A Bitstream-based, Scalable Video-Quality Model for HTTP Adaptive Streaming: ITU-T P.1203.1

Alexander Raake*, Marie-Neige Garcia†, Werner Robitza‡, Peter List‡, Steve Göring*, Bernhard Feiten‡
*Audiovisual Technology Group, Technische Universität Ilmenau, Germany, Email: alexander.raake@tu-ilmenau.de
†Assessment of IP-based Applications Group, Technische Universität Berlin, Germany
‡Telekom Innovation Laboratories, Deutsche Telekom AG, Germany

***Abstract*—The paper presents the scalable video quality model part of the P.1203 Recommendation series, developed in a competition within ITU-T Study Group 12 previously referred to as P.NATS. It provides integral quality predictions for 1 up to 5 min long media sessions for HTTP Adaptive Streaming (HAS) with up to HD video resolution. The model is available in four modes of operation for different levels of media-related bitstream information, reflecting different types of encryption of the media stream. The video quality model presented in this paper delivers short-term video quality estimates that serve as input to the integration component of the P.1203 model. The scalable approach consists in the usage of the same components for spatial and temporal scaling degradations across all modes. The third component of the model addresses video coding artifacts. To this aim, a single model parameter is introduced that can be derived from different types of bitstream input information. Depending on the complexity of the available input, one of four scaling-levels of the model is applied. The paper presents the different novelties of the model and scientific choices made during its development, the test design, and an analysis of the model performance across the different modes.**

Keywords*—*Video quality, Quality of Experience, Quality model, HTTP Adaptive Streaming, HAS, Standardization

I. Introduction

With HTTP-based Adaptive Streaming HAS, videos are typically split into segments of 1 to 15 s duration (see e.g. DASH, ISO/IEC 23009-1 and [1]). These segments are processed and encoded into multiple representations so as to achieve different net bitrates, referred to as the *adaptation set*. Typical HAS players request segments in the representations that are considered suitable for delivering the best possible quality under the current network throughput constraints. Here, different playout algorithms can be conceived; they typically try to yield a good compromise between avoiding playout-buffer underruns and stalling, low start-up delay, high average audio and video quality, and a low amount of quality switches [1], [2]. Popular examples of implementations are HLS (HTTP Live Streaming, Apple), MSS (Silverlight Smooth Streaming, Microsoft), HDS (HTTP Dynamic Streaming, Adobe) or standardized solutions such as DASH (Dynamic Adaptive Streaming, MPEG, ISO/IEC 23009).

In most real-life applications, video quality, adaptations and the impact due to stalling and initial delay can be considered as the key aspects of overall HAS QoE [1], [3], [4]. In the standardization activity within ITU-T Study Group 12 (SG12) Question Q.14/12, initially called P.NATS, an integral modular model framework for HAS quality monitoring was developed. Here, the QoE of an HAS session was modeled in terms of an integration of short-term audio- and video-quality scores with stalling and initial delay information. The P.NATS competition has recently lead to the adoption of the P.1203 Recommendation series, see Sec. III. In this paper, its short-term video quality module, ITU-T P.1203.1, is described. The paper primarily focuses on the model candidate initially submitted to the competition by the authors.

QoMEX2017 – Erfurt, Germany; 978-1-5386-4024-1/17/$31.00 ©2017 IEEE

The P.1203 model was developed within SG12 based on a large set of training and validation test databases (30 subjective tests in total). Model development targeted the prediction of integral quality ratings obtained from test subjects for 1 to 5 min long HAS-type viewing sessions, given by the test subjects on the 5-point Absolute Category Rating (ACR) scale (cf. ITU-T Rec. P.910). The long-term quality ratings are based on integrating the output from the underlying audio- and video-quality modules that provide per-one-second output scores on the 5-point scale. Since the ground-truth ratings were collected for long sequences, the test databases did not contain direct ground-truth user ratings for training and validating the short-term audio- and video-quality estimation modules (cf. Sec. III). Hence, the short-term video-quality model presented here was primarily developed based on a reverse-engineering of the integral quality scores collected for the 1 to 5 min media sessions, compensating for quality-switching, initial-delay or stalling-events.

