AVHD-AS/P.NATS Phase 2 Project

Standardization and Model Performance Report

Dated: 8 December, 2020

	P.NATS Phase 2 Validation Report	
Purpose:	Information	
Contact:	Shahid Mahmood Satti OPTICOM GmbH Germany	Tel: +49 9131 53020-0 Fax: +49 9131 53020-20 Email: ss@opticom.de
Contact:	Silvio Borer Rohde and Schwarz SwissQual AG Switzerland	Tel: +41 32 686 65 65 E-mail: Silvio.Borer@rohde-schwarz.com
Contact:	Jörgen Gustafsson Ericsson Research, L.M. Ericsson Sweden	Tel: +46 730 783282 Email: jorgen.gustafsson@ericsson.com
Contact:	Alexander Raake Technische Universität Ilmenau Germany	Tel: +49 3677 69-1468 E-mail: alexander.raake@tu-ilmenau.de

Keywords: P.NATS Phase 2, video quality models, competition, validation

Abstract: The present document provides the P.NATS Phase 2 validation report and summary of the standardization success of the new Recommendation Series P.1204. This new standard series has jointly been developed in the first cross-model-type competition of video quality models, in a fruitful collaboration between VQEG and ITU-T SG12, Question14.

Based on the validation results, five different models were planned to be standardized: Bitstream Mode 0, Bitstream Mode 1, Bitstream Mode 3, Hybrid no-reference Mode 0, Pixel-based Reduced / Full Reference. The document was prepared by the Q14/12 Co-Rapporteurs in collaboration with the VQEG AVHD Co-Chairs. At the ITU-T SG12 meeting held from 26th Nov. to 5th Dec. 2019 in Geneva, three out of the five finally targeted models were standardized: Bitstream Mode 3, Hybrid no-reference Mode 0, Pixel-based Reduced / Full Reference. Due to the fact that a merging of the winning candidates for the two further models, Bitstream Mode 0 and Mode 1, could not be achieved according to the agreed-upon requirements, no standards have resulted for these cases.

Summary

Study Group 12 is proud to inform that the collaboration with VQEG on "P.NATS Phase 2" has lead to the new ITU-T P.1204 series of Recommendations, "Video quality assessment of streaming services over reliable transport for resolutions up to 4K", which has recently been consented at the Nov./Dec. SG12 Meeting in Geneva. This work represents a first direct, successful collaboration between Question Q14 of ITU-T SG12 and VQEG. A detailed overview of P.1204 multi model series and performance of the models on short databases developed during the course of the project as well on open databases can be found in [IEEE P1204]. This report details the performance of winning models for different model categories for both short and long databases.

This Recommendation series describes a set of objective video quality models. These can be used standalone for assessing video quality for 5-10 sec long video sequences, providing a 5-point ACR-type Mean Opinion Score (MOS) output. In addition, they deliver per-1-second MOS-scores that together with audio information and stalling / initial loading data can be used to form a complete model to predict the impact of audio and video media encodings and observed IP network impairments on quality experienced by the end-user in multimedia streaming applications. The addressed streaming techniques comprise progressive download as well as adaptive streaming, for both mobile and fixed network applications.

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

The consent of the P.1204 model standards marks the first time that video-quality models of all relevant types have been developed and validated within the same standardization campaign. The respective “P.NATS Phase 2” model competition used a total of 13 video-quality test databases for training, and another 13 video-quality test databases for validation. With this comparatively high number of data (more than 5000 video sequences), the resulting standards deliver class-leading video-quality prediction performance.

The P.1204 recommendation series currently consists of the following sub-recommendations:

- **ITU-T P.1204**: “Video quality assessment of streaming services over reliable transport for resolutions up to 4K”. Introductory recommendation for the whole P.1204 series. *Consented at the December Study Group 12 meeting.*
- **ITU-T P.1204.3**: “Video quality assessment of streaming services over reliable transport for resolutions up to 4K with access to full bitstream information”. *Consented at the December 2019 Study Group 12 meeting.*
- **ITU-T P.1204.4**: “Video quality assessment of streaming services over reliable transport for resolutions up to 4K with access to full and reduced reference pixel information”. *Consented at the December 2019 Study Group 12 meeting.*
- **ITU-T P.1204.5**: “Video quality assessment of streaming services over reliable transport for resolutions up to 4K with access to transport and received pixel information”. *Consented at the December 2019 Study Group 12 meeting.*

The recently consented recommendations are available at the ITU-T Recommendation page.

Nine different proponents took part in the AVHD/P.NATS2 competition. The winning AVHD/P.NATS2 video models have been determined according to the agreed statistical evaluation procedure detailed in [IEEE P1204]. As mentioned in [IEEE P1204] a training database is given a weight of 0.1 while a validation database is given a weight of 0.9 for computing the average values. A weighted average RMSE together with the significance criteria detailed in[IEEE P1204] was employed to determine the winning models.

The winning models which were standardized are as follows.

1. Bitstream Mode 3:
    a. Deutsche Telekom with TU Ilmenau
2. Reduced Reference Model:
    a. Rohde & Schwarz SwissQual AG
3. Hybrid No Reference Mode 0:
    a. OPTICOM GmbH

Note 1: Average and per-database RMSE and Pearson Correlation performance of all standardized models is detailed in this report. In case a party did not agreed to reporting model performance on their database, the respective row is marked with X.

Model Categories:

- BSM0: Bitstream Mode 0 Model
- BSM1: Bitstream Mode 1 Model
- BSM3: Bitstream Mode 3 Model
- PXNR: Pixel-based No Reference Model
- PXRR: Pixel-based Reduced Reference Model
- PXFR: Pixel-based Full Reference Model
- HYN0: Hybrid No Reference Mode 0 Model
- HYN1: Hybrid No Reference Mode 1 Model
- HYR0: Hybrid Reduced Reference Mode 0 Model
- HYF0: Hybrid Full Reference Mode 0 Model

Not all proponents submitted all model types. “X” means a model type submitted by a certain proponent. ANNY denotes companies who are not part of the winning group and the names of the companies cannot be disclosed for legal reasons.

