container formats such as audio video interleave (AVI), matroska video (MKV), WebM, third generation partnership (3GP), and MPEG-2 transport stream (MPEG-2-TS). Note that the model is agnostic to the type of container format.

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 H.264] Recommendation ITU-T H.264 (V14) (2021), *Advanced video coding for generic audiovisual services*.

[ITU-T H.265] Recommendation ITU-T H.265 (V9) (2023), *High efficiency video coding*.

[ITU-T P.910] Recommendation ITU-T P.910 (2023), *Subjective video quality assessment methods for multimedia applications*.

[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*.

[ITU-T P.1203.2] Recommendation ITU-T P.1203.2 (2017), *Parametric bitstream-based quality assessment of progressive download and adaptive audiovisual streaming services over reliable transport – Audio quality estimation module*.

[ITU-T P.1203.3] Recommendation ITU-T P.1203.3 (2019), *Parametric bitstream-based quality assessment of progressive download and adaptive audiovisual streaming services over reliable transport – Quality integration module*.

[ITU-T P.1204.3] Recommendation ITU-T P.1204.3 (2020), *Video quality assessment of streaming services over reliable transport for resolutions up to 4K with access to full bitstream information*.

[ITU-T P.1204.4] Recommendation ITU-T P.1204.4 (2022), *Video quality assessment of streaming services over reliable transport for resolutions up to 4K with access to full and reduced reference pixel information*.

[ITU-T P.1204.5] Recommendation ITU-T P.1204.5 (2023), *Video quality assessment of streaming services over reliable transport for resolutions up to 4K with access to transport and received pixel information*.

3 Definitions

3.1 Terms defined elsewhere

This Recommendation uses the following terms defined elsewhere:

3.1.1 bitstream [ITUT H.264]: A sequence of bits that forms the representation of coded pictures and associated data forming one or more coded video sequences. Bitstream is a collective term used to refer either to a network abstraction layer (NAL) unit stream or a byte stream.

3.1.2 integral quality [b-ITU-T P.1203]: 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.

3.1.3 media adaptation [b-ITU-T P.1203]: 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.

3.1.4 media quality level [b-ITU-T P.1203]: A particular encoding setting applied to a video or audio stream.

3.1.5 model, model algorithm [b-ITU-T P.1203]: An algorithm with the purpose of estimating the subjective (perceived) quality of a media sequence.

3.1.6 sequence [b-ITU-T P.1203]: An audiovisual stream composed of multiple non-overlapping segments.

3.1.7 video chunk [b-ITU-T G.1022]: A contiguous set of samples for one track of a video.

3.2 Terms defined in this Recommendation

This Recommendation defines the following term:

3.2.1 mean opinion score (MOS): The mean of opinion scores, which are values on a predefined scale that subjects assign to their opinion of the performance of the telephone transmission system used either for conversation or for listening to spoken material.

NOTE – Paraphrased from clause 7 of [b-ITU-T P.800.1].

4 Abbreviations and acronyms

This Recommendation uses the following abbreviations and acronyms:

ACR Absolute Category Rating

AV1 AOMedia Video 1

AVI Audio Video Interleave

DASH Dynamic Adaptive Streaming over HTTP

GoP	Group of Pictures
HD	High Definition
HEVC	High-Efficiency Video Coding
HLS	HTTP Live Streaming
HTTP	Hypertext Transfer Protocol
IP	Internet Protocol
MKV	Matroska Video
MO	Mobile
MOS	Mean Opinion Score
MP4	MPEG-4 Part 14
MPEG	Moving Pictures Expert Group
MPEG-2-TS	MPEG-2 Transport Stream
OTT	Over-The-Top
PC	Personal Computer
PVS	Processed Video Sequence
QoE	Quality of Experience
QUIC	Quick UDP Internet Connections
RExt	Range Extension
RMSE	Root Mean Square Error
RTMP	Real-Time Messaging Protocol
RTP	Real-time Transport Protocol
TA	Tablet
TCP	Transmission Control Protocol
TV	Television
UDP	User Datagram Protocol
UHD	Ultra-High Definition
VP9	Video Payload type 9
VVC	Versatile Video Coding

5 Conventions

This Recommendation uses the following conventions:

– 4K: Video resolution of 4 096 × 2 160 or 3 840 × 2 160.

