Comparison of conditions for omnidirectional video with spatial audio in terms of subjective quality and impacts on objective metrics resolving power

Andréas Pastor, Pierre Lebreton, Toinon Vigier, Patrick Le Callet

COMPARISON OF CONDITIONS FOR OMNIDIRECTIONAL VIDEO WITH SPATIAL AUDIO IN TERMS OF SUBJECTIVE QUALITY AND IMPACTS ON OBJECTIVE METRICS RESOLVING POWER

Andréas Pastor[1], Pierre Lebreton[1, 2], Toinon Vigier[1], Patrick Le Callet[1, 3]

[1]Nantes Université, Ecole Centrale Nantes, CNRS, LS2N, UMR 6004, F-44000 Nantes, France
[2]CAPACITÉS SAS [3]Institut universitaire de France (IUF)

ABSTRACT

Omnidirectional media formats, particularly 360° videos with spatial audio, provide new immersive experiences and introduce a novel dimension to content consumption.

We explore the relationship between subjective data quality and metric performance evaluation in the context of Omnidirectional videos with spatial audio. While methodologies for 360° video quality assessment have been standardized and well-documented, previous efforts primarily focus on video with limited audio conditions, e.g., mono/stereo rendering. Moreover, the experimental test setup and subjective test methodologies impact data quality and the ability to use these data for objective quality metrics performance evaluation. Such a problem is key in the industry and the standardization activities, as codecs and quality models must be compared. Hence, the requirements on the ground truth data quality have to be clarified to allow proper conclusions.

In this paper, we compare two setups and three test methodologies to study how experiment discriminability changes with conditions and participant number. Then, we show how discriminability impacts the resolving power of quality metrics. We show that higher-performing metrics require higher-quality data to reveal their full potential. In doing so, we put into relation the experimental cost, data quality, and resolving power.

***Index Terms*—** subjective quality assessment, omnidirectional media format, 360 video, spatial audio, Ambisonic

1. INTRODUCTION

In recent years, 360° videos paired with spatial audio and experienced through head-mounted displays (HMDs) have earned significant engagement. These immersive experiences offer a new dimension to content consumption.

Subjective experiments exploring the quality of these media are essential to reveal the impact of these two modalities, i.e. audio and video, on the Quality of Experience (QoE). Methodologies for 360° video quality experiments have recently been standardized in ITU-T Rec. P.919 [1], and previous works [2–5] focusing exclusively on the video component. Spatial audio, e.g., ambisonics, brings a better sense of immersion in content by allowing users to look in a specific direction and hear audio that reflects their head movements and positions. Moreover, it enhances immersive experiences and user presence by fostering exploration and engagement within interactive contexts. Spatial audio is a guiding force for directing visual attention and aids in the cognitive analysis of scenes.

In this work, we reproduce part of the subjective quality assessment tests with trained assessors of [6] on 360° video with Higher/4^{th} Order Ambisonic (HOA) audio. It is the first dataset with Mean Opinion Scores (MOS) to support perceptual quality research on immersive Audio-Visual content. In our paper, the focus is first to collect subjective data in other conditions for Audio-Visual experiment evaluation. Compared to the condition used in [6], we use a *consumer-grade* setup where the 26-channel loudspeakers are replaced by Head Mounted Display (HMD) integrated headphones. We are getting closer to real use cases, and we can explore differences with a high-quality *reference* setup. Another difference is that we engage naive assessors for the task instead of the trained assessors. While trained assessors offer experience and expertise, they can be subject to developing a training bias and exploring the content in a non-ecological manner. Training used for [6] can be found in [7] and includes training sequences evaluated during the subjective test.

Based on these data, we will demonstrate how rendering setup and subjective test methodologies affect the quality of subjective data in terms of discriminability [6, 8–11]. This is an important aspect of codec development and standardization activities, as system performances must be evaluated and compared. To this aim, statistical testing that accounts for subjective data reliability is commonly used to compare differences between correlations, Root Mean Square Error (RMSE) [12], Spearman Rank Order correlation (SRCC) [13], Resolving Power [14–16]. However, this is an afterthought, and the subjective data quality requirement and how much one shall invest in subjective testing to allow discriminating between different quality estimation models has only been weakly studied. To address this point, we will take commonly used audio and video quality models: VMAF [17], ITU-T Rec. P.1203 [18–20], ITU-T Rec.