In this paper, the term *scalable* is used to refer to the main contribution of the presented model, namely to be the first bitstream-based video quality model that is *scalable* to the type of input information available to it. It is principally based on the same structure and coefficient sets across the different modes of operation of P.1203, and mainly only one primary compression-specific parameter is calculated in a mode-dependent manner. The paper highlights the various novelties of the model, the scientific choices made during the model development and subjective test design, and provides a performance analysis for the different modes.

In Sec. II we briefly review related work. We give an introduction to the P.1203 framework in Sec. III, with information on test databases and processing in Sec. IV. The video quality model is introduced in Sec. V, including a performance analysis. An outlook on ongoing work and standardization for HAS QoE is given in Sec. VI.

II. Related work

Different approaches to QoE monitoring were conceived in the past (e.g. [3], [5]–[7]). Among others, [1], [4] review different studies on the impact of video quality and quality switches versus stalling and initial delay. Accordingly, many approaches for predicting integral HAS QoE are based on the integration of short-term video-quality estimates with initial-delay- and stalling-degradation terms to an overall, temporally pooled QoE [1], [8], [9]. It was shown that when only quality switches and no stalling- or initial delay occur, already a simple arithmetic mean over short-term quality predictions can deliver first estimates of integral quality [1], [8], [9]. Here – as pointed out in [8] – the applied short-term video-quality module is of high importance for the accuracy of model predictions.

For video there are different means to achieve a given bitrate target for the representations of an adaptation set. The main approaches are:

- Spatial scaling or *upscaling*, typically leading to different degrees of blurriness of the video frames.
- Temporal scaling, leading to different amounts of jerkiness.
- Scaling by compression, related to coding artifacts such as blockiness, ringing or non-uniform blurriness due to de-blocking filters.

For a complete short-term video quality model, these individual types of degradations can be combined [10]–[12].

III. Overview of P.NATS / P.1203

The short-term video quality model presented in this paper is part of the overall P.1203 model framework, see Fig. 1. The P.1203 series comprises a total of four standard documents, with P.1203 as the entry document containing all general information. The three further Recommendations specify the three modules P_v (P.1203.1, *video quality estimation module*, this paper), P_a (P.1203.2, *audio quality estimation module*, see also [13]) and P_q (P.1203.3, *quality integration module*). The standard series can be applied to two viewing scenarios, (i) on PC- or TV-screens, and (ii) on smartphones.

Different levels of encryption of the media bitstream can be handled by P.1203. This is reflected by its four modes of operation, offering incremental access to input information:

0 *Mode 0* corresponds to the highest considered level of encryption and lowest computational requirements; only has access to codec, target bitrate, resolution, frame rate, and segment durations and sizes.
1 *Mode 1* adds audio and video frame sizes and durations, and video frame types (e.g., I-, and Non-I).
2 *Mode 2* allows access to up to 2% of bitstream information, to reduce computational complexity over full access.
3 *Mode 3* adds access to full bitstream.

At its different layers, the model framework has interfaces for different types of in- and output information. The Input I.01 is the actual packet stream (see Fig. 1) from which different types of information are derived, related to audio (I.11), video (I.13) and player-side (re-)buffering in terms of initial delay and stalling (I.14). In this paper, only I.13 is addressed. All input- and output-level information can be considered as diagnostic information, a key benefit of the parametric approach followed by P.1203.

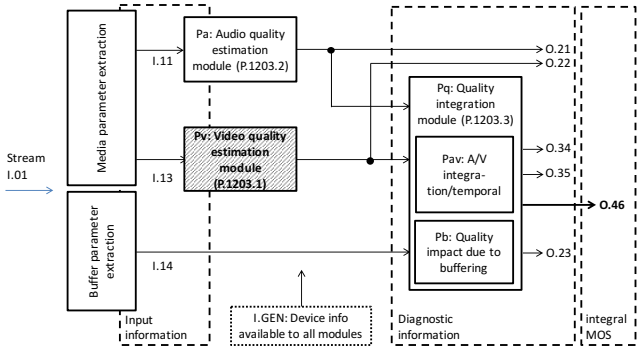

Fig. 1: Overview of P.1203 standard. The short-term video quality module P_v addressed in this paper is highlighted.

The P.1203 model provides different types of output information. The main output O.46, the *final media session quality score*, is given on a Mean Opinion Score (MOS) scale ranging from 1 to 5. O.22 is the short-term video quality per *output sampling interval*, a vector of per-one-second scores for each session on a 1–5 scale. O.21 is its corresponding audio quality output. The other outputs O.23, O.34, and O.35 can be used for diagnostics (see P.1203 for details).

