ITUPublications
Recommendations

International Telecommunication Union
Standardization Sector

Recommendation

ITU-T P.1204.5 (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 with access to transport and received pixel information

ITU-T P-SERIES RECOMMENDATIONS

Telephone transmission quality, telephone installations, local line networks

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

Recommendation ITU-T P.1204.5

Video quality assessment of streaming services over reliable transport for resolutions up to 4K with access to transport and received pixel information

Summary

Recommendation ITU-T P.1204.5 describes the hybrid no-reference video quality estimation model for monitoring the video quality for streaming using reliable transport (e.g., adaptive streaming (HAS) based on hypertext transfer protocol (HTTP) over the transmission control protocol (TCP) or quick user datagram protocol Internet connections (QUIC)). The estimate is validated for videos encoded with ITU-T H.264, ITU-T H.265, video payload type 9 (VP9) or AOMedia Video 1 (AV1) codecs at any resolution up to 4K/ultra-high definition-1 (UHD-1) resolution for personal computer (PC) monitors and television (TV) and up to 2 560 × 1 440 for mobile (MO) and tablet (TA) displays.

The ITU-T P.1204.x 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 these video quality models 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, combined further into a final per-session model output, an estimate of integral per-session quality (see, for example, [b-ITU-T P.1203], [b-ITU-T P.1203.2], [b-ITU-T P.1203.3]).

Recommendation ITU-T P.1204.5 was developed in collaboration with the Video Quality Experts Group (VQEG).

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

– Large-screen presentation 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.1204.5	2020-01-13	12	11.1002/1000/14158
1.1	ITU-T P.1204.5 (2020) Amd. 1	2021-01-07	12	11.1002/1000/14593
2.0	ITU-T P.1204.5	2023-10-29	12	11.1002/1000/15699

Keywords

Adaptive streaming, IPTV, mean opinion score, mobile video, mobile TV, monitoring, multimedia, over-the-top MOS, OTT, progressive download, QoE, TV, video.

* To access the Recommendation, type the URL https://handle.itu.int/ in the address field of your web browser, followed by the Recommendation's unique ID.

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.1204.5

Video quality assessment of streaming services over reliable transport for resolutions up to 4K with access to transport and received pixel information

1 Scope

This Recommendation describes the hybrid no-reference video quality assessment model 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 and adaptive streaming, for both mobile and fixed-network streaming applications. The video quality modules can also be used as a standalone video quality prediction model.

The model defined covers the use case of monitoring video quality where the video payload is fully encrypted and the pixel information is available, e.g., from the client side.

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 [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), video payload type 9 (VP9), AOMedia Video 1 (AV1), resolutions up to 4K/ultra-high definition-1 (UHD-1) and frame rates 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 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, [b-ITU-T P.1203.3] for sessions of 1 min to 5 min in duration).

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;
– Performance and quality assessment of live networks (including video encoding) 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 include the following.

– Over-the-top (OTT) services, as well as operator-managed video services (over the transmission control protocol (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 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 a 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 standalone document, the status of a Recommendation.

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

[ITU-T H.265] Recommendation ITU-T H.265 (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.1203.3] 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 (2023), *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 terms 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 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 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].

3.1.4 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.

3.1.5 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
AVC	Advanced Video Coding
AVI	Audio Video Interleave
CC	Content Complexity
CRF	Constant Rate Factor
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
IP	Internet Protocol
MKV	Matroska Video
MOS	Mean Opinion Score
MP4	MPEG-4 Part 14
MPEG	Moving Pictures Expert Group

MPEG2-TS	MPEG-2 Transport Stream
OTT	Over-The-Top
PC	Personal Computer
QUIC	Quick User datagram protocol Internet Connections
Rext	Range extension
RMSE	Root Mean Square Error
RTP	Real-time Transport Protocol
RTMP	Real-Time Messaging Protocol
SD	Standard Definition
TCP	Transmission Control Protocol
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 4096 × 2 160 or 3840 × 2 160;

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

– 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 – Application areas for which this Recommendation is applicable

Areas for which the model is applicable
In-service monitoring of video sent over reliable transport. Applicable to 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 – Application areas for which this Recommendation is not applicable

