IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received July 17, 2020, accepted September 14, 2020, date of publication October 21, 2020, date of current version November 3, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.3032080

Multi-Model Standard for Bitstream-, Pixel-Based and Hybrid Video Quality Assessment of UHD/4K: ITU-T P.1204

ALEXANDER RAAKE[1], (Member, IEEE), SILVIO BORER[2], (Member, IEEE), SHAHID M. SATTI[3], JÖRGEN GUSTAFSSON[4], (Member, IEEE), RAKESH RAO RAMACHANDRA RAO[1], STEFANO MEDAGLI[2], (Member, IEEE), PETER LIST[5], STEVE GÖRING[1], DAVID LINDERO[4], (Member, IEEE), WERNER ROBITZA[1], (Member, IEEE), GUNNAR HEIKKILÄ[4], (Member, IEEE), SIMON BROOM[6], CHRISTIAN SCHMIDMER[3], BERNHARD FEITEN[5], ULF WÜSTENHAGEN[5], THOMAS WITTMANN[3], MATTHIAS OBERMANN[3], AND ROLAND BITTO[3]

[1]Audiovisual Technology Group, Ilmenau University of Technology, 98693 Ilmenau, Germany
[2]Rohde & Schwarz SwissQual AG, 4528 Zuchwil, Switzerland
[3]Opticom GmbH, 91052 Erlangen, Germany
[4]Ericsson Research, L.M. Ericsson, 164 83 Stockholm, Sweden
[5]Deutsche Telekom AG, 10117 Berlin, Germany
[6]NETSCOUT, Ipswich IP1 1HN, U.K.

Corresponding authors: Alexander Raake (alexander.raake@tu-ilmenau.de), Silvio Borer (silvio.borer@rohde-schwarz.com), and Shahid M. Satti (ss@optiom.de)

The Article Processing Charge for this Open Access work was supported by the Open Access Publication Fund of Ilmenau University of Technology. The work presented in the paper was conducted as part of a video quality modeling competition conducted as a joint standardization project of Study Group 12 of the International Telecommunication Union (ITU-T) and the Video Quality Experts Group (VQEG). The work resulted in three video quality models, each specified in individual standard documents (''Recommendations'') and developed by a different team. Here, the standard ITU-T Rec. P.1204.3 was developed by the team R. Rao, P. List, S. Göring, W. Robitza, B. Feiten, U. Wüstenhagen and A. Raake. The standard ITU-T Rec. P.1204.4 was developed by S. Borer and S. Medagli. The standard ITU-T Rec. P.1204.5 was developed by S. M. Satti, C. Schmidmer, T. Wittmann, M. Obermann and R. Bitto.

ABSTRACT The paper presents a series of three new video quality model standards for the assessment of sequences of up to UHD/4K resolution. They were developed in a competition within the International Telecommunication Union (ITU-T), Study Group 12, in collaboration with the Video Quality Experts Group (VQEG), over a period of more than two years. A large video quality test set with a total of 26 individual databases was created, with 13 used for training and 13 for validation and selection of the winning models. For each database, video quality laboratory tests were run with at least 24 subjects each. The 5-point Absolute Category Rating (ACR) scale was used for rating, calculating Mean Opinion Scores (MOS) as ground-truth. To represent today's commonly applied HTTP-based adaptive streaming context, the test sequences comprise a variety of encoding settings, bitrates, resolutions and framerates for the three codecs H.264/AVC, H.265/HEVC and VP9, applied to a wide range of source sequences of around 8 s duration. Processing was carried out with an FFmpeg-based processing chain developed specifically for the competition, and via upload and encoding through exemplary online streaming services. The resulting data represents the largest, lab-test-based dataset used for video quality model development to date, with a total of around 5,000 test sequences. The paper addresses the three models ultimately standardized in the P.1204 Recommendation series, resulting in different model types and for different applications: (i) Rec. P.1204.3, no-reference bitstream-based, with access to encoded bitstream information; (ii) P.1204.4, pixel-based, using information from the reference and the processed video, and (iii) P.1204.5, no-reference hybrid, using both bitstream- and pixel-information without knowledge of the reference. The paper outlines the development process and provides holistic details about the statistical evaluation, test databases, model algorithms and validation results, as well as a performance comparison with state-of-the-art models.

INDEX TERMS Bitstream, full reference, http adaptive streaming (HAS), hybrid, pixel-based, QoE, reduced reference, video quality.

I. INTRODUCTION

The video quality achieved with a given encoding setting is of relevance for a variety of applications, such as video on demand, live streaming or audiovisual communication. For example, in services applying HTTP-based adaptive streaming (HAS), such as Video on Demand (VoD) or live streaming, the different video representations are typically realized by encoding each sequence at different resolutions and bitrates, reflecting a balance between target screen resolution and expected channel bandwidth (also referred to as ''bitrate ladders''), see e.g. [1]–[3]. Depending on its usage, a video bitrate ladder reflects aspects such as the optimal resolution and encoder setting for a given target bitrate, or the bitrate that is needed for a given resolution to reach a certain quality level.

For video-media services and applications, video quality represents an important component of the users' experience at large, the latter typically referred to as Quality of Experience (QoE). According to [4]–[6], QoE is ''the degree of delight or annoyance of the user of an application or service''. During a typical, HAS-based video streaming session, video quality may vary due to a time-varying network bandwidth characterized by quality switches, initial loading delay during the filling of the playout buffer when starting streaming, or stalling of the video playout when the buffer has run empty due to network problems. Considerations on a more holistic view of QoE for HAS-type or other streaming that includes long-term integration or effects such as initial loading and stalling may be found in [3], [7]–[19]. The present paper focusses on video quality, as a key element for video streaming QoE. The described models were designed for short-term video quality assessment of videos of around 10 sec duration. The primary focus of the models is the type of video used for HAS (e.g. MPEG-DASH or HTTP Live Streaming). For example, the models can be applied to analyze the quality of individual segments of HAS-type representations. Accordingly, reliable transport is assumed, using e.g. TCP or QUIC. It is noted that the models presented in this paper can principally be used also for assessing video quality for streams with unreliable transport, e.g. via plain UDP with RTP. Here, with the models described in this paper, the impact due to resolution re-scaling, framerate and encoding can be covered. Degradations due to packet loss resulting in slicing, freezing or some catching-up accelerations of the stream are not addressed by the models.

Due to its perceptual character, evaluating video quality ultimately requires feedback from users. Corresponding data have been collected during formal laboratory or crowd-sourcing tests [11], [12], [16], [20], [21], or were measured in terms of the viewing behaviour of users of a given service, e.g. in terms of whether users were stopping playback or take other actions in case of problems [13], [22]–[25]. When aiming for a sensitive assessment of encoding quality for high resolutions such as 4K UHD (3840 × 2160 pixels), laboratory tests with a controlled and 4K-appropriate viewing distance of 1.5 to 1.6 times the height of the screen (''1.5H'' or ''1.6H'') are recommended, see [26], [27]. As was shown in a number of studies, even in laboratory tests with high-quality screens and controlled viewing conditions that follow recommendations such as those in [27], [28], in many cases video quality can hardly be distinguished between HD and 4K UHD resolution, specifically depending on the initial quality of the source content used [29]–[33]. On the other hand, test contents in video quality tests [1] often are rather artificial and not representative of actual target applications such as VoD or live streaming. The role of content and its quality and representative character is discussed for example in [12], [29], [32], [34]. As a consequence, well-designed and well-conducted subjective tests are required, with a representative choice of contents for a valid determination of the video quality as experienced by end users.

Running such well-conceived subjective tests requires substantial human and material resources. Hence, for a systematic and automatic video quality assessment that is representative of human video quality ratings, instrumental prediction, that is, ''objective'' models are needed. Here, the suitability of a given model not only depends on the required prediction accuracy, but also on the targeted application and thus model input information and processing resources available. With a well designed and validated video quality model, a variety of applications may benefit, such as the aforementioned encoding-related bitrate ladder derivation, or a holistic streaming-service or network monitoring, as discussed further in Section VII.

Four basic categories of video quality models can be distinguished (see also [10], [35], [36]):

1) Metadata-based
2) Bitstream-based
3) Pixel-based
4) Hybrid

Metadata-based quality models (1) use information from the metadata layer such as the video codec used, image resolution, framerate and bitrate, which may be available from player logs or during the planning of a service. Metadata-based models can also be seen as lightweight variants of bitstream models that analyze only the metadata portion of the bitstream. An example is ITU-T Rec. P.1203.1,[2] ''Mode 0'' [37], [38]. Bitstream-based video quality models analyze the encoded video bitstream without decoding and do not require access to the original bitstream of the source signal.

The associate editor coordinating the review of this manuscript and approving it for publication was Gangyi Jiang.

[1] In the state-of-the-art literature, quality tests with human subjects are typically referred to as ''subjective tests'', and instrumental quality-prediction models as ''objective models''. This terminology is also adopted in this paper, in spite of some limitations.

[2] It is noted that the models described in ITU-T Rec. P.1203.1 [37] only provide a per-one-second video quality estimation on the 5-point ACR scale (MOS). In the absence of other degradations such as quality-switches due to changes in the representation, stalling or initial loading delay, a video quality estimate for short sequences of around 10 sec duration can principally be obtained e.g. by simple averaging of the per-one-second scores over time.

Examples are ITU-T Rec. P.1203.1 (Modes 1 and 3) [37], [38] for HAS-type streaming over TCP or QUIC, or P.1201.2 for IPTV over UDP that may show effects of packet loss [39]–[41]. Also ITU-T Rec. P.1204.3, which is addressed in the present paper as one of the three models, is an example of a bitstream-based model, with open-source software available from [42]. Further examples of bitstream-based models can be found in [10] and in Section II.

Pixel-based video quality models analyze the decoded frames of the video. Different variants can be distinguished:

- Full Reference (FR) models, which derive quality estimates from a comparison of the original content with the decoded, processed sequence under test. Examples range from Peak Signal-to-Noise Ratio (PSNR) [36] to Structural Similarity (SSIM) Index [43] and models such as Video Multi-Method Assessment Fusion (VMAF) [44] and several ITU recommendations, such as J.144 [45], J.247 [46], J.341 [47] – see Section II.
- Reduced Reference (RR) models, where "reduced" representations of the reference and the sequence to be evaluated are used. The new standard ITU-T Rec. P.1204.4 presented in this paper is a reduced-reference model. As was shown during the development of this standard, quality-prediction performance of this model is as good as with an FR-version of the same algorithm. Hence, in the remainder of this paper, P.1204.4 is referred to as "RR/FR". Further examples are mentioned in Section II.
- No Reference (NR) models, where the evaluation is performed without access to the reference content. Currently, no purely pixel-based NR model is known to provide sufficiently good prediction accuracy that could enable its usage in practical applications. In principle, both the bitstream-based and the hybrid video quality models presented in this paper are of the NR-type. More information on NR models is provided in Section II and [10].

Hybrid models are based on an evaluation of pixel information and additional bitstream or metadata information, as with the new standard ITU-T Rec. P.1204.5 presented in this paper. Further state-of-the-art hybrid models are outlined to in Section II.

The paper presents the results of a so far unique campaign to video quality model development: For the first time, bitstream-, pixel-based and hybrid models were developed, trained and validated on a large common subjective test dataset consisting of a total of 26 individual video quality tests, each with at least 24 subjects. The work on developing the video quality models was conducted in collaboration between Study Group 12 (SG12) of the International Telecommunication Union (ITU-T) and the Video Quality Experts Group (VQEG), referred to as the "P.NATS Phase 2" project. It followed up on the previous standardization project "P.NATS Phase 1" run in ITU-T SG12, leading to the standards series ITU-T Rec. P.1203, P.1203.1, P.1203.2 and P.1203.3 [37], [48]–[50].

The bitstream-based P.1203 is primarily targeted towards prediction of the integral quality of longer video streaming sessions between 1 min and 5 min duration, more in line with the idea of an overall session QoE rather than sheer video quality. The P.NATS Phase 1 model series comprises a short-term video quality component as well, P.1203.1 (see [37], [38]). However, so as to develop short-term video quality models with a degree of accuracy that would allow applications such as deriving fine-grained video-quality-based encoding ladders, ITU-T SG12 and VQEG launched the P.NATS Phase 2 project. While Phase 1 addressed bitstream-based models only, for Phase 2, a wider scope was envisaged, focusing on all relevant video quality model types that can enable high prediction accuracy: Bitstream-based, pixel-based (FR, RR) and hybrid.

The P.NATS Phase 2 standardization work has recently resulted in the new standard series ITU-T Rec. P.1204 [51], consisting of the bitstream-based NR model according to P.1204.3 [52], the pixel-based, RR/FR model ITU-T Rec. P.1204.4 [53] and the hybrid NR model ITU-T Rec. P.1204.5 [54].

The new P.1204 models presented in this paper target video resolutions up to 4K/UHD. They were trained and validated for three different video codecs, H.264, HEVC/H.265 and VP9, covering video framerates between 15 up to 60 fps, with different model variants for video presentation on PC or TV type screens, tablets and mobile phones. More details on the development procedure are presented in Sections III and IV.

In light of the target 4K/UHD resolution, the ground-truth data for model development and validation had to be based on a rigorous subjective laboratory testing approach. A dataset of 26 subjective video quality test databases were created for the competition, with a total of around 5,000 test sequences, each rated by at least 24 test subjects. Here, special emphasis was laid on selecting appropriate source sequences, coverage of a wide range of encoding settings, well-controlled presentation and rating conditions used in the cross-lab testing campaigns, based on dedicated approaches for data alignment such as common set sequences and common test conditions, a subsequent diligent checking of the individual test datasets with regard to subject bias and inter-rater agreement, and corresponding outlier detection and removal.

The paper, for the first time, summarizes the model development and standardization process in a scientific publication. The result of the according competition is a set of models applicable in a variety of contexts, enabling the choice from three highly accurate models, for example depending on the type of model input information that can be made available in a given application context. Hence, besides the large underlying subjective test dataset, the new P.1204 standard represents a unique combination of all relevant models, applicable to a wide range of encoding settings and formats. The analysis of the model prediction given in Section VI indicates the outstanding performance of all the three models, also in comparison to other metrics and models such as PSNR, SSIM and VMAF.

The key contributions of the paper can be summarized as follows:

1) Details on the ITU-T SG12 / VQEG "P.NATS Phase 2" standardization project are provided, including the statistical model evaluation criteria and procedure to determine the winning model candidates.
2) Description of the procedure to automatically generate a set of processed sequences to be rated in the P.NATS Phase 2 subjective tests. The procedure includes the creation of a dedicated processing chain to realize a variety of encodings and thus representations of video contents. The subsequent subjective tests resulted in a novel large proprietary subjective test database with a total of around 5,000 test sequences each rated by at least 24 test subjects that was established to train and validate the different model candidates targeted with the P.NATS Phase 2 work. The creation process and characteristics of the database are presented in detail in this paper for the first time. The novelty of the resulting database lies in the coverage of the three different codecs H.264, HEVC/H.265 and VP9, the inclusion of different resolutions, bitrates, framerates, and encoder settings, and the fact that all information is contained to enable that bitstream- and pixel-based models could be trained and validated on the same databases.
3) The resulting three different types of high-precision video quality models of the P.1204-series are presented in a scientific and harmonized form for the first time, outlining key algorithmic concepts.
4) A detailed model performance analysis is presented for the initially submitted model candidates as well as for the finally standardized models, using the P.NATS Phase 2 databases. Further, the performance of all models is compared to other models of similar kind, using the P.NATS Phase 2 database as well as additional open-source databases that enable a performance analysis across all three model types.

The paper is organized as follows: Section II provides an overview of the state-of-the-art, considering all the three model types addressed in this paper. The P.NATS Phase 2 competition run in collaboration between ITU-T SG12 and the VQEG Section III, including considerations such as the statistical model evaluation. In the subsequent Section IV, the training and validation databases are described, with details about source contents, processing chain and database characteristics. Then, Section V presents algorithmic descriptions of the bitstream-based (P.1204.3), pixel-based RR/FR (P.1204.4) and hybrid NR (P.1204.5) models, using a unified nomenclature for an aligned presentation. An in-depth model performance analysis is provided in Section VI, with performance indicators given for the initially submitted model candidates evaluated as it was done during the competition, performance data for the finally standardized models and a comparison with other metrics and models such as PSNR, SSIM, and VMAF, also including publicly available test databases.

II. RELATED WORK

A variety of bitstream-, pixel-based and hybrid video quality models have been reported in the literature over the past years. In this section, we focus on analysing some of these and present the need for novel approaches. For more comprehensive reviews and surveys on state-of-the-art video quality models, the authors refer to the various works provided, for example, by [36], [55]–[59]. A recent review of the HAS QoE modelling literature has been provided by Barman *et al.* in [10]. It primarily focusses on a more holistic modelling of HAS QoE, including audio and video quality as well as initial loading delay and stalling, as it can be done for example using the standard family ITU-T P.1203 [37], [48]–[50], see Sec. II-A. In turn, the present paper proposes new high-performance, short-term video quality models, solving some of the challenges mentioned in [10]. Correspondingly, this section primarily focuses on video quality models for video durations around 10 s.

A. BITSTREAM MODELS

Several bitstream-based no-reference models have been proposed in the literature for different use cases. The proposed models range from very simple curve-fitting-based bitstream models to more complex machine-learning-based ones. An earlier review of bitstream models for video quality prediction is presented by Joskowicz *et al.* in [60]. They conclude that the bitstream models show good results when compared with subjective quality ratings.

A more complex Mode 3 bitstream model for H.264/AVC-encoded videos using motion values, QP-values, frame types etc. is proposed by Keimel *et al.* [61]. The study shows good performance of this type of model, also in comparison to a number of full-reference models. For the case of IPTV (RTP/UDP or MPEG2-TS/RTP/UDP) with coding and packet-loss degradation, Raake *et al.* [62] and Garcia *et al.* in [40] propose two evolutions of packet-header-based bitstream-based models, for SD and HDTV resolution with H.264-type video encoding. The resulting video quality estimation can be integrated with audio quality [63] into an audiovisual quality estimation [64]. The complete audiovisual quality model for IPTV is standardized as ITU-T Rec. P.1201.2 [39], [41]. A complementary approach developed for RTP/UDP-based transmission and lower video resolutions, corresponding to typical mobile phone screens around 2010, was developed by Yamagishi *et al.* in [65] and has been standardized as ITU-T Rec. P.1201.1 [66]. Note that on the way towards the HAS-related standard ITU-T Rec. P.1203, the higher-resolution model in ITU-T Rec. P.1201.2 was extended to streaming with reliable transport, addressing HAS' predecessor "progressive download", based on the work presented by Hossfeld *et al.* and Garcia *et al.* in [67], [68] and [69], respectively.