P.1204.3 [21, 22], and ViSQOL [23] and show how subjective data discriminability affects our ability to compare these metrics. Subjective data discriminability will be studied across different setups and test methodologies, giving a unique point of view on the relationship between experimental cost and conclusions that can be drawn from the data. Contributions are as follows:

- We compare a pristine and degraded audio setup regarding discriminability.
- We study the cost of an experiment and its relation with discriminability.
- We study metric resolving power as a function of subjective discriminability.
- We propose a method to apply the no-reference parameter-based audiovisual quality estimation model ITU-T P.1203 mode 0 to 360° video evaluation.

2. QOE CONDITIONS

This section describes the selected test material, the test environment for Omnidirectional Videos with Spatialized Audio playback, and the test methodologies selected for the subjective evaluation conditions. We decided to compare with the "Audiovisual" test from [6], a test on 360° videos with HOA audio. We selected a subset of Sources (SRCs) and Process Video Sequences (PVS) of this dataset to compare in two new conditions.

2.1. Test material and environment

We choose from [6] the following sequences: *"CarWithChat"*, *"DoorOpen"*, *"HairDrying"*, *"DogBarking"*, *"RiverStream"*, *"CrossRoadNight"* and *"HandballMatch"*. These sequences were selected for their video spatial and temporal information diversity and audio characteristics: source nature, location, and movement. All the test sequences are in YUV420 color space and Equirectangular Projection (ERP) format. The video decoding and display are operated via Unity. For Audio playback, we use Reaper, with the Sparta AmbiBIN plugin [24], to render a binaural audio stream from the HOA bitstreams.

In [6], Visual stimuli were displayed in VR using a Samsung Odyssey+ Head Mounted Display (HMD) with a display resolution of 1440 × 1600 per eye, 110° horizontal Field of View (FOV). The HMD used during our experiments is the HTC Vive Pro eye, with the same 110° FOV and per-eye display resolution. In contrast with the original experiment, which used a 26-channel setup of 8040A Genelec loudspeakers, the audio is playback in the built-in HMD headphones. This difference will affect the perception of audio stimuli, as we will see in the analysis. But put us closer to a consumer-grade experience. During the experiment, the subjects were seated on a swivel chair to allow them to turn freely. The start position of each 360° video was reset to the ERP center at the start of each viewing, irrespective of final world positions. It ensures that all observers start at the same position.

2.2. Subjective testing methodologies

In [6], authors collected data using trained assessors. Their training procedure includes sequences evaluated during the subjective test part. More information about the assessors' training procedure can be found in [7]. In this first context, named "Reference condition", the subjective methodology was a modified version of Multiple Stimuli with Hidden Reference and Anchor (MUSHRA) to generalize from the SAMVIQ methodology [25] for video, and MUSHRA [26] for intermediate audio quality evaluation. In this paper, we want to reproduce with naive assessors the results obtained with trained assessors. Past works have focused on subjective methods comparison for mobile video viewing [27], 2D/3D videos [28], or 3D Graphics in Virtual Reality [8], and show that Degradation Category Rating (DCR) and Absolute Category Rating with Hidden Reference (ACR-HR) outperform SAMVIQ in the context of naive observers as it is longer and more complex to evaluate quality with it. These two methodologies are specified in ITU-T Rec. P.910 [29].

We design two contexts where naive assessors experience and rate audiovisual quality from HMD and build-in headphones: in the first context, an ACR-HR scale is used, namely "ACR-HR condition", and in the second context, a DCR scale, namely "DCR condition". For both conditions, we include a calibration phase before the start of the test to familiarize participants with the rating task and the degradation. We choose 3 PVS from *"ChamberMusic"* SRC for high, middle, and low audiovisual qualities.

In the ACR-HR condition, assessors are asked to rate the absolute quality of the PVS on a five-point ACR scale (1: Bad, 2: Poor, 3: Fair, 4: Good, 5: Excellent). This test has 49 PVS, a hidden reference + 6 PVS per SRC: video degradations are selected from (6K_QP22, 4K_QP0, 4K_QP28, 2K_QP0, 2K_QP34) and audio from (PCM, 64, 32, 16 kbps). PCM are the reference audio files at 1152 kbps/channel, 24bit, and 48kHz. The hidden references are 6K_QP0_PCM sequences. We balance the PVS selection on the audiovisual quality scale, using prior knowledge from "Reference Condition". The median session duration is around 19 minutes for this condition.