Ares 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, e.g., 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 ~7–10 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.</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/advanced video coding (AVC) (libx264), H.265/HEVC (libx265), VP9 (libvpx-vp9), AV1 (libaom-av1)</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.
AV1: Main (CPU-used 1, 6)</td></tr>
<tr><td rowspan="4">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 standard definition (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>High definition (HD)</td><td>720–1 080</td><td>500 kbit/s–15 Mbit/s</td><td>500 kbit/s–15 Mbit/s</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				
	Above HD	1 440–2 160	1.5 Mbit/s–45 Mbit/s	1.5 Mbit/s–20 Mbit/s
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 bits or 10 bits			
Chroma subsampling	YUV 4:2:0 and YUV 4:2:2 for H.264/H.265/VP9 YUV 4:2:0 for AV1			
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; 2160p, up to 60 frames/s. MO/TA: 1440p, 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 for H.264/H.265/VP9 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-Yamagishi]. NOTE 3 – For original content with a larger aspect ratio, letterboxing of up to 30% was allowed, that is, 1512 pixels height for video coded at 2160 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/MO, the model output is conservative and should be interpreted to correspond to a viewing distance of 1.5*H* to 1.6*H*.				

7 Building blocks

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

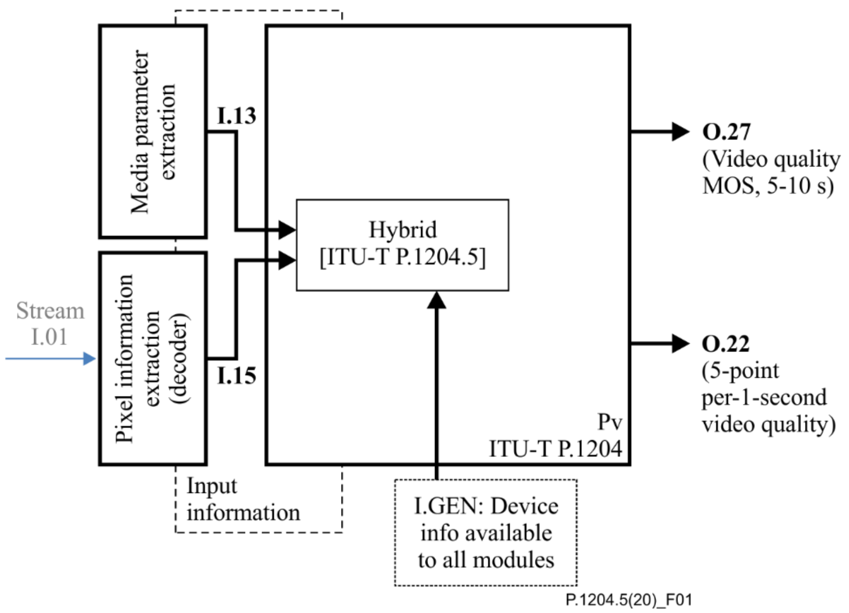

Figure 1 – Building blocks of the ITU-T P.1204 video quality model used in a standalone manner in the case of this Recommendation

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 model, 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 information, depending on its type (for details see Table 4):

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 inch;
- MO/TA: screen size 13 inch or smaller.

I.13 Video coding information.

I.15 Degraded video pixel information.

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

Table 4 describes the inputs I.GEN, I.13 and I.15.

Table 4 – I.GEN, I.13 and I.15 inputs description

ID	Description	Values	Frequency	Notation
I.GEN				
0	The resolution of the image displayed to the user	Number of pixels ($W \times H$) in displayed video	Per media chunk	*disRes*
1	The device type on which the media is played	"PC", "TV", "MO", "TA"	Per media chunk	*device*
I.13				
4	Video bitrate	Bitrate in kilobits per second	Per media chunk	*bitrate*
5	Video frame rate	Frame rate in frames per second	Per media chunk	*framerate*
6	Segment duration	Duration in seconds	Per media chunk	*duration*
7	Video encoding resolution	Number of pixels ($W \times H$) in transmitted video	Per media chunk	*codRes*
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	*codec, codecProfile*
I.15				

Table 4 – I.GEN, I.13 and I.15 inputs description

ID	Description	Values	Frequency	Notation
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 (using the bicubic upscaling method) to display resolution without buffering/stalling.	Per media chunk	*degVid*

Some examples of the inputs:

disRes: The video display resolution in number of pixels.