A. Model development process, Mixing-and-Matching

The P.NATS Phase 1 model was developed in a modelling competition with a total of seven participating companies, in the following referred to as the *P.NATS group*. Here, a novel approach – in terms of ITU-T SG12 processes – was applied, using so-called *mixing-and-matching*. Each proponent was allowed to submit one module candidate per each of the three modules P_a, P_v and P_q, and per each of the four modes, hence a maximum of $3 \cdot 4 = 12$ modules. The own modules were trained by each individual participant based on the set of training databases (Sec. IV) with the proprietary combination of the P_a, P_v and P_q modules for a given Mode i ($i \in [0, 3]$).

During the validation phase, the modules submitted for a given Mode i from all companies were combined into all possible combinations of P_a, P_v and P_q, to form different full P.NATS model candidates. Since the range of the per-one-second output of a given P_a (O.21) or P_v module (O.22) may not optimally fit the expected input for a given P_q module, linear mappings of the O.21 and O.22 scores were introduced as additional degrees of freedom. The O.46 output of the resulting "mixing-and-matching" candidates were compared against the ground-truth validation subjective test databases in terms of Root Mean Square Error (RMSE) according to ITU-T Rec. P.1401. In case that either a candidate's own complete model or a mixed-and-matched combined model could not be distinguished in a statistically significant manner from the best-performing model, it was considered as part of the winning candidate models for the given Mode (per module type P_a, P_v and P_q). A model for any higher Mode was only standardized if it performed significantly better than the models for the respective lower Modes.

In a subsequent merging phase, the P.NATS group decided to standardize one single (merged) P_q and P_a module each

TABLE I: Training (TR) and validation (VL) databases (DB). M: Mobile, PC: PC screen; # PVS: number of PVSs; dur: duration in min.

DB	M	PC	# PVS	dur.	DB	M	PC	# PVS	dur.
TR01	x	x	60	1	VL01	x		60	1
TR02	x	x	60	1	VL03		x	60	1
TR03	x	x	60	1	VL04		x	60	1
TR04	x	x	60	1	VL05	x		30	2
TR05		x	22	3	VL06		x	30	2
TR05M	x		22	3	VL07		x	30	2
TR06	x	x	22	3	VL08		x	30	2
TR07	x	x	22	3	VL09		x	22	3
TR08		x	14	5	VL10		x	22	3
TR09	x	x	14	5	VL11		x	22	3
					VL12	x		15	4
					VL13		x	15	4
					VL14		x	14	5

across all four Modes. For the video quality estimation module P_v, P.1203.1 describes a separate model for each Mode. The video quality model submitted by the authors and mainly described in this paper was part of the winning groups for all four Modes 0–3, and was selected as the basis for the P.1203.1 standard.

IV. Databases

A. Database Overview

The set of 30 databases created by the P.NATS group was split into 17 training and 13 validation databases, the latter created after the submission of the module candidates. The main factors varying between the databases comprised the length of sequences (1 to 5 min), the type of degradations primarily investigated (focus on quality switches or stalling events, frequency of quality switches vs. amplitude, etc.) and the display device used (PC/TV screens or mobile phones). A total number of 1064 audiovisual sequences (Processed Video Sequences, PVSes) were generated. Due to the split according to the display device these were partly rated on both devices or only one. Table I gives an overview of all databases.

B. Test Procedure

A common procedure was defined for all tests. Overall, a setup similar to ITU-T P.910 was followed, using the 5-point Absolute Category Rating scale in single-stimulus tests. The test sequences were designed in such a way that different conditions were paired with different source videos. In contrast to a full-factorial test design, where all conditions are applied to all source videos, no source video was repeated within the same test. The setup follows an "immersive" test paradigm [14], [15], aiming to increase the enjoyability for viewers and overall ecological validity of the test. With stimuli durations of 1 to 5 min, a repetition of contents was assumed to get boring for test participants. A variety of source contents was used. Requirements on the technical and semantic nature of the clips were defined: they needed to have pristine quality at Full HD resolution or more, carry at least 0.07 Bits per pixel of information (manually verified), contain as little dialog as possible (in order to allow for tests in different languages) and be enjoyable overall. To ensure reliability of the data, outlier detection was performed on the results. Any given subject's ratings was correlated with the MOS of the entire subject pool. If a per-subject Pearson correlation of less than $r = 0.70$ was found, the subject was rejected and the calculation was repeated. In total, a requirement of at least 24 valid subjects was imposed per each test.