In addition to the curve-fitting-based bitstream models, several machine-learning-based approaches have been reported in the literature. An approach based on Support

Vector Regression (SVR) was presented in [70], applicable to streaming over unreliable transport including packet loss. A model based on genetic programming-based symbolic regression was proposed by Staelens *et al.* in [71]. Mocanu et al. in [72] proposed a random neural network (RNN) no-reference bitstream model. Demirbilek *et al.* introduced a set of decision trees, deep learning and genetic programming based models [73]. These models were developed for H.263 or H.264 encoded videos and for non-reliable transmission, that is, including cases of packet loss and resulting in artifacts such as slicing. Since these effects are not present in HAS as addressed in this paper, the models are not directly applicable here. While encoding-type degradations are naturally included in these models as well, the underlying subjective tests used for model development are naturally biased towards packet-loss-type degradations. Moreover, different resolutions and framerates are typically not considered, further limiting the usage for today's streaming service quality assessment.

ITU-T Recommendation P.1203 [38], [48], [74] describes the first standardized QoE model for audiovisual HTTP-based adaptive streaming. The recommendation is divided into three modules, one each for audio quality [49], video quality [37] and quality integration [50]. The quality integration module [50] takes into account the per-one-second audio- and video quality output provided by the corresponding audio- and video quality modules, also considering corresponding quality switches, as well as information about the initial loading delay and stalling events. This standard explicitly handles the case of HAS, but is applicable only for H.264 encoded videos of up to 1080p resolution and framerate up to 30fps. An open-source implementation of the complete P.1203 model set is described in [74]. As mentioned in Section I, in the absence of quality-level switches, initial loading delay or stalling, the per-one-second video quality scores provided by the different bitstream-models described in ITU-T Rec. P.1203.1 [37], [38] can be integrated by simple averaging over time to video quality estimates for the short sequence durations addressed in this paper of around 8 to 10 sec. To take into account higher resolutions and framerates and also newer codecs, Ramachandra Rao *et al.* [75] proposed an extension to the the standardized P.1203 Mode 0 model. However, this extension is only based on two subjective tests with limited range of encoding settings.

Besides the standardized P.1203 series of models, several models have been proposed to predict video quality for the HAS-specific scenario [76]–[80].

In essence, although the presented models together are applicable in a wide range of scenarios, they suffer from the following drawbacks: (a) they were not developed to handle the case of higher resolutions (up to 4K/UHD-1), higher framerates (up to 60fps) and newer codecs such as MPEG-H HEVC/H.265 and VP9; (b) if applicable to higher resolutions and framerates and newer codecs, they are developed using a very limited number of quality test databases. To overcome these drawbacks, the bitstream model presented in this paper was developed, which is now standardized as ITU-T Rec. P.1204.3 [52]. Further details are provided in Sec. V-A.

B. PIXEL-BASED MODELS

Unlike bitstream models, pixel-based models use raw pixel data as model input to estimate video quality. Since these models do not require any knowledge of how the video was encoded, these types of models are agnostic to the underlying encoding or transmission technologies. As outlined in Sec. I, depending on the availability of the original undistorted, reference video, Full Reference (FR), Reduced Reference (RR) and No Reference (NR) models can be distinguished. FR models require complete access to the reference video. These models compute quality indicators using frame-by-frame comparison of the reference and degraded video. Examples of such metrics are PSNR (Peak Signal-to-Noise Ratio), SSIM (Structural Similarity) [43], [81], Netflix' VMAF (Video Multimethod Assessment Fusion) [44] and several ITU recommendations, such as J.144 [45], J.247 [46], J.341 [47].

Reduced Reference (RR) models extract a fixed, reduced set of features from the reference and from the processed video sequence, and compare these to estimate quality. Due to the limited access to the reference video, RR models were in the past less accurate than the FR models. Examples of such models include the ITU-T Recommendations J.246 [82], J.249 [83], J.342 [84]. Other examples are ST-RRED [85] and SpEED-VQA [86]. In their default versions, these include a higher amount of features extracted from the reference. Further, less complex variants were described that use one feature value per reference frame only and also show a lower prediction performance.

No Reference (NR) Models have no access to the reference video and use only the degraded pixel information to predict video quality. Examples of NR models include DIIVINE, BRISQUE, BLIINDS and NIQE [87]–[90]. In the absence of source information, such models are usually less accurate than the corresponding FR and RR counterparts [91], [92]. As a consequence, purely pixel-based NR models are not considered further in this paper, which targets higher-accuracy video quality models.

C. HYBRID MODELS

Hybrid models use video pixel information in combination with bitstream information for predicting video quality. Like pixel-based models, hybrid models can be classified into three main categories, depending on the availability and use of reference-video pixel information, namely, hybrid-FR, hybrid-RR and hybrid-NR models.

The use of bitstream information helps such models to improve prediction accuracy considerably compared to the traditional NR models. One example of a hybrid-NR model is the model presented by Yamagishi *et al.* in [93] which was developed for estimating video quality in the IPTV scenario. This model uses features derived from the received packet headers and pixel information such as spatial and temporal information to estimate video quality. This model was developed for H.264 encoded videos with resolution of 1440 × 1080 and framerate of 30 fps.

Further, Osamu *et al.* [94] also propose a hybrid no-reference model applicable to H.264 encoded video. This model uses the quantization parameter (QP) as the bitstream feature. A spatial and temporal image feature each were developed to be used as an input to the proposed model. The spatial image feature estimates the block distortion that is usually encountered in block-based encoding schemes. The temporal image feature used in the model is used to quantify the extent of the flickering artifact and hence a "flickering measure" was developed. These features were then integrated to estimate the video quality.

A hybrid no-reference model that takes into account packet loss rate information has been proposed by Farias *et al.* [95]. The model uses features to estimate blockiness and blurriness as the pixel features that is then integrated with the packet loss rate information for video quality estimation. Like with the models proposed in [93] and [94], this model is applicable to H.264 encoded videos. This necessitates the development of models that are applicable for videos encoded with modern video codecs such as H.265, VP9 and capable of handling higher resolutions like UHD-1 and framerates like 60fps.

In addition to these models, the J.343-series of ITU Recommendations contains hybrid models of all types, developed to measure the perceptual video quality for HDTV and multimedia applications. These models are applicable for H.264 encoded videos, so similarly to ITU-T Recs P.1201 and P.1202 address unreliable transport resulting in possible packet-loss artifacts. These models cannot be used for resolutions higher than fullHD, or framerates above 30 fps. The standardized P.1204.5 [54] model is a hybrid no-reference model developed specifically for the case of reliable transport, thus not taking into account degradations like packet loss. In addition this model is applicable to resolutions up to UHD-1 and framerates upto 60fps.

III. OVERVIEW OF THE COMPETITION

The video quality model development campaign was conducted as a joint-venture between the ITU-T Study Group 12 (SG12), Question Q14/12 and the Audiovisual HD (AVHD) project of VQEG,[3] under the name "AVHD-AS / P.NATS Phase 2", or simply "P.NATS Phase 2". Its predecessor, "P.NATS Phase 1", was finalized in late 2016 with the consent of the standards series ITU-T Rec. P.1203, P.1203.1, P.1203.2 and P.1203.3 [37], [48]–[50]. The P.1203-series addresses metadata- and bitstream-based models to predict integral quality scores for longer video streaming sessions between 1 min and 5 min duration. With the inclusion of audio and video quality as well as initial loading delay and stalling, the P.1203 predictions represent holistic QoE measurements.

The P.1203 model has a modular architecture, using a short-term (per-1-second) estimation of video (P.1203.1 [37], [38]), and audio quality (P.1203.2 [49]) and their integration with additional information on initial loading delay and stalling into an estimate of streaming session QoE (P.1203.3 [50]). More details about the P.1203 model series and an open-source implementation can be found in [38], [74], and an independent evaluation in [96].

The video quality module P.1203.1 [37], [38] was developed by primarily reverse-engineering the retrospective integral quality ratings obtained from the test subjects after watching 1 min up to 5 min long audiovisual streaming sequences that partly included quality switches, initial loading delay and stalling events. As a consequence, it was clear to the involved parties that the video quality module P.1203.1 itself was of sub-optimal prediction accuracy so as to enable more precise quality estimations suitable for applications such as a highly accurate bitrate ladder derivation or quality monitoring, or possibly a monitoring-based player optimization.

The P.NATS Phase 2 project was run as a competition between nine participating institutions ("proponents") developing candidate models. A set of different competition "disciplines" is represented by the different types of models that could be submitted to the competition: (i) Bitstream-based, (ii) pixel-based, namely RR, and FR, and (iii) hybrid, metadata- and pixel-based, NR.

During model development, the nine proponents jointly created a set of dedicated training databases. Before submission, the proponents could train their model candidates on the training dataset, consisting of 13 individual video quality test databases. After model submission, a second, new validation dataset of further 13 subjective test databases was established by the proponents. Each proponent contributed a pre-defined, roughly equal number of training and validation databases to the competition, following a common test protocol. In total, about 5,000 test sequences were rated by at least 24 test subjects each (with one exception, see Section IV). The P.NATS Phase 2 development and standardization process is outlined in more depth in the following.

1) LIST OF ACRONYMS

The following acronyms are used in the remainder of this paper to specify different components of the model training and validation databases.

- SRC (Sources): This refers to the original undistorted source material, also referred to as the reference video, that is subjected to different encodings.
- HRC (Hypothetical Reference Circuit): The various encoding conditions that the SRC is treated with is referred to as the HRCs.
- PVS (Processed Video Sequence): This refers to the encoded video that is shown to the subjects for rating the video quality.
- P2STR: All databases related to the training stage of the competition are identifiable with this tag.
- P2SVL: This tag is used to indicate the validation databases.

[3] www.vqeg.org

A. GENERAL PROCEDURE

The PNATS2/AVHD project was conducted in 5 project parts, namely:

1) Training database creation
2) Model training and submission
3) Validation database creation
4) Model verification/validation
5) Model merging/optimization

In the training-database creation part (1), a total of 13 training databases (5 with display on a PC-monitor, plus 4 on TV and 4 on mobile) were created by the nine proponents. The training part involved identifying source material, defining the encoding conditions (also known as hypothetical reference circuits, HRCs) and subjective test conduction. The source material for both training and validation databases were obtained from free-sources, i.e. sources with Open CC license and further ones available to individual proponents.

These databases were used for training the models during the model training part (2). A period of around 4 months was allocated for training all the proponents' models. In total, 35 model candidates have been submitted into the different competition categories. This paper focusses on the three finally standardized models. Following the P.NATS Phase 2 approach of only standardizing models that provide an actual added value in terms of prediction performance and/or model complexity, the RR model was ultimately standardized as ITU-T Rec. P.1204.4. The FR model developed by the same institution as the RR model, and following a similar philosophy in algorithm design, did not show a significantly better performance than the RR variant, so that only the latter was standardized.

Each proponent submitting models did so by uploading a virtual machine to a dedicated ITU Telecommunication Standardization Bureau (TSB) server, containing all their submitted models in a runnable format.

After model submission, preparation of video sources and creation of validation databases was carried out (3). This separation between the training and validation-database creation was chosen so as to make sure that the validation data was completely unknown to the models at the time of model development and submission. During this validation part of the competition, a total of 13 databases (number of tests per display type: 1 PC-monitor, 8 TV, 3 mobile, 1 tablet) were created by the contributing proponents. The resulting subjective scores were submitted to the ITU TSB, while the data needed to run the models and obtain predictions was shared among all proponents.

The subjective scores were disclosed to the individual proponents during the following model verification/validation part of the competition (4), upon request to ITU. Before sharing the subjective scores for the validation databases, a bug fixing of submitted models could be requested by proponents from the rest of the group. Such bugs were typically identified after proponents had run their models on the validation-database model input information (i.e. bitstream and / or pixel info). Following a well-defined bug-fixing procedure, issues such as parsing errors or obvious mistakes which could not alter the performance of the models were agreed upon as allowable fixes by all proponents. After a bug fix (if any), each proponent was asked to derive the predicted scores using their submitted models on the validation databases, without the knowledge of the subjective scores for these databases. The produced scores were uploaded to the ITU TSB server into dedicated folders only accessible to the given proponent.

At a verification/validation meeting held in Stockholm in late 2019, a verification of the submitted scores was carried out to make sure that these were indeed produced by the submitted models. This way, it was sought to prevent that the model scores on the validation databases were obtained with a model that was modified over the initially submited version. In particular, proponents were asked to reproduce scores under the supervision of one other proponent. The newly produced scores had to match the earlier submitted scores, to confirm the verification of the models.

Once all models were verified, the subjective test scores were disclosed to all proponents by ITU TSB. The predicted scores were then compared against the subjective test scores to compute the model performance for each of the submitted models. Based on the criteria described in Sec. III-B, "winning groups" were determined for each model category.

According to the rules set out for the competition, in the model merging/optimization phase (5), all winning models of a certain category were to be merged and optimized to create the finally standardized model for that category. For all three model types presented in this paper, only one model candiate each ended up in the corresponding winning group. As a consequence, no model merging was required.

The model coefficients were optimized based on the cross-validation strategy elaborated in Sec. VI. In total, 5 such splits were created, and for each split the models were re-optimized. Following the validation criteria laid out in Sec. III-B, a training weight of 0.1 and validation weight of 0.9 was used to compute the average RMSE for a given cross-validation optimization run. The coefficients for the model version that led to the least average RMSE were finally reported in the corresponding ITU-T P.1204 model standards.

B. STATISTICAL EVALUATION

This section details the procedure followed to determine the winning models/groups across the different model categories. The final statistical evaluation procedure consisted of

- Data cleaning and mapping
- Calculating performance in terms of model prediction error per database for each submitted model
- Definition of the minimum model acceptance requirements
- Model performance comparison
- Selection of winning models/groups

The following subsections describe in details each step of the statistical evaluation process.

1) DATA CLEANING AND MAPPING

Prior to computing model performance, an inspection of the subjective test data was performed to identify problematic model input cases. Examples of such cases are errors in the applied processing chain or settings, or the use of unsuitable source sequences. Also, issues found with the subjective test procedure were used to remove databases. In case that it could be assumed that a whole database was affected by non-allowable conditions, the respective database was to be removed from validation.

A common set of PVSs was specified to help with analysing database validity. Three SRCs of varying complexity were matched with the HRCs described in Table 3. This set could then be used to investigate, how the rank order and absolute scores differed between labs and tests. The analysis of these PVSs enabled to confirm that every test had a similar quality range with both high and low quality scores.

During the analysis, one of the training databases for PC playout, P2STR07, was found to not comply with the subjective test procedure agreed upon at the beginning of the competition (see Sec. IV), and was hence removed. This database consisted of a total of 183 PVSs.

For all other databases, any bias between the subjective tests was removed by applying a linear mapping (per database) to the objective scores before computing any of the performance evaluation metrics [97]. The mapping coefficients were optimized by maximizing model performance, as discussed in the following subsection.

2) PERFORMANCE MEASURE

The models were evaluated and optimized based on one single statistical metric, i.e., the root mean square error (RMSE), aggregated across all databases [97]. The calculation of the *RMSE* for a model v and database k can be expressed as

$$RMSE_{k,v} = \sqrt{\frac{1}{N_k - 2}\sum_{i=1}^{N_k}(s_i - \hat{s}_{v,i})^2}, \qquad (1)$$

where s_i is the subjective score for the ith sample in the considered test, the score $\hat{s}_{v,i}$ denotes the objective score of the model v for the ith sample, and N_k the number of samples in the test k. The use of the subtraction by 2 in the denominator reflects the linear mapping to the subjective scale described in Sec. III-B1.

For the model performance comparison, both training and validation databases were used but weighted with different coefficients: $w_{training}$ and $w_{validation}$ for training and validation databases, respectively:

$$w_{training} = 0.1 \text{ and } w_{validation} = 0.9. \qquad (2)$$

The evaluation of the models was based on their performance across all subjective experiments, included in the training (known) and validation (unknown) datasets. Therefore, for each model v the aggregated error across all the databases was computed as a weighted sum of the mean squared error per database,

$$p_v = \frac{1}{W}\sum_{k=1}^{M} w_k \cdot RMSE_{k,v}^2, \qquad (3)$$

where M represents the total number of (training and validation) databases, w_k the weight of each database given in (2), and $RMSE_{k,v}$ the root mean square error of model v for database k. The normalization constant W is given by $W = \sum_{k=1}^{M} w_k$. A large value for p_v represents poor performances, therefore, the best model is the one achieving lowest p_v value.

3) MINIMUM REQUIREMENT

As a minimum requirement for model performance, a simple baseline model was defined as a parametrized linear mapping of log(bitrate) to subjective MOS,

$$Q_{baseline} = a \cdot \log(bitrate + b) + c, \qquad (4)$$

where the coefficients a, b and c depend on the codec and on the target device. Thus, the six sets of coefficients (a, b, c), for the three codecs times the two target devices, were optimized on the corresponding samples of the training data. These coefficients were then fixed and used to determine the performance $p_{baseline}$ of the baseline model according to (3). Model candidates with an aggregated error $p_v \geq p_{baseline}$ did not satisfy minimum requirements and were removed from any further evaluation.

4) MODEL PERFORMANCE COMPARISON

All the models which pass the minimum requirement criteria qualify for this step. Model performances are not compared on absolute-RMSE basis, rather any difference in model performance was tested for statistical significance. The statistical significance test was applied to the aggregated error p_v. The aggregated error p_v is approximately χ^2-distributed according to the Welch-Satterthwaite approximation [98], with the degrees of freedom θ calculated by

$$\theta \approx \frac{(\sum_{k=1}^{M} w_k)^2}{\sum_{k=1}^{M} \frac{(w_k)^2}{\theta_k}}, \qquad (5)$$

where w_k represents the weight of the database k given in equation (2) and θ_k denotes the degrees of freedom of $RMSE_{k,v}^2$ and is given by $\theta_k = N_k - 2$, with N_k the number of samples in the database k. For the aggregated error p_v of model v, the statistical significance test takes the form

$$t_v = \max\left(0, \frac{p_v}{p_{v_{min}}} - F(0.95, \theta, \theta)\right) \qquad (6)$$

Here, v_{min} denotes the model with lowest error $p_{v_{min}}$ in the evaluation, $F(0.95, \theta, \theta)$ denotes the 0.95-quantile of the F-distribution with θ degrees of freedom [99]. If $t_v = 0$, the model v is considered to be statistically equivalent to the model v_{min}. In case that $t_v > 0$, the difference in performance between the model v_{min} and model v is called "statistically significant", or "significant" for short.

5) MODEL SELECTION PROCEDURE

The three proposed models are the result of the model selection procedure described in this section. For most model categories, multiple models were submitted to the competition. The model selection procedure was used to determine the best performing model candidate per model category.

First, all models were required to perform better than the baseline model, Sec. III-B3. Second, for each model category, the best model together with all statistically equivalent performing models were determined, according to Sec. III-B4. Third, more complex models,[4] in terms of model input, were required to perform significantly better than simpler models. With the present paper, it is intended to provide an overview of the competition and especially the three standardized models and their performance, omitting some of the more fine-grained details about what other models were submitted, etc. The interested reader can find some more information in [100], for example.

C. RESULT OF COMPETITION

As a result of the competition, each of the three models proposed in this paper, the bitstream model ITU Rec. P.1204.3, the reduced-reference pixel-based model P.1204.4, and the no-reference hybrid model P.1204.5 were the single best performing model in their category. In particular, none of the full-reference pixel-based models submitted to the competition performed significantly better than the reduced-reference model (P.1204.4) described in this paper. As a consequence, due to its equivalent performance, the P.1204.4 model is referred to as reduced-/full-reference model.