For instance, PC monitor and TV UHD screens have a video display resolution of:

$$disRes = 3840 \times 2160 = 8294400 \text{ pixels}$$

For instance, MO/TA screens have a video display resolution of:

$$disRes = 2560 \times 1440 = 3686400 \text{ pixels}$$

codRes: The video encoding resolution in pixels. The video resolution used to encode the video, for instance, $codRes = 854 \times 480 = 409920$ pixels.

7.3 Model output information

The model provides 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 is used, ranging from 1, meaning "bad", to 5, meaning "excellent", 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-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 Model algorithm

The video model defined in this Recommendation has one output, O.27. It provides a single output value on the five-point ACR scale (MOS) for a short video. The core model algorithm is described in clause 8.1.

8.1 Core model

8.1.1 Relative raw bitrate factor (*relRawBitrateRatio*)

For a parametric model, *bitrate* carries the most important information about the quality of the video. However, *bitrate* only makes sense together with information about the encoded chroma subsampling format (content complexity (CC) estimate). This is because the same video in YUV420 or YUV422 would need a slightly different bitrate in order to be encoded to the same quality. The same is also true for 8 bit or 10 bit reference videos.

relRawBitrateRatio is defined as:

$$relRawBitrateRatio = \frac{RawBitrateActual}{RawBitrate8bitYUV420} \quad (1)$$

$$RawBitrate8bitYUV420 = codRes * framerate * 1.5 * 8$$

$RawBitrateActual$ for different CC values:

- (yuv420p) 8 bit YUV420 = $codRes * framerate * 1.5 * 8.0$;
- (yuv422p) 8 bit YUV422 = $codRes * framerate * 2.0 * 8.0$;
- (yuv420p10le) 10 bit YUV420 = $codRes * framerate * 1.5 * 10.0$;
- (yuv422p10le) 10 bit YUV422 = $codRes * framerate * 2.0 * 10.0$.

Thus:

$$relRawBitrateRatio = f(CC) = \begin{cases} yuv420p & 1.0 \\ yuv422p & 2.0/1.5 \\ yuv420p10le & 10.0/8.0 \\ yuv422p10le & (10.0 * 2.0)/(8.0 * 1.5) \end{cases} \quad (2)$$

8.1.2 *codecProfile* to chroma subsampling type mapping

Separate *codecProfile* to chroma subsampling mappings are used for each codec. Note that this information may not be known to the model in all cases, in which case it can use a pre-assumed default chroma subsampling type:

H.264:

$$CC = g(codecProfile) = \begin{cases} ConstrainedBaseline & yuv420p \\ Main & yuv420p \\ Hi & yuv420p \\ Hi10 & yuv420p10le \\ Hi422 & yuv422p \\ others/unknown & yuv422p \end{cases}$$

H.265:

$$CC = g(codecProfile) = \begin{cases} Main & yuv420p \\ Main10 & yuv422p10le \\ Rext & yuv422p \\ others/unknown & yuv422p \end{cases}$$

VP9:

$$CC = g(codecProfile) = \begin{cases} 0 & yuv420p \\ 1 & yuv422p \\ 2 & yuv420p10le \\ 3 & yuv422p10le \\ others/unknown & yuv422p \end{cases}$$

AV1:

$$CC = g(codecProfile) = \begin{cases} Main & yuv420p \\ High & yuv420p10le \\ Professional & yuv422p10le \\ others/unknown & yuv420p \end{cases}$$

8.1.3 Adjusted log bitrate (*logBitrate*)

$$bitrateAdj = bitrate * \exp\left(-h_0 * (relRawBitrateRatio - 1)\right) \quad (3)$$

or

$$bitrateAdj = bitrate * \exp\left(-h_0 * \left(f\left(g(codecProfile)\right) - 1\right)\right) \quad (4)$$

and

$$logBitrate = \log10(bitrateAdj) \quad (5)$$

where h_0 is specified in Table 5.

Table 5 – *logBitrate* coefficients

	PC/TV	MO/TA
	h_0	h_0
H.264	1.1776641027814067e-09	0.5923649958216682
H.265	0.1648644781080738	0.6286917954823384
VP9	1.4370415811329779e-15	0.3595185885781488
AV1	9.999999999999999e-05	0.49999999999999994

8.1.4 Upscaling feature (*scaleFactor*)

scaleFactor defines the degree of upscaling after the decoding for display purposes.

$$scaleFactor = \max\left(\frac{disRes}{codRes}, 1\right) \quad (6)$$

The larger the *scaleFactor* value, the greater the upscaling performed at the display device.