Model	**BSM0**	**BSM1**	**BSM3**	**HYF0**	**HYN0**	**HYN1**	**HYN3**	**HYR0**	**PXFR**	**PXNR**	**PXRR**
ANNY	X	X		X	X	X			X	X	
ANNY	X										
ANNY	X										
OPTI	X	X		X	X	X			X		
RSSQ	X	X			X			X	X		X
DTTU	X	X	X	X	X				X	X	
ANNY	X										
ANNY	X			X				X	X		X
ANNY									X		

1. Introduction

The AVHD/P.NATS Phase 2 work item has reached the stage where the winning models have been identified in the validation phase of the work item. There are 11 model categories (3 bitstream, 3 pixel-based and 4 hybrid). Both pixel-based and hybrid models are submitted with the long term integration function, while bitstream-based models are submitted without the long term integration. For bitstream-based models, the plan is to consider integrating the new Pv modules with the integration approach from P.1203, i.e. P.1203.3 at a later stage.

At the Study Group 12 WP3 and Q14 meeting 02-04 September 2019 in Stockholm, the verification of the submitted models was carried out according to the P.NATS requirement specification. This included

- Installation of the models submitted to International Telecommunication Union (ITU) Study Group (SG) 12 TSB on the computers provided by Ericsson at the meeting location in Stockholm.
- Agreed bug fixes of the submitted models were implemented for the models running on these computers, too. This was overseen by the other proponents to ensure that only agreed bug fixes were made.
- Verification that the output scores provided by all proponents are identical to what is produced by the models submitted to ITU TSB. Around 10% of the scores, randomly selected, were compared. See the model verification section for details.
- The output scores from the models submitted prior to the disclosure of subjective scores for validation databases were compared to the scores from validation and training databases. The results were analysed with a script that implements the Statistical Validation Document to identify which models showed the best performance, that is, are the models that produces scores closest to the ground-truth (GT) data from the subjective tests.
- Further, it was validated whether all model types will be standardized, based on the hierarchical process outlined in the Requirement Specification of the AVHD/P.NATS Phase 2 competition and **[IEEE P1204]**.
- In addition, the winning groups for P.NATS Phase 2 were determined based on a first analysis shortly after the meeting.
- All the submitted scores (with and without bugfixes) have been compiled into a csv file. To avoid duplication of submitted scores, a clean csv has been created which contains the scores after applying all the bugfixes and removing the duplicates.
- There are a few cases for which the score verification is pending, see the section “model verification” for details. It is foreseen that these scores will be verified before the scores can be included in this report.

2. Overview of Databases

Training Databases:

DB ID	**Display**	**Display Res.**	**Display Size (inches)**	**Viewing Dist.**	**Video Length**	**Nr. of PVSs**	**Type**	**Proponent**
P2STR01	Mobile	2560x1440	5.1	5-7H	8sec	203	Short	ANNY
P2STR02	Mobile	2560x1440	5.1	5-7H	8sec	199	Short	RSSQ
P2STR03	Mobile	2560x1440	5.1	5-7H	8sec	200	Short	ANNY
P2STR04	PC Monitor	3840x2160	32	1.5H	8sec	199	Short	ANNY

DB ID	Display	Display Res.	Display Size (inches)	Viewing Dist.	Video Length	Nr. of PVSs	Type	Proponent
P2STR05	PC Monitor	3840x2160	31.5	1.5H	8sec	187	Short	ANNY
P2STR06	Mobile	2560x1440	5.1	5-7H	8sec	187	Short	ANNY
P2STR08	TV	3840x2160	65	1.5H	8sec	179	Short	OPTI
P2STR09	PC Monitor	3840x2160	55	1.5H	8sec	187	Short	DTTU
P2STR10	PC Monitor	3840x2160	55	1.5H	8sec	187	Short	DTTU
P2STR11	TV	3840x2160	75	1.5H	8sec	187	Short	ANNY
P2STR12	PC Monitor	3840x2160	31.5	1.5H	8sec	183	Short	RSSQ
P2STR13	TV	3840x2160	75	1.5H	8sec	187	Short	ANNY
P2STR14	TV	3840x2160	55	1.5H	8sec	179	Short	ANNY
P2LTR15	Mobile	2560x1440	5.1	5-7H	60sec	60	Long	OPTI
P2LTR17	Mobile	2560x1440	5.1	5-7H	180sec	20	Long	ANNY

Validation Databases:

DB ID	Display	Display Res	Display Size	Viewing Dist.	Nr. of HRCs	Nr. of PVSs	Type	Proponent
P2SVL01	TV	3840x2160	55	1.5H	8sec	185	Short	DTTU
P2SVL02	Mobile	2560x1440	5.1	5-7H	8sec	186	Short	ANNY
P2SVL03	Mobile	2560x1440	5.1	5-7H	8sec	186	Short	ANNY
P2SVL04	Mobile	2560x1440	5.1	5-7H	8sec	195	Short	RSSQ
P2SVL05	TV	3840x2160	65	1.5H	8sec	194	Short	OPTI
P2SVL06	TV	3840x2160	75	1.5H	8sec	191	Short	ANNY
P2SVL07	TV	3840x2160	65	1.5H	8sec	188	Short	OPTI
P2SVL08	PC-Monitor	3840x2160	37	1.5H	8sec	195	Short	ANNY
P2SVL09	TV	3840x2160	55	1.5H	8sec	191	Short	DTTU
P2SVL10	TV	3840x2160	55	1.5H	8sec	195	Short	ANNY
P2SVL11	TV	3840x2160	75	1.5H	8sec	195	Short	ANNY
P2SVL12	Tablet	2560x1440	10	5-7H	8sec	195	Short	ANNY
P2SVL13	TV	3840x2160	55	1.5H	8sec	187	Short	ANNY
P2LVL15	PC-Monitor	3840x2160	37	1.5H	60sec	59	Long	ANNY
P2LVL18	Mobile	2560x1440	5.1	5-7H	120sec	30	Long	RSSQ
P2LVL19	TV	3840x2160	55	1.5H	120sec	30	Long	DTTU

DB ID	Display	Display Res	Display Size	Viewing Dist.	Nr. of HRCs	Nr. of PVSs	Type	Proponent
P2LVL23	Mobile	2560x1440	5.1	5-7H	300sec	14	Long	ANNY

Details about Subjective Testing of Databases

Correlation values in the table below are mean, max and average of per-subject linear correlation with the MOS.