– Pa designates the audio quality estimation module, see [ITU-T P.1203.2].

– Pv designates the video quality estimation module (as specified in this Recommendation).

– Pq designates the quality integration module, see [ITU-T P.1203.3].

– Reliable transport: Reliable delivery with protocols guaranteeing no loss of information.

6 Areas of application

6.1 Application range for the models

Table 1 shows the application range of the model in this Recommendation on what the model has been developed for and Table 2 lists areas where it is not applicable. Table 3 lists test factors and coding technologies for which this Recommendation has been validated.

Table 1 – Areas for which this Recommendation is applicable

Areas for which the model is applicable
In-service monitoring of video sent over reliable transport. Both OTT services and operator-managed video services, use reliable delivery with protocols such as HTTP or HTTP2 over TCP/IP or QUIC, or RTMP over TCP/IP. Note that this model is agnostic to the type of container format
Performance and quality assessment of live networks (including video encoding) considering impairments due to encoding bitrate, encoding resolution, and encoding frame rate
Laboratory testing of video systems
Benchmarking of different service implementations
Benchmarking of different encoder implementations. Note that only the full or reduced reference pixel-based model type can be used for direct benchmarking of this type
Evaluation of transcoding solutions

Table 2 – Areas for which this Recommendation is not applicable

Areas for which the model is not applicable
In-service monitoring of video streaming using unreliable transport (e.g., real-time transport protocol / user datagram protocol RTP/UDP), where packet loss introduces visible quality degradations
Evaluation of visual quality of display / device properties
Evaluation of audio/video sync distortions
Evaluation of video codecs for which the model is not validated MPEG-I Part 3 [versatile video coding (VVC)], etc.)
Evaluation of the effects of noise, delay, colour correctness, or other content-production-related aspects

Table 3 – Test factors and coding technologies for which this Recommendation has been validated

Video test factors for which the model has been validated	
Video content	Movies and movie trailers, sports videos, documentaries, computer-generated graphics/games, etc.
Input video length	The video modules are trained and validated to produce one overall video-quality score for a chunk of 7-10 s and also provide the per-second scores. Optimal performance for ~8 s. Models are assumed to provide valid overall video-quality estimations for 5-10 s long sequences
Bitstream container	AVI, MP4, MKV, WebM
Encoder types (and implementation, see Note 1)	H.264/AVC (libx264), H.265/HEVC (libx265), VP9 (libvpx-vp9), AV1 (libaom-av1)

Table 3 – Test factors and coding technologies for which this Recommendation has been validated

