Perceptual evaluation on audio-visual dataset of 360 content

Randy Fela, Andreas Pastor, Patrick Le Callet, Nick Zacharov, Toinon Vigier, Soren Forchhammer

PERCEPTUAL EVALUATION ON AUDIO-VISUAL DATASET OF 360 CONTENT

Randy F Fela[1,4], Andréas Pastor[2], Patrick Le Callet[2], Nick Zacharov[3], Toinon Vigier[2], Søren Forchhammer[4]*

[1]SenseLab, FORCE Technology, 2970 Hørsholm, Denmark
[2]Nantes Université, Ecole Centrale Nantes, CNRS, LS2N, UMR 6004, F-44000 Nantes, France
[3]Meta Reality Labs – Meta, Paris, France
[4]DTU Electro, Technical University of Denmark, 2800 Kgs. Lyngby, Denmark

ABSTRACT

To open up new possibilities to assess the multimodal perceptual quality of omnidirectional media formats, we proposed a novel open source 360 audiovisual (AV) quality dataset. The dataset consists of high-quality 360 video clips in equirectangular (ERP) format and higher-order ambisonic (4^{th} order) along with the subjective scores. Three subjective quality experiments were conducted for audio, video, and AV with the procedures detailed in this paper. Using the data from subjective tests, we demonstrated that this dataset can be used to quantify perceived audio, video, and audiovisual quality. The diversity and discriminability of subjective scores were also analyzed. Finally, we investigated how our dataset correlates with various objective quality metrics of audio and video. Evidence from the results of this study implies that the proposed dataset can benefit future studies on multimodal quality evaluation of 360 content.

***Index Terms*—** omnidirectional media format, audiovisual dataset, 360 video, ambisonic, quality evaluation.

1. INTRODUCTION

Omnidirectional media formats (as 360 video with spatial audio) offers a more immersive experience than traditional audiovisual presentations, leading to an increasing level of adoption across multimedia service platforms [1]. To achieve a high-level user experience for multimedia services, perceived quality needs to be better understood, which is commonly evaluated through computational objective metrics validated by a series of subjective experiments. While multisensory evaluation is necessary to perform a higher degree of model prediction, a study in this integral quality is relatively unexplored due to the lack of multimodal dataset.

Public datasets of 360 video that contains subjective quality scores are IVQAD [2], VR-VQA [3], and VQA-ODV [4]. However, these datasets are limited partially due to the 4K video resolution [2] and/or sourced from streaming services with unknown quality control [3, 4]. Meanwhile, large spatial audio datasets can be found in projects of 3D-MARCo [5], EigenScape [6], and ARTE [7]. However, all mentioned datasets were mainly focused on a single modality and none of the spatial audio datasets performed audio quality evaluation.

To better understand the quality of audio-visual multimodal quality, a high quality multimodal dataset is required. For traditional multimedia applications such as 2D video and channel-based audio, several audiovisual datasets exist such as PLYM [8], TUM 1080p50 [9], VQEG [10], Vienna Made for Mobile [11], VTT [12], and INRS [13]. In comparison to traditional media formats, it is important to investigate immersive multimedia formats as they provide spatial information, allowing users an increased spatial experience. Several databases have been proposed and studied for different purposes such as attribute evaluation [14], developing media player [15], audio generation [16], and audiovisual attention [17]. However, there is a current shortage of high quality immersive audiovisual datasets and the absence of subjective quality scores.

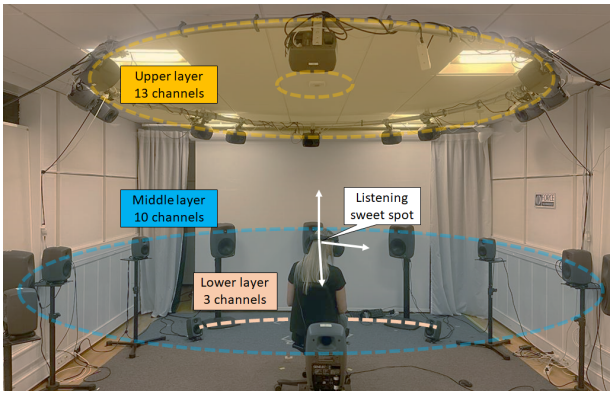

Fig. 1: Experimental setup for subjective experiments.

In this work, we present a dataset which consists of recorded 8K 360 video with higher/4^{th} order ambisonic (HOA) audio along with mean opinion scores collected from audio, video, and audiovisual subjective experiments. The focus of this paper is to present our investigation on subjective scores which include overall mean-CI analysis, SOS analysis benchmarked with existing datasets, and evolution of accuracy performance. After the investigation of subjective scores, a number of audio and video objective quality metrics were computed to investigate how these metrics perform in correlation to our subjective data, and to identify potential metrics towards the development of audiovisual quality metrics. To the best of our knowledge, this is the first recorded audiovisual dataset created to support perceptual quality research in immersive audiovisual content.

The remainder of this paper is organized as follows. Section 2 includes the description of the database, encoding and decoding steps to create processed sequences, and subjective evaluation procedures. We discuss our findings from subjective evaluations and objective quality metrics in Section 3. Finally, the conclusion is addressed in Section 4.

