Towards guidelines for subjective Haptic quality assessment: a case study on quality assessment of compressed haptic signals

Andréas Pastor, Patrick Le Callet

▶ **To cite this version:**

Andréas Pastor, Patrick Le Callet. Towards guidelines for subjective Haptic quality assessment: a case study on quality assessment of compressed haptic signals. 2023 IEEE International Conference on Multimedia and Expo (ICME), Jul 2023, Brisbane, Australia. pp.1667-1672, 10.1109/ICME55011.2023.00287 . hal-04132004

HAL Id: hal-04132004
https://hal.science/hal-04132004v1

Submitted on 19 Jun 2023

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

TOWARDS GUIDELINES FOR SUBJECTIVE HAPTIC QUALITY ASSESSMENT: A CASE STUDY ON QUALITY ASSESSMENT OF COMPRESSED HAPTIC SIGNALS

Andréas Pastor, Patrick Le Callet

Nantes Université, Ecole Centrale Nantes, CNRS, LS2N, UMR 6004, F-44000 Nantes, France

ABSTRACT

Modern systems are multimodal (e.g., video, audio, smell), and haptic feedback provides the user with additional entertainment and sensory immersion. Standard recommendation groups extensively studied and focused on video and audio subjective quality assessment, especially in signal transmission. In that context, subjective quality assessment and Quality of Experience (QoE) of Haptic signals is at its infant age. We propose further analyzing the collected data from a recent subjective quality assessment campaign as part of the MPEG haptic standardization group. In particular, we are addressing the following questions: 1) How the emerging field of haptic signal QoE can benefit from existing efforts of video and audio quality assessment standards? 2) How to detect possible outliers or characterize the rater's reliability? 3) How does the discriminability of haptic tests increases with the number of raters? Towards this goal, we question if traditional analysis as proposed for audio or video signal are suitable, as well as other state-of-the-art techniques. We also compare the discriminability of the haptics quality assessment tests with other modalities such as audio, video, and immersive content (360° contents). We propose recommendations on the number of raters required to meet the usual discriminability obtained for other perceptual modalities and how to process ratings to remove possible noise and biases. These results could feed future recommendations in standards such as BT500-14 or P.913 but for haptic signals.

***Index Terms*—** Haptic quality assessment, human perception, compression performance assessment, quality of experience

1. INTRODUCTION

While visual and audio Quality of Experiences (QoE) has been investigated thanks to decades of product development and consumer adoption at scale, other senses, such as smell and touch (haptic), are still at an infant age. Nevertheless, the rise of new use cases such as Metaverse offers unique opportunities for technologies that enhance the user experience by stimulating all senses and could boost their development. In particular, the future of a more immersive experience could come with the possibility of transmitting haptic feedback. There are different types of haptic feedback. Vibrotactile haptics, this technology is already used in game controllers, keyboards of our smartphones, or VR controllers where tiny motors create vibrations. Microfluidics haptics where liquid or air is compressed into smart-textile [1]. Ultrasonic mid-air haptics[1] where multiple ultrasound loudspeakers generate waves adding up into a focal point and feel like a "virtual touch" by users. Surface haptics [2] where the friction between a user's finger and a touchscreen is manipulated to create a tactile effect. Force control haptics [3] with levers, or other mechanical devices are used to exert force on a user's hands, limbs, or whole body.

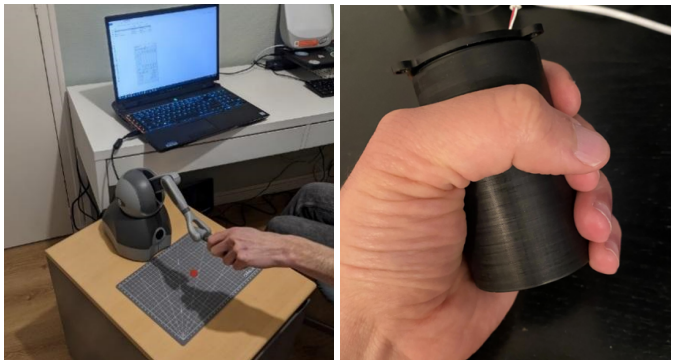

Fig. 1: Overall MUSHRA testing system setup for kinesthetic effects test platform (left) and vibrotactile short and long effects device grip (right): images from the Call for Proposal report [4].

