International Telecommunication Union

ITU-T

TELECOMMUNICATION
STANDARDIZATION SECTOR
OF ITU

P.1204.3

(01/2020)

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 with access to full bitstream information

Recommendation ITU-T P.1204.3

International Telecommunication Union

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.1204.3

Video quality assessment of streaming services over reliable transport for resolutions up to 4K with access to full bitstream information

Summary

Recommendation ITU-T P.1204.3 describes a bitstream-based mode 3 video quality model for monitoring the video quality for streaming using reliable transport (e.g., hypertext transfer protocol-(HTTP-) based adaptive streaming (HAS) over the transmission control protocol (TCP), quick user datagram protocol internet connections (QUIC)). The estimate is validated for videos encoded with H.264, H.265 or video payload type 9 (VP9) codecs at any resolution up to 4K/ultra-high definition (UHD) resolution for personal computer (PC) monitors and television (TV) and up to 2 560 × 1 440 for smartphone and tablet displays.

The ITU-T P.1204 series of Recommendations provide sequence-related (between 5 s and 10 s) and per-1-second video-quality estimation. In principle, the per-one-second outputs of this video-quality model can be used together with an audio 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.1204.3 was developed in collaboration with the Video Quality Experts Group (VQEG).

The ITU-T P.1204 series of Recommendations addresses three application areas:

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

– mobile streaming on handheld devices such as smartphones;

– presentation on tablet-type devices.

This Recommendation includes an electronic attachment with the Trees for final prediction announced in clause 8.2.

History

Edition	Recommendation	Approval	Study Group	Unique ID[*]
1.0	ITU-T P.1204.3	2020-01-13	12	11.1002/1000/14156

Keywords

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

[*] 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 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

Recommendation ITU-T P.1204.3

Video quality assessment of streaming services over reliable transport for resolutions up to 4K with access to full bitstream information

1 Scope

This Recommendation[1] describes a bitstream-based video quality model that can be used: stand-alone as a video quality prediction model; or together with audio and integration modules 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 and adaptive streaming, for both mobile and fixed network streaming applications.

This model is 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, e.g., from the client side. The model thus has a wide range of application, from encoding optimization over client-side quality of experience (QoE) assessment to network or service optimization or benchmarking purposes. The model in this Recommendation is bitstream based.

The model described here is 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 that may be handled in integration modules such as those of [ITU-T P.1203.3] and not in the video modules in this Recommendation. This Recommendation is able to handle various video codecs (i.e., H.264, H.265/high-efficiency video coding (HEVC) and video payload type 9 (VP9), resolutions up to 4K/ultra-high definition-1 (UHD-1) and frame rates up to 60 frames/s. In contrast to 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 model predicts 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 s to 10 s). In addition to the overall quality score, this video quality model produces a per-one-second quality score, suitable for diagnostics or integration into an integral quality score for longer sessions (see, for example [ITU-T P.1203.3] for 1 min to 5 min duration sessions).

The model 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 coded video media data being transmitted over an IP connection with certain performance and do not include 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) and 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:

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

[1] This Recommendation includes an electronic attachment with the Trees for final prediction announced in clause 8.2.

- performance and quality assessment of live networks (including codecs) considering the effect due to encoding bitrate, encoding resolution and encoding frame rate;
- laboratory testing of video systems;
- benchmarking of different service implementations.

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 TCP).
- Video over both mobile 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 using 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 (MPEG2-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 (2019), *Advanced video coding for generic audiovisual services*.

[ITU-T H.265] Recommendation ITU-T H.265 (2019), *High efficiency video coding*.

[ITU-T P.910] Recommendation ITU-T P.910 (2008), *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.1203.3] 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] Recommendation ITU-T P.1204 (2020), *Video quality assessment of streaming services over reliable transport for resolutions up to 4K.*

3 Definitions

3.1 Terms defined elsewhere

This Recommendation uses the following term defined elsewhere:

3.1.1 bitstream [ITU-T 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 NAL unit stream or a byte stream.

3.1.2 mean opinion score (MOS) [ITU-T P.1204]: 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].

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 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.