C. Processing

In order to generate realistic adaptive streaming conditions without playout issues, all PVSes were pre-rendered before the tests, based on the following processing steps for each source sequence: (1) Cutting of original source videos to length, scaling to 1920×1080 pixels. (2) Encoding into different representations (i.e., different bitrates, resolutions, frame rates) using x264 and a two-pass VBR setting. (3) Creation of MPEG-2 TS segments from given representations. (4) Decoding and stitching of sequences according to the given condition (i.e., simulating quality switches between different representations). (5) Insertion of re-buffering indicator (moving spinner) if required.

In this Phase 1 of P.NATS, variable bitrate encoding and a two-pass approach was used for attaining the highest possible quality. Segmentation was performed on the respective encoded representations. With such fixed bitrate target per representation and the taken approach, the encoding of a source sequence with a given content complexity lead to a content-dependent quality-impact. More recent approaches encode HAS segments individually and in parallel [16], [17], after having identified the minimum necessary bitrate for a given content and quality level. Hence, for future work it will be interesting to adapt bitstream-based parametric models to consider different encoding strategies.

V. Scalable Model: Video-Quality Module (Pv)

The three types of degradations addressed with the video quality model are those due to (i) upscaling (degradation D_u), (ii) temporal scaling (degradation term D_t) and (iii) compression (D_q). With our model, scalability is achieved by keeping the model's architecture the same across the different Modes. The model components that represent the impact due to spatial (D_u) and temporal scaling (D_t) remain identical across Modes, as well as the general structure of the quantization / compression-related component D_q. Only for the D_q component, one intermediate parameter $quant$ is calculated in a way that needs to be *scalable* with the respective Mode, and hence the available type of input information. For Mode 1, an additional term was introduced during the standard finalization, as it was found to improve the overall Mode 1 performance as compared to using the original P_v module submitted by the authors for this Mode.

A. Core model and mode-independent components

The P_v model is based on a core model that comprises the four different Modes. It includes the degradation terms for video up-scaling and temporal scaling. The degradation term for video compression artifacts is the only mode-dependent one. In the following, the core model and the mode-independent terms are described. The (mode-dependent) model component for handling compression-related degradations is described in Sec. V-B. A general overview of the P_v module is shown in Fig. 2.

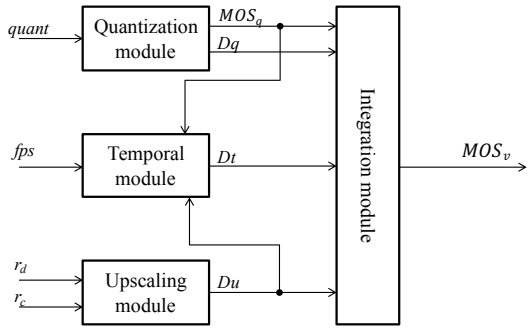

Fig. 2: P_v module, adapted from ITU-T Rec. P.1203.1.

The model uses a set of primary parameters derived from I.13 depending on the mode:

- $quant, \in [0,1]$: Parameter quantifying the quantization degradation. This is the only parameter that is calculated in a Mode-specific manner.
- br: Bitrate in kBit/s.
- fr: Video frame rate in frames per seconds.
- r_d: Video display resolution in total number of pixels (i.e., width $\times$ height).
- r_c: Video encoding resolution in total number of pixels.

The input information is processed by the model using a sliding window with a maximum length of 20 s (see Sec. 7.4 of P.1203). It ensured that during the competition no temporal integration (that would otherwise happen in P_q) would be performed by P_v. The model calculations described in the following are applied to all the available information within the sliding window, as long as it belongs to (contiguous) segments of the same representation of the adaptation set, called the *measurement window*.

As indicated above, the model output O.22 is provided on the 5-point ACR scale per each second of media. O.22 is referred to as MOS_v in the following. The per-one-second output can be pooled by simple averaging to deliver more traditional per-10-second scores, and thus use it as a standalone video quality model.