In the DCR experiment, assessors use a five-point impairment scale (1: Very annoying, 2: Annoying, 3: Slightly annoying, 4: Perceptible but not annoying, 5: Imperceptible). The first stimulus is always the SRC. The second stimulus is the impaired PVS. Due to the explicit reference impacting test duration, only 5 SRCs are evaluated in these tests, with 6 PVS per SRC. *"CrossRoadNight"* and *"HandballMatch"* sequences are not evaluated. The average viewing duration is 24 min for the 30 PVS in the DCR condition. We recruit naive assessors aged 19 to 64 (mean=33.8, std=12.9) with different nationalities and educational backgrounds. Before the start of the test, we tested every assessor's vision in terms of visual acuity with Snellen charts and their color perception with the Ishihara test, and hearing acuity was self-reported. Assessors

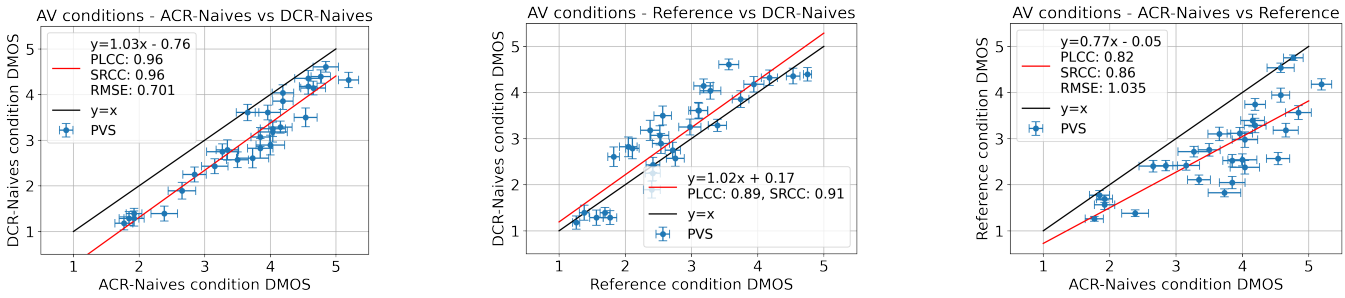

Fig. 1: Comparison of Audiovisual conditions on rating scale usage by assessors.

Table 1: Mean Content Ambiguity (CA) estimates from SUREAL MLE [30] on data of the three conditions.

Conditions	360° AudioVisual test
”Reference”	0.28
”DCR–Naives”	0.45
”ACR–Naives”	0.55

who did not meet the requirements were not recruited.

3. COMPARISON OF CONDITIONS

In this section, we compare the data collected in our two conditions with naive assessors against trained assessors of the ”Reference condition”.

3.1. Difficulty and subjectivity of the task

Advanced methods for outlier removal techniques like ”MLE” and ”MLE_CO_AP2” from SUREAL package[1] clean MOS. The ”MLE” algorithm jointly recovers subjective quality scores from noisy raw measurements, subjects’ Bias and Inconsistency, and a Content Ambiguity estimate for each SRC. By modeling the noise, these methods help to understand the influence of assessors and SRCs on MOS. The reference papers of SUREAL [30] show that the estimates of ”MLE” and ”MLE_CO_AP2”, compared to methods using a threshold like BT.500 [31] and P.913 [32] are more interpretable and robust to spammer behaviors.

The Content Ambiguity (CA) estimated by the ”MLE” model indicates how subjective it is to annotate the quality of a particular sequence. By averaging estimated CAs across all SRCs of a test, we obtain a Mean CA score, reflecting the subjectivity of a QoE task. We can compare this score across test conditions and conclude if a task is more or less subjective for assessors. For the SAMVIQ scale in ”Reference condition”, we first scale the assessor’s Opinion Score from 0–100 to 1–5 to fit the scale of ACR-HR and DCR. From table 1, we can see that the subjectivity of the task in our conditions with naive assessors using ACR or DCR scale is higher than for trained assessors. This shows how effective the training is on assessors using the SAMVIQ scale. Moreover, we can see that naive assessors using an ACR scale compared to a DCR scale face a more subjective task. This can be explained by the explicit reference in DCR easing the rating task.