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

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

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

3.1.8 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

None.

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
HAS	HTTP-based adaptive streaming
HD	High Definition
HEVC	High-Efficiency Video Coding
HLS	HTTP Live Streaming
HTTP	Hypertext Transfer Protocol
I-	Intra-predicted
IP	Internet Protocol

IQR	Interquartile Range
MKV	Matroska Video
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
QHD	Quad High Definition
QoE	Quality of Experience
QUIC	Quick User datagram protocol Internet Connections
Rext	Range extension
RMSE	Root Mean Square Error
RTMP	Real-Time Messaging Protocol
RTP	Real-time Transport Protocol
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;

– Pv designates the video quality estimation module (as specified in this Recommendation for the case of bitstream-based prediction, see [ITU-T P.1204] for alternative implementations such as pixel based and hybrid);

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

6 Areas of application

6.1 Application range for the model

Table 1 shows the application range of the model in this Recommendation based on what the model has actually 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, using 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.

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 (AOMedia Video 1(AV1), 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

<table>
<tr><th colspan="5">Video test factors for which the model has been validated</th></tr>
<tr><td>Video content</td><td colspan="4">Movies and movie trailers, sports videos, documentaries, computer generated graphics/games, etc.</td></tr>
<tr><td>Input video length</td><td colspan="4">The video modules were trained and validated to produce one overall video-quality score for a chunk of ~7–9 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.</td></tr>
<tr><td>Bitstream Container</td><td colspan="4">AVI, MP4, MKV, WebM</td></tr>
<tr><td>Encoder types (and implementation, see Note 1)</td><td colspan="4">H.264/AVC (libx264), H.265/HEVC (libx265), VP9 (libvpx-vp9)</td></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.</td></tr>
<tr><td rowspan="5">Video resolution and bitrate</td><th>Resolution definition</th><th>Video height range</th><th>Personal computer/ television (PC/TV)</th><th>Mobile/tablet (MO/TA)</th></tr>
<tr><td>Below SD</td><td>180-270</td><td>—</td><td>90 Kbps-1 Mbps</td></tr>
<tr><td>SD</td><td>360-540</td><td>150 Kbps-4 Mbps</td><td>150 Kbps-4 Mbps</td></tr>
<tr><td>HD</td><td>720-1 080</td><td>500 Kbps-15 Mbps</td><td>500 Kbps-15 Mbps</td></tr>
<tr><td>Above HD</td><td>1 440-2 160</td><td>1.5 Mbps-45 Mbps</td><td>1.5 Mbps-20 Mbps</td></tr>
</table>

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 aspect ratio	16:9, see Note 2
Group of pictures (GoP)	Variable. Average GOP length can be between 0.5 s and chunk duration
Bit-depth	8 bit or 10 bit
Chroma subsampling	YUV 4:2:0 and YUV 4:2:2
OTTs	Online providers that offer video on demand and video encoding as a service. It should be noted that the models are applicable for similar OTTs.
Display resolution and frame rate	PC/TV; 2 160 p, up to 60 frames/s. MO/TA: 1 440 p, up to 60 frames/s.
Viewing distances	PC/TV: 1.5*H* to 3*H* (*H*: Screen height), see Note 3 MO/TA: 4*H* to 6*H*
NOTE 1 – During training and validation, FFmpeg 3.2.2 was used with x264 snapshot 20170202-2245, x265 v2.2, libvpx 1.6.1. NOTE 2 – 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 3 – It is noted that for PC/MO, the model output is conservative and should be interpreted to correspond to a viewing distance of 1.5H to 1.6H.	