First, an overall degradation D is calculated from the upscaling degradation D_u, the temporal degradation D_t and the quantization degradation D_q. All degradation values are expressed on a scale from 0 to 100, following the impairment-principle underlying the "Transmission Rating Scale" of the E-model (ITU-T Rec. G.107, 2015). On this internal scale, it is assumed that individual degradations can better be summed up, also since in subjective tests the 5-point scale typically saturates at its end-points.

$$D = \max\left(\min\left(D_q + D_u + D_t, 100\right), 0\right) \tag{1}$$

Here, the max- and min-operations ensure that $D \in [0, 100]$. From the overall degradation, the quality Q_v on a scale from 0 to 100 is obtained according to

$$Q = 100 - D \tag{2}$$

The output O.22, MOS_v, is then calculated as

$$MOS_v = \begin{cases} \hat{MOS}_q & \text{if } D_u = 0 \ \& \ D_t = 0 \\ \text{RfromMOS}\left(Q\right) & \text{otherwise} \end{cases} \tag{3}$$

The rationale behind Eq. (3) can better be understood when the quantization degradation D_q is introduced:

$$D_q = 100 - \text{RfromMOS}\left(\hat{MOS}_q\right) \tag{4}$$

$\hat{MOS}_q$ corresponds to the estimated quality obtained when only compression degradation is present. With the case-wise calculation in Eq. (3) we avoid a double-conversion between the 100-point and the 5-point "MOS" scale.

Here, RfromMOS corresponds to the S-shaped conversion from the 5-point ACR scale to the model internal 100-point scale, similar to the calculations within the E-model ITU-T Rec. G.107 (for reasons of space, the equations are not included here; see Annex E of P.1203.1, also of the inverse function MOSfromR). In essence, it decompresses the saturation of the 5-point-scale at the end points. The coding-related quality $\hat{MOS}_q$ is calculated as

$$\hat{MOS}_q = q_1 + q_2 \cdot \exp\left(q_3 \cdot quant\right), \tag{5}$$

and subsequently clipped to $[1, 5]$ (simple $\min$, $\max$ operations). The coefficients are set to fixed values, with $q_1 = 4.66$, $q_2 = -0.07$ and $q_3 = 4.06$ (cf. P.1203.1). How the parameter $quant$ is calculated in a Mode-specific (i.e. "scalable") manner is described in Sec. V-B.

The upscaling degradation term D_u is calculated as

$$D_u = u_1 \cdot \log_{10}\left(u_2 \cdot (scaleFactor - 1) + 1\right), \tag{6}$$

with

$$scaleFactor = \max\left(\frac{r_d}{r_c}, 1\right) \tag{7}$$

The two coefficients are set to fixed values, namely $u_1 = 72.61$ and $u_2 = 0.32$ (see P.1203.1).

The temporal degradation D_t addressing the impact due to frame rate is specified here in a slightly simplified manner as compared to P.1203.1, due to the paper space limitations:

$$D_t = \begin{cases} \frac{(100 - D_q - D_u) \cdot (t_1 - t_2 \cdot fr)}{t_3 + fr} & \text{if } fr < 24 \\ 0 & \text{otherwise} \end{cases} \tag{8}$$

In P.1203, the coefficients are set to $t_1 = 30.98$, $t_2 = 1.29$ and $t_3 = 64.65$. Based on the P.NATS tests, the frame rate of $fr = 24$ fps is assumed to be the threshold beyond which no jerkiness and hence no temporal-scaling degradation is perceived.

B. Mode-dependent calculations, quant, and Mode 1

At the time of submitting the P_v modules, the only parameter varying between Modes 0 to 3 in our model candidate was $quant$, with the rest of the model being Mode-independent. Hence, $quant$ is the *scalable* component of the model. During the merging procedure, the model was adapted in collaboration with the P.NATS group. During this collaboration step, the initial Mode 1 model was extended by the P.NATS group, adding a component to the $quant$-related calculation that is based on the frame-sizes for specific frame types. This way, the Mode 1 model better accounts for different content complexity. For Mode 0, frame-type and -size information is not available, and hence no such extension is possible. As content-specific information is implicitly contained in the Mode 2 and Mode

3 P_v modules (see below), an extension by frame-type and -size information does not introduce an improvement for these higher Modes.

In the following, we will briefly outline the $quant$ calculation for the different Modes.