<table>
<tr><th colspan="5">Video test factors for which the model has been validated</th></tr>
<tr><td>Encoder profiles</td><td colspan="4">H.264 (MPEG-4 Part 10): Constrained baseline, Main, Hi, Hi10, Hi422.
H.265: Main, Main10, range extension (RExt).
VP9: 0, 1, 2, 3.
AV1: Main (cpu-used 1, 6)</td></tr>
<tr><td rowspan="5">Video resolution and bitrate</td><td>Resolution definition</td><td>Video height range</td><td>Personal computer/ television (PC/TV)</td><td>Mobile/tablet (MO/TA)</td></tr>
<tr><td>Below SD</td><td>180-270</td><td>—</td><td>90 Kbit/s- 1 Mbit/s</td></tr>
<tr><td>SD</td><td>360-540</td><td>150 Kbit/s- 4 Mbit/s</td><td>150 Kbit/s- 4 Mbit/s</td></tr>
<tr><td>HD</td><td>720-1 080</td><td>500 Kbit/s- 15 Mbit/s</td><td>500 Kbit/s- 15 Mbit/s</td></tr>
<tr><td>Above HD</td><td>1 440-2 160</td><td>1.5 Mbit/s - 45 Mbit/s</td><td>1.5 Mbit/s - 20 Mbit/s</td></tr>
<tr><td>Video aspect ratio</td><td colspan="4">16:9, see Note 2</td></tr>
<tr><td>Group of pictures (GoP)</td><td colspan="4">Variable. Average GOP length can be between 0.5 s and chunk duration</td></tr>
<tr><td>Bit-depth</td><td colspan="4">8 bit or 10 bit</td></tr>
<tr><td>Chroma subsampling</td><td colspan="4">YUV 4:2:0 and YUV 4:2:2 for H.264/H.265/VP9
YUV 4:2:0 for AV1</td></tr>
<tr><td>OTT(s)</td><td colspan="4">Online providers that offer video on demand and video encoding as a service. It should be noted that the models are applicable for similar OTT(s)</td></tr>
<tr><td>Display resolution and frame rate</td><td colspan="4">PC/TV; 2 160p, up to 60 frames/s.
MO/TA: 1 440p, up to 60 frames/s.</td></tr>
<tr><td>Viewing distances</td><td colspan="4">PC/TV: 1.5H to 3H (H: Screen height), see Note 3
MO/TA: 4H to 6H</td></tr>
<tr><td colspan="5">NOTE 1 – During training and validation of H.264/H.265/VP9, FFmpeg 3.2.2 was used with x264 snapshot 20170202-2245, x265 v2.2, libvpx 1.6.1.
NOTE 2 – During training and validation of AV1, FFmpeg 4.2.2 was used with libaom-av1 library [b-IEICE-Trans-NTT].
NOTE 3 – For original content with a larger aspect ratio, letterboxing of up to 30% was allowed, that is 1 512 pixels height for video coded at 2 160 pixels height. Video content with 1.89:1 aspect ratio (e.g., cinema 4K) may also be used.
NOTE 4 – It is noted that for PC or MO, the model output is conservative and should be interpreted to correspond to a viewing distance of 1.5H to 1.6H.</td></tr>
</table>

6.2 Model types

The model types are specified in Tables 4, 5, and 6, which also provide more information on input. Meta-data is defined here as being the header information and information on the I.GEN interface as defined in clause 7.1.

Table 4 – ITU-T P.1204 types of bitstream-based model

Model type and Recommendation	Input (see clause 7.1 for a complete list of inputs)	Complexity
Transport, see Note 1 For further study	Meta-data	Low
Video frame-level, see Note 2 For further study	Meta-data and frame header information	Low
Bitstream, see Note 3 [ITU-T P.1204.3]	Meta-data and any information from the video bitstream	High
NOTE 1 – Corresponding to mode 0 in [b-ITU-T P.1203]. NOTE 2 – Corresponding to mode 1 in [b-ITU-T P.1203]. NOTE 3 – Corresponding to mode 3 in [b-ITU-T P.1203].		

Table 5 – ITU-T P.1204 types of pixel-based model

Model type and Recommendation	Type related inputs (see clause 7.1 for a complete list of inputs)	Complexity
Reduced reference (RR). Can be used as a full reference (FR) model [ITU-T P.1204.4]	Side information derived from the reference video Degraded video	High

Table 6 – ITU-T P.1204 types of hybrid model

Model type and Recommendation	Type-related inputs (see clause 7.1 for a complete list of inputs)	Complexity
Hybrid no reference (NR) with transport information [ITU-T P.1204.5]	Degraded video Meta-data information	Low to high

7 Building blocks

The module layout of the [ITU-T P.1204] model is depicted in Figure 1 and Figure 2.