IV. DATABASES CREATION

In this section, details about each step of the database creation part of the competition are provided. The database creation stage involved content selection, HRC design, the encoding pipeline to create the resultant processed video sequences (PVS) and the final distribution of these PVSs into different databases. Content selection and HRC design steps were conducted in parallel to use the time optimally, and a final mapping of HRCs to the content complexity was done using a content complexity measure described later.

A. CONTENT SELECTION

The subjective tests used in the process of creating the P.1204.3-5 recommendations were performed with SRC clips of around 8 s duration. 4K *Source footage* from both openly available internet sources and some provided to the project in kind by proponents (Yonsei University, TU Ilmenau and Ericsson AB) was collected to create a large pool to draw from. 1440p *Source footage* was allowed for databases intended to run on Mobile or Tablet. All these videos were individually reviewed and screened for impairments such as shaky scenes, regions of non-pristine picture quality etc. The *Source footage* parts deemed to be of appropriate quality were then cut into source files (SRC) according to the information specified in the manually created *Scenecut file* for each corresponding *Source footage*. The cutting was done with FFmpeg using the `-copy` video codec option to capture the correct frames into a new file.

Each resulting SRC was further manually reviewed by each proponent to ensure the best content clarity and, in case problems were identified, either a recut was performed or the corresponding SRC was rejected. For example, SRCs with at least a scene cut in the first and last 2 seconds were rejected. Approval from at least three proponents was needed to consider an SRC to be valid for being included in subjective testing.

The collection of *Source footage* for validation was performed only after the model submission, to ensure that proponents had no prior knowledge of the validating contents.

The selected footages encompass a vast variety of possible contents, i.e. natural scenes, movies, dynamic scenes, animations, video games etc. 3 SRCs were used both in the training and validation phase to generate the "common set PVSs" (see section IV-B). One further SRC from the training phase was re-used in validation with different test conditions. The number of unique footages and SRCs for both the training and validation phases is reported in Table 1.

TABLE 1. Number of unique footages and SRC files used in the training (TR) and validation (VL) phase, and according footage framerates in frames per second (fps).

		TR	VL	TOT
Footages	50/60 fps	27	20	43 (4 common TR/VL)
	24/25/30 fps	32	97	129
	Total	59	117	172 (4 common TR/VL)
SRC files	50/60 fps	203	79	278 (4 common TR/VL)
	24/25/30 fps	138	294	432
	Total	341	373	710 (4 common TR/VL)

All SRCs were characterized in terms of spatial and temporal complexity, using the spatial and temporal information measures SI and TI, respectively, as specified in ITU-T Rec. P.910 [28]. The mean SI and TI values per SRC used in the training and validation tests are shown in Fig. 1.

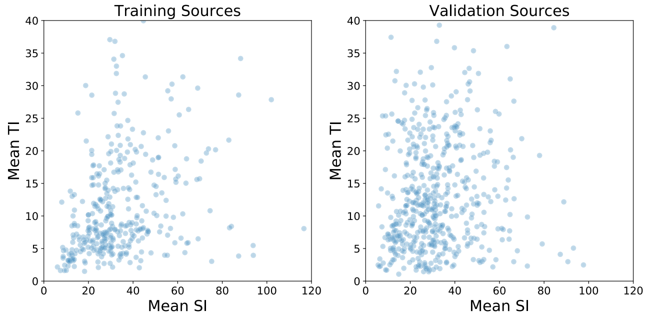


FIGURE 1. SI-TI of all the sources used in training and validation.

[4] Here, complexity means that either additional sources of information are required (e.g. a pixel-based NR model vs. a pixel-based hybrid NR model), or referring to complexity of input information of similar type, e.g. reduced reference vs. full reference, with FR being more complex.

B. HRC DESIGN

In this section, details about the HRC design process and hence test conditions are provided.

At first, the codec parameter ranges were agreed upon among all the proponents. Since the application areas of the models developed are wide-spread, the parameter ranges cover the typical encoding settings used in adaptive streaming applications, and extend even beyond. In Table 2, all parameter ranges are listed that were used for the three video encoders across all the subjective tests. Fig. 2 depicts the bitrate ranges for each encoder.

TABLE 2. Parameter ranges for video encoders.

Parameter	Range
Video Codec	H.264, H.265, VP9
Encoded Resolution	TV/Monitor: 640 × 360 – 3840 × 2160, Mobile/Tablet: 426 × 240 – 2560 × 1440
Framerate	15, 24, 25, 30, 50, 60 frames per seconds
Presets	H.264/H.265: online, i.e. Youtube, Bitmovin or Vimeo; medium, ultrafast, fast, veryfast, slower, slow, veryslow. VP9: speed presets 0, 1, 2, 3, 4
GOP Size	Auto, 2, 5 seconds
Encoder Implementation	H.264: libx264 (ffmpeg), H.265: libx265 (ffmpeg), VP9:libvpx-vp9 (ffmpeg), YouTube, Bitmovin, Vimeo
Chroma Subsampling	YUV420, YUV422
Bit-depth	8,10 bits
Encoding Types	1-pass, 2-pass (with and without min max bitrate constraints), Constant rate factor (CRF) encoding. Unknown encoding recipes employed by YouTube, Vimeo, Bitmovin
Bitstream Container	mp4, webm, mkv

Framerate up-sampling and resolution upscaling, where the encoded framerate and resolution is higher than the reference video framerate and resolution, was not part of our test matrix. HRCs were designed using a top-down approach, where the above parameter ranges were spanned using a number of test conditions. Then these test conditions were split into individual databases by making sure that each database contained roughly equal representations of different video codecs, encoded resolutions and framerates. The bitrate for different encoding resolutions was randomly sampled from the specified ranges. For YouTube, Bitmovin and Vimeo encodings, defined as "online conditions" in Table 2, the SRCs were uploaded to the respective service, and the encoded video bitstreams were downloaded. For YouTube and Vimeo, no encoding parameters were allowed to be specified. For Bitmovin, it is possible to exactly specify the input parameters. However, it was completely unknown, how exactly the actual video encoding was performed for these services. All databases include 5 common HRCs. Each common condition was mapped to 3 common SRCs, resulting in 15 common PVSs. The idea with this "common set" used in the tests across the different labs is to find out whether all databases were roughly aligned in terms of the resulting quality ratings and hence scale usage. The encoding parameters for the common set are detailed in Table 3.

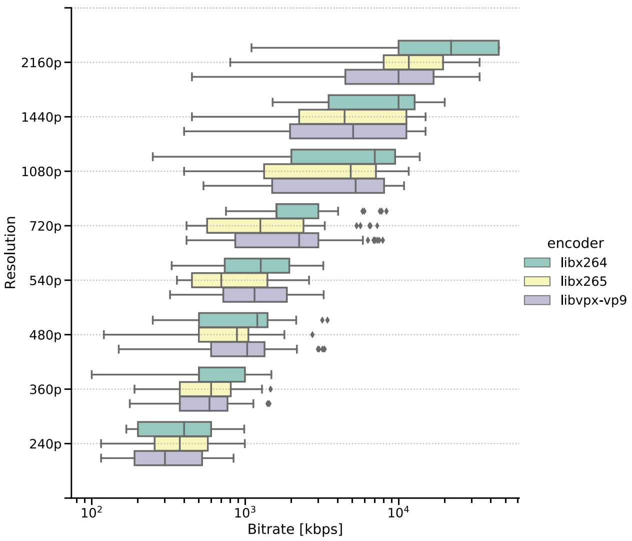


FIGURE 2. Bitrate range for each encoder–resolution pair.

To account for the difference in the target resolution of the considered display devices, namely, PC/TV and Mobile (MO) / Tablet (TA), implicitly comprising also different subject expectations for quality on these different device categories, the highest and lowest anchors were adjusted accordingly. Since the display resolution of the MO/TA category was 2560 × 1440, the highest anchor HRC was HRC0484 and not HRC0571 as it was used for PC/TV, for which the coding resolution is 3840 × 2160. The lowest-quality anchor for MO/TA was chosen as HRC0001, with an encoding resolution of 426 × 240 and encoding framerate of 15 fps. For the PC/TV case, the lowest-quality anchor was HRC0115, with an encoding resolution of 640 × 360 and encoding framerate of 24/25/30 fps, to account for typical real-life conditions and the higher expectation of quality on these devices.

To balance SRCs in terms of content complexity, a coding-specific complexity measure was conceived. To this aim, CRF encoding with the H.264 codec was used, encoding all the SRCs at a fixed CRF value of 30. The resulting bitrate was used to categorize SRCs into four different complexity classes. For each HRC, 2 alternative values for bitrate were specified as *low/high* value. The actual bitrate of a given PVS took into account the complexity class of the corresponding SRC: The *low* value was assigned to sources with complexity 0 or 1, while *high* was assigned to sources with complexity class 2 or 3.

TABLE 3. Common HRCs for the PC-Monitor/TV case. Video codec is H.264 for all common conditions.

HRC-ID	Resolution	Bitrate (kbps)	FPS	MOS Range	
				PC/TV	MO/TA
HRC0001	240p	100/200	15	-	1.167 - 2.476
HRC0115	360p	300/500	24/25/30	1.160 - 2.917	1.792 - 3.571
HRC0388	720p	800/1600	50/60	1.500 - 3.917	2.833 - 4.542
HRC0436	1080p	3500/7000	50/60	2.958 - 4.833	3.833 - 4.810
HRC0484	1440p	6000/10000	50/60	3.333 - 4.875	4.083 - 4.762
HRC0571	2160p	30000/45000	50/60	3.667 - 5.000	-

C. DATABASES

A total of 13 training and 13 validation databases were created as part of the competition. Each database contains between 180-203 PVSs, each of 7 to 9 s duration. Subjective tests were performed on four different display devices, namely, PC-Monitors (31.5-37 inch size), TV (55-75 inch size), Mobile (Samsung Galaxy S7, 5.1 inch) and Tablet (10 inch size). For the PC-Monitor and TV tests, the viewing distance was 1.5H [101], where H denotes the height of the display. The display resolution for PC-Monitor/TV tests was 4K/UHD-1 (3840 × 2160 pixel). For mobile and tablet databases, the viewing distance was 5-7H [101]. All subjective tests were conducted in compliance with ITU.P910 [28]. Subjects were handed written instructions common to all test labs, and shown training videos to provide an understanding of the test. Each test was roughly an hour long, including the breaks. A minimum of 24 valid subjects were required for each test. Outlier detection was based on Pearson Correlation (PCC) of individual subjects with all others, using a threshold of 0.75 below which subjects were considered as outliers. The details of individual databases in terms of the number of PVSs, display type, number of subjects, average correlation over all subjects and the average confidence interval are provided in Tables 4 and 5.

TABLE 4. Training database details. ("DB-ID": Database ID. "Display" used for playout. "N": number of subjects. Avg. correl.: Average correlation of individual subjects with mean. "Avg. CI": Average confidence interval of mean. "PVSs": Number of PVSs in test.).

DB-ID	Display	N	Avg. correl.	Avg. CI	PVSs
P2STR01	Mobile	26	0.82	0.29	203
P2STR02	Mobile	24	0.87	0.27	199
P2STR03	Mobile	30	0.87	0.23	200
P2STR04	PC	26	0.91	0.24	199
P2STR05	PC	26	0.84	0.27	187
P2STR06	Mobile	24	0.82	0.25	187
P2STR08	TV	24	0.89	0.26	179
P2STR09	PC	25	0.86	0.25	187
P2STR10	PC	34	0.86	0.21	187
P2STR11	TV	24	0.89	0.25	187
P2STR12	PC	24	0.85	0.28	183
P2STR13	TV	25	0.87	0.25	187
P2STR14	TV	24	0.84	0.24	179

TABLE 5. Validation database details. ("DB-ID": Database ID. "Display" used for playout. "N": number of subjects. Avg. correl.: Average correlation of individual subjects with mean. "Avg. CI": Average confidence interval of mean. "PVSs": Number of PVSs in test.).

DB-ID	Display	N	Avg. Correl	Avg. CI	PVSs
P2SVL01	TV	30	0.82	0.25	185
P2SVL02	Mobile	24	0.82	0.26	186
P2SVL03	Mobile	21*	0.82	0.30	186
P2SVL04	Mobile	24	0.88	0.28	195
P2SVL05	TV	25	0.87	0.28	194
P2SVL06	TV	24	0.89	0.26	191
P2SVL07	TV	25	0.86	0.26	188
P2SVL08	PC	27	0.82	0.29	195
P2SVL09	TV	28	0.81	0.28	191
P2SVL10	TV	26	0.86	0.21	195
P2SVL11	TV	24	0.87	0.27	195
P2SVL12	Tablet	24	0.84	0.20	195
P2SVL13	TV	26	0.84	0.25	187

* Extra subjects were removed from this database due to file copying bugs. Database was kept since correlation and CI was deemed ok after extensive analysis.

Some training PVSs were screened out due to bad content or wrong encoding settings. The total number of training and validation PVSs after the screening process was respectively 2464 and 2483.

D. VIDEO PROCESSING

An FFmpeg-based processing chain was developed to conveniently go from the selected SRCs and HRC-setting files to the PVSs intended to be viewed in the subjective tests. To make the processing as repeatable as possible without having all parties to buy the same hardware, an Ubuntu 16.04 virtual machine (VM) image was shared. This image was prepared with a specific build of FFmpeg 3.2.2 that could handle both 8-bit and 10-bit video for all combinations of H.264, H.265, and VP9 encoding/decoding. It also included specific versions for the other software and libraries necessary for running the processing chain. The FFmpeg lossless codec ffv1 was used as an intermediate codec for all modifications that were not codec-specific. An overview of the Processing Chain is shown in the flow-chart in Fig. 3.

TABLE 6. **Proportions of different parameters in validation databases.**

Parameter Type	Parameter Value	Proportion
Codec	H.264	34%
	H.265	33%
	VP9	33%
Coding Type (per codec)	1-pass encoding	20%
	2-pass encoding	65%
	crf encoding	7%
	online services	8%
Encoding Presets	ultrafast,veryfast,fast	10%
	medium	75%
	veryslow,slow,slower	15%
Frame rate	15 fps	3%
	24/25/30 fps	81%
	50/60 fps	16%
Resolution	240p	4%
	360p	10%
	480p	10%
	540p	10%
	720p	13%
	1080p	18%
	1440p	20%
	2160p	15%

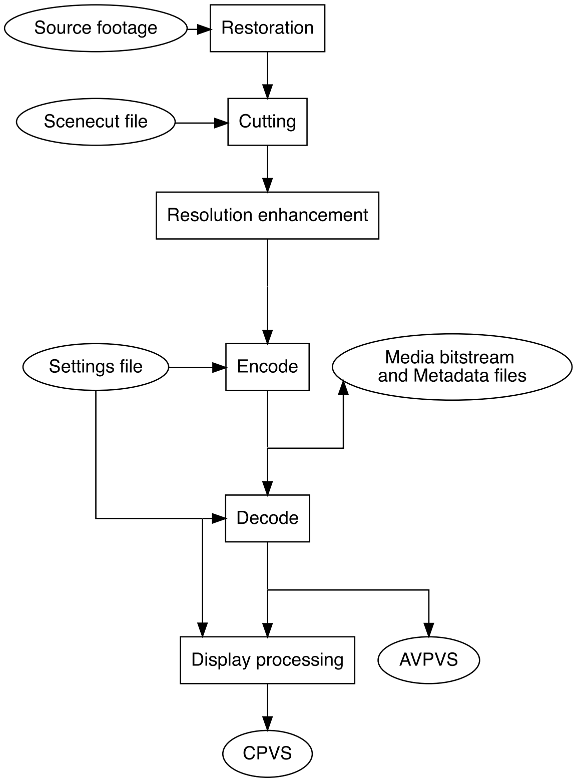


FIGURE 3. **Processing chain flowchart.**

To process the set of HRC and SRC combinations that comprise a database, a *Settings file* had to be created in a pre-defined YAML format. This file contains information about the HRCs such as encoder settings, adaptation levels and durations, stalling duration and so on. Only codec, pixel depth, framerate- and resolution-related parameters were part of the HRCs in the tests for the P.NATS Phase 2 competition. No stalling or explicit bitrate adaptation was used, even though the processing chain has the capabilities to automatically produce PVSs with such degradations. The.yaml-file also describes how these HRCs should be combined with the available SRCs and what, if any, post-processing should be performed to create playable video output files.

Based on these inputs, the processing chain then creates a set of FFmpeg commands to encode, decode, add stalling events, and, if necessary, concatenate the decoded video sequences. These commands are put in a queue and are processed in series or in parallel, depending on the available hardware, to create bitstream *videoSegment* files, decoded video files referred to as *AVPVS*s, and meta data information files describing quality-change events, stalling events and media frame sizes (*.qchanges-files*, *.buff-files* and *.afi/.vfi-files*).[5] Following this, *CPVS* files are generated from the AVPVS to create a video file that is not further upscaled or changed in any way by the display it is played on. This last step, *Display processing*, is done to minimize the effect of the different TV and PC-display brands' proprietary upscaling algorithms. All *CPVSs* intended for PC/TV were output with a resolution of 3840 × 2160 and 60 frames per second, while the *CPVS* for Mobile/Tablet were in 2560 × 1440 pixel resolution with the same frame rate, matching the resolution of the display used in each test. PC/TV *CPVS* used a rawvideo or v210 codec, depending on whether it was an 8-bit or 10-bit video. The playout software for PC/TV supported both .mkv and .avi containers. The Mobile/Tablet player [102] could not play out rawvideo without stuttering or frame loss, so a very high quality H.264 setting was used instead. The *CPVS* were encoded with libx264 in FFmpeg using `-crf 15 -preset fast -profile:v high` settings.

If a video was supposed to be processed by online services (YouTube/Vimeo/Bitmovin), the *SRC* was uploaded using SFTP or manual upload, depending on the service. Some services did not leave any choice for different encoding parameters, while other presented a number of quality levels. The intended encoded video was downloaded and renamed as a valid *videoSegment* file. This enabled the processing chain to generate all the metadata, *AVPVS* and *CPVS* files even for cases for which the encoding was not performed by the processing chain itself.

V. MODEL DESCRIPTION

A detailed description of the three standardized models, namely, the bitstream-based NR model (ITU-T Rec. P.1204.3), the pixel-based FR/RR model (P.1204.4) and the hybrid, meta-data and pixel-based NR model (P.1204.5) is provided in this section.

At the start of the P.NATS Phase 2 standardization project, the design of the P.1204 models was chosen so as to principally be compatible with the modular P.1203 model architecture [38], [48], [74]. Accordingly, besides video quality

[5] It is reminded that for the short-term video quality models presented in this paper, no quality changes, stalling or initial loading delay were used, in contrast to what was done during the development of the longer-sequence ITU-T Rec. P.1203 standard family.

estimates for sequences of between 5 to 10 sec duration as the primary model output, all models also provide per-1-second video quality scores on a 5-point scale.