8.1.5 Temporal feature (*framerateFactor*)

The maximum frame rate the model was tested for is 60 frames/s. The coded frame rate can be less than 60 frames/s. *framerateFactor* is defined as

$$framerateFactor = \max\left(\frac{60.0}{framerate}, 1\right) \quad (7)$$

The larger the *framerateFactor* value, the greater the frame rate up-conversion during playback.

8.1.6 Source complexity feature (*contentFactor*)

The model uses a codec-based source complexity estimation method. Using the constant rate factor (CRF) coding recipe of the libvpx-vp9 encoder, *degVid* is encoded at a certain quality *Q* to an encoded file *degVidEncoded*, where *Q* is an unknown quality value resulting from the CRF encoding of the *degVid* at a CRF value of 32.

To determine the *contentFactor* for H.264/H.265/VP9 codec the degraded video sequence is encoded using the following command line:ffmpeg –i degVid.avi –pix_fmt yuv420p –an –c:v libvpx-vp9 –crf 32 –b:v 0 degVidEncoded.mp4For the static build of the ffmpeg implementation used, see [b-ffmpeg_3.4], which uses the source code [b-libvpx-vp9 v1.6.1].

To determine the *contentFactor* for AV1 codec the degraded video sequence is encoded using the following command line:

```
ffmpeg –i degVid.avi –pix_fmt yuv420p –an –c:v libaom-av1 –crf 32 –b:v 0
degVidEncoded.mp4
```

For the static build of the ffmpeg implementation used, see [b-ffmpeg_5.1.2].

The *contentFactor* is determined as follows:

$$norm\_crf\_bitrate = \frac{sizebytes(degVidEncoded)*1000}{framerate*T*disRes} \quad (8)$$

where *sizebytes* is the size in bytes including the MP4 overhead.

The source complexity is computed as a function of $norm_{crf_{bitrate}}$

$$srcComplexity = 7.273 * log10(norm\_crf\_bitrate) \quad (9)$$

and

$$contentFactor = \mathrm{c}_1 * srcComplexity + \mathrm{c}_2 \quad (10)$$

where the constants c_1 and c_2 are specified in Tables 6 and 7.

Table 6 – *contentFactor* coefficients for PC/TV

	PC/TV	
	$\mathbf{c_1}$	$\mathbf{c_2}$
H.264	0.026020856130385718	0.18771981049276384
H.265	0.321901099557003	–0.9339240842451443
VP9	0.027131654431210638	–0.07758026781152491
AV1	0.027724803351637916	-0.15229669418176808

Table 7 – *contentFactor* coefficients for MO/TA

	MOS/TA	
	$\mathbf{c_1}$	$\mathbf{c_2}$
H.264	0.03304059217693778	0.5191195117506
H.265	0.054392293564817444	–0.4752924970529189
VP9	0.01703446988358945	–0.09703179546863315
AV1	0.018967755729372333	-0.15196435191178395

8.1.7 Feature integration

The constants a, b and c are computed as:

$$\mathrm{a} = \mathrm{a}_0 - \mathrm{a}_\mathrm{s} * \log10(u_\mathrm{a} * (scaleFactor - 1) + 1) - a_\mathrm{f} * framerateFactor - a_\mathrm{c} * contentFactor \quad (11)$$

$$\mathrm{b} = \mathrm{b}_0 - \mathrm{b}_\mathrm{s} * \log10(u_\mathrm{b} * (scaleFactor - 1) + 1) + \mathrm{b}_\mathrm{f} * framerateFactor + \mathrm{b}_\mathrm{c} * contentFactor \quad (12)$$

$$\mathrm{c} = \mathrm{c}_0 - \mathrm{c}_\mathrm{s} * \log10(u_\mathrm{c} * (scaleFactor - 1) + 1) - \mathrm{c}_\mathrm{f} * framerateFactor + \mathrm{c}_\mathrm{c} * contentFactor \quad (13)$$

with the constraint on a minimum value of b as in Equation 14:

$$\mathrm{b} = \max(0, \mathrm{b}) \quad (14)$$

The three computed constants along with another constant k_0 can be combined to a score value *S* using Equation 15.

$$S = \mathrm{a} * \left(\frac{1-\exp(-k_0*(logBitrate-\mathrm{c}))}{1+\exp(-\mathrm{b}*(logBitrate-\mathrm{c}))}\right) \quad (15)$$

All the model constants used in Equation 11 and constants used in the derivation of a, b and c are specified in Tables 8 and 9.