Training Databases

Database	Nr. of Subjects	Min. Sub. Correl.	Max. Sub. Correl.	Avg. Sub. Correl.	Min. Conf. Interval	Max. Conf. Interval	Avg, Conf. Interval
P2STR01	26	0.71	0.91	0.82	0.16	0.44	0.29
P2STR02	24	0.79	0.93	0.87	0.00	0.41	0.27
P2STR03	30	0.76	0.93	0.87	0.12	0.36	0.23
P2STR04	26	0.82	0.95	0.91	0.00	0.40	0.24
P2STR05	26	0.78	0.91	0.84	0.00	0.41	0.27
P2STR06	24	0.76	0.88	0.82	0.08	0.38	0.25
P2STR08	24	0.80	0.95	0.89	0.08	0.39	0.26
P2STR09	25	0.81	0.90	0.86	0.00	0.41	0.25
P2STR10	34	0.39	0.93	0.86	0.00	0.32	0.21
P2STR11	24	0.81	0.94	0.89	0.11	0.37	0.25
P2STR12	24	0.76	0.91	0.85	0.11	0.41	0.28
P2STR13	25	0.77	0.93	0.87	0.00	0.38	0.25
P2STR14	24	0.76	0.88	0.84	0.11	0.54	0.24
P2LTR15	22	0.66	0.91	0.78	0.16	0.46	0.33
P2LTR17	27	0.75	0.97	0.87	0.12	0.38	0.25

Validation Databases

Database	Nr. of Subjects	Min. Sub. Correl.	Max. Sub. Correl.	Avg. Sub. Correl.	Min. Conf. Interval	Max. Conf. Interval	Avg, Conf. Interval
P2SVL01	30	0.77	0.90	0.82	0.11	0.40	0.25
P2SVL02	24	0.75	0.91	0.82	0.15	0.39	0.26
P2SVL03	21	0.70	0.89	0.82	0.13	0.44	0.30
P2SVL04	24	0.76	0.92	0.88	0.14	0.43	0.28
P2SVL05	25	0.82	0.92	0.87	0.08	0.43	0.28
P2SVL06	24	0.80	0.94	0.89	0.00	0.39	0.26

Database	Nr. of Subjects	Min. Sub. Correl.	Max. Sub. Correl.	Avg. Sub. Correl.	Min. Conf. Interval	Max. Conf. Interval	Avg, Conf. Interval
P2SVL07	25	0.80	0.92	0.86	0.11	0.43	0.26
P2SVL08	27	0.72	0.90	0.82	0.00	0.43	0.29
P2SVL09	28	0.75	0.90	0.81	0.08	0.39	0.28
P2SVL10	26	0.81	0.92	0.86	0.08	0.35	0.21
P2SVL11	24	0.77	0.91	0.87	0.00	0.39	0.27
P2SVL12	24	0.79	0.93	0.84	0.08	0.30	0.20
P2SVL13	26	0.80	0.88	0.84	0.08	0.38	0.25
P2LVL15	29	0.71	0.93	0.82	0.13	0.44	0.30
P2LVL18	24	0.72	0.94	0.88	0.00	0.47	0.25
P2LVL19	31	0.71	0.95	0.85	0.11	0.31	0.25
P2LVL23	26	0.77	0.98	0.87	0.17	0.35	0.28

3. Model Verification

The model submission and verification was divided into several steps, each to ensure that the scores submitted by each proponent was actually produced by the submitted model.

An encrypted virtual machine (VM) image together with produced scores for the training databases and SHA256 checksum of the VM-image were submitted by each proponent in January 2019.

Model scores for the validation databases were shared before the interim meeting in Stockholm in August. This was also before any subjective scores for the validation databases were shared.

During the interim meeting in Stockholm, the host Ericsson provided computational resources and storage to facilitate running all submitted models to verify the submitted scores. The model inputs, video files and metadata, was either downloaded from a proponent file storage location or re-created in this computational environment. Using md5 checksums, all files were checked to match what had been used when creating the videos used in subjective testing.

Each proponent downloaded its VM-image from the ITU-T ftp and extracted it under supervision of another proponent. This was done in "proponent-pairs" so that the proponents could watch each other.

Any bug fixing allowed was also done at this stage, under the same supervision.

All models were run on a subset of all data, pseudo-randomly selected with at least one sample per database (for short databases). Some of the pixel-based models needed a lot of time to finish so it was not feasible to run them on all data. The parametric models (bitstream mode 0 and 1) were however fast enough to run on all samples in all short databases.

A verification script compared the freshly produced scores with the already submitted scores, treating everything that differed on a smaller scale than three decimals as equal. Some models had a few scores that didn't match and if these models were in the winning group, it was decided to do a later verification of these scores in the following ITU-T SG12 meeting.

4. Model(s) Performance

In this section, RMSE of all submitted models for each training and validation database is reported. The number highlighted in bold in each row indicates the best model for that database. If in a row no number is bold, an anonymous model for which the numbers are not specified in that table gives better RMSE for that database. Note that the numbers are reported after a final per-database mapping between the model output and the subjective scores of a database. This linear mapping is used to account for scale and bias variations between different databases.

Baseline Model Performance

As described in [IEEE P1204], a simple baseline model (log of bitrate) was trained on the training and validated on the validation data.