7 Model algorithm and output

7.1 Building blocks in relation ITU-T P.1204 model context

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

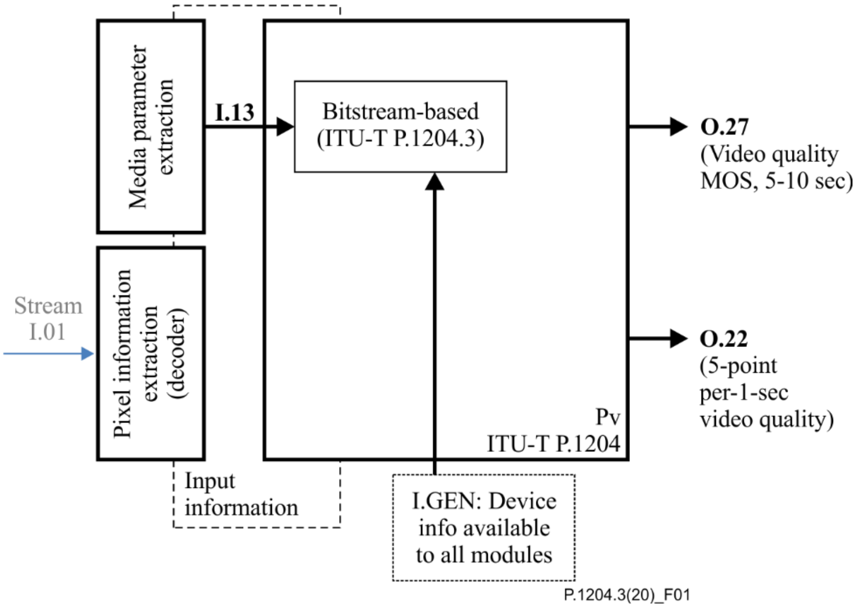
Media parameter extraction
I.13
Bitstream-based (ITU-T P.1204.3)
O.27
(Video quality MOS, 5-10 sec)
Stream I.01
Pixel information extraction (decoder)
O.22
(5-point per-1-sec video quality)
Pv
ITU-T P.1204
Input information
I.GEN: Device info available to all modules
P.1204.3(20)_F01

Figure 1 – Building blocks of the bitstream-based video quality model of this Recommendation (Pv$_{P.1204.3}$) and input information processing

7.2 Model input interfaces

The model receives the following input information:

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

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

I.13: Video coding information

7.3 Specification of inputs I.GEN, I.13

See Table 4.

Table 4 – I.GEN and I.13 inputs description (see Note 1)

ID	Description	Values	Frequency	Used in this Recommendation
I.GEN				
0	The resolution of the image displayed to the user	Number of pixels ($W \times H$) in displayed video	Per media chunk	Yes
1	The device type on which the media is played	"PC", "TV", "MO", "TA"	Per media chunk	Yes
2	Device display size	Display size (diagonal in inches)	Per media chunk	Yes
3	Relative viewing distance in multiple of display height	Relative viewing distance	Per media chunk	Yes
I.13				
4	Video bitrate	Bitrate in kilobits per second	Per media chunk	Yes
5	Video frame rate	Frame rate in frames per second.	Per media chunk	Yes
6	Segment duration	Duration in seconds	Per media chunk	Yes
7	Video encoding resolution	Number of pixels ($W \times H$) in transmitted video	Per media chunk	Yes
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.	Per media chunk	Yes
9	Video frame number	Integer, starting at 1, denoting the frame sequence number in encoding order	Per video frame	No
10	Video frame duration	Duration of the frame in seconds	Per video frame	Yes
13	Video frame size	The size of the encoded video frame in bytes	Per video frame	Yes
14	Type of each picture	See Note 2. "I"/"P"/"B" for this Recommendation	Per video frame	Yes

Table 4 – I.GEN and I.13 inputs description (see Note 1)

ID	Description	Values	Frequency	Used in this Recommendation
15	Video bitstream	Encoded video bytes for the frame	Per video frame	Yes
16	Video pixel format	8-bit or 10-bit, together with 4:2:2 or 4:2:0 chroma subsampling.	Per media chunk	Yes
NOTE 1 – This table will also address ITU-T P.1204.1 and ITU-T P.1204.2 once these have been approved by ITU-T. NOTE 2 – Other values are under study.				

7.4 Model output information

The video module defined in this Recommendation had two outputs, O.22 and O.27. It provides output values on the five-point ACR scale (MOS).

8 Model architecture of this Recommendation

The general model structure is shown in Figure 2. The model consists of two parts, namely, parametric and machine-learning parts. The machine-learning part of the model is based on random forests. The overall prediction is a weighted sum of the parametric and machine-learning part predictions.