[1] SUREAL: https://github.com/Netflix/sureal/tree/master/sureal

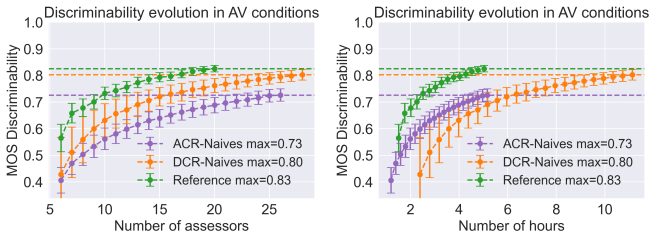

Fig. 2: MOS discriminability evolution for the three 360° AudioVisual conditions, as a function of assessors’ number and data collection duration.

3.2. Usage of the scale

In Figure 1, we present an analysis of the MOS obtained in the three test conditions. We fit a linear function (in red) and extract the slope and the intercept. We analyze these coefficients to see how assessors perceive the stimuli differently and how they use the rating scales. The black line translates the ”one-to-one” relationship. In the Figure 1 left plot, we see a strong agreement between assessors in our conditions, with high correlation scores. Due to explicit references in the DCR condition helping detect differences, DMOS scores are, on average, lower for DCR than for the ACR scale. In the two other plots, the correlation is lower, translating that stimuli are perceived differently from naive assessors in consumer-grade conditions to trained assessors in a reference environment.

3.3. Discriminability analysis

In [6, 8–10], the authors suggested examining the MOS discriminability evolution with increasing numbers of assessors to show how well a subjective methodology can recover accurate MOS scores and compare subjective methodologies efficiency [33]. A two-sample Wilcoxon test is applied to all the possible pairs of MOS in a dataset to test the proportion of significantly different ones. We plot the evolution of this ratio in function of the assessors’ number. The analysis is restricted to the 30 PVS DMOS common in the DCR–Naive, ACR-HR–Naive, and SAMVIQ–Trained conditions.

The results of the three 360° Audio-Visual conditions are presented in Figure 2 with a 95% confidence interval over 10,000 permutations. We can see that the most discriminative condition is the ”Reference” one. The DCR–Naives condition can achieve similar discriminability with more assessors. 28 DCR–Naives assessors achieve the discriminability of 17 SAMVIQ–Trained assessors in ”Reference” setup.

A factor that could explain the gap is the training followed by ”Reference” condition assessors and the difference between the quality of the ”Reference” and ”Consumer-Grade” setups. ”Reference” setup with a 26-channel loudspeaker

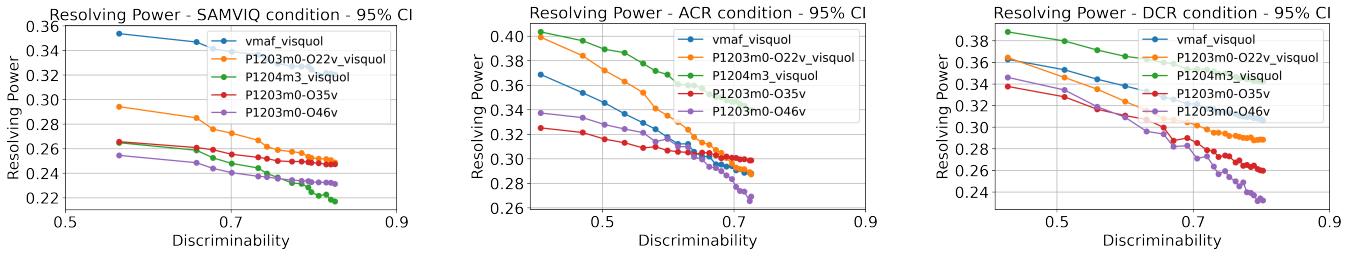

Fig. 3: Audiovisual objective quality metrics Resolving Power [15] for the 3 conditions in the function of the discriminability

Table 2: Objective metrics SRCC performance on data of the three audiovisual conditions: bold for best and italic for worst.