It is noted that this continuous score can be considered as a *memoryless instantaneous score*, related with but conceptually different from the instantaneously rated quality as it has been assessed, for example, in [8], [9], [17]–[19]. In such studies, test subjects typically rate quality on a continuous scale with a slider, following perceived quality over time. A corresponding test method is SSCQE (Single Stimulus Continuous Quality Rating), see ITU-R BT.500 [27]. Here, ratings are dependent on the quality at previous times of the same viewing session, and hence comprise aspects of human memory.

The *memoryless instantaneous score*, provided per-1-second by the P.1204.X models – and also their FHD bitstream-based predecessor P.1203.1 – do not include these memory effects, for a reason. As they are quasi memoryless, they can be used continuously regardless of the prior history of quality in a given session. With a model that predicts instantaneously rated quality, there is no time-shift invariance, since memory will differ depending on when the viewing is considered to have started. Instead, with the chosen per-1-sec scores, memory and longer-term integration can be addressed at a later stage by a suitable quality integration module, such as P.1203.3 [14], [50], [74], possibly together with according per-1-second audio-quality data, as well as initial loading delay and stalling information.

An illustration of the three P.1204 models and their corresponding input information is shown in Figure 4.

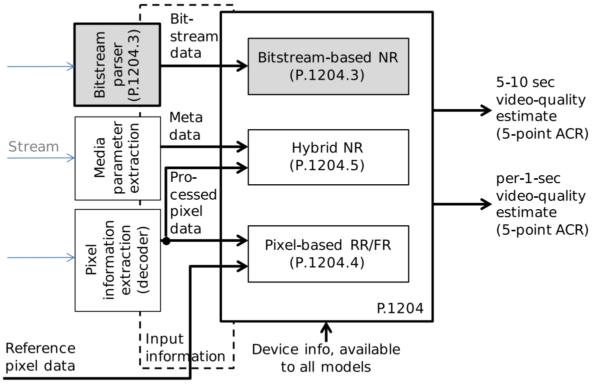

Bitstream parser (P.1204.3)
Media parameter extraction
Pixel information extraction (decoder)
Stream
Bit-stream data
Meta data
Pro-cessed pixel data
Bitstream-based NR (P.1204.3)
Hybrid NR (P.1204.5)
Pixel-based RR/FR (P.1204.4)
P.1204
5-10 sec video-quality estimate (5-point ACR)
per-1-sec video-quality estimate (5-point ACR)
Reference pixel data
Input information
Device info, available to all models

FIGURE 4. Model outline for the three different P.1204 model variants P.1204.3, P.1204.4 and P.1204.5 and their respective input information.

As can be seen from the diagram, information about the device used is available to all three model types (i.e., PC/TV, tablet, mobile). Further, the P.1204.3 bitstream model uses input information obtained from parsing the encoded bitstream. The P.1204.3 model algorithm and the bitstream information that the model requires are summarized in Section V-A. An open-source implementation including the bitstream parser is available, see [42]. The P.1204.4 pixel-based RR/FR model requires both the processed-pixel and reference-pixel information as input. Details about the model algorithm are given in Section V-B. The hybrid NR model P.1204.5 uses video metadata such as the codec used, resolution, framerate and bitrate together with the processed-pixel information as input. The algorithm of the hybrid model is described in detail in Section V-C.

A. BITSTREAM-BASED MODEL: P.1204.3

The bitstream model P.1204.3 consists of two parts, a "parametric" model part based on arithmetic functions mapping input parameters to quality, and a machine learning model part. The two parts are described in detail in the following sections.

1) PARAMETRIC PART – CORE MODEL

The parametric part of the model, also referred to as "Core Model", follows the principle of degradation-based modeling, as used for example in ITU-T Rec. P.1203.1 [37], [38].

The general idea is that video quality can be modelled as the subtraction of different video degradations from a quality-value for a pristine presentation. Three different degradations are considered in this model: quantization degradation D_q, upscaling degradation D_u and temporal degradation D_t. All degradation values are expressed on a scale from 0 to 100, following the impairment principle underlying the "Transmission Rating Scale" of the so-called E-model, a planning tool for speech-quality assessment [103]. This mapping from the 5-point ACR scale to the 100-point scale is performed to compensate for the known compression of the 5-point ACR scale at its ends, which is due to, among others, the avoidance of extreme ratings by subjects (see e.g. [104]).

a: QUANTIZATION DEGRADATION: D_q

Quantization degradation relates to the observable coding degradations that are introduced due to the chosen quantization settings during the encoding process and is usually visible as blockiness or deblocking-filter-related blurring to the end-user. The Core Model handles D_q separately per codec.

First, the variable *quant* is defined as a function of the quantization parameter by

$$quant = \frac{QP_{non-Iframes}}{QP_{max}}, \tag{7}$$

where $QP_{non-Iframes}$ is the average of the Quantization Parameter (QP) for all non-I frames for an entire segment, and QP_{max} is the maximum quantization parameter. The number of codec categories is extended from the initial three (H.264, H.265, VP9) to five, by including the bit-depth information and splitting H.264 and H.265 into 8- and 10-bit variants. Here, QP_{max} is codec- and bit-depth-dependent, using 51 for the 8-bit variant of H.264 and H.265, 63 for the 10-bit variant of H.264 and H.265, and 255 for VP9. The calculation results in a scaled value $quant \in (0, 1]$. This value of *quant* is used to estimate an intermediate quality value resulting from encoding mos_q, using a parametrized exponential function,

$$mos_q = a + b \cdot \exp(c \cdot quant + d). \tag{8}$$

with $mos_q \in [1, 5]$.

Finally, mos_q is converted to a degradation D_{q_raw}, using the inverse-S-shape mapping function *RfromMOS* to map the 5-point ACR scale to a 100-point scale, similar to the one used in the E-model, see ITU-T Rec. G.107 [103].

$$D_{q\_raw} = 100 - RfromMOS(mos_q) \tag{9}$$

The final D_q value is the result of clipping D_{q_raw} to the range [0, 100],

$$D_q = \max(\min(D_{q\_raw}, 100), 0). \tag{10}$$

b: UPSCALING DEGRADATION: D_u

Besides the one for coding degradation, the Core Model comprises a component for resolution upscaling degradation. In general, an upscaling degradation results from upscaling the distorted video to the screen resolution during playback, which can be perceived by an end-user as blurriness. In the real-world streaming scenario, upscaling is typically performed by the player software, where streaming resolutions lower than the target screen resolution typically are a result of the adaptive streaming of bandwidth-dependent representations. In the model development process, this degradation was assumed to be codec-independent.

First, the factor f_{scale} is calculated as the ratio of the number of pixels N_{coding} at coding resolution to the number of pixels $N_{display}$ at display resolution,

$$f_{scale} = N_{coding}/N_{display}, \tag{11}$$

with $N_{display} = 3840 \times 2160$ for PC/TV display and $N_{display} = 2560 \times 1440$ for mobile/tablet. N_{coding} is the number of pixels of the encoded video. The factor f_{scale} is always limited to $f_{scale} \in (0, 1]$. Next, the upscaling degradation D_{u_raw} is calculated based on the scaling factor f_{scale} by

$$D_{u\_raw} = x \cdot \log(y \cdot f_{scale}) \tag{12}$$

and then clipped to the range [0, 100] by

$$D_u = \max(\min(D_{u\_raw}, 100), 0). \tag{13}$$

Here log denotes the natural logarithm, and x and y are device-specific coefficients determined during model training.

c: TEMPORAL DEGRADATION: D_t

The third degradation type considered by the Core Model is based on lower framerate representations as a possible means of streaming adaptation and subsequent adjustment to the used display, which may be perceivable as jerkiness. Similar to upscaling D_u, we handle this in a codec-independent fashion.

First, a frame rate factor $c_{framerate} \in (0, 1]$ is calculated as the ratio of *coding* frame rate fps_{coding} to the fixed *display* frame rate $fps_{display} = 60$,

$$c_{framerate} = \frac{fps_{coding}}{fps_{display}}. \tag{14}$$

Next, the temporal degradation D_{t_raw} is computed based on the frame rate factor by

$$D_{t\_raw} = z \cdot \log(k \cdot c_{framerate}) \tag{15}$$

and then clipped to the range [0, 100] using

$$D_t = \max(\min(D_{t\_raw}, 100), 0). \tag{16}$$

Here, z and k are device-specific coefficients.

d: PREDICTION AND MODEL COEFFICIENTS

The quality prediction $Q_{p,0-100}$ of the parametric part on the [0, 100]-scale is given by subtraction of all three degradations from the maximum quality,

$$Q_{p,0-100} = 100 - (D_q + D_u + D_t). \tag{17}$$

The final prediction $Q_{parametric}$ is given by a further rescaling to a 5-point MOS-scale, the details of which can be found in [52].

During training of the model, the subjective scores were linearly mapped to a 4.5-point scale from the 5-point scale in order to avoid information loss due to the *RfromMOS* and *MOSfromR* computations, since both of these mapping functions assume that the highest MOS that can be reached is 4.5. As a final step, the predictions on the 4.5-point scale were mapped back to the full 5-point scale range using a simple linear transformation, the details of which can be found in [52].

The coefficients for both the PC/TV and mobile/tablet cases are reported in the corresponding ITU-T standard ITU-T Rec. P.1204.3 and in the open-source model implementation,[6] see [42], [52].

2) MACHINE-LEARNING-BASED VIDEO QUALITY MODEL

The second part of the model uses a machine learning approach to estimate video quality. This part of the model is used mainly to estimate the ‘‘residual’’, that is, the part of the MOS that the parametric Core Model part is unable to predict. Hence, the target for the training of the machine learning part of the model is the residual

$$R_{target} = MOS - Q_{parametric}. \tag{18}$$

Random Forest (RF) regression is used to predict the residual. Two different RF models are trained, one for PC/TV and mobile/tablet cases. The model output is the predicted residual R_{pred}.

Features such as the average motion per frame, motion in the x-direction (horizontal motion) and frame sizes with

[6] https://github.com/Telecommunication-Telemedia-Assessment/bitstream_mode3_p1204_3

frame types are used in addition to the features of the parametric, Core Model part. The rationale behind this is that the parametric part is not able to fully incorporate spatio-temporal content complexity of the video sequences. Further, encoding-specific choices for certain bitstream representations cannot completely be captured by QP, framerate and resolution alone. The RF model also uses the parametric part's prediction $Q_{parametric}$ as an additional feature. These features are aggregated according to different functions and used as input to the random forests. These aggregations are presented in Table 7. The Random Forest model used 20 trees with a fixed depth of 8. The final Random Forest quality prediction $Q_{randomforest}$ is given by

$$Q_{randomforest} = Q_{parametric} + R_{pred}. \tag{19}$$

Hence, it is the addition of the predicted residual value R_{pred} to the parametric prediction $Q_{parametric}$.

TABLE 7. **Aggregated features for RF model.**

Aggregated Feature	Type
Framerate	float
Resolution $(width \times height)$ of the distorted video	int
Codec (H.264, H.264_10bit, H.265, H.265_10bit, VP9)	boolean
$Q_{parametric}$	float
Mean bitrate per segments	float
Maximum frame size	int
Kurtosis of the non-I frame sizes	float
Standard deviation of frame size of non-I frame in bits	float
Quant $(Quant = \frac{QP_{non-Iframes}}{QP_{max}})$	float
IQR of the average QP of non-I frames	float
IQR of the minimum QP per frame	float
Kurtosis of the average QP of non-I frames	float
Mean of the average QP of non-I frames	float
Standard deviation of maximum QP of non-I frames	float
Kurtosis of the average motion per frame over all frames in a segment	float
Minimum standard deviation of motion in the x-direction (horizontal motion) per frame	float

3) OVERALL VIDEO QUALITY PREDICTION

The final predicted quality Q of the model is then the convex linear combination of the prediction $Q_{parametric}$ from the parametric part and the prediction $Q_{randomforest}$ from the machine learning part,

$$Q = w \cdot Q_{parametric} + (1 - w) \cdot Q_{randomforest} \tag{20}$$

In the presented model, equal weights, thus $w = 0.5$, are assigned to both of the predictions, shown in Eq. 20.

In addition to the per-segment scores, the model also predicts the per 1-sec scores. The specific details of the per 1-sec score calculation can be found in the corresponding standard [52].

B. PIXEL-BASED MODEL: P.1204.4

This section describes the reduced-reference pixel-based model P.1204.4. A reduced-reference model is a special form of full-reference model. In a full-reference model, quality Q of a test video v – called degraded video – is estimated by a function G depending on the degraded video v and on the reference video v_{ref},

$$Q = G(v, v_{ref}). \tag{21}$$

In the reduced-reference case, the function G depends on the reference through features f_{ref} of the reference v_{ref} only. The features are extracted by the reference-feature extraction function ϕ,

$$f_{ref} = \phi(v_{ref}), \tag{22}$$

and there is a restriction on the size of the features. The quality of the degraded video is estimated by function G' by

$$Q = G'(v, f_{ref}). \tag{23}$$

The reference features f_{ref} are sometimes called the side information, as in an operational setup this information can be transmitted over a side-channel to the measurement device.

The following description contains the main ideas of the reduced-reference model. The full details can be found in ITU-T Rec. P.1204.4 [53], together with the values of constants and parameters used in this description.

1) OVERVIEW

The general computation steps are presented here slightly simplified to outline the overall ideas. For the video frames of the test video and the reference, a multi-resolution pyramid of the Y-component is computed. For each resolution, an edge representation is determined. Local patch statistics based on this edge representation are computed, where the local patches are local both in space and orientation. Based on patch statistics, relative feature values are determined: the feature value of the test video is measured relative to the reference feature. Features computed per video frame are converted to a common scale with values in [0,1], measuring degradations, $D_0, D_1, ..$, such that larger values correspond to stronger degradations and lead to lower quality. This conversion uses S-shaped parametrized transformations $S_{par} : \mathbb{R}^+ \rightarrow [0, 1]$, mapping values from the positive real numbers to the unit interval. Aggregated, the quality Q is given by a multiplication of the form

$$Q = \prod_i (1 - D_i) \tag{24}$$

to account for interactions between different degradations. Besides a temporal degradation accounting for low frame rates, main degradations are spatial degradations based on a common edge feature described in the next paragraph.

2) EDGE REPRESENTATION

Let Y denote the Y component of a video frame for a given resolution, a matrix, with the indices denoted by i, j in the following. Features based on edge orientation and strength are computed. To reduce the complexity of the algorithm, edges are computed using the simplest possible filter: by difference of adjacent pixels. The resulting pixel difference is compressed using an inverse tangent function. A normalized edge representation is computed in the following way: Let $R[i,j]$ denote the edge strength at a spatial position (i,j) of the frame and $\phi[i,j]$ the orientation, the angle of the edge. Let $S[i,j]$ denote the average edge strength at the two positions at a fixed distance Δ of the point (i,j) on a line perpendicular to the edge.

The normalized edge strength Z is computed as the exceeding of the center edge strength above the lateral average $S[i,j]$, relative to the sum of center and lateral edge strength,

$$Z[i,j] = \frac{\max(0, R[i,j] - S[i,j])}{c + R[i,j] + S[i,j]}. \qquad (25)$$

Here, the strictly positive value of c avoids a division by zero. The lateral inhibition by S is twofold, in the numerator by subtraction, and by inclusion in the denominator.

3) PATCH STATISTIC

Based on the normalized oriented edge statistic computed at different resolutions, local patch statistics are computed. Patches are determined in a continuous way using a partition of unity in the spatial domain, and a partition of unity in orientation. For each of these local patches, a statistic is computed. In more detail, a partition of unity is a family of positive continuous [0, 1]-valued functions $(\theta_k)_{k=0,,,L-1}$ for some integer L having $\sum_{k=0}^{L-1} \theta_k = 1$. A family of patches (P_{mnk}) is computed using the partition of unity (Ψ_{mn}) in the spatial domain and the partition (θ_k) in orientation, i.e. a partition of unity on the unit circle. For orientation index k, and location indices m, n, a local patch P is computed by multiplication of the spatial partition, the orientation partition and the edge strength

$$P_{mnk}[i,j] = \Psi_{mn}[i,j] \cdot Z[i,j] \cdot \theta_k(\phi[i,j]). \qquad (26)$$

A patch statistic s_{mnk} is computed as the average over all values of P_{mnk} above a fixed quantile q,

$$s_{mnk} = \sum_{i,j} P_{mnk}[i,j], \qquad (27)$$

where the sum runs over all indices i, j with $P_{mnk}[i,j] > q$. The patch statistics are also called patch features. The value of q depends on the resolution at which the patch statistic is computed. The values of s_{mnk} are determined by the strongest edges of similar orientation at a close location. In particular, at high resolutions, there is a relation between the values of s_{mnk} and the sharpness, or the loss of sharpness due to up-scaling of the video.

Hence, at highest resolution, the sharpness statistic s_{sharp} is computed as the average over all patch statistic values above the $q = 0.95$ quantile, independent of spatial location and orientation.

These patch statistics s_{mnk}, computed at a fixed medium resolution, the sharpness statistic, together with the frame timestamps constitute extracted features of the video sequence.

For the reference video, the patch statistics, the sharpness statistics, together with the display time of each frame correspond to the extracted features f_{ref} of equation (22). These features can be computed based on the reference only. Thus, for a fixed reference, these features need to be computed just once and can be stored. All degraded videos having the same reference can be evaluated by using only the stored features of the reference. These features take at most 32kB for each second of reference video duration.

4) QUALITY PREDICTION

Relating patch statistics of the degraded video to those of the reference allows estimation of degradations. Missing details, blurriness of the test video show up in patch statistics having lower relative values. On the other hand, blockiness, deformed details as a result of strong compression can lead to an increase in patch statistics values. In particular, it can change the orientation of strong edges locally due to deformed details or blockiness. Thus, the orientation sensitivity of the patch statistics is important to measure an increase and decrease of relative patch features at the same time. The perception of degradations due to missing details and blurriness can be quite different from deformed details and blockiness. Therefore, the relative patch features are decomposed into a positive and a negative part. Either degradation part is mapped with a different S-transformation onto the quality scale, whose product according to equation (24) determines the overall quality.

Quality prediction is based on four spatial degradation measures: increase and decrease in patch feature values at a fixed medium resolution are the first two. Based on patch feature values at the highest resolution, sharpness is computed, and a decrease and increase in sharpness are the other two degradation measures. In more detail, the decrease of sharpness statistic s_{sharp} of the test video relative to the sharpness statistic r_{sharp} of the reference, is computed as

$$s_{rel\_sharp} = \min\left(1, \frac{s_{sharp} + c_s}{r_{sharp} + c_s}\right), \qquad (28)$$

where a constant $c_s > 0$ avoids a division by zero. Similarly, a fourth degradation measure determines the increase in the relative sharpness statistic. These degradation measures correspond to $D_1, ..D_4$ in equation (24).

This presentation is simplified, as perceptually and in the model, the estimated degradation is a function on the amount and spatial distribution of edges. In particular, a relative degradation close the border of the frame is weighted less than in the center, as attention is rarely driven to the border area. Further, a weighting based on motion and luminance is included.