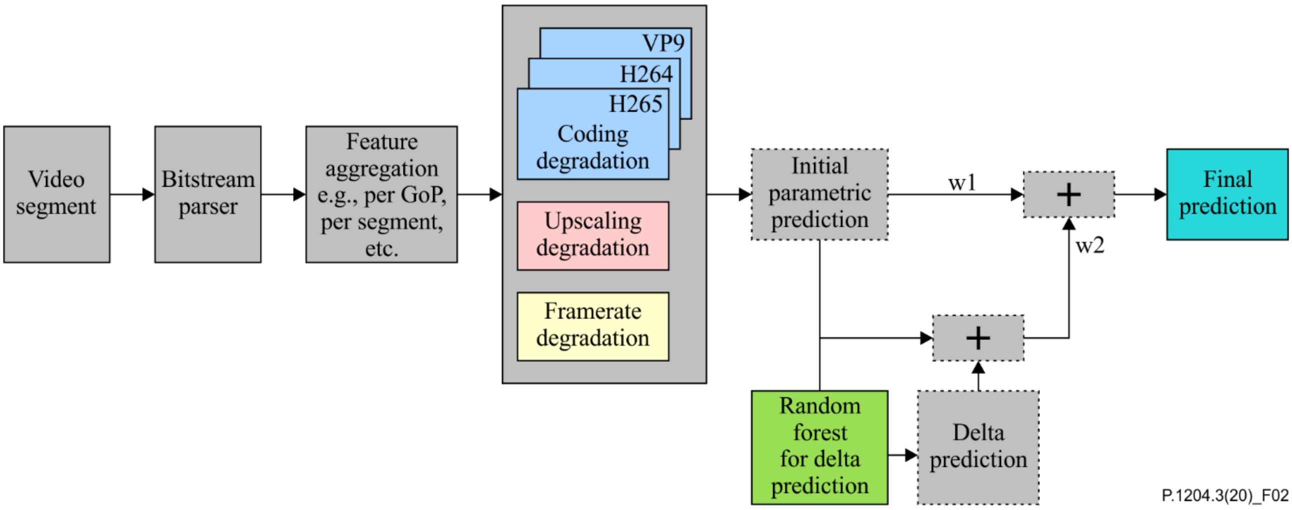

Figure 2 – General model structure

The model has one output with values on the five-point ACR scale (MOS). The parametric part of the algorithm is the core model. The parametric part of the model $M_{\text{parametric}}$ is based on the principle of degradation-based modeling. In the proposed approach, three different degradations are identified that may affect the perceived quality of a given video. The general concept is that the higher the degradation, the lower the quality of the video.

The three degradations that affect that quality of a given video are as follows.

- Quantization degradation. This relates to the coding-related degradations that are introduced in videos based on the quantization settings selected. This degradation can be perceived by the end-user as blockiness and other artefacts. The types of artefact and their strength are codec dependent, as different codecs introduce different distortions based on the selected quantization settings.

- Upscaling degradation. This relates to the degradation introduced due mainly to the encoded video being upscaled to the higher display resolution during playback, thereby resulting in blurring artefacts. These are the same for all codecs, as the display resolution is the only influencing factor for this degradation. It is further assumed that the upscaling algorithm is constant and independent of the codec used, which is the case in real world streaming, where upscaling is performed by the player software or display device used.
- Temporal degradation: This relates to the degradation introduced due to playing out the distorted video at a reduced frame rate compared to the display's native frame rate, thereby resulting in jerkiness. This is the same for all codecs, as the video frame rate is the only influencing factor for this degradation.

Of the three degradations, only quantization degradation is codec dependent.