For Mode 0, $quant$ is calculated as:

$$quant = a_1 + a_2 \cdot \ln\left(a_3 + \ln\left(br\right) + \ln\left(br \cdot bpp + a_4\right)\right), \quad (9)$$

with bpp the number of Bits per pixel, calculated as

$$bpp = \frac{br}{r_c \cdot fr}, \quad (10)$$

and $a_1 = 11.998$, $a_2 = -3$, $a_3 = 41.248$ and $a_4 = 0.1318$. The shape of this formula may appear surprising, with its double $\ln$ term. It assumes that there is a general relation between bitrate and MOS of the form (see e.g. [16]):

$$MOS = a + b \cdot \log\left(br + d\right). \quad (11)$$

By inverting this formula and considering that framerate and resolution also play a role for the resulting bitrate, Eq. (9) was found to lead to the best fit of the overall model predictions O.46 (note: long-term quality) with the ground-truth test results. This is achieved with the additional bpp-term that explicitly includes resolution and framerate. Also, this term makes the model more robust to resolution and framerate settings not included during model development. Details on how the bitrate br can be calculated from segment size and information about audio and potential Transport Stream or packetization header information can be found in P.1203.1.

For Mode 1, the same formula as in Eq. (9) is used for calculating $quant$, however with different values for the coefficients a_j. Further, the bitrate can now be calculated from the video frame sizes, providing more accurate estimates, with better temporal resolution (cf. P.1203.1, Annex B). For Mode 1, further changes over Mode 0 were standardized in P.1203.1. Here, to account for content-specific effects, the quantization degradation MOS_q is modified over Eq. (5) to

$$MOS_{q,Mode1} = MOS_q + \mathrm{sigmoid}\left(k_0, s_x, m_x, r_i\right), \quad (12)$$

with k_0, s_x and m_x as the coefficients of the sigmoid function

$$\mathrm{sigmoid}\left(k_0, s_x, m_x, r_i\right) = k_0 - \frac{k_0}{1 + e^{-s_x(r_i - m_x)}}, \quad (13)$$

and

$$r_i = \overline{I}/\overline{I_n}. \quad (14)$$

Here, $\overline{I}$ is the average size of the I-frames in the measurement window, and $\overline{I_n}$ the respective average size of the non-I-frames. Introducing the I-frame ratio r_i and the sigmoid-term for calculating $MOS_{q,Mode1}$ makes the model more sensitive to the spatio-temporal complexity of different contents, and leads to a significantly better performance than Mode 0.

For Mode 2, the initial calculation of MOS_q as in Eq. (5) is used again. Here, only $quant$ is adapted to this particular Mode, calculated as:

$$quant = \frac{\overline{QP_{PB}}}{51}. \quad (15)$$

Here, in essence, $\overline{QP_{PB}}$ is the Quantization Parameter (QP) value found in the bitstream for all non-I-frames, that is, P- and B- frames. In Mode 2, only 2% of each frame's payload information can be used, which means that for small frames there are cases where no full macroblocks can be parsed. At worst this may lead to measurement windows where no QP information can be extracted for the considered frames. Then, O.22 for the respective instance is calculated by falling back to Mode 1 or 0. If however only few frames in the measurement window cannot be parsed with the 2% limit, the QP info from parsable neighboring non-I-frames is used instead. It is noted that due to the ability of jumping from frame header to frame header during parsing, the 2% refer to the information in individual frames. For detailed information including pseudo-code for the parameter extraction cf. Annex C of P.1203.1.

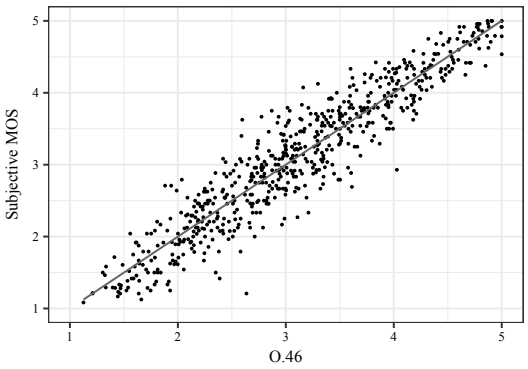

Fig. 3: Performance of overall P.1203 model for Mode 3 on PC/TV databases.