Besides relative degradations estimated based on patch statistics, there is a degradation measure d_0 determining the impact of low framerates, as a function of display time of each frame and motion in the video sequence. As framerates below 24 fps are rare nowadays, the impact of this last ''jerkiness''-type degradation measure is minor. Each degradation is computed per-frame: the product of equation (24) computes a per-frame quality in the range [0, 1]. This per-frame quality is non-linearly aggregated to an overall video quality. This non-linear aggregation takes into account that low quality can have a stronger impact on the overall quality than what is achieved by a linear aggregation. Finally, the overall quality is rescaled to the MOS range [1, 5]. In addition to the overall quality, the model outputs a per 1-second score, which is the average per-frame quality over the 1-sec interval.

Model parameters were optimized for two different viewing conditions: a viewing condition using a small relative viewing distance representing a TV set or PC monitor setting, and a viewing condition representing a mobile use case with a smartphone display. The model can provide predictions for intermediate viewing distances by interpolation within the core model.

C. HYBRID MODEL: P.1204.5

Next, the hybrid no-reference model ITU-T Rec. P.1204.5 will be described [54]. The input for the hybrid model includes

- raw pixels as seen by the test subjects: i.e., decoded and up-scaled video *degVid*
- bitstream metadata information: type of encoder (H.264, H.265 or VP9), encoded video bitrate, encoded video resolution, encoded video framerate and the display resolution

The performance of the hybrid model was assessed with respect to three models, namely, the baseline model, the best pixel-based no-reference model working using the pixels of the decoded and upscaled video, the best bitstream model of the corresponding category (in this case Metadata Mode 0 model).

The hybrid no-reference model presented in this section has a 4-parameter logistic a-like function which for a given encoder maps average bitrate based feature x to an intermediate quality prediction S, where x is computed for each video segments of fixed resolution and framerate.

$$S = a \cdot \left(\frac{1 - \exp(-d \cdot (x - c))}{1 + \exp(-b \cdot (x - c))} \right). \tag{29}$$

Note that the above function without the term $(1 - \exp(-d \cdot (x - c)))$ is exactly the logistic function, where the constants a, b and c determine the saturation point, decay rate and shift of the quality curve with respect to x. The additional term $(1 - \exp(-d \cdot (x - c)))$ is introduced to add a faster decay of the curve towards lower values of x, where the constant d determines the decay factor of this additional decay term.

The constants a, b and c of the above equation are further functions of the three quantities, namely, the framerate, encoded resolution and the content complexity.

1) DEFINITION OF x

For a metadata-only model, bitrate carries the most important information about the quality of the video. However, bitrate only makes sense together with the information of encoder used and the encoded chroma subsampling format. This is because different encoders offer different compression efficiency and different chroma formats, due to their different size of the raw color information, may yield slightly different encoded bitrates. Let bitrate be defined in kilobits per second, then x is defined as:

$$x = \log_{10}(bitrate \cdot \exp(-h_0 \cdot (r - 1))), \tag{30}$$

where r has a different value for each chroma subsampling format. Precisely, r have values 1.0, 2/1.5, 5/4 and 5/3 for YUV420-8buit, YUV422-bit, YUV420-10bit and YUV422-10bit chroma subsampling modes, respectively. $h_0 > 0$ is a codec-specific constant. In other words, the raw bitrate is adjusted depending on the actual chroma format of *degVid*. Additionally, $\log_{10}$ is used to compress the range of the adjusted bitrate values.

2) IMPACT OF ENCODED FRAMERATE ON QUALITY

a in Eq.29 is an increasing function of the framerate. This is because high framerate yields a smoother representation of motion and hence a higher quality compared to low framerates. However, higher framerate means more frames to be encoded, which in turn means higher encoded bitrate. Hence, the quality curve shifts slightly to the right for high framerates. In other words, c increases with framerate. On the other hand, quality decay rate with regard to the bitrate increases for lower framerates, because low framerate brings more jerkiness in the represented motion, and hence b is a decreasing function of the framerate. The above understanding of the trend of the quality curve as functions of framerate (fps) can be formulated as:

$$a' = a_0 - a_f \cdot \left(\frac{60}{fps} \right) \tag{31}$$

$$b' = b_0 + b_f \cdot \left(\frac{60}{fps} \right) \tag{32}$$

$$c' = c_0 - c_f \cdot \left(\frac{60}{fps} \right), \tag{33}$$

where $a_f > 0$, $b_f > 0$ and $c_f > 0$ are codec-specific constants. a_0, b_0 and c_0 are codec-specifc initial values.

3) IMPACT OF ENCODED RESOLUTION ON QUALITY

The quality curve for a higher resolution saturates at a higher MOS and at higher bitrate values, so like the framerate case, a and c are also increasing functions of the encoded resolution. However, unlike framerate, quality decay reduces for lower resolution, i.e., the quality versus bitrate curve for a lower resolution is generally flatter compared to a higher resolution. Hence we can say that b is an increasing function of the encoded resolution. The above understanding of the trend of the quality curve as functions of encoded resolution can be formulated as

$$a'' = a' - a_s \cdot \log_{10}(u_a \cdot (f_{scale} - 1)) \tag{34}$$

$$b'' = b' - b_s \cdot \log_{10}(u_b \cdot (f_{scale} - 1)) \tag{35}$$

$$c'' = c' - c_s \cdot \log_{10}(u_c \cdot (f_{scale} - 1)), \tag{36}$$

where all constants in the above equations are codec-specific positive constants. The factor f_{scale} is defined as:

$$f_{scale} = \max\left(\frac{W_d \cdot H_d}{W_e \cdot H_e}, 0\right), \tag{37}$$

where $W_d \cdot H_d$ and $W_e \cdot H_e$ define the display and encoded resolutions, respectively.

4) IMPACT OF CONTENT COMPLEXITY ON QUALITY

Content complexity perhaps plays the most important role in determining the saturation points, decay rate and the shift of the quality curve of Eq.29. A simple content, for example, involving talking heads, is much easy to compress compared to a more complex content involving high motion or fast camera movement.

Traditionally, the content complexity is categorized using spatial information (SI) and temporal information (TI) features [28]. These measures require the availability of the original reference video to categorize the source complexity. Being no-reference, the standardized hybrid model only uses the pixels of the decoded signal which will have all the distortions, hence such a SI/TI characterization will not be accurate.

Moreover, these measures do not reflect the spatial and temporal complexity from the encoders point of view. For example, if we consider a video capturing only the translation motion of an object, TI will reflect temporal activity. However, for encoders it is still a low temporal complexity scene, as the motion compensation can perfectly capture the simple translation motion of the object. Similarly, fairly regular spatial features in a video image can be easily predicted using the intra prediction components in the encoder, while SI may suggest a higher spatial activity for such frames. So it is important that an encoder-consistent view of the content complexity is employed to make a quality prediction of encoded videos.

The standardized P.1204.5 hybrid model employs a VP9-based content complexity characterization feature. Using the constant rate factor (CRF) coding recipe of the VP9 codec, the *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 *degVid* at CRF value of 32. The bitrate of the resulting *degVidEncoded* is normalized with respect to framerate and resolution to create a content complexity feature $C_{complexity}$. The idea is that with a higher content complexity, videos will require higher bitrate to encode to the quality Q. Similarly, a lower-complexity content will require lower bitrate to achieve Q. This way, the VP9 codec can be used as a tool to obtain an encoder-consistent view of the content complexity.

It is known that the quality of a high-complexity source decays fast with regard to the bitrate compared to a low-complexity source. This is because complex videos are more susceptible to blocking artifacts compared to low-complexity videos. Hence, b is an increasing function of the source complexity. Since a higher-complexity video requires more bits to achieve the same quality than a low complexity video, c is an increasing function of the source complexity. As for the saturation point a is concerned, the higher the content complexity, the lower the saturation point. The above understanding of the trend of the quality curve as functions of content complexity can be formulated as

$$a = a'' - a_k \cdot C_{complexity} \tag{38}$$

$$b = b'' + b_k \cdot C_{complexity} \tag{39}$$

$$c = c'' + c_k \cdot C_{complexity}, \tag{40}$$

where $a_k > 0$, $b_k > 0$ and $c_k > 0$ are codec specific constants.

Equations 31 to 40 can be additively combined to yield values of a, b and c, which can then be used to compute the quality S for a certain video codec using the Eq.29.

TABLE 8. Linear mapping coefficients for device separation.

Device	m	g
PC Monitor	0.967	0.153
TV	1.051	-0.187
Mobile	0.942	0.146
Tablet	1.080	-0.330

5) IMPACT OF DISPLAY DEVICE ON QUALITY

The standard model has two sets of model coefficients, one set for the PC-Monitor/TV displays and the other for Tablet/Mobile displays. This is logical as subjects may assess the quality differently on different devices. Quality assessment on PC-Monitor and TV was quite consistent, hence these devices were not dealt with separately at the coefficient level. The same is true for the Tablet and Mobile display type. A final linear mapping accounts for slight variation in quality prediction between PC-Monitor and TV, and the Tablet and Mobile cases. Q, where $1.0 \leq Q \leq 5.0$, is the actual model prediction output.

$$S_d = m \cdot S + g \tag{41}$$

$$Q = \min(5, \max(1, S_d)), \tag{42}$$

where S_d denotes the device-based mapped quality. The table below reports the slope m and offset g values for the linear mappings for different devices.

In addition to per-segment score, the model also produces per 1-sec scores, which are directly derived from the per-segment score. The specific details of the per 1-sec score calculation can be found in the corresponding standard [54].

VI. MODEL PERFORMANCE

In this section, the prediction performance of each of the three models is presented. To evaluate the models, two different categories of databases were considered, namely the competition databases and open databases. The competition databases consist of the training and validation databases developed during the course of the P.NATS Phase 2 project within ITU-T/VQEG. The open databases are publicly available ones, which can be further categorized into two types

- Databases developed during the competition: For these databases the HRCs are developed with a similar design philosophy as the P.NATS Phase 2 databases, and can be used to evaluate other models in comparison, in contrast to the proprietary standardization databases.
- Completely independent databases, which are available from external sources. The HRCs of such databases can be designed with focus on a particular aspect of the application area.

The evaluation on complementary open databases is done to ensure that the model-performance evaluation is reproducible.

As a first step to evaluate the performance of the models, the P.NATS Phase 2 training and validation databases are used. To start, the output of the baseline model is plotted in comparison to the mean subjective scores (MOS) in the scatter plot shown in Fig. 5. For each database, a linear mapping was used to map the model output to the subjective scores, to normalize the scale of the subjective databases, following [97]. Further details on this normalization step are given in Sec. III-B. The figure shows the mapped model output with respect to the MOS for the 13 validation databases.

The $x = y$ line depicts the ideal prediction line, for the theoretical case of perfect agreement between model output and subjective MOS scores for each tested video. The indicated right boundary line corresponds to under-predictions of the subjective scores by 1 MOS. Similarly, the left boundary line corresponds to over-predictions of 1 MOS. It can clearly be seen that the baseline model has a significant number of points falling away from the ideal prediction line. The prediction is particularly bad for lower MOS values. Fig. 5b depicts the probability distribution function (PDF) of the prediction error. For the computation of the PDF, a bin size of 0.05 is used. Note that the prediction error for the baseline model is not symmetric. The PDF indicates that the baseline model over-predicts quality when compared to the MOS. The over-prediction is particularly high for lower MOS values – see the model prediction for the MOS range 1.0 to 2.5 in Fig 5a. This means that despite the per-database mapping, the baseline model does not have a neutral, unbiased scale for MOS prediction.

Figures 5c, 5e, and 5g depict the scatter plots for the winning bistream, pixel-based RR and Hybrid model candidates, recpectively. As discussed in Sec. III-A, the initially submitted model candidates were optimized before final standardization. Only the points for the 13 validation databases are shown in the scatter plots. For all three winning candidates, a large majority of points lie close to the ideal prediction line. There are some outlier cases for each of the three models. However, in general the prediction is significantly better compared to the baseline model. Additionally, the points are roughly equally spread along the two sides of the ideal prediction line. This can be confirmed by the roughly symmetric nature of the prediction error PDF plots of the three models shown in 5d, 5f, 5h. Moreover, from the three scatter plots it is evident that the models have a fairly neutral model scale for prediction of normalized MOS quality. Like for the baseline model, a MOS normalization was performed using a per-database linear mapping (based on [97], see Sec. III-B).

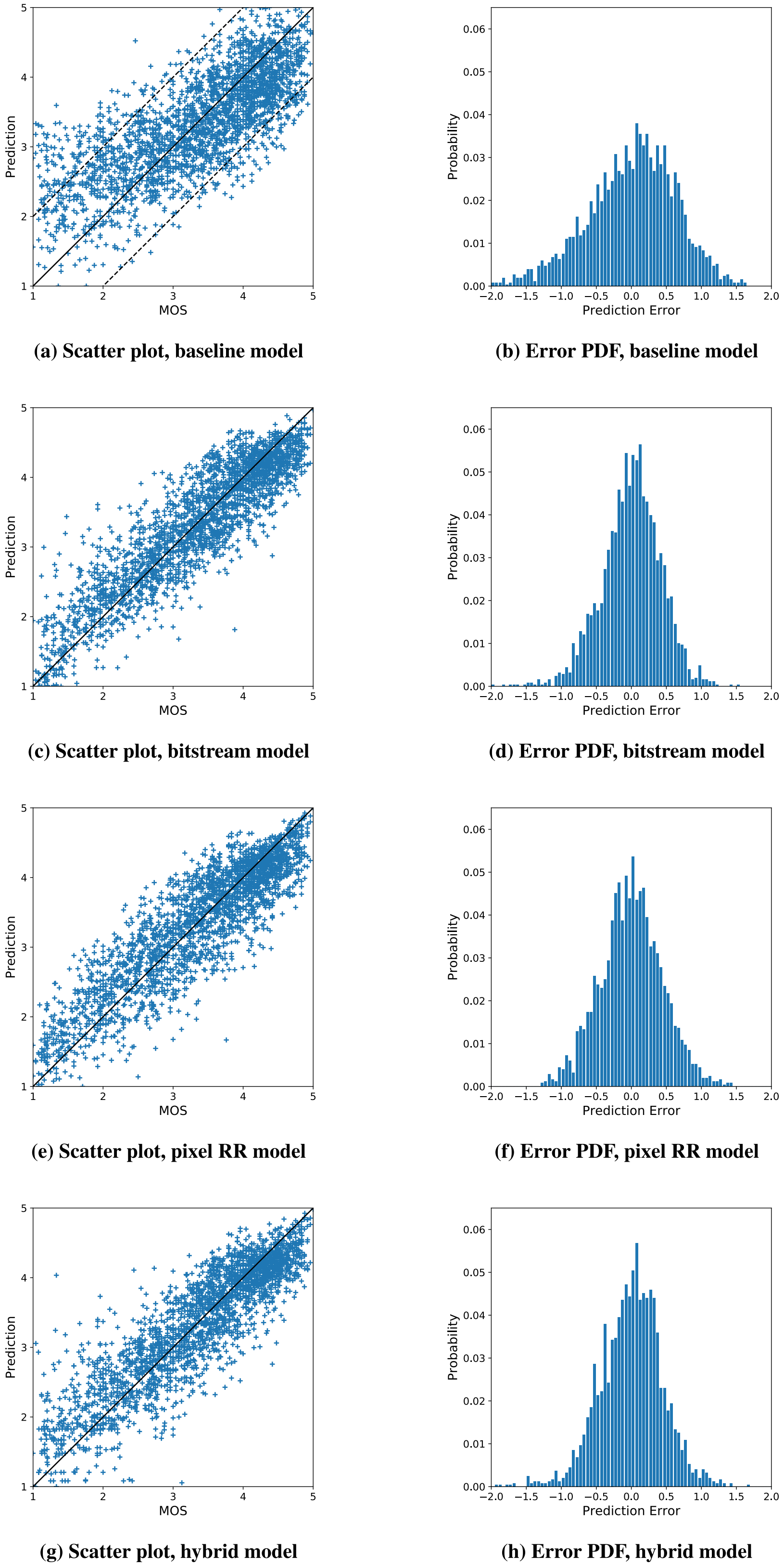
After the model-validation phase, which determined the winning models for each model category, the model coefficients for the three winning models for the three model categories were then re-optimized based on a cross-validation strategy (cf. Sec. III-A). Note that the submitted models were trained on the training databases identified by the prefix "P2STR" and validated on the databases identified by the prefix "P2SVL", see Tables 4 and 5.

The model re-optimization was done using a 5-fold cross validation. First, from the 26 databases, five splits of databases were created, each split containing 13 training and 13 validation databases. The following procedure was used to define the splits:

- Firstly, a level of prediction difficulty for each database was determined using the average RMSE of three models for that database. A lower average RMSE means the database is easy to predict while a high average RMSE means the database is difficult to accurately predict quality.
- Following this, 5 sets of 50 : 50 training-validation split were determined by ensuring that (a) splits have least similarity with each other, i.e., minimum overlap of databases between different splits, (b) for each split, the overall prediction difficulty of training databases is not very different from the one for the validation databases. (a) ensures that coefficients for models trained on different cross-validation splits are different, while (b) ensures that the trained models will generalize well for validation databases.
- For each split, databases of different display types (TV/PC-Monitor and Mobile/Tablet) have a balanced representation in the training and validation sets.

Model re-optimization was performed for each of the 5 cross-validation splits. The procedure outlined in Sec. III-B was used to compute the aggregated RMSE for each split. The coefficients corresponding to the best performing splits (the ones with the least aggregated RMSE for the respective model) have been reported in the final standard documents [52]–[54].

Table 9 reports the aggregated RMSE of the three submitted models and their standardized versions on the validation database set and for all databases. When computing the aggregated RMSE for all (both training and validation) databases, a 0.1/0.9 training/validation weighting is used, as explained in Sec. III-B. Note that for the submitted and standardized models, the actual training and validation databases are different. As indicated above, for the submitted models, the training databases are indicated by "P2STR" (Table 4), and the



(a) Scatter plot, baseline model

(b) Error PDF, baseline model

(c) Scatter plot, bitstream model

(d) Error PDF, bitstream model

(e) Scatter plot, pixel RR model

(f) Error PDF, pixel RR model

(g) Scatter plot, hybrid model

(h) Error PDF, hybrid model

FIGURE 5. Scatter plots of MOS vs the predicted scores and error probability distribution function (PDF) for the baseline model and the three winning model candidates.

validation databases by "P2SVL" (Table 5). For the standardized models, the training and validation databases were determined by the respective cross-validation split, as described above. Note that all three models have comparable RMSE figures when comparing the submitted and the finally optimized/standardized versions. This confirms that the three models, already in their submitted versions, provided stable predictions. The model optimization via cross-validation only resulted in a slight improvement in the performance of each model. Since for each model the RMSE of the optimized version on the validation databases ("VL") is comparable to the RMSE for the training ("TR"), it can be ensured that the standardized models generalize well to unknown cases.