8.1 Parametric part – The core model

Determination of the quantization degradation:

$$quant = \frac{QP_{\text{non-I-frames}}}{QP_{\max}} \tag{1}$$

where

$QP_{\text{non-I-frames}}$ is the average of the QP for other than intra-predicted (I-) frames for an entire segment;

$QP_{\max}$ is codec and bit-depth dependent:

– for H.264/H.265 8 Bit $QP_{\max} = 51$,

– for H.264/H.265 10 Bit $QP_{\max} = 63$,

– for VP9 8 or 10 Bit $QP_{\max} = 255$;

$quant \in [0, 1]$.

$$mos_q = a + b * \exp(c * quant + d) \tag{2}$$

$$D_{q\_raw} = 100 - RfromMOS(mos_q) \tag{3}$$

$$D_q = \max\left(\min\left(D_{q\_raw}, 100\right), 0\right) \tag{4}$$

where *RfromMOS* is defined in Annex A.

NOTE – The $RfromMOS$ and $MOSfromR$ computations involve information loss due to the fact that these two functions assume that the highest MOS that can be reached is 4.5, thereby resulting in clipping on the MOS-scale for ratings higher than 5. To avoid this information loss, all the subjective data used to train the model is compressed to the 4.5-scale by a simple linear transformation, and the model is trained on this data. Therefore, the resulting coefficients predict the initial prediction on a 4.5-scale. To obtain the prediction on the original five-point scale, the initial prediction is scaled back to the five-scale using the inverse linear transformation.

Determination of the upscaling degradation:

$$scale\_factor = \frac{coding\_res}{display\_res} \tag{5}$$

where

$display\_res = (3840 * 2160)$ for PC/TV and $(2560 * 1440)$ for MO/TA;

$coding\_res$ is the resolution at which the video is encoded $(height * width)$;

$scale\_factor \in [0, 1]$.

$$D_{u\_raw} = x * \log(y * scale\_factor) \tag{6}$$

$$D_u = \max\left(\min\left(D_{u\_raw}, 100\right), 0\right) \tag{7}$$

Determination of the frame rate degradation:

$$D_{t\_raw} = z * \ln(k * (framerate\_scale\_factor) \tag{8}$$

where

$$framerate\_scale\_factor = \frac{coding\_framerate}{60}$$

$framerate\_scale\_factor \in [0, 1]$

$$D_t = \max\left(\min\left(D_{t\_raw}, 100\right), 0\right) \tag{9}$$

Parametric part related final MOS:

$$M_{\text{parametric}} = 100 - (D_q + D_u + D_t) \tag{10}$$

$$M_{\text{parametric}} = MOSfromR(M_{\text{parametric}}) \tag{11}$$

$$M_{\text{parametric}} = scaleto5(M_{\text{parametric}}) \tag{12}$$

where $scaleto5$ is defined in Annex A.

Scaling is done as the coefficients are trained by compressing the subjective scores to a scale of 4.5 to avoid the information loss that can be introduced by the *RfromMOS* and *MOSfromR* calculations, as noted in this subclause.

8.1.1 Model coefficients

The model has access to the entire bitstream as input.

- Coding degradation. It is codec- and bit-depth-dependent. This results in five sets of coefficients, one each for H.264-8bit, H.264-10bit, H.265-8bit, H.265-10bit and VP9 codecs.
- $QP_{\max}$ is 51 for H.264-8bit, H.265-8bit; 63 for H.264-10bit, H.265-10bit; 255 for VP9.

The model coefficients are listed in Tables 5, 6, 7 and 8.

Table 5 – Mode 3 – PC/TV

Codec	*a*	*b*	*c*	*d*
H.264	4.4344	−1.7058	4.9654	−4.1203
H.264-10bit	4.6467	−0.8091	5.9835	.−4.4398
H.265	4.3789	−1.0208	5.7572	−4.5625
H.265-10bit	4.5458	−0.866	6.1116	−3.3828
VP9	4.3404	−0.9961	4.5282	−3.9641

Table 6 – Mode 3 – MO/TA

Codec	*a*	*b*	*c*	*d*
H.264	4.4365	−1.4909	5.4251	−4.5198
H.264-10bit	4.5399	−0.414	6.2249	−4.2599
H.265	4.3089	−0.6685	6.0551	−4.6974
H.265-10bit	4.9999	−2.6821	1.5069	−1.7664
VP9	4.4024	−1.2504	2.9268	−3.0087

Table 7 – Resolution upscaling

End-device	*x*	*y*
PC/TV	−9.5497	1.1999
MO/TA	−8.4690	1.1999

Table 8 – Frame rate upscaling

End-device	*k*	*z*
PC/TV	4.1696	−8.3084
MO/TA	4.2701	−6.3648

8.2 Machine-learning-based part of the model

The proposed random forest model estimates a residual prediction, i.e., the difference between the real video quality score obtained from subjective tests during model training and the prediction of the parametric part of the model, which uses only *QP* and the separate components addressing upscaling and temporal degradation due to the given frame rate. This difference can be explained by the contribution of features to the overall quality score, which are not available in the parametric model part.