At the same time, efforts in haptic signal compression and its evaluation are still very recent, and more work is needed to make it a mature technology in this aspect. Audio and Video Quality Assessment domains can influence and help toward guidelines to evaluate haptic signal quality. For instance, the following works discuss methodologies to evaluate the quality of different modalities: spatial audio and video for 360° audiovisual contents [5], local spatio-temporal distortions in videos [6].

Recommendation standards are also present to give guidelines on methodologies to use, how many raters to select, and methods to screen outliers: International Telecommunication

[1] https://www.ultraleap.com/haptics/

Union (ITU) provides on these topics BT.500 [7], P.910 [8], P.913 [9], and BS.1534-3 [10].

This paper uses data collected during the Haptics Phase 1 Call for Proposals [11] on "Basic Haptics" requirements and application scenarios. It's the first study in the research community to characterize and evaluate the codec performance for haptic signals.

Two problems attempt to be solved in this article. First, the rater's reliability on haptic feedback and how can annotations provided by a rater be trusted. This is considered sensitive since it can be a novel experience for the rater to be exposed to such systems. Only experts with prior experience with haptics and quality assessment were considered for this research work. The second point explored is the discriminability of the Mean Opinion Scores (MOS) collected during the subjective tests. In other words, how can we ensure that the data collected is as precise as possible to be later used to improve systems and train objective metrics?

The remainder of this paper is organized as follows. Section 2 includes the description of the database, the haptic devices used, and the subjective evaluation procedures. We discuss in Section 3 our findings on the rater's reliability and different outlier methodologies applied to clean the data. We present an analysis of the subjective scores discriminability in Section 4. Finally, the conclusion and recommendations are presented in Section 5.

2. SUBJECTIVE QUALITY ASSESSMENT DATASET

This section describes the dataset produced by the MPEG haptic standardization group. The subjective tests were conducted for both vibrotactile and kinesthetic test signal modalities.

2.1. Test design: Haptic End device, SRC and HRC

During this project phase, different hardware configurations have been evaluated and used to playback the haptic stimuli.

The first Haptic End device is the Vibrotactile Test Platform: Foster Electric Co., Ltd provided the actuator (model 576865) for the two Vibrotactile tests, one on long and the other on short vibrotactile stimuli. The usable dynamic range of this actuator is 65–300 Hz. In [11], Short effects haptics are defined as with a simple envelope, Attack Decay Sustain Release (ADSR), where the envelope is the salient characteristic of the effect, ranging from 20 ms to 1000 ms. Short effects haptics are characterized by placement, and their subcomponents envelopes are the salient characteristics of the effect, ranging from 1000 ms to over 5000 ms.

The second Haptic End device is the Kinesthetic Test Platform: the 3DSystem Geomagic Touch provides 3 Degrees of Freedom (DoF) force-feedback and inputs 6 DoF positions through a handheld stylus.

The first modality evaluated is Vibrotactile Haptics with two tests named "Vibrotactile Short Effects" and "Vibrotactile Long Effects" on the Vibrotactile Test Platform. 8 Sources (SRCs) were used in the "Vibrotactile Short Effects" (VSE) test, and 11 SRCs in the "Vibrotactile Long Effects" (VLE) test. 4 Hypothetical Reference Circuits (HRCs) are used to compress and generate the Processed Haptics Signals (PHS) to evaluate subjectively.

The second modality evaluated is on Kinesthetic Haptics, with one test named "Kinesthetic Effects" (KE) on the Kinesthetic Test Platform. 11 SRCs were used with 2 HRCs in this test.