The baseline model was trained by RSSQ and OPTI and both trainings resulted in slightly different model performance of the baseline model. The per-database RMSE values are reported below:

Training Databases

	RSSQ Baseline	**OPTI Baseline**
P2STR01	0.645348	0.6613
P2STR02	0.679607	0.6863
P2STR03	0.656022	0.6685
P2STR04	0.53546	0.5283
P2STR05	0.605308	0.6009
P2STR06	0.534559	0.5392
P2STR08	0.645324	0.6419
P2STR09	0.563504	0.5664
P2STR10	0.504882	0.5048
P2STR11	0.667147	0.6659
P2STR12	0.602265	0.6039
P2STR13	0.640846	0.6339
P2STR14	0.562899	0.5631

Validation Databases

	RSSQ Baseline	OPTI Baseline
P2SVL01	0.570975	0.5467
P2SVL02	0.626397	0.6013
P2SVL03	0.61609	0.6076
P2SVL04	0.679009	0.6683

	RSSQ Baseline	OPTI Baseline
P2SVL05	0.659014	0.62
P2SVL06	0.690877	0.682
P2SVL07	0.586261	0.5306
P2SVL08	0.611092	0.6117
P2SVL09	0.579497	0.5756
P2SVL10	0.675522	0.6786
P2SVL11	0.565662	0.5634
P2SVL12	0.589819	0.5976
P2SVL13	0.671878	0.5783

Average RMSE Baseline RSSQ: 0.610
Average RMSE Baseline OPTI: 0.607

The model with a lower RMSE value will be taken as the baseline for model comparison.

In the following, the performance results for the different candidate models are summarized, per model type (and mode for the bitstream models).

4.1 Bitstream Models

Bitstream Mode 0 and Bitstream Mode 1

Both mode 0 and mode 1 resulted in multiple winning models. For the standardization of mode 0 and mode 1 models, it was required to merge the winning models to a single merged model. However, due to various commercial and legal disagreements between the involved parties the merging task could not be initiated at the time of writing this report. In addition, a consensus by which the performance of the winning models can be described in this report was also not reached. For that reason model performance of mode 0 and mode 1 will not be described in this section.

Bitstream Mode 3

For bitstream mode 3 (BSM3) only one model was submitted. It performed significantly better compared to the winning BSM1 models, hence it has been standardized.

Training Databases

	DTTU BSM3
P2STR01	0.2915
P2STR02	0.3356
P2STR03	0.3059
P2STR04	0.2718
P2STR05	0.375
P2STR06	0.3321
P2STR08	0.311
P2STR09	0.2993
P2STR10	0.2939
P2STR11	0.2838
P2STR12	0.3235
P2STR13	0.294
P2STR14	0.3674

Validation Databases

	DTTU BSM3
P2SVL01	0.382
P2SVL02	0.4165
P2SVL03	0.4055
P2SVL04	0.4258
P2SVL05	0.4828
P2SVL06	0.4241
P2SVL07	0.3776
P2SVL08	0.4504
P2SVL09	0.4072
P2SVL10	0.5475
P2SVL11	0.401
P2SVL12	0.4233
P2SVL13	0.4334

4.3 Pixel-based Models

No-Reference Model

For no-reference models (PXNR), 2 models were submitted. No result for PXNR models are specified here. Both models performed worse than the baseline, hence PXNR has not been standardized. The description and performance numbers for the best performing PXNR model in the competition can be found in [PSTR-PXNR-ITUT-Technical-Report].

Reduced-Reference Model

For reduced reference models (PXRR) 2 models were submitted. Results of the winning model are tabulated below.

Training Databases

	RSSQ PXRR
P2STR01	0.3684
P2STR02	0.3596
P2STR03	0.37
P2STR04	0.347
P2STR05	0.4441
P2STR06	0.428
P2STR08	0.4543
P2STR09	0.3674
P2STR10	0.3605
P2STR11	0.4082
P2STR12	0.4144
P2STR13	0.3862
P2STR14	0.4269
P2LTR15	0.3626
P2LTR17	0.5164

Validation Databases

	RSSQ PXRR
P2SVL01	0.4685
P2SVL02	0.5078
P2SVL03	0.4307
P2SVL04	0.4179

P2SVL05	0.4546
P2SVL06	0.4525
P2SVL07	0.4601
P2SVL08	0.4265
P2SVL09	0.4196
P2SVL10	0.4968
P2SVL11	0.4484
P2SVL12	0.4448
P2SVL13	0.4004
P2LVL15	0.4447
P2LVL18	0.4323
P2LVL19	0.3374
P2LVL23	0.7143

Full-Reference Model

For full-reference models (PXFR) 6 models were submitted. For PXFR models the overall performance is not significantly better than the winning PXRR model Hence PXFR was not standardized.

4.4 Hybrid Models

Hybrid No-Reference Mode 0

For hybrid no-reference mode 0 models (HYN0) four models were submitted. Results for the winning model are tabulated below.

Training Databases

	OPTI HYN0
P2STR01	0.4118
P2STR02	0.5456
P2STR03	0.4805
P2STR04	0.3637
P2STR05	0.5028
P2STR06	0.4355
P2STR08	0.4234
P2STR09	0.3985

P2STR10	0.3272
P2STR11	0.4432
P2STR12	0.4662
P2STR13	0.3634
P2STR14	0.4094
P2LTR15	0.3184
P2LTR17	0.3654

Validation Databases

	OPTI HYN0
P2SVL01	0.3761
P2SVL02	0.4712
P2SVL03	0.436
P2SVL04	0.5036
P2SVL05	0.4398
P2SVL06	0.4586
P2SVL07	0.4187
P2SVL08	0.4057
P2SVL09	0.4647
P2SVL10	0.5586
P2SVL11	0.426
P2SVL12	0.4957
P2SVL13	0.4148
P2LVL15	0.5702
P2LVL18	0.5431
P2LVL19	0.4758
P2LVL23	0.6705

Hybrid No-Reference Mode 1

For hybrid no-reference mode 1 models (HYN1) two models were submitted. For HYN1 models the overall performance is not significantly better than the winning HYN0 model. Hence, HYN1 was not standardized.