Metrics	Reference	ACR–Naives	DCR–Naives
VMAF_VISQUOL	*0.846*	0.855	0.870
P1203m0-O22v_VISQUOL	**0.882**	0.857	0.879
P1204m3_VISQUOL	0.877	*0.689*	*0.731*
P1203m0-O35v	0.872	0.882	**0.920**
P1203m0-O46v	0.870	**0.892**	0.895

versus binaural rendering can affect the quality and perception of reference spatial audio stimuli playback. Requiring more assessors to achieve the same discriminability is also linked to the Mean Content Ambiguity score presented in the last section: task subjectivity. ACR–Naives condition has the highest mean CA score and requires 26 assessors to equal 16 DSIS–Naives assessors and 10 SAMVIQ–trained assessors from the "Reference" setup.

Lastly, regarding experimental effort and data collection duration, ACR seems to be, at first, an efficient method to discriminate and obtain accurate MOS scores. But, we can see that using a DCR scale with an explicit reference per stimuli helps to achieve high discriminability in the long run. In the "Reference" condition, assessors are trained before the start of the data collection. This factor is not taken into account and can vary with the training process. It is, therefore, difficult to conclude on its benefit.

4. OBJECTIVE METRICS RESOLVING POWER

Metric resolving power has been defined in ITU-T Rec. J.149 [15], ITU-R Rec. BT.1676 [14] and in [16]. It represents the lowest quality difference a metric can measure that has a statistical difference from the subjective ratings point of view. This characterizes the meaningfulness of quality differences measured by prediction models. However, resolving power depends on metric accuracy and also on the dataset quality that is used for evaluating the metric performance. This section will show how much performance can be claimed for different models depending on subjective data quality.

Five audiovisual quality prediction models are considered: **vmaf_visquol:** VMAF [17] video quality scores combined with ViSQOL [23] audio quality scores using the linear combination defined in the audiovisual quality integration module of ITU-T Rec. P.1203.3[2] [19, 20, 34]. **P1204m3_visquol:** ITU-T P.1204[3] mode 3 bitstream video quality estimation model [21, 22] combined with ViSQOL like vmaf_visquol.

In addition, three parameter-based models based on ITU-T Rec. P.1203 mode 0 [18] are proposed. ITU-T Rec. P.1203 mode 0 only uses video bitrate, resolution, frame rate, and audio bitrate values to perform quality predictions. To apply this model to 360 videos, the headset FOV (110°) is used to scale resolution and bitrate values to represent what is seen in the headset. For example, a 6K (6144x3072) ERP with a 110° viewport leads to 1878x1877 pixels seen into the HMD. Video bitrate is also scaled by $(360\times180)/(110\times110)\approx5.35$. Doing so, a per-viewport P.1203 mode 0 quality scores are computed and reported as **P1203m0-O35v** and **P1203m0-O46v**. O35 is the quality score after audiovisual and temporal pooling, while O46 is the final ITU-T P.1203 mode 0 prediction [34]. Finally, **P1203m0-O22v_visquol** is proposed to replace the parameter-based audio quality estimation from the ITU model by ViSQOL.

Results are presented in Figure 3. For the three conditions, increasing the discriminability, hence the quality of subjective data improves the resolving power of quality metrics and shows the importance of having reliable MOS. Moreover, linking with results presented in table 2, **P1203m0-O46v** performs relatively well regarding SRCC across the different datasets. Its low resolving power also suggests it. Lower resolving power is better. Poorly performing metrics like VMAF_VISQUOL on "Reference" conditions data and P1204m3_VISQUOL on ACR–Naives and DCR–Naives have accordingly high resolving power, and it is not changing much with higher discriminative subjective data. Finally, orders between metrics can change with increasing discriminability, as shown with **P1203m0-O35v** and **P1203m0-O22v_VISQUOL**.

5. CONCLUSION

In this work, we compare the efficiency of subjective methodologies to assess the quality of 360° videos with HOA audio across different setups. We show that high discriminability obtained by trained assessors can be replicated with naive assessors and DCR methodology. Increasing discriminability enables to achieve of better resolving power of quality metrics. High-performing metrics tend to benefit more from high discriminability to express their best-resolving power. As well, the order of metrics can change across discriminability values. Finally, our proposal of a new parameter-based quality estimation model adapted to viewport resolution has

[2]https://github.com/itu-p1203/itu-p1203

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

shown competitive results with Full-Reference approaches.