Three bitrates were used, 2, 16, and 64 kbits, to generate the PHS of each test. It resulted in 104 PHS being evaluated in the VSE test, 143 in the VLE test, and 70 in the KE test.

2.2. Subjective testing methodology

The subjective protocol applied during the three tests uses multiple stimuli with a hidden reference without an anchor (modified MUSHRA) to generalize the standard evaluation methodology found in SAMVIQ [12] for video and MUSHRA [10] for intermediate audio quality evaluation, more details can be found in ITU P.913 standard [9]. Each assessor was asked to rate their overall perceived quality of the PHS on a continuous rating scale between 0 and 100. The number of PHS to evaluate on each annotation trial was set to 5 for VSE and VLE tests (i.e., a hidden reference + 4 PHS) and set to 3 for KE tests (i.e., a hidden reference + 2 PHS).

Experts were recruited from the 5 laboratories participating in the joint effort: Immersion Corporation, Canada (IMMR), University of Southern California, USA (USC), Pohang University of Science and Technology, South Korea (POST), Kyung-Hee University, South Korea (KHU), and the University of Nantes, France (UNF).

Table 1 summarizes the experimental effort of the different laboratories. A total of 36 unique people participated in the first VSE test. This test is divided into three separate test sessions, each lasting, on average, 10 minutes. 95 sessions were recorded across all the laboratories, with 33 sessions for the first test sessions, 30 for the second, and 32 for the last one. Not all experts participated in all 3 test sessions.

For the second test VLE, 37 experts took part in the study. 92 sessions were recorded across 3 test sessions: 28 for the first, 29 for the second, and 35 for the last. Finally, in the test on "Kinesthetic Effects", 94 sessions were collected: 37, 31, and 31 sessions across the different 3 test sessions.

3. RATER'S RELIABILITY

In this section, we present the analysis performed on the rater's reliability through outlier detection with the methods from standards and the state-of-the-art. We also present how "Content Ambiguity" estimated from these methods can be

Test	Session	IMMR	USC	POST	KHU	UNF	Total
VSE	1	8	8	12	5	-	33
	2	7	8	10	5	-	30
	3	8	8	12	4	-	32
VLE	1	9	8	9	2	-	28
	2	10	8	8	3	-	29
	3	11	8	11	5	-	35
KE	1	-	9	16	3	9	37
	2	-	9	10	1	11	31
	3	-	9	10	2	10	31
Total	-	53	75	98	30	30	286

Table 1: Subject count per test session and laboratories.

used to provide information on the difficulty of evaluating the quality of haptic feedback by experts.

3.1. Processing of subjective scores

There are multiple ways to convert subjective scores into Mean Opinion Scores (MOS) and, at the same time, clean the data by reducing the effect of outliers. The first solution is to compute the MOS as the average of subjective scores given by subjects. BT.500 [7] and P.913 [9] standards present two outlier rejection methods. The latter removes subject bias before applying BT.500 outlier rejection. More advanced methods compute MOS using Maximum Likelihood Estimation (MLE) and estimate statistics for contents and subjects' behaviors. We will focus on the ones proposed in [13, 14] namely "MLE" and "MLE_CO_AP2". These methods are available as a package: SUREAL[2].

The first one, "MLE" jointly recovers subjective quality scores from noisy raw measurements, subjects' Bias and Inconsistency, and a Content Ambiguity estimate for each of the SRCs. The second method, "MLE_CO_AP2" uses a solving procedure named Alternated Projection to estimate subjective quality scores, subjects' Bias, and Inconsistency.

In the "MLE" method, the raw opinion scores are modeled as a random variable $X_{e,s}$ as follows:

$$\begin{aligned} X_{e,s} &= x_e + B_{e,s} + A_{e,s}, \\ B_{e,s} &\sim \mathcal{N}(b_s, v_s^2), \\ A_{e,s} &\sim \mathcal{N}(0, a^2_{c:c(e)=c}) \end{aligned} \tag{1}$$

where x_e the estimate of the model for the stimuli e, $B_{e,s}$ models the bias b_s of subject s and its inconsistency v_s. $A_{e,s}$ reflects the content ambiguity $a_{c:c(e)=c}$ associated with the reference content c of the stimuli e under test.