Hybrid Reduced-Reference Mode 0

For hybrid reduced-reference mode 0 models (HYR0) two models were submitted. For HYR0 models the overall performance is not significantly better than the winning PXRR model. Hence, HYR0 models was not standardized.

Hybrid Full-Reference Mode 0

For hybrid full-reference mode 0 models (HYF0) four models were submitted. For HYF0 models the overall performance is not significantly better than the winning PXFR model. Hence, HYF0 model was not standardized.

5. Winning Groups

The tables below tabulate the average RMSE numbers of validated models.

Validation based only on Short Databases

Model Type	**SignificanceThreshold RMSE**	**Winning Model(s)**
Baseline		0.607
BSM3	0.434	0.421 (DTTU)
PXNR		No winning model
PXRR	0.458	0.444 (RSSQ)
PXFR		No winning model
HYN0	0.466	0.452 (OPTI)
HYN1		No winning model
HYR0		No winning model
HYF0		No winning model

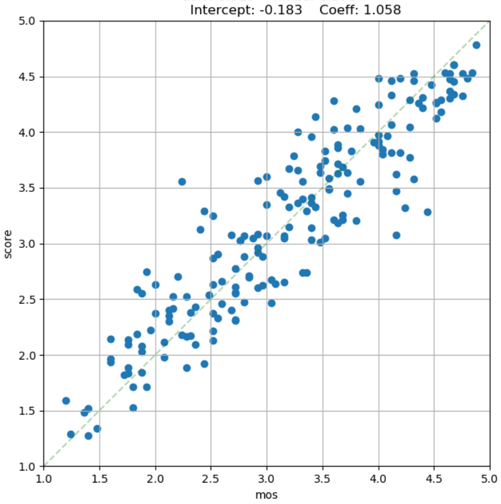
Validation based on Short and Long Databases

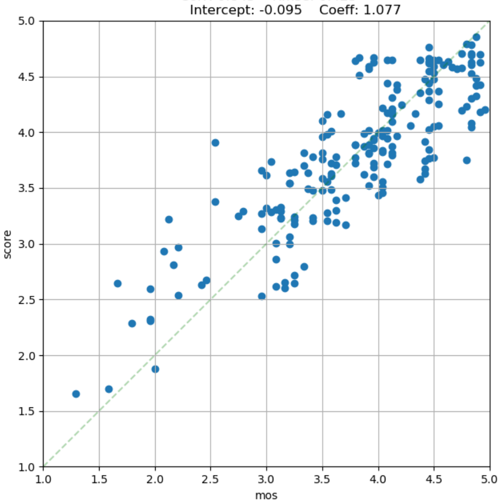
Model Type	Significance Threshold RMSE	Winning Model(s)
Baseline		0.607
PXNR		No winning model
PXRR	0.478	0.457 (RSSQ)
PXFR		No winning model
HYN0	0.500	0.478 (OPTI)
HYN1		No winning model
HYR0		No winning model
HYF0		No winning model

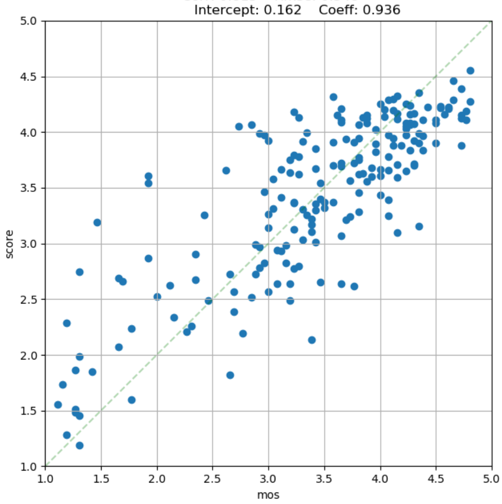
Note: For PXNR, PXFR, HYN1, HYR0 and HYF0 categories no model was winning as either the model did not meet the minimum RMSE requirement or none of submitted model of this category was statistically better than the winning model(s) of a lower complexity category. It is further noted that the outcome of the Bitstream Mode 0 and Mode 1 model competition parts was taken out here, since the involved parties could not agree on a common handling.

6. Scatter Plots for Winning Models

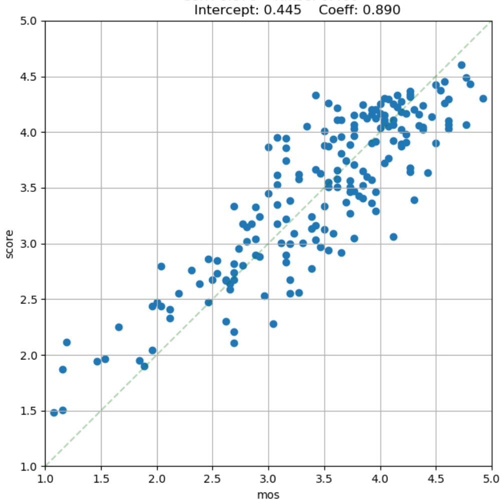
For winning models of BSM3, PXRR and HYN0 model categories, scatter plots for 3 short databases, resulting in the least, median and highest RMSE are reported.