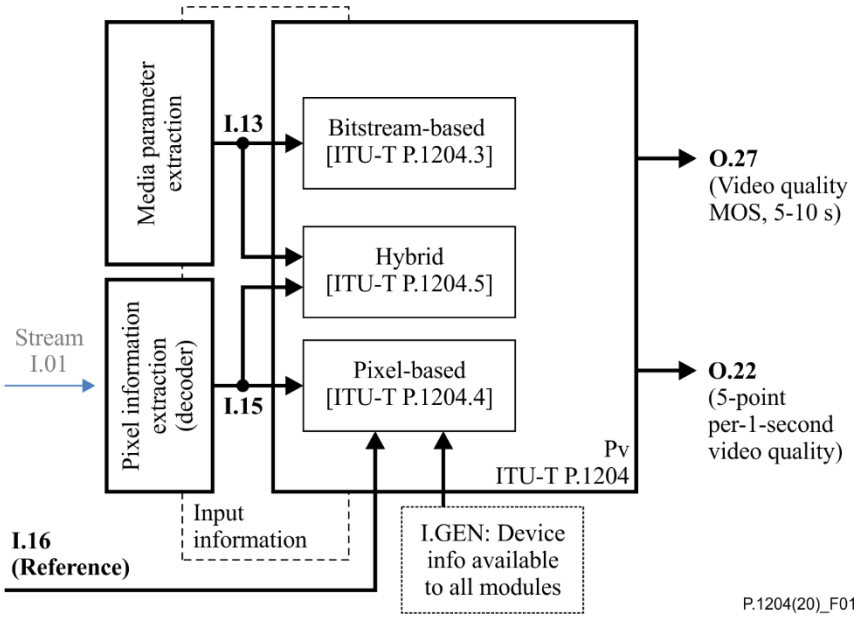

Figure 1 – Building blocks of the video quality models in this Recommendation; the input information used by the different types of model is indicated