It has been shown in the reference paper of SUREAL [13, 14] that the estimates of "MLE" and "MLE_CO_AP2" are more interpretable, compared to thresholds proposed in BT.500 and P.913, and robust to outlier/spammer or partial spammer behaviors (see reference papers for more details).In other words, the "MLE_CO_AP2" model improves classical MOS calculation by removing the inconsistency and bias from subjects. "MLE" instead of removing only the inconsistency and bias from raters, also estimates the content ambiguity, which constitutes valuable information for further analysis. In [15], a probabilistic graphical annotation model is proposed to infer the underlying ground truth and to model the annotator's behavior. We also reported the results of the post-screening rule from the ITU-R BS.1534-3 (MUSHRA) standard [10]: "If a subject scores the hidden reference below 90% for more than 15% of the test items, then all the scores of that subject for that test are discarded.".

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

Model	VSE test	VLE test	KE test
BT.500 [7]	-	-	-
P.913 [9]	'3'	-	-
Li [15]	'psub4', 'sub8', '3', 'psub7', '5'	'sub8', 'sub14' 'sub4'	'4', '6', 'sub02' 'sub17', 'sub12', 'sub03', 'sub25'
post-screening ITU-R BS.1534-3 [10]	'3', 'psub4', 'sub8', 'psub7', '8' 'sub4'	'sub7', 'sub14' 'sub4', 'sub6' 'psub9'	'4', '6', '7' 'sub17', 'sub7', 'sub18', 'sub5'
MLE [13]	'sub8', '3' '1', '5', '2'	'sub8', 'sub14' 'sub4'	'sub07' 'sub27' '7'
MLE_CO_AP2 [13]	'sub8', '2' '1', '5', '3'	'sub8', 'sub14' 'sub4'	'sub07', '7' 'sub27'

Table 2: Outliers found and rejected by BT.500 and P.913 compared to the top 3 or 5 subjects with highest inconsistency estimates from Li, MLE, and MLE_CO_AP2.

In table 2, we summarized the results from the different outlier rejection techniques. We can notice that BT.500 and P.913 never reject participants (except P.913 on the VSE test with rater '3'). The "Li" method rejects 3 participants in the VLE test. These participants correspond to the one with the highest inconsistency from the "MLE" and "MLE_CO_AP2" methods.

For the VSE test, 3 out of the 5 outliers detected by the "Li" method are also in the top 5 estimated highest inconsistency from the "MLE" and "MLE_CO_AP2" methods. For the KE test, outliers found by the "Li" method don't correspond to the ones of the "MLE" and "MLE_CO_AP2" methods.

It appears that it is necessary to use more advanced screening technics to remove biases and noise from outliers since procedures from audio and video standards retrieve either no outlier (i.e., BT.500 and P.913) or screen a lot of outliers (i.e., BS.1534-3). On top of that, the interpretability of the estimates from the state-of-the-art methods can lead to a clearer understanding of the rater's behavior, avoid complete removal of all the ratings of a rater, and can provide support for interesting analysis, as we will present in the next section.

3.2. Content Ambiguity and task difficulty

The "MLE" method from SUREAL provides a Content Ambiguity per SRCs of a dataset. These estimates can give insight into the difficulty of annotating contents, as they

Datasets	Mean CA	STD of Mean CA
VSE test	8.62	5.08
VLE test	6.05	2.15
KE test	8.04	3.86
AV video [5]	5.88	1.26
AV audio [5]	6.45	1.75
AV audiovisual [5]	9.12	0.91
SiSEC08 [16]	2.26	2.23
SiSEC18 [16]	7.22	2.88
SASSEC [16]	4.54	2.66
IRCCYN [17]	5.01	3.90
IRCCYN2 [17]	7.50	3.49

Table 3: Mean Content Ambiguity (CA) estimates from SUREAL MLE from datasets evaluated with SAMVIQ.