Table 8 – Feature Integration constants, PC/TV

Constant	PC/TV			
	H.264	H.265	VP9	AV1
a_0	5.677728847992967	5.03853891104581	4.859699233665362	4.999999999999999
b_0	3.4712005807048745	2.0993542290664227	2.6541304260526557	1.9622389633887367
c_0	2.326478357956036	2.8334365643929855	2.9399953618001136	2.9872409840441514
a_s	1.8350235211981674	2.558825165003877	2.3476224402785877	5.717534474637609
b_s	1.4141232302855393	0.5098792603744106	7.255415776808229e-11	9.999999999999999e-05
c_s	0.23475280755478767	0.22681818096833914	0.2873320369663877	0.04997627866562337
u_a	0.1778191362520981	0.08444039691348859	0.12643591444328875	0.020601186106930385
u_b	0.156900730863524	1.5410279574057658e-36	0.004818194829532265	0.330282384409527
u_c	42.406080941967936	2.0059093997172757	2.0509739990614357	69.89607767078054
a_f	0.39159165912177857	0.2525211972777661	0.15581905716465846	0.2973292141251956
b_f	2.6729710558144443e-28	2.6688343545615205e-21	6.690412679884795e-15	1.3736245971496305e-37
c_f	0.29490002469830306	0.21402618037698756	0.20483793964560515	0.382830506764624
a_c	1.6943267545826664e-13	0.0431077938951142	1.668359219633742e-14	7.951961674350778e-38
b_c	7.0362956885089e-14	0.43792733573736864	4.093588017285955	2.320340266589841
c_c	3.678498383915767	0.358852205906036	4.3023537324911105	6.052262005021103
k_0	1.4419774585129321	2.9400708635994275	2.9195734718894553	1.751244787657414

Table 9 – Feature integration constants, MO/TA

Constant	MO/TA			
	H.264	**H.265**	**VP9**	**AV1**
a_0	5.268960765324393	5.0474497689434275	4.984684538764142	4.968727251068815
b_0	3.970252547227931	1.26707140012788e-21	5.2136891589367425	1.2894001352986943e-18
c_0	0.955861731604233	2.884571319491612	2.7840703793378223	2.709056174062231
a_s	4.36888019813821	3.0455666232932663	5.803265994082781	4.16057739925183
b_s	2.1125548778844156	0.00017290708274250087	1.4701594292800126	1.9584330069917135e-11
c_s	0.40383887688983744	0.10996363240734348	0.21040175571457492	0.39999999588661567
u_a	0.024553971967259326	0.04988189636286348	0.01833878302910475	0.02684399919409856
u_b	0.5557309759968077	5.020735385579775	25.189492746842372	26.733809678612673
u_c	1.4393665855340954	3.351799514986455	4.425914043223159	0.020277979706128196
a_f	0.23654971807507216	0.2118845114345596	0.20658178681704242	0.2710149081970915
b_f	8.69531265907939e-37	3.1098630749524796	0.9720701616151223	1.7192436462133898
c_f	0.19146906019485413	0.1515064042031239	0.14910953368910074	0.25260824307933305
a_c	0.26458342387745737	7.844661892720165e-36	1.9881820627248652e-24	1.4751833641256406e-23
b_c	1.4427813426296531e-33	1.5165682395521835e-10	0.0017425312678303107	3.43156521514303e-18
c_c	2.953357298372877	2.0316300541234864	6.80531487679437	10.24111816313156
k_0	2.7475799851849545	2.20751587008015	2.5709237715026094	1.8913833959565682

8.1.8 Device-based linear mapping

A final linear mapping is used to map the *S* score to the predicted MOS for H.264/H.265/VP9 codecs:

$$\widehat{MOS} = \min(\max(m_1 * S + m_2, 1), 5) \qquad (16)$$

where m_1 and m_2 are specified in Table 10.

Table 10 – Device-based linear mapping coefficients

	m_1	m_2
PC monitor	0.967	0.153
TV	1.051	–0.187
MO	0.942	0.146
TA	1.080	–0.330

NOTE – For AV1 codec unity mapping $\boldsymbol{m_1 = 1}$ and $\boldsymbol{m_2 = 0}$ is used.