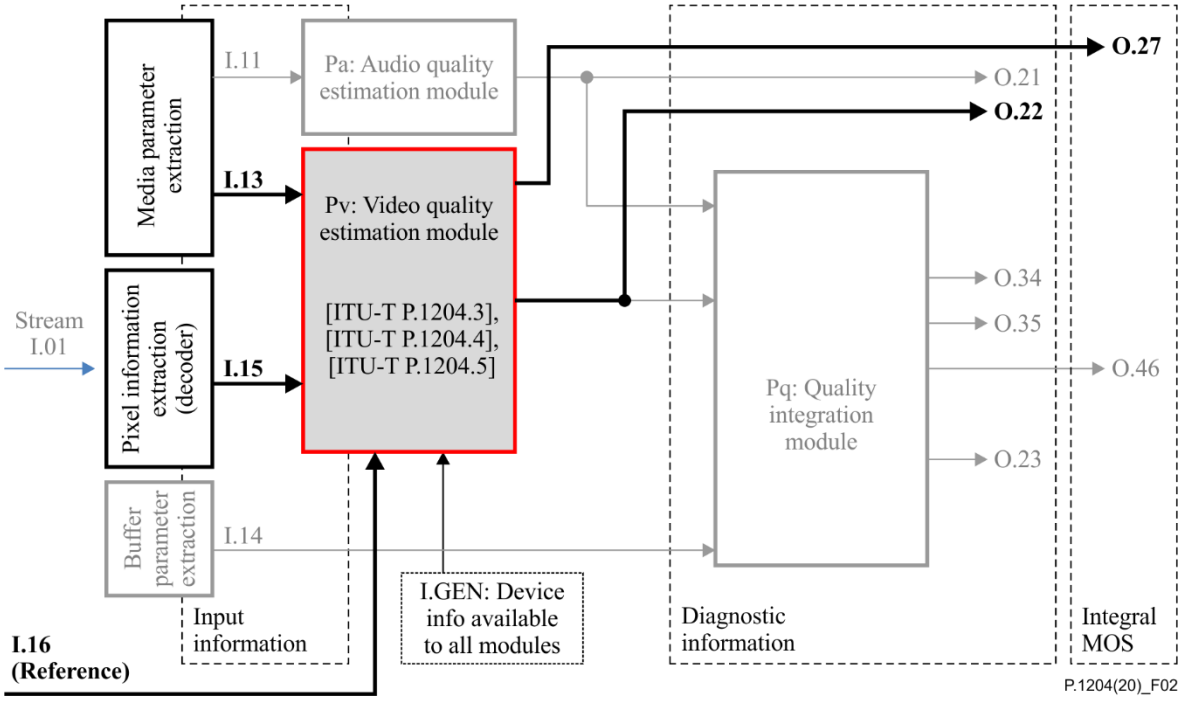

Figure 2 – Building blocks of the [ITU-T P.1204] video quality module in the context of longer-term integration. Here, all greyed-out components reflect an implementation in line with the audiovisual-quality and long-term integration according to [b-ITU-T P.1203]

7.1 Model input interfaces

The [ITU-T P.1204] model will receive media information and prior knowledge about the media stream or streams. For the different types of models, the following inputs may be extracted or estimated in different ways, which lie outside the scope of this Recommendation although they may be added in future annexes. The model receives the following input information, depending on its type (for details see Table 7):

I.GEN: Display resolution and device type. The device type is defined as follows:

- PC/TV: screen size 24 inches or larger and less than or equal to 100 inches.
- MO/TA: screen size 13 inches or smaller.

I.11: Audio coding information. Not used in this Recommendation, kept to maintain consistency with [b-ITU-T P.1203] for an example.

I.13: Video coding information.

I.14: Initial loading delay and stalling event information. Not used in this Recommendation, kept to maintain consistency with [b-ITU-T P.1203] for an example.

I.15: Degraded video pixel information.

I.16: Reference video pixel information.

7.2 Specification of inputs I.GEN, I.13, I.15 and I.16

See Table 7.

Table 7 – I.GEN, I.13 and I.14 input description

ID	Description	Values	Frequency	Available to models (ITU-T P.1204.X)
I.GEN				
0	The resolution of the image displayed to the user	Number of pixels ($W \times H$) in the displayed video	Per media chunk	All
1	The device type on which the media is played	"PC", "TV", "MO", "TA"	Per media chunk	All
2	Device display size	Display size (diagonal in inches)	Per media chunk	All
3	Relative viewing distance in multiple display height	Relative viewing distance	Per media chunk	All
I.13				
4	Video bitrate	Bitrate in kilobits per second	Per media chunk	3, 5
5	Video frame rate	Frame rate in frames per second.	Per media chunk	3, 5
6	Segment duration	Duration in seconds	Per media chunk	3, 5
7	Video encoding resolution	Number of pixels ($W \times H$) in the transmitted video	Per media chunk	3, 5
8	Video codec and profile	H.264 (MPEG-4 Part 10): Constrained Baseline, Main, Hi, Hi10, Hi422. H.265: Main, Main10, RExt. VP9: 0, 1, 2, 3 AV1: Main (cpu-used 1, 6)	Per media chunk	3, 5

Table 7 – I.GEN, I.13 and I.14 input description

ID	Description	Values	Frequency	Available to models (ITU-T P.1204.X)
9	Video frame number	Integer, starting at 1, denoting the frame sequence number in encoding order	Per video frame	3
10	Video frame duration	Duration of the frame in seconds	Per video frame	3
11	Frame presentation timestamp	The frame presentation timestamp	Per video frame	3
12	Frame decoding timestamp	The frame decoding timestamp	Per video frame	3
13	Video frame size	The size of the encoded video frame in bytes	Per video frame	3
14	Type of each picture	See Note. "I"/"P"/"B" for ITU-T P.1204.3	Per video frame	3
15	Video bitstream	Encoded video bytes for the frame	Per video frame	3
16	Video pixel format	8-bit or 10-bit, together with 4:2:2 or 4:2:0 chroma subsampling	Per media chunk	3
I.15				
26	Degraded video	The raw pixels (YUV file including metadata required for parsing; width, height, frame rate, and pixel format) of the processed video, i.e., the video decoded and upscaled to display resolution without buffering or stalling. The frame information in I.16 and I.15 is synchronized, i.e., no frame misalignments are present	Per media chunk	4, 5
I.16				
27	Reference video information	The reference-side information extraction module takes as input the reference video and outputs the side information file. The reference model side channel bandwidth limit is 256 kbit/s. Thus, the side information of the reference model for a video sequence v is stored in a file with a size at most 256/8 * t_v kB, where t_v is the duration of video v in seconds.	Per media chunk	4
NOTE – Other values are under study.				