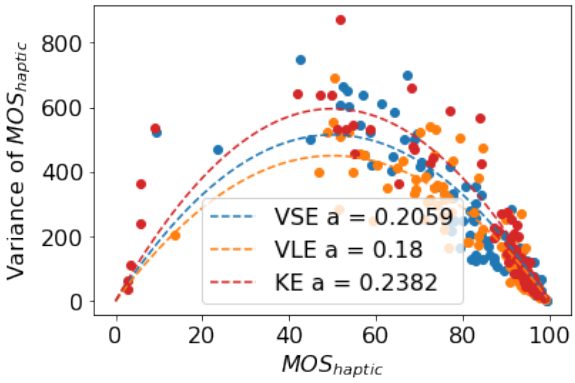

Fig. 2: SOS analysis for the three haptics datasets.

model the noise due to the subjectivity of the SRCs, which is linked to the type of modality (e.g., audio, video, or haptic). We gathered this information from various datasets; from the audio domain: "AV audio" from [5], and "SiSEC08", "SiSEC18", and "SASSEC" from [16]. We also used "IRCCYN" and "IRCCYN2" [17] and "AV video" [5] as representants for video and 360° video datasets. Finally, "AV audiovisual" [5] is a dataset where the audio and video quality of 360° stimuli are evaluated simultaneously. This type of evaluation can be considered highly subjective since it is a multiple-sensory experience with a high cognitive load on the subject. We selected these datasets since the SAMVIQ rating methodology uses a scale from 0 to 100.

In table 3, we reported the mean Content Ambiguity (mCA) across all the SRCs of each dataset: as a reminder, the "VSE" test has 8 SRCs, and each SRC has an estimated Content Ambiguity, the mean and standard deviation on these 8 values is reported in the table.

Two audio, a video, and the 360° audiovisual datasets have higher mCA than the "VLE" dataset. Similarly, for the "VSE" and "KE" datasets, the 360° audiovisual dataset has a higher mCA score, and they are in similar ranges as the "SiSEC18" and "IRCCYN2" datasets considering the standard deviation size.

We can conclude that the two haptic modalities evaluated in these three tests have high but reasonable mean Content Ambiguity scores compared to other multimedia modalities. The difficulty of annotating the quality of haptic feedback is comparable to other audio and video quality assessment tasks.

4. DISCRIMINABILITY ANALYSIS

In this section, we applied SOS analysis from [18] and discriminability analysis proposed in [19] to derive information on the quality of the data from these subjective tests.

4.1. SOS Analysis

By investigating perceptual quality scores and user rating diversity, we can explore the quality of the collected data. Therefore, we employed the Standard deviation of Opinion Scores (SOS) hypothesis, which postulates a quadratic relationship between the MOS and SOS[2], which depends only on one parameter α. We modified the equation formulated in [18] for Absolute Category Rating (ACR) use case 1 – 5 to our continuous rating scale 0 – 100, as done in [5]

$$\sigma^2(MOS) = \alpha(MOS - 0)(100 - MOS) \qquad (2)$$

In figure 2, we reported the parameter α on each dataset. The range [0.18; 0.23] is comparable to the α scores reported in [18] for video streaming subjective studies and smaller than the scores reported for web surfing [0.23; 0.28] or cloud gaming [0.27; 0.34] subjective evaluation datasets.

4.2. Subjective scores discriminability analysis

In [19, 5], the authors suggested examining the evolution of subjective scores discriminability with an increasing number of assessors to show how well a subjective methodology can retrieve accurate MOS scores. A two-sample Wilcoxon test is performed on all the possible pairs of stimuli of the dataset. The statistical test is applied between two haptic stimuli estimated MOS. An estimated MOS is computed from N randomly selected raters out of the M possible ones, and we plot the evolution of the percentage of significantly different pairs with an increasing value of N. A p_{value} of 0.05 is used to compute the percentage of significantly different pairs. The number of possible pairs for experiments "Vibrotactile Short Effects", "Vibrotactile Long Effects", and "Kinesthetic Effects" is 5356, 10153, and 2415 pairs, respectively. The result of the three haptic datasets is presented in Figure 3 with a 95% confidence interval over 100 simulations.