DTTU, BSM3
P2SVL07
Corr: 0.919   Rmse: 0.378
Intercept: -0.183   Coeff: 1.058

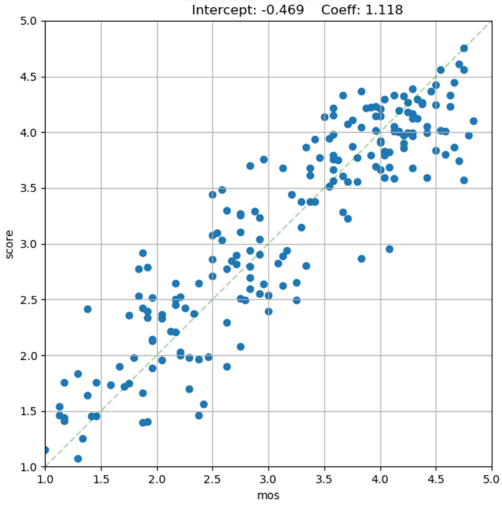

DTTU, BSM3
P2SVL12
Corr: 0.845   Rmse: 0.423
Intercept: -0.095   Coeff: 1.077


DTTU, BSM3
P2SVL10
Corr: 0.815   Rmse: 0.548
Intercept: 0.162   Coeff: 0.936

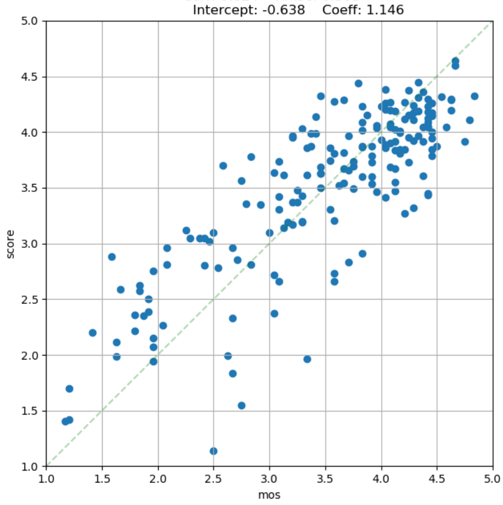
RSSQ, PXRR
P2SVL13
Corr: 0.870   Rmse: 0.400
Intercept: 0.445   Coeff: 0.890

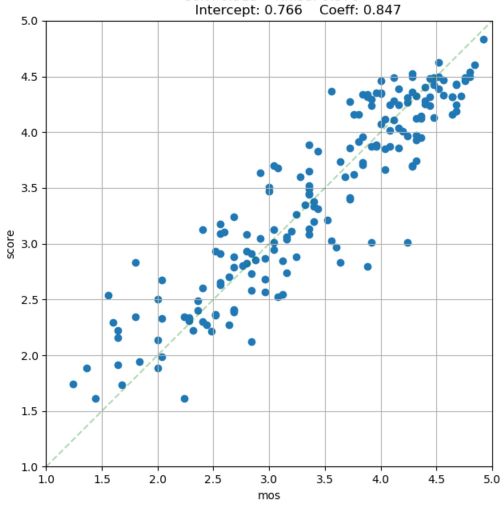


RSSQ, PXRR
P2SVL11
Corr: 0.905   Rmse: 0.448
Intercept: -0.469   Coeff: 1.118

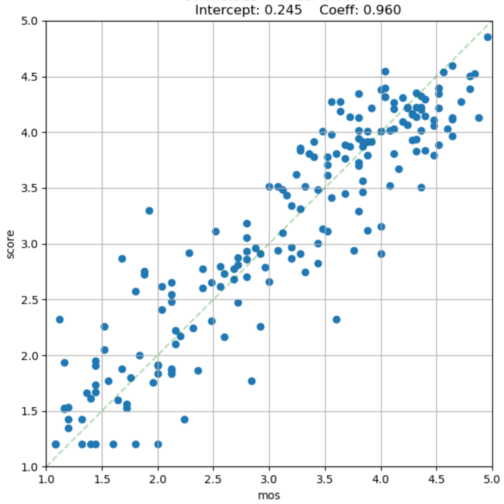


RSSQ, PXRR
P2SVL02
Corr: 0.824   Rmse: 0.508
Intercept: -0.638   Coeff: 1.146

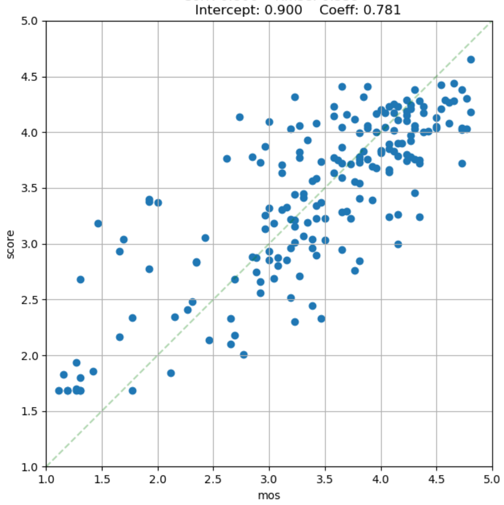




OPTI, HYN0
P2SVL01
Corr: 0.913   Rmse: 0.376
Intercept: 0.766   Coeff: 0.847



OPTI, HYN0
P2SVL05
Corr: 0.919   Rmse: 0.440
Intercept: 0.245   Coeff: 0.960



OPTI, HYN0
P2SVL10
Corr: 0.806   Rmse: 0.559
Intercept: 0.900   Coeff: 0.781

7. Linear Correlation for Winning Models

(Note: This is for information only, and was not used to determine the winning group models.)

Training Databases

	DTTU BSM3	**RSSQ PXRR**	**OPTI HYN0**
P2STR01	0.94	0.91	0.88
P2STR02	0.95	0.94	0.86
P2STR03	0.95	0.92	0.86
P2STR04	0.97	0.96	0.95
P2STR05	0.92	0.89	0.85
P2STR06	0.93	0.88	0.87
P2STR08	0.96	0.91	0.92
P2STR09	0.95	0.92	0.91
P2STR10	0.95	0.93	0.94
P2STR11	0.96	0.93	0.91
P2STR12	0.94	0.91	0.88
P2STR13	0.96	0.93	0.93
P2STR14	0.91	0.88	0.89
P2LTR15		0.89	0.92
P2LTR17		0.83	0.92