7.3 Model output information

The Pv modules provide one score per second and one overall video quality score for the chunk under consideration.

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, a quality scale of 1 to 5 is used, where "1" means "bad" quality, and "5" means "excellent" quality, as specified in [ITU-T P.910].

The [ITU-T P.1204] Pv-model outputs are as follows:

- O.22: Video coding quality per second
  - per-1-second scores provided per chunk and on a quality scale of 1 to 5
- O.27: Final video session quality score
  - single score for the chunk, on a quality scale of 1 to 5.

8 Overview of databases used for model development

For model development and validation, 26 databases were created in total. Each database consists of a set of processed video sequences (PVSs). The source videos of each database were of a duration of 6 s to 9 s. Each source video was repeated at most six times within a database. The number of PVSs in each database was chosen to be around 200. In total, 5 002 PVSs were used.

The videos were encoded with H.264, H.265, VP9 or AV1 using the libx264, libx265, libvpx or libaom-av1 codec implementations, respectively. In addition, some videos were encoded via frequently used online streaming services. The videos were encoded using one of the profiles given in item 8 under I.13 in Table 7.

For each database, an ACR [ITU-T P.910] subjective test was performed to collect ratings on the 5-point scale. Out of the 26 subjective tests, 12 were performed using a UHD PC monitor for playback, six using a UHD TV set, and seven using an MO with a 5-inch to 6-inch display. One test was performed on a 10-inch tablet.

Out of the 26 databases, 13 were initially shared for model development and training. The remaining 13 databases were used for model selection and validation.

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_k, summed over all databases and normalized,

$$p = \frac{1}{N}\sum_{k=1}^{M} w_k \times \mathrm{MSE}_k \quad (1)$$

where the weight $w_k = 0.1$ if the database k is part of the initially shared databases, and $w_k = 0.9$ otherwise. The total number of databases M is 26, and the normalization constant N is given by $N = \sum_{k=1}^{M} w_k$.

9 Description of ITU-T P.1204 model algorithms

Detailed descriptions of the individual modules can be found in the respective Recommendations, and their annexes: [ITU-T P.1204.3] for the bitstream-based video quality model; [ITU-T P.1204.4] for the pixel-based video quality model; and [ITU-T P.1204.5] for the hybrid no-reference video quality model.

Appendix I

Performance figures

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

In this appendix, the root mean square errors (RMSEs) of Pv models are reported for all codecs except AV1. Note that the numbers are reported after a final per-database mapping between the model output and the subjective scores of a database. This linear mapping is used to account for scale and bias variations between different databases.

Table I.1 – Validation performance of Pv model: The submitted model is the model trained on the exchanged training databases and frozen before the creation of validation data. Models were retrained using a five-fold cross-validation approach, with their validation performance listed to show the stability of the performance indicating no over-fitting

Full bitstream	Submitted model	0.421				
	Five-fold cross-validation	0.394	0.407	0.402	0.413	0.401
Pixel-based reference	Submitted model	0.444				
	Five-fold cross-validation	0.418	0.418	0.425	0.442	0.429
Hybrid transport-level	Submitted model	0.452				
	Five-fold cross-validation	0.451	0.440	0.441	0.443	0.457

The re-training of the submitted model was performed on five different splits. The splits were defined on the database level. The following is the procedure that was followed to determine the splits.

– All training and validation databases were merged to obtain in total 26 different short databases (18 PC/TV and eight MO/TA).