The curves show that the discriminability with few assessors is the same on the three datasets, around 0.35. However,

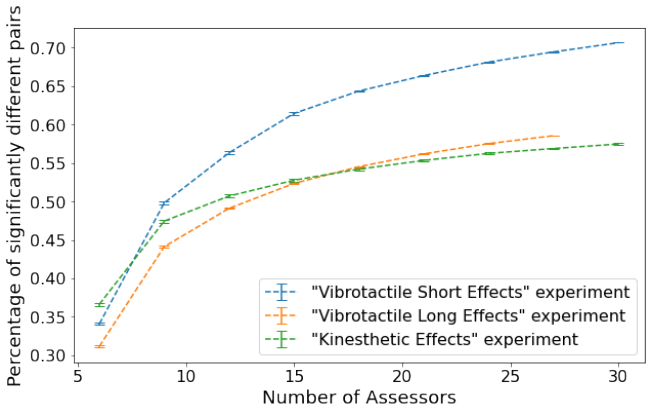

Fig. 3: The evolution of the percentage of significantly different pairs (y-axis) with an increasing number of assessors (x-axis) for the three experiments. The curves represent mean percentages, and the error bars represent 95% confidence intervals over 100 simulations.

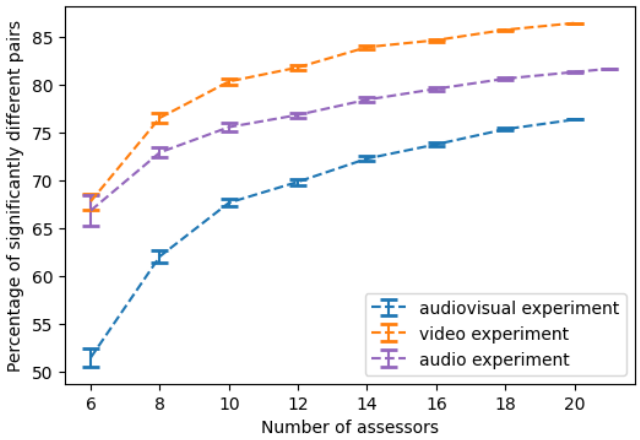

Fig. 4: The evolution of the percentage of significantly different pairs (y-axis) with an increasing number of assessors (x-axis) for the three experiments presented in [5]. Over 100 simulations, the curves represent mean percentages, and the error bars represent 95% confidence intervals.

by increasing the number of assessors, the discriminability rate of "Vibrotactile Short Effects" increase faster than the two other tests, as shown by the gap between curves. This could show that "Vibrotactile Short Effects" stimuli are easier to separate and evaluate by experts. Another comment relates to the "plateau effect" visible on each curve after 15, for VLE and KE tests, to 20 subjects, for the VSE test. This indicates that collecting more opinion scores will result in a small increase for the ratio of significantly different pairs.

In Figure 4, we reported the discriminability presented in [5], where the number of assessors follows standard audio recommendation [10]: "data from no more than 20 assessors are often sufficient for drawing appropriate conclusions from the test". We can observe a similar "plateau effect" after 15 raters on video and audio modalities of this work, in line with the recommendation of the standard.

From subjective assessment standard [9], the recommendation is "At least 24 subjects must be used for experiments conducted in a controlled environment". This number seems to be fitted for the modality with Vibrotactile Short Haptic feedback, where the discriminability is still increasing after 20 raters. For Vibrotactile Long and Kinesthetic types of Haptic feedback, the discriminability is already achieved at 20 raters, and 15 to 18 could be a better number.

5. CONCLUSION

In this work, we presented existing methods from the audio and video quality assessment domains and investigated how they perform for the quality assessment of compressed haptic signals. These analyses serve as a baseline towards guidelines for subjective haptic quality assessment.