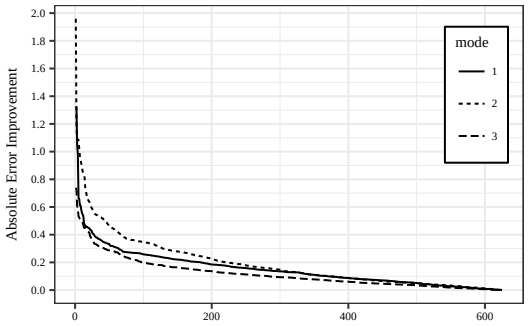

Fig. 4: Improvement in absolute error over PVSs for each Mode m, as compared to Mode $m-1$. Differences are shown in order of decreasing improvements for PC/TV databases. Hence, PVS-indices (x-axis) do not match between modes.

In Mode 3, the full bitstream is available to the model. Here, valid information on QP can be parsed for all frames in the measurement window. For Mode 3, MOS_q is calculated according to Eq. (5) and $quant$ according to Eq. (15), as for Mode 2. Since it is expected that in Mode 3 all QP information can be extracted, there is no need for a fallback to a lower Mode or for artificial copying of QP information from valid neighbouring frames. Since it has access to much more information, Mode 3 performs significantly better than Mode 2, and hence than Modes 1 and 0.

C. Performance

Fig. 3 depicts the prediction accuracy of the final, merged P.1203 P_v and P_q modules for Mode 3. Fig. 4 shows the

TABLE II: Performance results in MSE, RMSE and Pearson Correlation ("Cor.") for all modes and devices on all databases.

Mode	Device	MSE	RMSE	Cor.
0	mobile	0.201	0.449	0.869
	PC	0.252	0.502	0.852
1	mobile	0.163	0.404	0.895
	PC	0.203	0.450	0.883
2	mobile	0.144	0.380	0.908
	PC	0.157	0.396	0.911
3	mobile	0.127	0.356	0.920
	PC	0.107	0.327	0.940

improvement for different PVSs from mode to mode, that is, the Absolute Error of a given Mode m compared to the one for the respective lower Mode $m-1$. The RMSE and Pearson correlation between the model outputs O.46 and the subjective MOS for all PC and mobile-screen databases are shown in Table II.

From the table and plots, the significantly improved performance from mode to mode can be observed, where the model is able to capture variations in quality with increasing precision by probing the bitstream in more and more detail. As discussed earlier in this paper, only one single P_a and P_q module have been standardized in P.1203.2 and P.1203.3, respectively, unchanged between Modes. These standardized P_a and P_q modules were used for the performance analysis presented in this paper. Since they remained unchanged between modes, any overall performance improvement can entirely be ascribed to the P_v module, indicating the performance improvement contributed by the different P_v Modes.

In order to compensate for rating differences between databases, a per-database first-order linear regression adjustment was applied between O.46 and MOS, restricting O.46 to the range 1–5. Therefore, the average performance scores show a slightly better values than obtained for a general mapping. Nonetheless, the general performance is quite high even for the lowest Mode 0, considering that it is a parametric model. The remaining outliers seen in the scatter plot (Fig. 3, left) can be explained by the reverse-engineering approach applied for P_v model development, where no explicit short-term video quality scores were available from the official P.NATS training and validation databases.

VI. Conclusion and Outlook

The paper describes the video-quality model part (P_v) of P.NATS Phase 1, a bitstream-based model framework for predicting integral quality of HAS media sessions from 1 to 5 min duration (P.1203). Also since P.1203 and its P_v-model are bitstream-based, it will perform best on the processing and encoding settings used during model development. However, with its modular and scalable architecture, and the relatively low number of model coefficients used, the P_v module can be adjusted to other encoding or adaptation-set processing settings. A follow-up standardization work item has recently been started as a collaboration between ITU-T SG12 and VQEG (https://www.its.bldrdoc.gov/vqeg/vqeg-home.aspx), referred to as "AVHD-AS / P.NATS Phase 2". One goal is to develop signal- and bitstream-based, P_v-type short-term video quality models that provide per-one-second as well as per-10-second video quality scores. Alongside to bitstream-based models applicable to a wider range of encoding settings, full-, reduced-, no-reference, and hybrid models are addressed. The scope is extended beyond HD towards UHD-1, including encoding with H.264 as well as H.265 and VP9. The presented P_v module is planned to be used as a baseline, with re-trained coefficients.

Acknowledgment

The authors wish to thank the ITU-T Q.14/12 group for the constructive and fruitful collaboration when developing P.1203.