In Table 10, the model performance of the submitted versions of the three models described in this paper is compared against FR models commonly used in the literature, namely PSNR, SSIM and VMAF. For each model, a per-database mapping is used to map the objective scores to the subjective MOS before computing the performance metrics. For VMAF and the three models described in this paper, a linear mapping is used, while for PSNR and SSIM, a 3rd-order polynomial mapping is used, as PSNR/SSIM are known to show a non-linear relationship to subjective quality scores. As the main performance criterion, the RMSE is employed in this paper also for comparison with other than the standardized models, reflecting the criterion used for model-performance evaluation in the P.NATS Phase 2 competition. In addition, values for Pearson correlation are provided as indicative information, reflecting the common practice in video quality model evaluation.

For the computation of Pearson correlation, remapped scores from all validation databases were pooled together. Here, all MOS values from different experiments were first combined to a joint set, then used to calculate correlations. Note that this is unlike the derivation of the values given in Table 9, where the RMSE per database was first computed, and then a weighted aggregation of RMSE values was performed. For both performance metrics reported in Table 10 (left part, "All HRCs"), the proposed models outperform PSNR, SSIM and VMAF. As can be seen from the results, VMAF performs better than PSNR and SSIM, which is expected. The lower performance of VMAF compared to the proposed models can be partly explained by the fact that the validation set includes frame-rate reduction HRCs, and VMAF lacks a feature to handle such cases. This can be confirmed when considering the complementary values in Table 10, columns denoted by "HRCs using SRC fps", obtained by recomputing the two performance figures for VMAF and the other models for a subset of cases that do not simulate frame rate reduction, that is, only consider cases where the SRC and HRC framerates are the same. The performance figure for VMAF on this subset (right two columns) is better compared to the full set, while for submitted models roughly show the same performance as on all data. It is worth pointing out that frame rate reduction scenarios are quite common in actual video streaming services. Just to give an example, a 60 fps 4K upload to YouTube will yield HD quality level with 30 fps.

TABLE 9. Aggregated RMSE on validation and on all databases (training and validation databases according to (3)) of the models submitted to the competition, and the standardized (re-trained) versions of the models.

Model	Validation DBs	All DBs
Submitted Bistream Model	0.429	0.421
P.1204.3 Standard	0.397	0.394
Submitted Pixel RR Model	0.448	0.444
P.1204.4 Standard	0.415	0.418
Submitted Hybrid NR Model	0.451	0.452
P.1204.5 Standard	0.442	0.440

The RMSE on individual validation databases is shown in the subplots of Fig. 6 for the three models P.1204.3, P.1204.4, and P.1204.5 as a deviation from the mean RMSE. In each subplot, the RSME values for PSNR, SSIM and VMAF are added for comparison. In general, the databases vary in terms of quality-prediction difficulty, and hence model efficiency can be different across databases. Moreover, since the three models use different types of input information and follow different modeling strategies, it can happen that one model performs better on one database than other models. Note that P.1204.3 and P.1204.5, which do not have access to the reference, have quite similar per-database RMSE distributions around the mean, while for P.1204.4, the RMSE distribution is slightly different. For database 10 ("P2SVL10"), P.1204.4 performs much better than the other two models.

A. EVALUATION ON OPEN DATABASES

A performance of the models on the aforementioned open databases is presented in the following. For this purpose, two different datasets, namely, AVT-VQDB-UHD-1 [91] and MCML [92] are considered. To evaluate the model on the AVT-VQDB-UHD-1 database, only samples for which the source video was available were considered. Due to limited digital rights for some sources, not all sequences could be made available. This resulted in considering 432 out of 756 samples for this part of the evaluation. The resolutions that were used in this dataset range from 240p to 2160p and framerates from 15 fps to 60 fps. Three codecs, namely, H.264, H.265 and VP9 were used to encode the videos. libx264, libx265 and libvpx were the encoder implementations used for H.264, H.265 and VP9 respectively. This database consists of four different subjective tests that are denoted as Test 1, a Test 2, Test 3 and Test 4 in Tables 11 and 12.

The four sets use similar conditions (HRCs) as in the P.NATS Phase 2 databases. Contrary to the P.NATS Phase 2 databases, the four sets use a full-matrix design with a smaller number of source videos, which can explain the large variation in RMSE values among the sets. These databases were developed during the competition and use the same

TABLE 10. Overall model performance of different models on P.NATS Phase 2 validation databases only (the ones with prefix "P2SVL"). Left: All HRCs. Right: Only HRCs where the HRC framerate corresponds to that of the SRC.

Model	All HRCs			HRCs using SRC fps		
	RMSE	Pearson	Spearman	RMSE	Pearson	Spearman
PSNR	0.716	0.630	0.615	0.688	0.625	0.609
SSIM	0.648	0.609	0.704	0.580	0.665	0.725
VMAF	0.611	0.761	0.773	0.548	0.794	0.790
P.1204.3	0.422	0.899	0.883	0.429	0.891	0.875
P.1204.4	0.441	0.889	0.872	0.440	0.884	0.864
P.1204.5	0.448	0.885	0.880	0.447	0.880	0.871

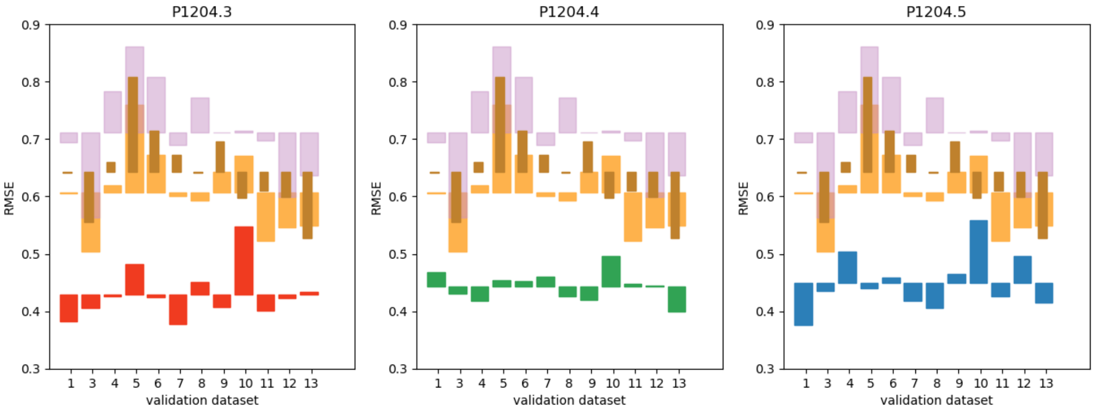


FIGURE 6. Model prediction error (RMSE) per validation dataset. Plotted is the prediction error for the submitted models P.1204.3 (red, left), P.1204.4 (green, middle), P.1204.5 (blue, right), and on all three subplots PSNR (purple), VMAF (orange), and SSIM (brown). For each model, the bars show the deviation from the mean prediction error. It can be seen that the prediction error for the models P.1204.x is lower than the prediction error of VMAF and PNSR.

TABLE 11. Details of the additional databases used for model validation.

	Test 1	Test 2	Test 3	Test 4	MCML
Sources	6	6	6	6	10
Codecs	3 (H.264, H.265, VP9)	2 (H.264, H.265)	2 (H.265, VP9)	1 (H.264)	3 (H.264, H.265, VP9)
Resolution	4 (360p, 720p, 1080p, 2160p)	4 (360p, 720p, 1080p, 2160p)	4 (360p, 720p, 1080p, 2160p)	6 (360p, 480p, 720p. 1080p, 1440p, 2160p)	2 (1080p, 2160p)
Framerate	1 (60 fps)	1 (60 fps)	1 (60 fps)	4 (15, 24, 30, 60 fps)	1 (30 fps)
PVSs	180	192	192	192	250
Participants	29	24	26	25	25
Display	65" (Panasonic)	55" (LG OLED)	55" (LG OLED)	55" (LG OLED)	84" (LG 84LM9600)

HRC design philosophy as the P.NATS Phase 2 databases, that is, a similar processing chain and FFmpeg-based encoding algorithms.

As a completely independent database, the MCML databases by Cheon *et al.* [92] is considered for model evaluation. This database consists of 250 samples (240 compressed and 10 reference videos) that are used for evaluation. It should be noted that the samples span only two resolutions namely, FHD and 4K UHD with a framerate of 30fps. This database uses different encoder implemenatations than the ones used for the P.NATS Phase 2 databases. For the case of H.264/AVC, the JM reference software version 18.5 was used, while for H.265, the HM reference software version 10.0 was used. The libvpx software version 1.3.0 was used for VP9 encoding and decoding. More detailed information of these two datasets is provided in Table 11.

TABLE 12. Model validation on additional databases – RMSE figures.

Model	Test 1	Test 2	Test 3	Test 4	MCML
VMAF	0.459	0.448	0.588	0.631	0.340
P.1204.3	0.270	0.222	0.328	0.501	0.378
P.1204.4	0.341	0.334	0.380	0.420	0.322
P.1204.5	0.239	0.458	0.327	0.371	0.395

VII. MODEL APPLICATIONS

There are a variety of application areas for the standardized models presented here. Criteria for classification can be found, for example, in [35]. For the models presented in this paper, the applications can be categorized in terms of a set of non-orthogonal factors, such as (1) the target service to be assessed, e.g. on-demand streaming, live-broadcast, interactive, real-time communication etc., (2) the goal of the assessment, such as encoding ladder derivation or holistic service or network monitoring, (3) the implementation of the assessment approach, considering the locations of the quality model and of the probe for input data acquisition along the distribution chain, (4) whether the assessment takes place during service operation in a non-intrusive, that is passive manner, or off-line, as active (intrusive) measurement, (5) the target quality-criterion being assessed, i.e. in the case of this paper short-term video quality or an integrated quality reflecting the QoE of a longer session.

In principle, all three models can be applied to a variety of cases, with somewhat differing implications for the actual implementation. In the paper, the models are described in an integrative way that comprises the feature extraction and quality estimation parts. Generally, implementaions are conceivable where these parts are distributed and done in different phyiscal or topological places, with the quality-estimation itself and the measurement probe for model input data acquisition implemented in different locations. Accordingly, different ‘‘modes of operation’’ may be distinguished. Similar to [35], a 2-letter code can be used to describe the selected approach, one each for the probe and the model locations. Considering that today’s streaming is typically operated in end-to-end encrypted sessions, the following discussion does not include within-network monitoring (based on encrypted traffic). Hence, for both probe and model, the possible locations are: (H) Head-end server, in case that the service-provider is involved in the measurements or provides quality-related information as side information; (C) client, which may be the case if any of the involved entities is running a measurement based on data obtained at the streaming client; (B) both, where the respective component is distributed across head-end server or client. A few likely combinations of probe and model placement are given in the following. It is noted that further combinations can be conceived.

HH Probe and model are located at the server site. Possible applications here are encoding-ladder derivation or encoder optimization. To this aim, in principle any of the three models can be used. An RR/FR model may have the advantage that it may be more robust against variations of encoder settings. This assumption has to be substantiated by further research, though.

CC Both model input information acquisition and the model are run in the client. This is possible for NR models that have access to all required types of input information. Depending on the level of access enabled to bitstream and/or pixel information, the bitstream-based NR model P.1204.3 or the hybrid NR model P.1204.5 may be used.

BC Some model input information is provided from the head-end, some from the client, and the model is located in the client. An example is the provision of reference-information to an RR or FR model such as P.1204.4 running in the client, via a side channel. Or, short-term quality information for the current segment may be provided from the head-end server to an NR-model located in the client via a side channel, either for short-term quality calculations using P.1204.3 or P.1204.5, or for longer-term session QoE assessment together with a quality-integration component such as P.1203.3 [50].

BH Similar to ‘‘BC’’, where the model input information is partly provided from the client, partly from the head-end. Here, the model is located in the head-end server. Any of the BC use cases are similarly possible here. However, a dedicated example may be quality-monitoring by an over-the-top (OTT) service provider, whereby reference, encoded-bitstream or processed-signal information are acquired at the head-end server site, and client information is used to indicate which segments are being played out during streaming. This case could be realized with any of the three models presented in this paper, possibly in conjunction with a quality integration component such as P.1203.3 [50].

In the following, exemplary possible applications are briefly discussed per model type.

A. APPLICATIONS BITSTREAM-BASED MODEL P.1204.3

The required input information for the bitstream model is readily available at the head-end site. Consequently, it can be used for bitrate ladder derivation (HH) or, in conjunction with additional information from the client side about the played out segments, for more holistic service monitoring (BH). Similarly, the model can be used for real-time quality derivation at the head-end, delivered as side information to the client side for such a more holistic service monitoring (BC). When bitstream information is made available at the client during decoding, also purely client-side monitoring can be realized (CC). Since the bitstream model is computationally much less complex than a decoder, real-time implementations are easily conceivable.

B. APPLICATIONS RR/FR MODEL P.1204.4

The reduced-reference model has three computational parts: extraction of the features of the reference video, extraction of the features of the transmitted video, and the score prediction based on these two sets of features. As full-reference model, it can be used to e.g. evaluate a codec’s performance, or estimate a bitrate ladder (see e.g. HH-mode above). An additional operational setup for a reduced-reference model is to compute the reference features at the head-end and transmit these over a side channel to the client, to predict the scores on the client side (BC). It is also possible to extract the features on the client side, and transmit these back to the head-end, where the reference features are computed and the score prediction takes place (BH). In such a setup, it can serve as a monitoring solution. For the evaluation of a fixed reference, the reference features, which are very small in size compared to the size of the reference video, can be directly installed on the client side (specific implementation of BC). The computational complexity of the reduced reference model is kept low by design. It is much lower than encoding a video at medium settings, thus real-time applications are possible.

C. APPLICATIONS HYBRID NR MODEL P.1204.5

Extracting segment-level parametric information (like video bitrate, codec, resolution and framerate) can be done by parsing the bitstream header in real-time (i.e., as the segments are decoded and played out on the screen). For source-complexity measurement, screen capturing solutions can be used to capture the frames. These capturing solutions can be applied to dump the frames of a played out video segment in CRF encoding format. This way, the hybrid model's source complexity feature can be extracted on a per-segment basis in real-time. These aspect makes the P.1204.5 model suitable for CC type video quality monitoring applications. Note that the per-segment content complexity feature can already be computed offline at the server side and transmitted along a side channel to the client to realize BC type applications. Or, the played out segment information can be relayed back to the server to realize a BH type of applications with the P.1204.5 model.

VIII. DISCUSSION

It was shown that the three models all are of very high prediction performance across a number of databases. The authors acknowledge, that due to the standardization framework that lead to the three models, specific encoder implementations have been dominant during training and standardization-related validation. However, performance was shown to be similarly good also for other test databases, which the models were either not trained on, or which were completely unknown.

In comparison to other typical models such as PSNR, SSIM and VMAF, it was shown that the new standards series can achieve highly competitive performance. Considering the fact that none of these models comprise a dedicated component for the case of frame-rate reduction to lower than 24 fps, model performance was analyzed also for a reduced set of test cases of higher frame rates. Here, too, the three new models underlined their competitive performance.

When inspecting performance on specific databases such as P2SVL10, a somewhat lower prediction performance was found especially for the two NR-models, the bitstream-based and the hybrid. This can be explained with the partly uncommon encoding cases included in these specific tests. For example, with the automatic generation of HRCs, a number of cases with ''ultrafast 2-pass encoding'' have been applied. Since these cases were not present during training, especially the initially submitted models did not cater as well for the resulting degradations as they did for the more common ones. The RR/FR model can better handle this case, since it is based on a comparison of a degraded sequence with the reference. In real-life settings, this encoding approach is likely to never be used, since the two comprised approaches actually contradict each other.

For the performance comparison with the other metrics and models PSNR, SSIM and VMAF, it needs to be mentioned that these do not comprises a specific framerate or ''jerkiness''-related feature. Hence, in Sec. III-B4, the comparison was carried out by considering only the cases for which the HRC framerate was not different from the SRC framerate. While especially VMAF performs better in this case than on the full dataset, overall the three new models still clearly perform better than the state-of-the-art ones.

Hence, for practical usage scenarios with the encoding settings common today, all three models may be applied. Especially due to their high prediction accuracy, the models can be employed also in case of demanding tasks such as bitrate ladder derivation, as well as for a variety of other applications.

IX. CONCLUSION AND OUTLOOK

This paper presents the details of the P.NATS Phase 2 competition that resulted in the P.1204 series of Recommendations for video quality prediction for sequences of up to 4K/UHD resolution. Further, the paper provides and evaluation of the models on open databases, showing the strong performance also in com parison to other models. An overview of the competition encompassing the competition procedure, statistical evaluation of the models and the determination of the winning groups are presented. The descriptions of the three standardized models, namely, bitstream (P.1204.3), pixel-based reduced reference (P.1204.4) and hybrid no-reference (P.1204.5) indicate key algorithmic modelling concepts. The models were analyzed to be the best among the submitted models for the respective model categories in the so-called ITU-T ''PNATS Phase 2'' competition, where 9 proponent companies and research institutions had submitted models. Extensive model training, validation and optimization phases were carried out to yield stable model coefficients.

As shown in the paper, the models demonstrate a neutral prediction scale with regard to the subjective video quality scores used for validation, as well as a symmetrically distributed prediction error. The models were first evaluated on the PNATS Phase 2 databases. Here, it was found that the prediction performance for all three standardized models is significantly superior in comparison to the most widely used open source full-reference metrics PSNR, SSIM and VMAF, for both mobile and TV display type viewing. To ensure the reproducibility of the performance analysis of the three new models and also their applicability to different encoder configurations, the models were evaluated on open datasets. Here, too, a high prediction performance could be shown, also in comparison with the best performing state-of-the-art model VMAF.

The general application scope of the standardized models is that of HAS/DASH-type video streaming video quality and QoE prediction. In particular, the models can be used for short-term video segment quality evaluation of up to 10 s duration, or to determine per-1-second video-quality scores as part of a more holistic QoE evaluation of up to 5 min long streaming sessions, together with an integration module such as ITU-T Rec. P.1203.3. The three new short-term video quality models cover a wide range of settings, for encoding with either H.264, H.265/HEVC or VP9, and a variety of video encoding resolutions from 240p to 4K/UHD-1. Based on the good prediction performance, the paper describes a number of possible application scenarios for the new models.

As future work, the new model standards can be extended for different formats such as HDR, higher resolutions (UHD-2/8K) and framerates (> 60 fps). Moreover, the applicability of the models for different related use cases such as gaming- and 360°-video quality assessment will be investigated. A further logical extension will be to develop a more optimally tailored long-term integration model, beyond the existing ITU-T Rec. P.1203.3, to best combine the short-term video-quality predictions of the new P.1204 standard series with DASH/HAS-specific impairments such as quality switching and stalling.

ACKNOWLEDGMENT

The authors would like to thank the ITU-T team, especially M. Adolph from ITU-T's Telecommunication Standardization Bureau, as well as all colleagues from the Q14 group, especially C. Lee (Yonsei University, South Korea), S. Argyropoulos (StreamOwl, Greece), and L. Amour and S. Souihi (both UPEC, France) from the proponent parties for their continued co-operation over the duration of the competition that resulted in the P.1204 series of Recommendations. They would also like to thank to the Video Quality Experts Group (VQEG), the collaboration partners on this project, for their inputs over the course of the work, especially Margaret Pinson and Kjell Brunnström. Further, they would also like to thank the great assistance in terms of offering source material by Netflix (USA), NTIA (USA), Sky (U.K.), and ARRI (Germany), and in terms of providing exemplary encodings by Bitmovin (Austria).