Appendix I

Performance figures

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

In this section, 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

Hybrid No reference mode 0 model	*Submitted model*	0.452				
	Fivefold cross-validation	0.451	0.440	0.441	0.443	0.441
NOTE – The *submitted model* is the model trained on the exchanged training databases and frozen before creation of validation data. Models were retrained using a fivefold cross-validation approach, with their validation performance listed to show the stability of the performance indicating no over-fitting.						

The following steps were carried out:

All training and validation databases were merged to obtain 26 different short databases (18 PC/TV and 8 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.

The final selected model is the model from cross-validation set 2 (RMSE 0.440). This is because the model resulting from this split has the best overall performance.

Appendix II

Long-term integration module (Pq) for ITU-T P.1204.5

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

The current approach for long-term integration described in this appendix is to be considered as an intermediate solution. It has been developed based on a total of 6 databases (2 for training and 4 for validation), while the short-term video quality model (Pv) described in the normative part of this Recommendation was developed using 26 databases (13 for training and 13 for validation).

As a consequence of the relatively low number of long-sequence databases used, the integration module presented in this appendix does not form an integral part of the Recommendation, and is considered for information. It is planned to be superseded in the future by a more comprehensive integration module trained and validated on a higher number of databases, which is currently under development.

The streaming parameters ranges used in the training and validation tests of the long-term integration model presented in this appendix are summarized in Table II.1.

Table II.1 – Parameter ranges used in the tests for the development of the ITU-T P.1204.5 long-term integration model

Video sequence duration	60 seconds – 5 minutes
Initial-loading delay	0-30 seconds
Total stalling duration	0-26 seconds
Number of stalling events	0-5
Total number of quality level switches	0-39

The ranges of parameter settings of other test factors related to each of the segments in a video session are summarized in Table 3 of ITU-T P.1204.5. It should be noted that audio quality was not varied in any of the 6 tests used to train and validate the long-term integration model presented in this appendix. For the PC/TV case, the audio from the source video was encoded using 16-bit PCM with 2 channels and 48 kHz sampling frequency. For the case of mobile/tablet, the AAC codec was used for audio with a bit rate of 512 kbits/s.

II.1 Model input

The model must receive the following input signals regardless of the mode of operation:

– O.21: audio quality per output sampling interval, as specified in clause 7 of [ITU-T P.1204];
– O.22: video quality per output sampling interval, as specified in ITU-T P.1204.5 – see clause 7.3;
– I.14: stalling events, as described in [ITU-T P.1204] clause 7;
– I.GEN: device type (either of "PC", "TV", "Mobile" or "Tablet"), as specified in [ITU-T P.1204].

II.2 Model output

The Pq model outputs the following information:

– O.23: perceptual buffering indication;
– O.34: audiovisual segment coding quality per output sampling interval;
– O.35: final audiovisual coding quality score;

– O.46: final media session quality score.

The values O.23, O.35 and O.46 will be output once per session.

The value O.34 will be output once per output sampling interval.

II.3 Model description

II.3.1 Parameters

The parameters are as follows:

1 Initial-loading time in seconds – $initialLoadingLen$. This can be extracted from I.14, i.e., the stalling duration value at playout time 0. If not specified in I.14 the default value is 0.0.

2 Total stalling length in seconds – $totalBuffLen$. This can be computed from I.14, i.e., the sum of all stalling values not including $initialLoadingLen$. If not specified in I.14 the default value is 0.

3 Number of stalling – $numStalls$. This can be computed from I.14, i.e., the number of all stalling events (not including the initial-loading event). If not specified in I.14 the default value is 0.

4 Total buffering-free video duration in seconds – T. This is the length of the O.22 list.

5 Time between end of last buffering to the end of video in seconds – $timeSinceLastBuff$. For a stalling-free video this is equal to T.

II.3.2 Computation of O.34

The audiovisual quality score O_{i} for the *i*th output sampling interval is derived from the linear combination of *i*th *O*.21 (i.e., a_{i}) and the *i*th *O*.22 (i.e., v_{i}) score as follows:

$$O_{\mathrm{i}} = 0.05 * a_{\mathrm{i}} + 0.95 * v_{\mathrm{i}} \qquad \text{(II-1)}$$

Note that it is assumed that audio signal is always present, i.e., *O*21 and *O*22 arrays have equal length, and audio is coded with high quality, i.e., audio MOS of 4.5 or above. Let $O.34 = [O_1, O_2, O_3, \dots, O_{\mathrm{T}}]$ denote an array of *O*.34 scores.