Different statistical aggregations of the features are computed and used as the input to the random forest model. In addition to the content-related features, the random forest model explicitly takes into account the prediction from the parametric part of the model as further input. The final random forest-based prediction is the summation of the prediction of the parametric part and the predicted residual.

$$M_{\text{randomForest}} = M_{\text{parametric}} + Residual \tag{13}$$

8.2.1 Random forest features

This clause lists the features used in the random forest model. To generate features for the random forest model that are not directly available from the model input, aggregations of input data may be performed. The features are listed in Table 9.

Table 9 – Features

Aggregated feature	Type	Feature index in code
Minimum standard deviation of motion in the *x*-direction (horizontal motion) per frame	float	x[0]
Maximum frame size in bytes	int	x[1]
Mean bitrate per segment in kilobits per second	float	x[2]

Table 9 – Features

Aggregated feature	Type	Feature index in code
Frame rate	float	x[3]
Resolution ($width * height$) of the distorted video	int	x[4]
H.264 See Note	boolean (0=False, 1=True)	x[5]
H.264_10bit See Note	boolean (0=False, 1=True)	x[6]
H.265 See Note	boolean (0=False, 1=True)	x[7]
H.265_10bit See Note	boolean (0=False, 1=True)	x[8]
Interquartile range (IQR) of the average quantization parameter of non-I-frames	float	x[9]
IQR of the minimum quantization parameter per frame	float	x[10]
Kurtosis of the average motion per frame over all frames in a segment	float	x[11]
Kurtosis of the average quantization parameter of non-I-frames	float	x[12]
Kurtosis of the non-I frame sizes	float	x[13]
Mean of the average quantization parameter of non-I-frames	float	x[14]
$M_{\text{parametric}}$	float	x[15]
$Quant$ ($Quant = \frac{QP_{\text{non-I-frames}}}{QP_{\text{max}}}$)	float	x[16]
Standard deviation of frame size of non-I frame in bits	float	x[17]
Standard deviation of maximum QP of non-I frames	float	x[18]
VP9 See Note	boolean (0=False, 1=True)	x[19]
NOTE – A binary feature, e.g., in the case of [ITU-T H.264], this value is true (1) if the video is encoded with[ITU-T H.264], otherwise false (0)		

The random forest model uses 20 trees with a fixed depth of eight. The individual trees are transformed as functions and a function ***predict*** for aggregation (mean value of all individual tree predictions) of these trees for final prediction are added in the software attachment to this Recommendation as Python code, providing two different versions for MO/TA and PC/TV.

The feature index in code column in Table 9 refers to how the aggregated features are indexed in the function ***predict*** with respect to the feature vector x.

NOTE – Depending on the final random forrest model used, feature indices used by the trees can be different and are selected in the corresponding ***predict*** function.

8.3 Final prediction

The final prediction of quality is the weighted average of the prediction from the parametric part and the random forest part.

$$Q = w_1 * M_{\text{parametric}} + w_2 * M_{\text{randomForest}} \tag{14}$$

Here, $w_1 = 0.5$ and $w_2 = 0.5$; both parts get equal importance in the final score.