References

[1] M. Seufert, S. Egger, M. Slanina, T. Zinner, T. Hossfeld, and P. Tran-Gia, "A Survey on Quality of Experience of HTTP Adaptive Streaming," *IEEE Communication Surveys & Tutorials Communication Surveys & Tutorials*, vol. 17, no. 1, pp. 469–492, 2015.

[2] X. Yin, A. Jindal, V. Sekar, and B. Sinopoli, "A Control-Theoretic Approach for Dynamic Adaptive Video Streaming over HTTP," in *2015 ACM Conference on Special Interest Group on Data Communication (SIGCOMM)*, 2015.

[3] A. Balachandran, V. Sekar, A. Akella, S. Seshan, I. Stoica, and H. Zhang, "Developing a Predictive Model of Quality of Experience for Internet Video Categories and Subject Descriptors," *ACM SIGCOMM Computer Communication Review*, vol. 43, no. 4, pp. 339–350, 2013.

[4] M.-N. Garcia, F. De Simone, S. Tavakoli, N. Staelens, S. Egger, K. Brunnström, and A. Raake, "Quality of Experience and HTTP Adaptive Streaming: A Review of Subjective Studies," in *Proc. QoMEX*, 2014, pp. 141–146.

[5] H. Nam, K.-H. Kim, and H. Schulzrinne, "QoE Matters More Than QoS: Why People Stop Watching Cat Videos," in *IEEE International Conference on Computer Communications*, 2016.

[6] C. Alberti, D. Renzi, C. Timmerer, C. Mueller, S. Lederer, S. Battista, and M. Mattavelli, "Automated QoE evaluation of dynamic adaptive streaming over HTTP," in *Proc. QoMEX*, 2013.

[7] R. Schatz, T. Hoßfeld, and P. Casas, "Passive youtube QoE monitoring for ISPs," in *Innovative Mobile and Internet Services in Ubiquitous Computing (IMIS), 2012 Sixth International Conference on*. IEEE, 2012, pp. 358–364.

[8] M.-N. Garcia, W. Robitza, and A. Raake, "On The Accuracy of Short-Term Quality Models for Long-Term Quality Prediction," in *Proc. QoMEX*, Costa Navarino, 2015.

[9] M. Seufert, M. Slanina, S. Egger, and M. Kottkamp, "'To pool or not to pool': A comparison of temporal pooling methods for HTTP adaptive video streaming," in *Proc. QoMEX*, jul 2013, pp. 52–57.

[10] Y. Chen, K. Wu, and Q. Zhang, "From QoS to QoE: A Tutorial on Video Quality Assessment," *IEEE Communication Surveys & Tutorials*, vol. 17, no. 2, 2015.

[11] Y.-F. Ou, Y. Xue, and Y. Wang, "Q-star: a perceptual video quality model considering impact of spatial, temporal, and amplitude resolutions," *IEEE Transactions on Image Processing*, vol. 23, no. 6, pp. 2473–2486, 2014.

[12] J.-S. Lee, F. De Simone, T. Ebrahimi, N. Ramzan, and E. Izquierdo, "Quality assessment of multidimensional video scalability," *IEEE Communications Magazine*, vol. 50, no. 4, pp. 38–46, April 2012.

[13] M.-N. Garcia, A. Raake, and B. Feiten, "Parametric audio quality model for IPTV services – ITU-T P. 1201.2 audio," in *Proc. QoMEX*. IEEE, 2013, pp. 194–199.

[14] M. H. Pinson, M. Sullivan, and A. Catellier, "A New Method for Immersive Audiovisual Subjective Testing," in *VPQM*, Chandler, 2014.

[15] W. Robitza, M.-N. Garcia, and A. Raake, "At Home in the Lab: Assessing Audiovisual Quality of HTTP-based Adaptive Streaming with an Immersive Test Paradigm," in *Proc. QoMEX*, Costa Navarino, 2015.

[16] J. De Cock, Z. Li, M. Manohara, and A. Aaron, "Complexity-Based Consistent-Quality Encoding in the Cloud," in *IEEE International Conference on Image Processing (ICIP)*, 2016.

[17] Y. C. Lin, H. Denman, and A. Kokaram, "Multipass encoding for reducing pulsing artifacts in cloud based video transcoding," in *IEEE International Conference on Image Processing (ICIP)*, 2015, pp. 907–911.