Validation Databases

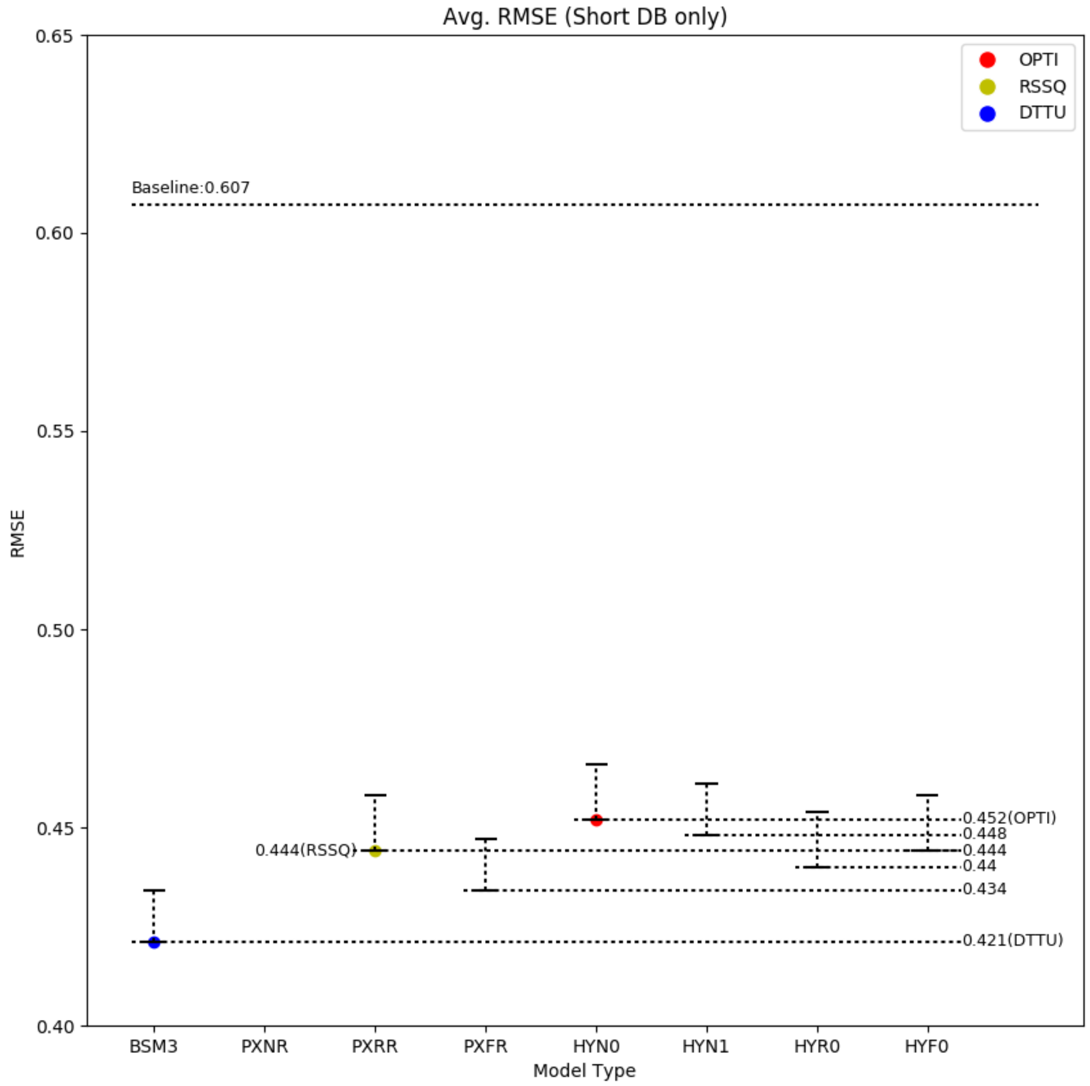
	DTTU BSM3	**RSSQ PXRR**	**OPTI HYN0**
P2SVL01	0.91	0.86	0.91
P2SVL02	0.89	0.82	0.84
P2SVL03	0.87	0.86	0.85
P2SVL04	0.92	0.93	0.89
P2SVL05	0.90	0.91	0.92
P2SVL06	0.93	0.92	0.91
P2SVL07	0.92	0.88	0.90
P2SVL08	0.88	0.90	0.91
P2SVL09	0.89	0.88	0.85
P2SVL10	0.81	0.84	0.81
P2SVL11	0.92	0.91	0.91

	DTTU BSM3	RSSQ PXRR	OPTI HYN0
P2SVL12	0.85	0.82	0.75
P2SVL13	0.85	0.87	0.87
P2LVL15		0.89	0.81
P2LVL18		0.92	0.87
P2LVL19		0.94	0.87
P2LVL23		0.89	0.81

Since bitstream models are only evaluated for short databases, correlation values for long databases are not reported in the above table.

8. Choice of Model for Standardization

The criteria for model selection are detailed in **[IEEE P1204]**. The idea is that a model needs to perform significantly better compared to the baseline model and the models of lower complexity in order to be considered for standardization.