A final adjustment to the prediction in Equation 14 is added to compensate for differences in subjective ratings due to the heterogeneity of tests across different laboratories for the training and validation databases. The final per-media chunk prediction $O.27$ is then given by

$$O.27 = a * Q + b \tag{15}$$

Here, $a = 1.036$ and $b = -0.1457$

8.4 Per-second score prediction

In addition to the overall video quality score, the model also outputs the per-second scores. The per-second video quality score ($O.22$) is calculated as follows:

$$O.22 = \frac{\mathrm{mean}\left(QP_{\mathrm{non-I,per-seg}}\right)}{\mathrm{mean}\left(QP_{\mathrm{non-I,per-sec}}\right)} * Q \tag{16}$$

where

$QP_{\mathrm{non-I,per-seg}}$ is the average QP of all non-I frames in a segment;

$QP_{\mathrm{non-I,per-sec}}$ is the average QP of all non-I frames for each second;

Q is the per-segment video quality score as described in Equation (14).

Annex A

Helper function definitions

(This annex forms an integral part of this Recommendation.)

MOSfromR can be expressed as follows:

```
function MOSfromR(Q):
    MOS_MAX = 4.5
    MOS_MIN = 1.0

    if Q >= 100:
        return MOS_MAX
    if Q <= 0:
        return MOS_MIN

    return (
        MOS_MIN
        + ((MOS_MAX – MOS_MIN) * Q / 100)
        + Q * (Q – 60) * (100 – Q) * 0.000007
    )
```

RfromMOS can be expressed as follows:

```
function RfromMOS(MOS):
    x = (18566 – 6750 * MOS)
    if MOS > 4.5:
        MOS = 4.5

    if x < 0:
        num = 15 * sqrt(-903522 + 1113960 * MOS – 202500 * MOS * MOS)
        den = 6750 * MOS – 18566
        fra = num / den
        h = (pi – arctan(fra)) / 3
    else:
        num = 15 * sqrt(-903522 + 1113960 * MOS – 202500 * MOS * MOS)
        den = 18566 – 6750 * MOS
        fra = num / den
        ar = arctan(fra)
        h = arctan(num / den) / 3
    R = 20.0 * (8 – sqrt(226) * cos(h + pi / 3)) / 3
    return
```

*scaleto*5 can be expressed as follows:

```
function scaleto5(x):

    input_start = 1
    input_end = 4.5
    output_start = 1
    output_end = 5

    if x >= 4.5:
        return 5
    return output_start + ((output_end – output_start) /
(input_end – input_start)) * (
        x – input_start
    )
```

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. 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 creation of validation data. Models were retrained using a five-fold cross-validation approach, with their validation performance listed to show stability of the performance indicating no over-fitting.

Bitstream mode 3	Submitted model	0.421				
	Five-fold cross-validation	**0.394**	0.407	0.402	0.413	0.401

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 average prediction error over all models.
- A 50:50 training:validation split was determined randomly, but respecting the level of difficulty. In total, five different splits were defined. Each split had a balanced distribution of databases based on 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.

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.911] Recommendation ITU-T P.911 (1998), *Subjective audiovisual quality assessment methods for multimedia applications*.

[b-ITU-T P.1201.1] Recommendation ITU-T P.1201.1 (2012), *Parametric non-intrusive assessment of audiovisual media streaming quality – Lower resolution application area*.

[b-ITU-T P.1201.2] Recommendation ITU-T P.1201.2 (2012), *Parametric non-intrusive assessment of audiovisual media streaming quality – Higher resolution application area*.

[b-ITU-T P.1202] Recommendation ITU-T P.1202 (2012), *Parametric non-intrusive bitstream assessment of video media streaming quality*.

[b-ITU-T P.1202.1] Recommendation ITU-T P.1202.1 (2012), *Parametric non-intrusive bitstream assessment of video media streaming quality – Lower resolution application area*.

[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-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.*

[b-ITU-T P.1204.4] Recommendation ITU-T P.1204.4 (2020), *Video quality assessment of streaming services over reliable transport for resolutions up to 4K with access to full and reduced reference pixel information.*

[b-ITU-T P.1204.5] Recommendation ITU-T P.1204.5 (2020), *Video quality assessment of streaming services over reliable transport for resolutions up to 4K with access to transport and received pixel information.*

[b-ITU-T P.1401] Recommendation ITU-T P.1401 (2020), *Methods, metrics and procedures for statistical evaluation, qualification and comparison of objective quality prediction models*.

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, 2020