For future audiovisual quality metrics development, we should consider more the discriminability of subjective data along with metrics performance evaluation.

6. REFERENCES

[1] ITU-T Rec. P.919, "Subjective test methodologies for 360º video on head-mounted displays," 2020.

[2] Ashutosh Singla, Stephan Fremerey, Frank Hofmeyer, Werner Robitza, and Alexander Raake, "Quality assessment protocols for omnidirectional video quality evaluation," *Electronic Imaging*, vol. 2020, no. 11, pp. 69–1, 2020.

[3] Majed Elwardy, Yan Hu, Hans-Jürgen Zepernick, Thi My Chinh Chu, and Veronica Sundstedt, "Comparison of acr methods for 360° video quality assessment subject to participants' experience with immersive media," in *2020 14th International Conference on Signal Processing and Communication Systems (ICSPCS)*. IEEE, 2020, pp. 1–10.

[4] Jesus Gutierrez, Pablo Perez, Marta Orduna, Ashutosh Singla, Carlos Cortes, Pramit Mazumdar, Irene Viola, Kjell Brunnström, Federica Battisti, Natalia Cieplińska, et al., "Subjective evaluation of visual quality and simulator sickness of short 360 videos: Itu-t rec. p. 919," *IEEE transactions on multimedia*, vol. 24, pp. 3087–3100, 2021.

[5] Marta Orduna, Pablo Pérez, Jesús Gutiérrez, and Narciso García, "Methodology to assess quality, presence, empathy, attitude, and attention in 360-degree videos for immersive communications," *arXiv preprint arXiv:2103.02550*, 2021.

[6] Randy F Fela, Andréas Pastor, Patrick Le Callet, Nick Zacharov, Toinon Vigier, and Soren Forchhammer, "Perceptual evaluation on audio-visual dataset of 360 content," in *2022 IEEE International Conference on Multimedia and Expo Workshops (ICMEW)*, 2022, pp. 1–6.

[7] Randy Frans Fela, Nick Zacharov, and Søren Forchhammer, "Assessor selection process for perceptual quality evaluation of 360 audiovisual content," *Journal of the Audio Engineering Society*, vol. 70, no. 10, pp. 824–842, 2022.

[8] Yana Nehmé, Jean-Philippe Farrugia, Florent Dupont, Patrick Le Callet, and Guillaume Lavoué, "Comparison of subjective methods for quality assessment of 3D graphics in virtual reality," *ACM Transactions on Applied Perception (TAP)*, vol. 18, no. 1, pp. 1–23, 2020.

[9] Andréas Pastor and Patrick Le Callet, "Towards guidelines for subjective haptic quality assessment: A case study on quality assessment of compressed haptic signals," in *2023 IEEE International Conference on Multimedia and Expo (ICME)*, 2023, pp. 1667–1672.

[10] Andréas Pastor and Patrick Le Callet, "Perceptual annotation of local distortions in videos: tools and datasets," in *Proceedings of the 14th Conference on ACM Multimedia Systems*, 2023, pp. 458–462.

[11] Andréas Pastor, Lukáš Krasula, Xiaoqing Zhu, Zhi Li, and Patrick Le Callet, "Comparison of subjective methodologies for local perception of distortion in videos and impact on objective metrics resolving power," working paper or preprint, Feb. 2024.

[12] ITU-T Rec. P.1401, "Methods, metrics and procedures for statistical evaluation, qualification and comparison of objective quality prediction models," 2020.

[13] Babak Naderi and Sebastian Möller, "Transformation of mean opinion scores to avoid misleading of ranked based statistical techniques," in *2020 Twelfth International Conference on Quality of Multimedia Experience (QoMEX)*, 2020, pp. 1–4.

[14] ITU-R Rec. BT.1676, "Methodological framework for specifying accuracy and cross-calibration of video quality metrics," 2004.

[15] ITU-T Rec. J.149, "Method for specifying accuracy and cross-calibration of video quality metrics (VQM)," 2004.

[16] Lukáš Krasula, Karel Fliegel, Patrick Le Callet, and Miloš Klíma, "On the accuracy of objective image and video quality models: New methodology for performance evaluation," in *2016 Eighth International Conference on Quality of Multimedia Experience (QoMEX)*, 2016, pp. 1–6.