II.3.3 Computation of O.35

1 First a sliding window of size 30 is used with a step size of 1 over the $O.34$ array to yield overlapping pieces of *O*.34 scores. Each piece contains 30 scores. The window sliding is stopped if the scores contained in the window are less than 30. For each piece a normalized histogram of the *O*.34 quality scores is computed using the bin boundaries given below:

$$\text{histBins1} = (1.0, 1.5), (1.5, 2.5), (2.5, 3.5), (3.5, 4.5), (4.5, 5.0)$$

Let $qualHistogramList = [h_0, h_1, h_2, \dots, h_{\mathrm{N}}]$ denote the array of histograms, where h_{i} denotes the histogram resulting from the ith shift of the sliding window, $0 \le i \le N$. The pseudo code for computing the histogram can be found at the end of this section.

2 A complementary array of quality changes $\Delta O.34 = [\Delta O_1, \Delta O_2, \Delta O_3, \dots, \Delta O_{T-1}]$ is prepared using the difference between successive *O*.34 scores, i.e., $\Delta O_{\mathrm{i}} = O_{\mathrm{i+1}} - O_{\mathrm{i}}$ for $1 \le i \le T - 1$. Like (1), a sliding window of size 30 is used to create normalized histograms based on $\Delta O.34$ array, using the bin boundaries.

$$\text{histBins2} = (-4.5, -3.5), (-3.5, -2.5), (-2.5, -1.5), (-1.5, -0.5), (-0.5, 0.5), (0.5, 4.0)$$

Similar to (1), let $qualChangeHistogramList = [g_0, g_1, g_2, \dots, g_{\mathrm{N-1}}]$ denote the array of quality change histograms. The pseudo code for computing the histogram can be at the end of this section.

3 Let h_{ij} and g_{ij} denote the jth element of histograms of h_i and g_i, respectively. Create an aggregate array $F = [f_0, f_1, f_2, \ldots, f_{N-1}]$ by combining $qualHistogramList$ and $qualChangeHistogramList$ as follows:

$$f_i = \sum_{j=1}^{5} a_j * h_{ij} + \sum_{j=1}^{6} b_j * g_{ij} \quad \text{(II-2)}$$

where $0 \le i \le N - 1$.

The values of histogram related constants are given in Table II.2.

Table II.2 – Values of histogram related constants

a_1	1.7036144962372886
a_2	1.6281208003842298
a_3	2.14625868168416
a_4	3.154522195465948
a_5	3.1811440812907144
b_1	-12.892854165904497
b_2	-6.205923716980252
b_3	-2.477111070479436
b_4	-0.9875867258584734
b_5	0.778247340510056
b_6	0.4101562929016858

4 Compute a feature list and linearly weight the features to compute the O.35 score

$$L = [\min(F), \max(F), \mathrm{median}(F), \mathrm{mean}(F), f_{N-1}]$$

$$O.35 = \sum_{j=1}^{5} w_j * l_j \quad \text{(II-3)}$$

where l_j denotes the jth feature of the feature list L.

The values of O.35 related constants are given in Table II.3.

Table II.3 – Values of *O*.35 related constants

w_1	0.29508584543387967
w_2	0.00146837942360000
w_3	0.00118943982340000
w_4	0.35482926488923905
w_5	0.34742707042988136

Pseudocode for histogram computation

```
function hist = histogram (x, bins):
  h = An array of zeros with size one less than bins
  for every value x_i of x:
    for every bin b of bins:
      h[b] += max(0, 1 - abs((bins[b + 1] + bins[b]) / 2 – x_i))
  hist = normalize each element of h by sum(h)
```

II.3.4 Computation of *O*.46

First by using a four-dimensional exponential function, we create an intermediate variable $InitLoadAndStallImpact$, which combines the effect of initial-loading, stalling and recency due to the last occurred stall. The values of initial-loading and stalling related constants are given in Table II.4.

$$InitLoadAndStallImpact = \exp(-s_1 * numStalls) * exp\left(-s_2 * \frac{initialLoadingLen}{T}\right) * \exp\left(-s_3 * \frac{totalBuffLen}{T}\right) * \exp\left(-s_4 * \frac{T - timeSinceLastBuff}{T}\right) \quad \text{(II-4)}$$