REFERENCES

[1] C. Chen, S. Inguva, A. Rankin, and A. Kokaram, "A subjective study for the design of multi-resolution abr video streams with the VP9 codec," *Electron. Imag.*, vol. 2016, no. 2, pp. 1–5, 2016.

[2] J. De Cock, Z. Li, M. Manohara, and A. Aaron, "Complexity-based consistent-quality encoding in the cloud," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Sep. 2016, pp. 1484–1488.

[3] C. G. Bampis, Z. Li, I. Katsavounidis, T.-Y. Huang, C. Ekanadham, and A. C. Bovik, "Towards perceptually optimized end-to-end adaptive video streaming," 2018, *arXiv:1808.03898*. [Online]. Available: http://arxiv.org/abs/1808.03898

[4] P. Le Callet, S. Möller, and A. Perkis, Eds., *Qualinet White Paper on Definitions of Quality of Experience*, 1st ed. Lausanne, CH-Switzerland: COST Action IC, 2012.

[5] A. Raake and S. Egger, "Quality and quality of experience," in *Quality of Experience. Advanced Concepts, Applications and Methods*, S. Möller and A. Raake, Eds. Springer, 2014.

[6] "Vocabulary for performance, quality of service and quality of experience," Int. Telecommun. Union, Geneva, Switzerland, Tech. Rep. ITU-T Rec. P.10/G.100, 2017.

[7] C. G. Bampis, Z. Li, A. K. Moorthy, I. Katsavounidis, A. Aaron, and A. C. Bovik, "Study of temporal effects on subjective video quality of experience," *IEEE Trans. Image Process.*, vol. 26, no. 11, pp. 5217–5231, Nov. 2017.

[8] C. G. Bampis, Z. Li, and A. C. Bovik, "Continuous prediction of streaming video QoE using dynamic networks," *IEEE Signal Process. Lett.*, vol. 24, no. 7, pp. 1083–1087, Jul. 2017.

[9] C. G. Bampis, Z. Li, I. Katsavounidis, and A. C. Bovik, "Recurrent and dynamic models for predicting streaming video quality of experience," *IEEE Trans. Image Process.*, vol. 27, no. 7, pp. 3316–3331, Jul. 2018.

[10] N. Barman and M. G. Martini, "QoE modeling for HTTP adaptive video streaming—A survey and open challenges," *IEEE Access*, vol. 7, pp. 30831–30859, 2019.

[11] M.-N. Garcia, F. De Simone, S. Tavakoli, N. Staelens, S. Egger, K. Brunnstrom, and A. Raake, "Quality of experience and HTTP adaptive streaming: A review of subjective studies," in *Proc. 6th Int. Workshop Qual. Multimedia Exper. (QoMEX)*, Sep. 2014, pp. 141–146.

[12] 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. 7th Int. Workshop Qual. Multimedia Exper. (QoMEX)*, May 2015, pp. 1–6.

[13] W. Robitza and A. Raake, "(Re-)actions speak louder than words? A novel test method for tracking user behavior in Web video services," in *Proc. 8th Int. Conf. Qual. Multimedia Exper. (QoMEX)*, Jun. 2016, pp. 1–6.

[14] W. Robitza, M.-N. Garcia, and A. Raake, "Modular HTTP adaptive streaming QoE model–candidate for ITU-T P. 1203 ('P. NATS')," in *Proc. 9th Int. Conf. Qual. Multimedia Exper. (QoMEX)*, May 2017, pp. 1–6.

[15] W. Robitza, A. Ahmad, P. A. Kara, L. Atzori, M. G. Martini, A. Raake, and L. Sun, "Challenges of future multimedia QoE monitoring for Internet service providers," *Multimedia Tools Appl.*, vol. 76, no. 21, pp. 22243–22266, Nov. 2017.

[16] 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 Commun. Surveys & Tuts.*, vol. 17, no. 1, pp. 469–492, 1st Quart., 2014.

[17] S. Tavakoli, K. Brunnström, K. Wang, B. Andrén, M. Shahid, and N. Garcia, "Subjective quality assessment of an adaptive video streaming model," *Proc. SPIE*, vol. 9016, Feb. 2014, Art. no. 90160K.

[18] S. Tavakoli, K. Brunnström, J. Gutiérrez, and N. García, "Quality of experience of adaptive video streaming: Investigation in service parameters and subjective quality assessment methodology," *Signal Process., Image Commun.*, vol. 39, pp. 432–443, Nov. 2015.

[19] S. Tavakoli, S. Egger, M. Seufert, R. Schatz, K. Brunnstrom, and N. Garcia, "Perceptual quality of HTTP adaptive streaming strategies: Cross-experimental analysis of multi-laboratory and crowdsourced subjective studies," *IEEE J. Sel. Areas Commun.*, vol. 34, no. 8, pp. 2141–2153, Aug. 2016.

[20] T. Hoßfeld, M. Seufert, M. Hirth, T. Zinner, P. Tran-Gia, and R. Schatz, "Quantification of YouTube QoE via crowdsourcing," in *Proc. IEEE Int. Symp. Multimedia*, Dec. 2011, pp. 494–499.

[21] T. Hossfeld, C. Keimel, M. Hirth, B. Gardlo, J. Habigt, K. Diepold, and P. Tran-Gia, "Best practices for QoE crowdtesting: QoE assessment with crowdsourcing," *IEEE Trans. Multimedia*, vol. 16, no. 2, pp. 541–558, Feb. 2014.

[22] F. Dobrian, A. Awan, D. Joseph, A. Ganjam, J. Zhan, V. Sekar, I. Stoica, and H. Zhang, "Understanding the impact of video quality on user engagement," *Commun. ACM*, vol. 56, no. 3, pp. 91–99, Mar. 2013.

[23] P. Lebreton, K. Kawashima, K. Yamagishi, and J. Okamoto, "Study on viewing time with regards to quality factors in adaptive bitrate video streaming," in *Proc. IEEE 20th Int. Workshop Multimedia Signal Process. (MMSP)*, Aug. 2018, pp. 1–6.

[24] P. Lebreton and K. Yamagishi, "Study on user quitting rate for adaptive bitrate video streaming," in *Proc. IEEE 21st Int. Workshop Multimedia Signal Process. (MMSP)*, Sep. 2019, pp. 1–6.

[25] S. Takahashi, K. Yamagishi, P. Lebreton, and J. Okamoto, "Impact of quality factors on users' viewing behaviors in adaptive bitrate streaming services," in *Proc. 11th Int. Conf. Qual. Multimedia Exper. (QoMEX)*, Jun. 2019, pp. 1–6.

[26] "The present state of ultra-high definition television," Int. Telecommun. Union, Geneva, Switzerland, Tech. Rep. ITU-R Rec. BT.2246-6, 2017.

[27] "Methodologies for the subjective assessment of the quality of television images," Int. Telecommun. Union, Geneva, Switzerland, Tech. Rep. ITU-R Rec. BT.500-14, 2019.

[28] "Subjective video quality assessment methods for multimedia applications," Int. Telecommun. Union, Geneva, Switzerland, Tech. Rep. ITU-T Rec. P.910, 1999.

[29] K. Berger, Y. Koudota, M. Barkowsky, and P. Le Callet, "Subjective quality assessment comparing UHD and HD resolution in HEVC transmission chains," in *Proc. 7th Int. Workshop Qual. Multimedia Exper. (QoMEX)*, May 2015, pp. 1–6.

[30] G. Van Wallendael, P. Coppens, T. Paridaens, N. Van Kets, W. Van den Broeck, and P. Lambert, "Perceptual quality of 4K-resolution video content compared to HD," in *Proc. 8th Int. Conf. Qual. Multimedia Exper. (QoMEX)*, Jun. 2016, pp. 1–6.

[31] R. Sotelo, J. Joskowicz, M. Anedda, M. Murroni, and D. D. Giusto, "Subjective video quality assessments for 4K UHDTV," in *Proc. IEEE Int. Symp. Broadband Multimedia Syst. Broadcast. (BMSB)*, Jun. 2017, pp. 1–6.

[32] S. Göring, J. Zebelein, S. Wedel, D. Keller, and A. Raake, "Analyze and predict the perceptibility of UHD video contents," *Electron. Imag.*, vol. 2019, no. 12, pp. 1–215, 2019.

[33] P. A. Kara, W. Robitza, N. Pinter, M. G. Martini, A. Raake, and A. Simon, "Comparison of HD and UHD video quality with and without the influence of the labeling effect," *Qual. User Exper.*, vol. 4, no. 1, Dec. 2019, doi: 10.1007/s41233-019-0027-3.

[34] M. H. Pinson, L. Janowski, and Z. Papir, "Video quality assessment: Subjective testing of entertainment scenes," *IEEE Signal Process. Mag.*, vol. 32, no. 1, pp. 101–114, Jan. 2015.

[35] A. Raake, J. Gustafsson, S. Argyropoulos, M.-N. Garcia, D. Lindegren, G. Heikkilä, M. Pettersson, P. List, and B. Feiten, "Ip-based mobile and fixed network audiovisual media services (–current approaches for monitoring)," *IEEE Signal Process. Mag.*, vol. 28, no. 6, pp. 68–79, Oct. 2011.

[36] S. Winkler and P. Mohandas, "The evolution of video quality measurement: From PSNR to hybrid metrics," *IEEE Trans. Broadcast.*, vol. 54, no. 3, pp. 660–668, Sep. 2008.

[37] "Parametric bitstream-based quality assessment of progressive download and adaptive audiovisual streaming services over reliable transport—Video quality estimation module," Int. Telecommun. Union, Geneva, Switzerland, Tech. Rep. ITU-T Rec. P.1203.1, 2019.

[38] A. Raake, M.-N. Garcia, W. Robitza, P. List, S. Göring, and B. Feiten, "Scalable video quality model for ITU-T P.1203 (aka P.NATS) for bitstream-based monitoring of HTTP adaptive streaming," in *Proc. QoMEX*, 2017.

[39] "Parametric non-intrusive assessment of audiovisual media streaming quality—Higher resolution application area," Int. Telecommun. Union, Geneva, Switzerland, Tech. Rep. ITU-T Rec. P.1201.2, 2012.

[40] M. N. Garcia and A. Raake, "Frame-layer packet-based parametric video quality model for encrypted video in IPTV services," in *Proc. 3rd Int. Workshop Qual. Multimedia Exper.*, Sep. 2011, pp. 102–106.

[41] M.-N. Garcia, P. List, S. Argyropoulos, D. Lindegren, M. Pettersson, B. Feiten, J. Gustafsson, and A. Raake, "Parametric model for audiovisual quality assessment in IPTV: ITU-T Rec. P.1201.2," in *Proc. IEEE 15th Int. Workshop Multimedia Signal Process. (MMSP)*, Sep. 2013, p. 1201.

[42] R. R. R. Rao, S. Goring, P. List, W. Robitza, B. Feiten, U. Wustenhagen, and A. Raake, "Bitstream-based model standard for 4K/UHD: ITU-T P.1204.3—Model details, evaluation, analysis and open source implementation," in *Proc. 12th Int. Conf. Qual. Multimedia Exper. (QoMEX)*, May 2020, p. 1204.

[43] Z. Wang, L. Lu, and A. C. Bovik, "Video quality assessment based on structural distortion measurement," *Signal Process., Image Commun.*, vol. 19, no. 2, pp. 121–132, Feb. 2004. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0923596503000766

[44] Z. Li, A. Aaron, I. Katsavounidis, A. Moorthy, and M. Manohara. (2017). *Toward a Practical Perceptual Video Quality Metric*. [Online]. Available: http://techblog.netflix.com/2016/06/towardpractical-perceptual-video.html

[45] "Objective perceptual video quality measurement techniques for digital cable television in the presence of a full reference," Int. Telecommun. Union, Geneva, Switzerland, Tech. Rep. ITU-T Rec. J.144, 2004.

[46] "Objective perceptual multimedia video quality measurement in the presence of a full reference," Int. Telecommun. Union, Geneva, Switzerland, Tech. Rep. ITU-T Rec. J.247, 2008.

[47] "Objective perceptual multimedia video quality measurement of HDTV for digital cable television in the presence of a full reference," Int. Telecommun. Union, Geneva, Switzerland, Tech. Rep. ITU-T Rec. J.341, 2012.

[48] "Parametric bitstream-based quality assessment of progressive download and adaptive audiovisual streaming services over reliable transport," Int. Telecommun. Union, Geneva, Switzerland, Tech. Rep. ITU-T Rec. P.1203, 2016.

[49] "Parametric bitstream-based quality assessment of progressive download and adaptive audiovisual streaming services over reliable transport—Audio quality estimation module," Int. Telecommun. Union, Geneva, Switzerland, Tech. Rep. ITU-T Rec. P.1203.2, 2017.

[50] "Parametric bitstream-based quality assessment of progressive download and adaptive audiovisual streaming services over reliable transport—Quality integration module," Int. Telecommun. Union, Geneva, Switzerland, Tech. Rep. ITU-T Rec. P.1203.3, 2020.

[51] "Video quality assessment of streaming services over reliable transport for resolutions up to 4k," Int. Telecommun. Union, Geneva, Switzerland, Tech. Rep. ITU-T Rec. P.1204, 2020.

[52] "Video quality assessment of streaming services over reliable transport for resolutions up to 4k with access to full bitstream information," Int. Telecommun. Union, Geneva, Switzerland, Tech. Rep. ITU-T Rec. P.1204.3, 2020.

[53] "Video quality assessment of streaming services over reliable transport for resolutions up to 4k with access to full and reduced reference pixel information," Int. Telecommun. Union, Geneva, Switzerland, Tech. Rep. ITU-T Rec. P.1204.4, 2020.

[54] "Video quality assessment of streaming services over reliable transport for resolutions up to 4k with access to transport and received pixel information," Int. Telecommun. Union, Geneva, Switzerland, Tech. Rep. ITU-T Rec. P.1204.5, 2020.

[55] S. Chikkerur, V. Sundaram, M. Reisslein, and L. J. Karam, "Objective video quality assessment methods: A classification, review, and performance comparison," *IEEE Trans. Broadcast.*, vol. 57, no. 2, pp. 165–182, Jun. 2011.

[56] U. Engelke and H.-J. Zepernick, "Perceptual-based quality metrics for image and video services: A survey," in *Proc. Next Gener. Internet Netw.*, May 2007, pp. 190–197.

[57] M.-N. Garcia, S. Argyropoulos, N. Staelens, M. Naccari, M. Rios-Quintero, and A. Raake, "Video streaming," in *Quality of Experience*. Springer, 2014, pp. 277–297.

[58] W. Lin and C.-C. Jay Kuo, "Perceptual visual quality metrics: A survey," *J. Vis. Commun. Image Represent.*, vol. 22, no. 4, pp. 297–312, May 2011.

[59] M. Vranješ, S. Rimac-Drlje, and K. Grgić, "Review of objective video quality metrics and performance comparison using different databases," *Signal Process., Image Commun.*, vol. 28, no. 1, pp. 1–19, Jan. 2013.

[60] J. Joskowicz, R. Sotelo, and J. C. Lopez Arado, "Comparison of parametric models for video quality estimation: Towards a general model," in *Proc. IEEE Int. Symp. Broadband Multimedia Syst. Broadcast.*, Jun. 2012, pp. 1–7.

[61] C. Keimel, J. Habigt, M. Klimpke, and K. Diepold, "Design of no-reference video quality metrics with multiway partial least squares regression," in *Proc. 3rd Int. Workshop Qual. Multimedia Exper.*, Sep. 2011, pp. 49–54.

[62] A. Raake, M.-N. Garcia, S. Moller, J. Berger, F. Kling, P. List, J. Johann, and C. Heidemann, "T-V-model: Parameter-based prediction of IPTV quality," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, Mar. 2008, pp. 1149–1152.

[63] M. N. Garcia, A. Raake, and B. Feiten, ''Parametric audio quality model for IPTV services–ITU-T P.1201.2 audio,'' in *Proc. 5th Int. Workshop Qual. Multimedia Exper. (QoMEX)*, Jul. 2013, pp. 194–199.

[64] M. N. Garcia, R. Schleicher, and A. Raake, ''Impairment-factor-based audiovisual quality model for IPTV: Influence of video resolution, degradation type, and content type,'' *EURASIP J. Image Video Process.*, vol. 2011, no. 1, 2011, Art. no. 629284.

[65] K. Yamagishi and S. Gao, ''Light-weight audiovisual quality assessment of mobile video: ITU-T Rec. P.1201.1,'' in *Proc. IEEE 15th Int. Workshop Multimedia Signal Process. (MMSP)*, Sep. 2013, pp. 464–469.

[66] ''Parametric non-intrusive assessment of audiovisual media streaming quality—Lower resolution application area,'' Int. Telecommun. Union, Geneva, Switzerland, Tech. Rep. ITU-T Rec. P.1201.2, 2012.

[67] T. Hoßfeld, R. Schatz, E. Biersack, and L. Plissonneau, ''Internet video delivery in YouTube: From traffic measurements to quality of experience,'' in *Data Traffic Monitoring and Analysis*. Springer, 2013, pp. 264–301.

[68] T. Hossfeld, D. Strohmeier, A. Raake, and R. Schatz, ''Pippi Longstocking calculus for temporal stimuli pattern on YouTube QoE,'' in *Proc. 5th Workshop Mobile Video MoVid*, 2013, pp. 37–42.

[69] M. N. Garcia, D. Dytko, and A. Raake, ''Quality impact due to initial loading, stalling, and video bitrate in progressive download video services,'' in *Proc. 6th Int. Workshop Qual. Multimedia Exper. (QoMEX)*, Sep. 2014, pp. 129–134.

[70] S. Argyropoulos, A. Raake, M.-N. Garcia, and P. List, ''No-reference video quality assessment for SD and HD H.264/AVC sequences based on continuous estimates of packet loss visibility,'' in *Proc. 3rd Int. Workshop Qual. Multimedia Exper.*, Sep. 2011, pp. 31–36.

[71] N. Staelens, D. Deschrijver, E. Vladislavleva, B. Vermeulen, T. Dhaene, and P. Demeester, ''Constructing a no-reference H.264/AVC bitstream-based video quality metric using genetic programming-based symbolic regression,'' *IEEE Trans. Circuits Syst. Video Technol.*, vol. 23, no. 8, pp. 1322–1333, 2013.

[72] D. C. Mocanu, J. Pokhrel, J. P. Garella, J. Seppänen, E. Liotou, and M. Narwaria, ''No-reference video quality measurement: Added value of machine learning,'' *J. Electron. Imag.*, vol. 24, no. 6, Dec. 2015, Art. no. 061208.

[73] E. Demirbilek and J.-C. Grégoire, ''Machine learning-based parametric audiovisual quality prediction models for real-time communications,'' *ACM Trans. Multimedia Comput., Commun., Appl.*, vol. 13, no. 2, pp. 1–25, May 2017.