[17] Netflix, "VMAF v0.6.1 Model," https://github.com/Netflix/vmaf.

[18] ITU-T Rec. P.1203, "Parametric bitstream-based quality assessment of progressive download and adaptive audiovisual streaming services over reliable transport," 2017.

[19] Alexander Raake, Marie-Neige Garcia, Werner Robitza, Peter List, Steve Göring, and Bernhard Feiten, "A bitstream-based, scalable video-quality model for HTTP adaptive streaming: ITU-T P.1203.1," in *Ninth International Conference on Quality of Multimedia Experience (QoMEX)*, Erfurt, May 2017, IEEE.

[20] Werner Robitza, Steve Göring, Alexander Raake, David Lindegren, Gunnar Heikkilä, Jörgen Gustafsson, Peter List, Bernhard Feiten, Ulf Wüstenhagen, Marie-Neige Garcia, Kazuhisa Yamagishi, and Simon Broom, "HTTP Adaptive Streaming QoE Estimation with ITU-T Rec. P.1203 – Open Databases and Software," in *9th ACM Multimedia Systems Conference*, Amsterdam, 2018.

[21] ITU-T Rec. P.1204.3, "Video quality assessment of streaming services over reliable transport for resolutions up to 4k with access to full bitstream information," 2020.

[22] Rakesh Rao Ramachandra Rao, Steve Göring, Werner Robitza, Alexander Raake, Bernhard Feiten, Peter List, and Ulf Wüstenhagen, "Bitstream-based model standard for 4k/uhd: Itu-t p.1204.3 – model details, evaluation, analysis and open source implementation," in *2020 Twelfth International Conference on Quality of Multimedia Experience (QoMEX)*, Athlone, Ireland, May 2020.

[23] Andrew Hines, Eoin Gillen, Damien Kelly, Jan Skoglund, Anil Kokaram, and Naomi Harte, "ViSQOLAudio: An objective audio quality metric for low bitrate codecs," *The Journal of the Acoustical Society of America*, vol. 137, no. 6, pp. EL449–EL455, 2015.

[24] Leo McCormack and Archontis Politis, "SPARTA & COMPASS: Real-time implementations of linear and parametric spatial audio reproduction and processing methods," in *2019 AES International Conference on Immersive and Interactive Audio*. Audio Engineering Society, 2019.

[25] F Kozamernik, V Steinmann, P Sunna, and E Wyckens, "SAMVIQ—A new EBU methodology for video quality evaluations in multimedia," *SMPTE motion imaging journal*, vol. 114, no. 4, pp. 152–160, 2005.

[26] ITU-R Rec. BS.1534-3, "Method for the subjective assessment of intermediate quality levels of coding systems," 2015.

[27] Toshiko Tominaga, Takanori Hayashi, Jun Okamoto, and Akira Takahashi, "Performance comparisons of subjective quality assessment methods for mobile video," in *2010 Second international workshop on quality of multimedia experience (QoMEX)*. IEEE, 2010, pp. 82–87.

[28] Taichi Kawano, Kazuhisa Yamagishi, and Takanori Hayashi, "Performance comparison of subjective assessment methods for 3d video quality," in *2012 Fourth International Workshop on Quality of Multimedia Experience*, 2012, pp. 218–223.

[29] ITU-T Rec. P.910, "Subjective video quality assessment methods for multimedia applications," 2008.

[30] Zhi Li and Christos G Bampis, "Recover subjective quality scores from noisy measurements," in *2017 Data compression conference (DCC)*. IEEE, 2017, pp. 52–61.

[31] ITU-R Rec. BT.500-13, "Methodology for the subjective assessment of the quality of television pictures," 2012.

[32] ITU-T Rec. P.913, "Methods for the subjective assessment of video quality, audio quality and audiovisual quality of internet video and distribution quality television in any environment," 2016.

[33] Andréas Pastor et al., ""Discriminability-Experimental Cost" tradeoff in subjective video quality assessment of codec: DCR with EVP rating scale versus ACR-HR," working paper or preprint, Dec. 2023.

[34] ITU-T Rec. P.1203.3, "Parametric bitstream-based quality assessment of progressive download and adaptive audiovisual streaming services over reliable transport - quality integration module," 2019.