We show that more advanced outlier screening methods are necessary to cure a haptic dataset. These methods can model the behavior of annotators and give interpretable estimates (i.e., annotator bias and inconsistency). We also investigated the difficulty of evaluating haptic quality for subjects by comparing with multiple datasets across different multimedia modalities Content Ambiguity. As a result, the Content Ambiguity of Haptic SRCs is on par with other modalities SRCs. This indicates that the task and methodology used in this work are suited.

We also suggest relying on 18 raters when evaluating Vibrotactile Long or Kinesthetic type of haptic feedback modalities and 21 for Vibrotactile Short haptic feedbacks: good numbers of assessors to evaluate haptic feedback quality and obtain a satisfying degree of discriminability.

6. REFERENCES

[1] Joo Chuan Yeo, Jiahao Yu, Zhao Ming Koh, Zhiping Wang, and Chwee Teck Lim, "Wearable tactile sensor based on flexible microfluidics," *Lab on a Chip*, vol. 16, no. 17, pp. 3244–3250, 2016.

[2] Cagatay Basdogan, Frédéric Giraud, Vincent Levesque, and Seungmoon Choi, "A review of surface haptics: Enabling tactile effects on touch surfaces," *IEEE transactions on haptics*, vol. 13, no. 3, pp. 450–470, 2020.

[3] Mike Sinclair, Eyal Ofek, Mar Gonzalez-Franco, and Christian Holz, "Capstancrunch: A haptic vr controller with user-supplied force feedback," in *Proceedings of the 32nd annual ACM symposium on user interface software and technology*, 2019, pp. 815–829.

[4] MPEG134, "Submissions and evaluation procedures for haptics cfp – phase 1," 2021, MDS20228 WG02 N0071.

[5] Randy F Fela, Andréas Pastor, Patrick Le Callet, Nick Zacharov, Toinon Vigier, and Soren Forchhammer, "Percep-

tual evaluation on audio-visual dataset of 360 content," in *2022 IEEE International Conference on Multimedia and Expo Workshops (ICMEW)*, 2022, pp. 1–6.

[6] Andréas Pastor, Lukáš Krasula, Xiaoqing Zhu, Zhi Li, and Patrick Le Callet, "On the accuracy of open video quality metrics for local decision in av1 video codec," in *2022 IEEE International Conference on Image Processing (ICIP)*. IEEE, 2022, pp. 4013–4017.

[7] ITU Recommendation BT.500-14, "Methodologies for the Subjective Assessment of the Quality of Television Images," 2019.

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

[9] 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.

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

[11] MPEG134, "Call for proposals on the coded representation of haptics – phase 1," 2021, MDS20227 WG02 N00070.

[12] 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.

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

[14] Zhi Li, Christos G. Bampis, Lucjan Janowski, and Ioannis Katsavounidis, "A simple model for subject behavior in subjective experiments," *CoRR*, vol. abs/2004.02067, 2020.

[15] Jing Li, Suiyi Ling, Junle Wang, and Patrick Le Callet, "A probabilistic graphical model for analyzing the subjective visual quality assessment data from crowdsourcing," in *Proceedings of the 28th ACM International Conference on Multimedia*, 2020, pp. 3339–3347.

[16] Thorsten Kastner and Jürgen Herre, "An efficient model for estimating subjective quality of separated audio source signals," in *2019 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*. IEEE, 2019, pp. 95–99.

[17] Stéphane Péchard, Romuald Pépion, and Patrick Le Callet, "Suitable methodology in subjective video quality assessment: a resolution dependent paradigm," in *International Workshop on Image Media Quality and its Applications, IMQA2008*, 2008, p. 6.

[18] Tobias Hoßfeld, Raimund Schatz, and Sebastian Egger, "SOS: The MOS is not enough!," in *2011 third international workshop on quality of multimedia experience*. IEEE, 2011, pp. 131–136.

[19] 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.