[74] W. Robitza, S. Göring, A. Raake, D. Lindegren, G. Heikkilä, J. Gustafsson, P. List, B. Feiten, U. Wüstenhagen, M.-N. Garcia, K. Yamagishi, and S. Broom, ''HTTP adaptive streaming QoE estimation with ITU-T rec. P. 1203: Open databases and software,'' in *Proc. 9th ACM Multimedia Syst. Conf.*, Jun. 2018, pp. 466–471.

[75] R. R. Ramachandra Rao, S. Göring, P. Vogel, N. Pachatz, J. J. Villamar Villarreal, W. Robitza, P. List, B. Feiten, and A. Raake, ''Adaptive video streaming with current codecs and formats: Extensions to parametric video quality model ITU-T P.1203,'' *Electron. Imag.*, vol. 2019, no. 10, pp. 314–321, 2019.

[76] H. T. T. Tran, N. P. Ngoc, A. T. Pham, and T. C. Thang, ''A multi-factor QoE model for adaptive streaming over mobile networks,'' in *Proc. IEEE Globecom Workshops (GC Wkshps)*, Dec. 2016, pp. 1–6.

[77] H. T. T. Tran, T. Vu, N. P. Ngoc, and T. C. Thang, ''A novel quality model for HTTP adaptive streaming,'' in *Proc. IEEE 6th Int. Conf. Commun. Electron. (ICCE)*, Jul. 2016, pp. 423–428.

[78] K. D. Singh, Y. Hadjadj-Aoul, and G. Rubino, ''Quality of experience estimation for adaptive HTTP/TCP video streaming using H.264/AVC,'' in *Proc. IEEE Consum. Commun. Netw. Conf. (CCNC)*, Jan. 2012, pp. 127–131.

[79] J. De Vriendt, D. De Vleeschauwer, and D. Robinson, ''Model for estimating qoe of video delivered using HTTP adaptive streaming,'' in *Proc. IFIP/IEEE Int. Symp. Integr. Netw. Manage. (IM)*, May 2013, pp. 1288–1293.

[80] J. Xue, D.-Q. Zhang, H. Yu, and W. Chen, ''Assessing quality of experience for adaptive HTTP video streaming,'' in *Proc. IEEE Int. Conf. Multimedia Expo Workshops (ICMEW)*, Jul. 2014, pp. 1–6.

[81] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, ''Image quality assessment: From error visibility to structural similarity,'' *IEEE Trans. Image Process.*, vol. 13, no. 4, pp. 600–612, Apr. 2004.

[82] ''Perceptual visual quality measurement techniques for multimedia services over digital cable television networks in the presence of a reduced bandwidth reference,'' Int. Telecommun. Union, Geneva, Switzerland, Tech. Rep. ITU-T Rec. J.246, 2008.

[83] ''Perceptual video quality measurement techniques for digital cable television in the presence of a reduced reference,'' Int. Telecommun. Union, Geneva, Switzerland, Tech. Rep. ITU-T Rec. J.249, 2010.

[84] ''Objective multimedia video quality measurement of hDTV for digital cable television in the presence of a reduced reference signal,'' Int. Telecommun. Union, Geneva, Switzerland, Tech. Rep. ITU-T Rec. J.342, 2011.

[85] R. Soundararajan and A. C. Bovik, ''Video quality assessment by reduced reference spatio-temporal entropic differencing,'' *IEEE Trans. Circuits Syst. Video Technol.*, vol. 23, no. 4, pp. 684–694, Apr. 2013.

[86] C. G. Bampis, P. Gupta, R. Soundararajan, and A. C. Bovik, ''SpEED-QA: Spatial efficient entropic differencing for image and video quality,'' *IEEE Signal Process. Lett.*, vol. 24, no. 9, pp. 1333–1337, Sep. 2017.

[87] A. Mittal, A. K. Moorthy, and A. C. Bovik, ''No-reference image quality assessment in the spatial domain,'' *IEEE Trans. Image Process.*, vol. 21, no. 12, pp. 4695–4708, Dec. 2012.

[88] A. Mittal, R. Soundararajan, and A. C. Bovik, ''Making a 'completely blind' image quality analyzer,''' *IEEE Signal Process. Lett.*, vol. 20, no. 3, pp. 209–212, Nov. 2012.

[89] M. A. Saad, A. C. Bovik, and C. Charrier, ''Blind image quality assessment: A natural scene statistics approach in the DCT domain,'' *IEEE Trans. Image Process.*, vol. 21, no. 8, pp. 3339–3352, Aug. 2012.

[90] A. K. Moorthy and A. C. Bovik, ''A two-step framework for constructing blind image quality indices,'' *IEEE Signal Process. Lett.*, vol. 17, no. 5, pp. 513–516, May 2010.

[91] R. R. Ramachandra Rao, S. Goring, W. Robitza, B. Feiten, and A. Raake, ''AVT-VQDB-UHD-1: A large scale video quality database for UHD-1,'' in *Proc. IEEE Int. Symp. Multimedia (ISM)*, Dec. 2019, pp. 1–8.

[92] M. Cheon and J.-S. Lee, ''Subjective and objective quality assessment of compressed 4K UHD videos for immersive experience,'' *IEEE Trans. Circuits Syst. Video Technol.*, vol. 28, no. 7, pp. 1467–1480, Jul. 2018.

[93] K. Yamagishi, T. Kawano, and T. Hayashi, ''Hybrid Video-Quality-Estimation model for IPTV services,'' in *Proc. GLOBECOM IEEE Global Telecommun. Conf.*, Nov. 2009, pp. 1–5.

[94] Osamu, S. Naito, S. Sakazawa, and A. Koike, ''Objective perceptual video quality measurement method based on hybrid no reference framework,'' in *Proc. 16th IEEE Int. Conf. Image Process. (ICIP)*, Nov. 2009, pp. 2237–2240.

[95] M. C. Q. Farias, M. M. Carvalho, H. T. M. Kussaba, and B. H. A. Noronha, ''A hybrid metric for digital video quality assessment,'' in *Proc. IEEE Int. Symp. Broadband Multimedia Syst. Broadcast. (BMSB)*, Jun. 2011, pp. 1–6.

[96] S. Satti, C. Schmidmer, M. Obermann, R. Bitto, L. Agarwal, and M. Keyhl, ''P.1203 evaluation of real OTT video services,'' in *Proc. 9th Int. Conf. Qual. Multimedia Exper. (QoMEX)*, May 2017, p. 1203.

[97] ''Methods, metrics and procedures for statistical evaluation, qualification and comparison of objective quality prediction models,'' Int. Telecommun. Union, Geneva, Switzerland, Tech. Rep. ITU-T Rec. P.1401, 2014.

[98] J. Neter, M. Kutner, C. Nachtsheim, and W. Wasserman, *Applied Linear Statistical Models*. New York, NY, USA: McGraw-Hill, 1996.

[99] R Core Team. (2019). *A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria. [Online]. Available: https://www.R-project.org/

[100] S. Satti, S. Borer, A. Raake, J. Gustafsson. (Oct. 2019). *AVHD/P.NATS Phase 2 Project*. Video Quality Experts Group. [Online]. Available: https://www.its.bldrdoc.gov/vqeg/vqeg-home.aspx and ftp://vqeg.its.bldrdoc.gov/Documents/VQEG_Shenzhen_Oct19/VQEG_AVHD_2019_October_AVHD-AS_P.NATS_overview.pptx

[101] ''Methodology for the subjective assessment of the quality of television pictures,'' Int. Telecommun. Union, Tech. Rep. ITU-T. RECOMMENDATION ITU-R BT.500-13, 2014.

[102] W. Robitza. (2018). *Subjective Player*. [Online]. Available: https://github.com/slhck/SubjectivePlayer

[103] ''The E-model: A computational model for use in transmission planning,'' Int. Telecommun. Union, Geneva, Switzerland, Tech. Rep. ITU-T Rec. G.107, 2015.

[104] S. Möller, *Assessment and Prediction of Speech Quality in Telecommunications*. Springer, 2000.

ALEXANDER RAAKE (Member, IEEE) studied electrical engineering and physics at Aachen (RWTH) and Paris (ENST/Télécom), with a subsequent research stay at EPFL, Lausanne, Switzerland. He received the Dr.-Ing. degree from the Faculty of Electrical Engineering and Information Technology, Ruhr-Universität Bochum, in January 2005. He was appointed Head of the Audiovisual Technology Group as a Full Professor with TU Ilmenau in 2015. From 2009 to 2015, he held an Assistant Professor and then an Associate Professor position at TU Berlin, heading the Assessment of IP-based Applications group at TU Berlin's An-Institut T-Labs, a joint venture between Deutsche Telekom AG and TU Berlin. From 2005 to 2009, he was a Senior Scientist with the Quality and Usability Lab of T-Labs, TU Berlin. From 2004 to 2005, he was a Postdoctoral Researcher with LIMSI-CNRS, Orsay, France. He has authored a book on the *Speech Quality of VoIP* (Wiley 2006). His research interests include speech, audio and video communication, quality of experience, audiovisual and multimedia services and networks, and human perception and cognition. Since 1999, he has been involved in the standardization activities of the International Telecommunication Union (ITU-T) on performance, quality of service (QoS), and quality of experience (QoE), where he also acts as a Co-Rapporteur for question Q.14/12 on monitoring models for audiovisual services.

SILVIO BORER (Member, IEEE) received the diploma degree in mathematics from the University of Zurich, and the Ph.D. degree from the Laboratory of Computational Neuroscience, EPFL, Lausanne. He is currently the Team Leader Video Analysis with Rohde & Schwarz SwissQual AG. He is active in standardization projects at the ITU-T. He is also a member of the board of VQEG. He is also the Vice Chair of the Audiovisual HD Quality project. His research interests include video quality in communication systems and quality of experience.

SHAHID M. SATTI received the Ph.D. degree in multimedia communication from the Free University of Brussels (VUB), Belgium. He worked as a Postdoctoral Researcher with the Department of Electrical Engineering, VUB, from 2012 to 2013. As a Senior Video Quality Research Engineer at OPTICOM GmbH, he has been active in video quality standardization at ITU-T since 2013. He contributed to ITU-T P.1203.X and P.1204.X series of standards. He has also been the Chair of the Audio-visual HD Quality (AVHD) project at the Video Quality Expert Group (VQEG) since 2018. His specializations involve video encoding, video quality modeling, machine learning, optimizations, and statistical analysis.

JÖRGEN GUSTAFSSON (Member, IEEE) received the M.Sc. degree in computer science from Linköping University. He is currently a part of the Ericsson AI Research leadership, heading research teams in the areas of machine learning and artificial intelligence. He has experience from several Swedish national research projects together with academia and other industry companies. He has been a Co-Rapporteur of ITU-T Study Group 12 Question 14 since many years, and the technical focus is on AI, machine learning, and quality of experience. He is also an inventor of several patents.

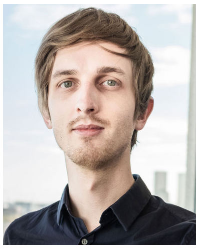
RAKESH RAO RAMACHANDRA RAO received the M.Sc. degree in communications engineering from RWTH Aachen University, in 2017, with a focus on image content analysis and millimeter wave transmission systems. He has been working as an Electrical Engineer with the Audiovisual Technology (AVT), TU Ilmenau, since 2017. Before joining AVT, he worked as an Intern with HEAD acoustics, where he worked on reference-based noise estimation. His research interests include video quality analysis and modeling. His specializations include image content analysis and video quality analysis and modeling.

STEFANO MEDAGLI (Member, IEEE) received the M.Sc. degree in telecommunication engineering from the University of Napoli Federico II, in 2015. He is currently working as a Research Engineer with Rohde & Schwarz SwissQual AG. Previously, he worked as a Synthetic Aperture Radar Image Processing Engineer with CerICT, and as a Radar Signal Processing and an Electromagnetic Research Engineer in a joint project involving TU Delft and Thales air System SAS. His current research interests include video quality in communication systems and data analysis. His specializations are electromagnetics, digital signal processing, and statistics and communication systems.

PETER LIST received the degree in physics and the Ph.D. degree in applied physics from the University of Frankfurt/Germany, in 1985 and 1989, respectively. Since 1990, he has been in various positions in Research and Development of Deutsche Telekom AG. For many years, he attended the standardization bodies for video compression in ITU and ISO/MPEG. He is currently with Deutsche Telekom AG.

STEVE GÖRING received the B.Sc. and M.Sc. degrees in computer science from TU Ilmenau, in 2008 and 2013, respectively. He is currently working as a Computer Scientist with the Audiovisual Technology Group, TU Ilmenau. Before, he started 2016 at the Audiovisual Technology Group, he was working with the Bauhaus University Weimar in Big Data Analytics group. His research interests include Weimar was improving search engines (using axiomatic re- ranking approaches), argumentation analysis, and analyzing large unstructured datasets using machine learning approaches. His current research interests include data analysis problems for video quality models and video streams. His specializations are data analytics/machine learning, video quality, and distributed communication/information systems.

DAVID LINDERO (Member, IEEE) received the M.Eng. degree from the Luleå University of Technology, in 2007. He immediately started working at Ericsson Research after this, first focusing on speech quality assessment and statistical modeling. Later projects have been in video quality, VR, and deep learning areas. He has been active in ITU-T SG12 standardization work since 2009, developing video quality estimation models and evaluation procedures.



WERNER ROBITZA (Member, IEEE) received the diploma degree in computer science from the University of Vienna, in 2013. He worked as a Research Assistant with the University of Vienna in 2009. He was a Staff Researcher with TU Ilmenau, Germany, in 2018. From 2014 to 2018, he was a Researcher with the Telekom Innovation Laboratories of Deutsche Telekom AG and TU Berlin. He has been actively involved in ITU-T's video quality standardization work since 2014. He is currently a Ph.D. Researcher Associate with the Ilmenau University of Technology. He is also the CEO and the Co-Founder of AVEQ GmbH, a company based in Vienna, Austria, which offers video streaming and web quality monitoring solutions. His research interests include quality of experience for multimedia applications and user behavior aspects for Web TV services, with a focus on subjective video quality testing and video quality measurement tools.

GUNNAR HEIKKILÄ (Member, IEEE) received the M.Sc. degree in computer science from the Luleå University of Technology, Sweden. He is currently a Senior Specialist in performance measurements with Ericsson Research. Since 1996, he has been focused on user experience quality assessment and measurements, including standardization in ITU-T, 3GPP, ETSI, and CTA. He joined Ericsson in 1987 and has previously worked with control system software for military defense radar systems, and with software design for synchronous digital hierarchy optical fiber transmission systems.

SIMON BROOM received the M.Eng. degree in electronic systems engineering from the University of York, in 1995. After graduating, he joined BT to work on developing objective models to predict voice transmission quality and was first involved in ITU-T SG12 standards, actively contributing to various speech and video quality related standards. In 2001, he joined Psytechnics, a BT spin-off company, to develop VoIP quality assessment software. Psytechnics were subsequently acquired by NETSCOUT in 2011, where he continues his interest and involvement in ITU-T SG12 video quality assessment projects and manages NETSCOUT's subjective testing facility in Ipswich, U.K.

CHRISTIAN SCHMIDMER studied electronic engineering at the University of Erlangen. After achieving his M.S. degree (Diploma), he spent five years as a Scientist at the Audio Department of the famous Fraunhofer Institute for Integrated Circuits in Erlangen (the home of mp3), mostly dedicated to the research of psychoacoustics and the development of perceptual measurement tools as well as audio codecs, contributing to the development of mp3. In 1997, he joined OPTICOM as the CTO and a Co-Owner. OPTICOM's core business is the development and management of IPR for voice, audio, and video quality measurement. He is active in standardisation bodies, such as ITU, VQEG, and ETSI. He is the author of many scientific publications and frequently presented papers at conferences and workshops. He is one of the main developers behind the recommendations ITU-R BS.1387 / PEAQ (Perceptual Evaluation of Audio Quality), ITU-T P.563 / 3SQM (no-reference voice quality assessment), and ITU-T P.863 / POLQA (full reference voice quality assessment).

BERNHARD FEITEN received the Dr.-Ing. degree in electronic engineering from Technische Universität Berlin in the field of psychoacoustics and audio bit rate reduction. He worked as an Assistant Professor with Technische Universität Berlin in the field of communication science, digital signal processing, and computer music at the Elektronisches Studio. Since 1996, he has been with Deutsche Telekom AG, now Technology & Innovation, working as a Senior Expert and a Project Manager for innovative multimedia services, quality of experience, and network analytics. His research and development activities comprise audio and video coding quality, broadcasting applications, high quality Internet media distribution and streaming, and QoE monitoring and optimization.

ULF WÜSTENHAGEN studied technical acoustics at the Technical University of Dresden and became Graduate Engineer with room acoustic and subjective acoustics. Afterwards, he was working in several fields especially in subjective assessment for audio and video services. He was involved in development of ITU standards for measurement and evaluation of Quality of Experience. Later, he was involved in the planning and set-up of Deutsche Telekom's IPTV service. He has been with Deutsche Telekom AG since 1990. He is currently dealing with quality evaluation for audio and video applications in fixed and mobile IP networks.

THOMAS WITTMANN received the degree of Dipl.-Ing. (Univ.) in electrical engineering from the Friedrich-Alexander-Universität Erlangen-Nürnberg. He has been working with Vierling Electronics as a Software Developer from 1997 to 2005. One of his projects was a telecommunication measurement system that estimates the speech quality with the algorithms TOSQA and PESQ (ITU-T P.862). Away from the test and measurement industry, he worked with BHS Corrugated Maschinen- und Anlagenbau as a Software Developer from 2005 to 2009. BHS manufactures and installs machinery for the production of corrugated cardboard. Since 2009, he has been working with OPTICOM GmbH as a Software Developer. He contributed to OPTICOM's products PESQ (ITU-T P.862) and POLQA (ITU-T P.863) and implemented OptiPlay. OptiPlay is a video player, which can play back uncompressed UHD video up to a framerate of 60 fps on a specific video card (Blackmagic Decklink) and is used for subjective video tests. His specializations include digital signal processing, video encoding, and cryptography.

MATTHIAS OBERMANN received the M.S. degree (Diploma) in electrical engineering from the University of Erlangen. He joined OPTICOM in 2009 to work on various topics of objective quality of experience assessment, including voice and video quality. He was involved in the standardization at ITU-T (P.863) and ITU-R (J.343). He is currently a Project Leader with OPTICOM GmbH for bit stream based quality assessment products. His research interests include deep packet inspection, video streaming, and encoding.

ROLAND BITTO received the Dipl.-Ing. degree in electrical engineering from the Friedrich-Alexander University of Erlangen-Nürnberg, Germany, in 1992. After three years in industry, he joined the Fraunhofer Institute for Integrated Circuits as a Research Scientist. His work there was dedicated to the research of psychoacoustics, the development of perceptual measurement tools and audio codecs, also contributing to mp3 and aac. In 2000, he joined OPTICOM, where he is focusing on development on voice-, audio-, and video quality measurements metrics. He is one of the main contributors to the recommendations ITU-R BS.1378 PEAQ, ITU-T P.563 3SQM and ITU-T a247, and ITU-T343 PEVQ.