Table II.4 – Values of initial-loading and stalling related constants

s_1	0.08768743173928367
s_2	0.7167602031580045
s_3	0.06981494241303295
s_4	0.30959519998764706

$InitLoadAndStallImpact$ is the factor by which the buffering-free quality *O*.35 is reduced, i.e.,

$$Q = 1 + (O.35 - 1) * InitLoadAndStallImpact$$

The intermediate aggregated quality value Q is mapped to the final *O*.46 using a linear mapping and a limiting function which constraints the output value between 1 and 5.

$$O.46 = \min(5.0, \max(1.0, m * \mathrm{Q} + c))$$

where constants m and c depend on the display device type – see Table II.5.

Table II.5 – Values of display devices related constants

Display device	m	c
TV/PC monitor	1.11	−0.232
Mobile/tablet	1.0	−0.25

II.3.5 Computation of *O*.23

The perceptual buffering indication *O*.23 is calculated based only on the impacts of initial-loading, buffering and the media length as follows:

$$O.23 = 1 + 4 * InitLoadAndStallImpact$$

II.4 Performance figures

In this clause, the aggregated RMSE of the model is given. Aggregated RMSE is defined as:

$$RMSE = \frac{1}{W} \sum_{\mathrm{k}=1}^{M} w_{\mathrm{k}} \cdot RMSE_{\mathrm{k}} \quad \text{(II-5)}$$

where *M* represents the total number of (training and validation) databases, w_k and $RMSE_k$ espectively the weight and RMSE for database k r, and $W = \sum_{\mathrm{k}=1}^{M} w_{\mathrm{k}}$. Training and validation databases have different weights:

$$w_{\text{training}} = 0.1$$

$$w_{\text{validation}} = 0.9$$

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.6 – Performance of the Pq described above, based on the submitted version of ITU-T P.1204.5

RMSE	**Training**		**Validation**	
	No. of databases	***No. of Samples***	***No. of databases***	***No. of Samples***
0.553	2	82	4	134

Table II.7 – Performance of the Pq described above, based on the standardized ITU-T P.1204.5

RMSE	**Training**		**Validation**	
	No. of databases	***No. of Samples***	***No. of databases***	***No. of Samples***
0.529	2	82	4	134

Appendix III

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 III.1 – Validation performance of Pv model (AV1 only) on 6 unknown validation databases

Hybrid **No reference mode 0 model**	Validated model	0.442

- Initially the pixel-based reference model was validated on 6 validation databases, which achieved high prediction efficiency (avg. RMSE 0.362).
- Then 8 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 8 training databases were computed using the validated ITU-T P.1204.4 [b-ITU-T P.1204.4] model, and was used as ground truth for the following retraining.
- Bitstream and hybrid model were trained on 8 training databases and model coefficients were frozen.
- 6 unknown validation databases (the same databases which were used for pixel model validation) with subjective MOS were used to validate the bitstream and hybrid models.

Note on Source complexity feature (*contentFactor*)

When AV1 was used to compress the video to simulate the compression distortion, the VP9 based source complexity estimate of eq 8–10 can yield suboptimal results due to affine motion feature of AV1. Instead, a *contentFactor* based on the AV1 codec shall be used.

Using the AV1 based *contentFactor* lead to the following average RMSE for the six validation databases.

Table III.2 – Re-optimized performance of Pv model (AV1 only) on 6 databases using AV1 based *contentFactor*

Hybrid **No reference mode 0 model**	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-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-ffmpeg_3.4] FFmpeg developers (2000–2017). ffmpeg version 3.4, built with gcc 7.2.0 (GCC). https://web.archive.org/web/20200914211304/https://ffmpeg.zeranoe.com/builds/win64/static/ffmpeg-3.4-win64-static.zip

[b-ffmpeg_5.1.2] FFmpeg developers (2000-2017). ffmpeg version 5.1.2, built with gcc 12.1.0 (GCC). https://www.gyan.dev/ffmpeg/builds/packages/ffmpeg-5.1.2-full_build.7z

[b- Yamagishi] K. Yamagishi, N. Egi, N. Yoshimura, and P. Lebreton (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.

[b-libvpx-vp9 v1.6.1] Github (2020), libvpx-vp9 v1.6.1 source code. https://github.com/webmproject/libvpx/releases/tag/v1.6.1

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

Published in Switzerland
Geneva, 2023