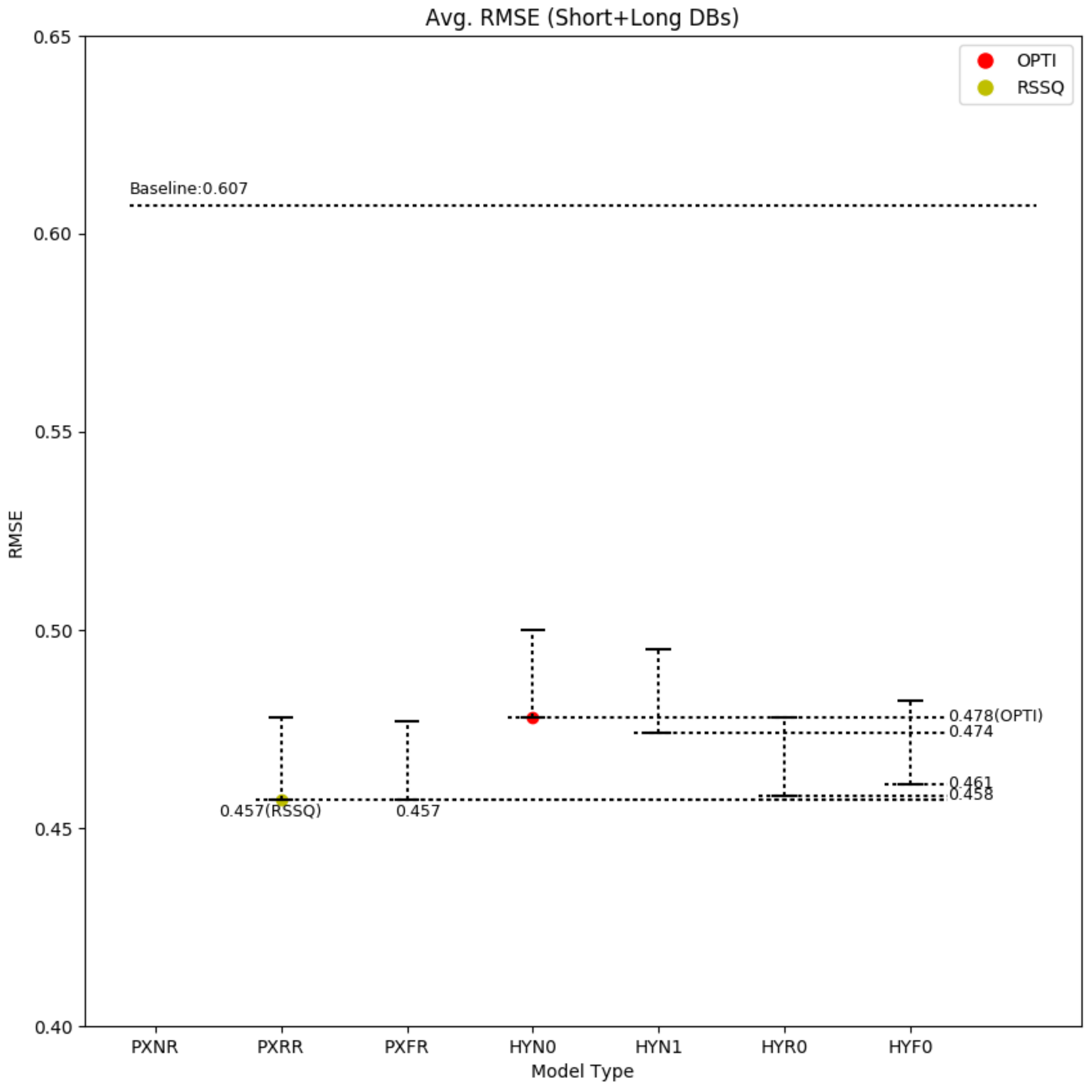


For short database validation, the best model along with the significance threshold for each model type is plotted in the above figure. For a model category all models falling within the significance interval (the dotted vertical lines) are considered as winning models.

1. None of the PXNR submissions (not shown in the figure) fulfill the minimum performance requirements.
2. The significance interval of PXFR is partially overlapped by the significance interval of PXRR, hence we do not have a Full Reference model that significantly outperforms the best PXRR model. Hence, the PXRR model can be considered to include both the PXRR and PXFR categories, which is planned to be explained in the standard accordingly.
3. The significance interval of HYN1 is partially overlapped by the significance interval of HYN0, hence we do not have a HYN1 model that significantly outperforms the best HYN0 model. Hence, HYN0 can be used instead of HYN1 to predict the quality of short videos.

4. The significance interval of HYR0 is partially overlapped by the significance interval of HYN0, hence we do not have a HYR0 model that significantly outperform the best HYN0 model. Hence, HYN0 can be used instead of HYR0 to predict the quality of short videos.
5. The significance interval of HYF0 is partially overlapped by the significance interval of HYN0, hence we do not have a HYF0 model that significantly outperform the best HYN0 model. Hence, HYN0 can be used instead of HYF0 to predict the quality of short videos.

The other 3 model categories BSM3, PXRR and HYN0 have significance intervals that are separate (see Fig. below). These model categories are considered for standardized.



For joint short and long database validation we only consider pixel-based and hybrid models. As explained above, this is due to the fact that according to the ToR, bitstream models were submitted without long term integration function.

1. Similar to the short database case, none of the PXNR submissions fulfill the minimum performance requirements for short+long database validation..
2. Similar to the short database case, the significance interval of PXFR is completely overlapped by the significance interval of PXRR, hence we do not have a Full Reference model that significantly outperforms the best PXRR model. Hence, PXRR can be used instead of PXFR to predict the quality of short and long videos.
3. Similar to the short database case, the significance interval of HYN1 is partially overlapped by the significance interval of HYN0, hence we do not have a HYN1 model that significantly outperform the best HYN0 model. Hence, HYN0 can be used instead of HYN1 to predict the quality of short and long videos.
4. The significance interval of HYR0 is partially overlapped by the significance interval of HYN0, hence we do not have a HYR0 model that significantly outperforms the best HYN0 model. Hence, HYN0 can be used instead of HYR0 to predict the quality of short and long videos.
5. The significance interval of HYF0 is partially overlapped by the significance interval of HYN0, hence we do not have a HYF0 model that significantly outperforms the best HYN0 model. Hence, HYN0 can be used instead of HYF0 to predict the quality of short and long videos.

9. Conclusions

The AVHD/PNATS Phase 2 competition jointly carried out by ITU-T SG12 and VQEG has successfully been finalized. Model performance for all submitted models is detailed in this report. Based on the set of predefined criteria, winning models were determined. For BSM0 and BSM1, more than one model is found to be in the winning group. For BSM3, PXRR and HYN0 models, a single winner for each category was determined.

For the three cases where only one model was in the winning groups and were to be standardized according to the rules laid out for the P.NATS Phase 2 competition, three new standards have been consented at the ITU-T SG12 Meeting in Geneva in Dec. 2019, founding the new standard series P.1204.

The respective models are: Bitstream Mode 3 (P.1204.3), Pixel-based Full Reference / Reduced Reference (P.1204.4) and Hybrid (P.1204.5). This report details the model performance of the winning models for these 3 models as they were submitted to the competition. The actual standards also report the optimized model performance.

For BSM0, BSM1 categories merged models could not be developed due to disagreement between the parties.