– A level of difficulty of prediction for each database was determined based on the average prediction error over all models.

– A 50:50 training:validation split was determined randomly, but respecting the level of the difficulty. In total, five different splits were defined. Each split had a balanced distribution of databases based on the difficulty in both the training and validation.

– The 50:50 split was separately performed for PC/TV and MO/TA cases.

– The final model coefficients correspond to the best performing split.

Appendix II

Performance figures

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

In this appendix, the root mean square errors (RMSEs) of Pv models are reported for the AV1 codec. Note that the numbers are reported after a final per-database mapping between the model output and the subjective scores of a database. This linear mapping is used to account for scale and bias variations between different databases.

Table II.1 – Validation performance (Average RMSE) of Pv model (AV1 only) on six unknown validation databases

Pixel-based reference	Validated model	0.362
Hybrid transport-level	Validated model	0.442

- Initially the pixel-based reference model was validated on six validation databases, which achieved high prediction efficiency (average RMSE 0.362).
- Then eight training databases for bitstream and hybrid model were taken from a known set of databases and re-encoded using AV1 codec.
- Subjective quality estimates for the eight training databases were computed using the validated P.1204.4 model and were used as ground truth for the following retraining.
- Bitstream and hybrid models were trained on eight training databases, and model coefficients were frozen.
- Six unknown validation databases (the same databases that were used for pixel model validation) with subjective MOS were used to validate bitstream and hybrid models.

Table II.2 – Re-optimized performance (Average RMSE) of hybrid no-reference Pv model (AV1 only) on six databases [ITU-T P.1204.5]

Hybrid transport-level	Re-optimized model	0.417

Bibliography

[b-ITU-T G.1022] Recommendation ITU-T G.1022 (2016), *Buffer models for media streams on TCP transport*.

[b-ITU-T P.800.1] Recommendation ITU-T P.800.1 (2016), *Mean opinion score (MOS) terminology*.

[b-ITU-T P.1203] Recommendation ITU-T P.1203 (2017), *Parametric bitstream-based quality assessment of progressive download and adaptive audiovisual streaming services over reliable transport*.

[b-IEICE-Trans-NTT] Yamagishi, K., Egi, N., Yoshimura, N., and Lebreton, P. (2021), *Derivation Procedure of Coefficients of Metadata-based Model for Adaptive Bitrate Streaming Services*, IEICE Transactions on Communications, vol. E104.B, no. 7, pp. 725-737. https://www.jstage.jst.go.jp/article/transcom/E104.B/7/E104.B_2020CQP0002/_article/-char/ja/

SERIES OF ITU-T RECOMMENDATIONS

Series A	Organization of the work of ITU-T
Series D	Tariff and accounting principles and international telecommunication/ICT economic and policy issues
Series E	Overall network operation, telephone service, service operation and human factors
Series F	Non-telephone telecommunication services
Series G	Transmission systems and media, digital systems and networks
Series H	Audiovisual and multimedia systems
Series I	Integrated services digital network
Series J	Cable networks and transmission of television, sound programme and other multimedia signals
Series K	Protection against interference
Series L	Environment and ICTs, climate change, e-waste, energy efficiency; construction, installation and protection of cables and other elements of outside plant
Series M	Telecommunication management, including TMN and network maintenance
Series N	Maintenance: international sound programme and television transmission circuits
Series O	Specifications of measuring equipment
Series P	**Telephone transmission quality, telephone installations, local line networks**
Series Q	Switching and signalling, and associated measurements and tests
Series R	Telegraph transmission
Series S	Telegraph services terminal equipment
Series T	Terminals for telematic services
Series U	Telegraph switching
Series V	Data communication over the telephone network
Series X	Data networks, open system communications and security
Series Y	Global information infrastructure, Internet protocol aspects, next-generation networks, Internet of Things and smart cities
Series Z	Languages and general software aspects for telecommunication systems

Printed in Switzerland
Geneva, 2023