*The work was performed whilst the author was at FORCE Technology - SenseLab. Currently the author is at Meta Reality Labs.

Fig. 2: Equirectangular video previews of the proposed dataset.

2. MATERIALS AND EVALUATIONS

In this section, we describe HOA–SSR database, encoding parameters to create processed audio/video sequences, and evaluation of objective quality metrics and subjective data.

2.1. 360 audiovisual stimuli

This audiovisual (AV) dataset consists of 12 recorded audiovisual scenes selected from HOA–SSR database which contains immersive AV contents with unique characteristics i.e. nature–mechanical, indoor–outdoor, static–dynamic, traffic–quiet, impulsive–steady, and speech–music. The video scenes were recorded using an Insta360 Pro2, a professional spherical 360 camera, that consists of 6 camera lenses to capture every angle of a scene at once. The final video format for the dataset was provided in .mov container with 8K resolution (7680x3840), 30fps, 8-bit color depth, and in YUV422. The audio signals were recorded using the em32 Eigenmike, a spherical microphone with an array of 32 microphones. The output of these recordings was in a raw 32-channel ambisonic A-format then processed in 4th order ambisonic B-format AmbiX (25 channels) in ACN ordering and SN3D normalization. All audio files were in PCM 1152 kbps/channel, 24bit and 48kHz. In terms of spatial characteristics of a microphone, a previous study reported the highest directional accuracy of em32 compared to all other high order sound field microphones [18, 19]. The AV dataset used in this study as illustrated in Figure 2 is publicly available upon request[1].

In order to cover a use case with cinematic VR video, additional 4 AV stimuli were provided from joint work of Vtopia360[2] and VRTonung[3]. All videos were in the same quality and format as the HOA–SSR and the audio signals were recorded by using an ORTF-3D microphone[4] mixed into 24 channel NHK layouts and provided in ADM format. While in principle, the ORTF-3D provided superior localization accuracy [20], it was only compared to first order ambisonic format. Higher-order ambisonic format will increase the spatial resolution, hence improving localization accuracy. In our study, the use of these two types of recording techniques was considered equivalent since only internal quality (e.g. bitrate) was evaluated without any comparison of recording technique and assessment of attribute quality.

2.2. Stimulus preparation: Encoding and decoding

From the original raw format YUV422, all video sources were downscaled to a resolution of 6144x3072 and in YUV420 format due to the maximum limit of our playback system. We used libx265 (H.265/HEVC) in FFmpeg 4.4[5] to encode the source videos into three resolutions i.e. 6K (6144x3072), 4K (3840x1920), 2K (1920x1080) and 4 QPs (0, 22, 28, 34), resulting in 12 encoding parameters and 192 video stimuli in total. Meanwhile, the ambisonic audio sources were encoded from 32-channel A-format into 25-channel 4th order ambisonic in ambiX. All audio was encoded using FFmpeg with AAC-LC encoder into four different bitrates/channels (16kbps, 32kbps, 64kbps, PCM/reference) resulting in 64 audio stimuli in total. Due to the limitation of the channel number in the AAC encoder, the audio channels were split into six groups of 4-channel and 1 mono channel prior to encoding and re-grouped thereafter. Ambisonic audio files were decoded by using the All-Round Ambisonic Decoding (AllRAD) algorithm as proposed in [21] into 26 multichannel loudspeaker setups that follow the standard in [22]. AllRAD provides energy preservation across direction and average localization sharpness. Only decoding part was required for NHK audio format. For the experiments, 20s length out of 1 minute original duration was selected based on spectral frequency profile.

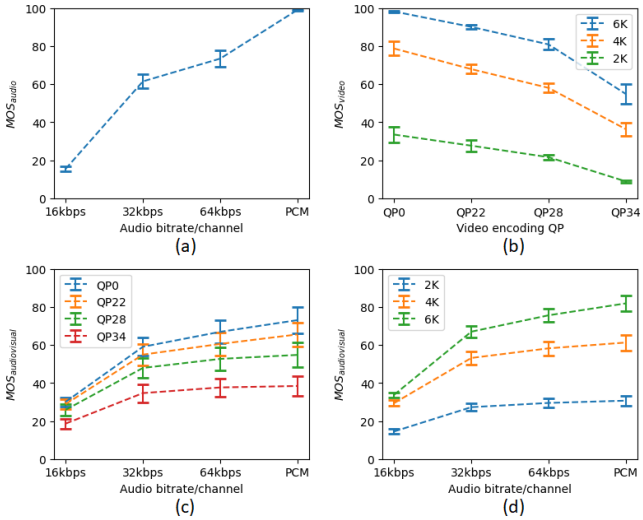

Fig. 3: Mean opinion score (CI 95%) of (a) audio, (b) video, and (c-d) audiovisual (AV).

2.3. Subjective evaluations

Twenty-one selected and trained assessors (13 males, 8 females) with age range 22 – 37 years (mean = 27.9, std = 4.0) were invited to

[1] https://bit.ly/HOA-SSR-Dataset
[2] https://vtopia360.com/
[3] https://www.vrtonung.de/en/
[4] https://schoeps.de/en/products/surround-3d/ortf-3d.html
[5] https://github.com/GyanD/codexffmpeg/releases/tag/4.4

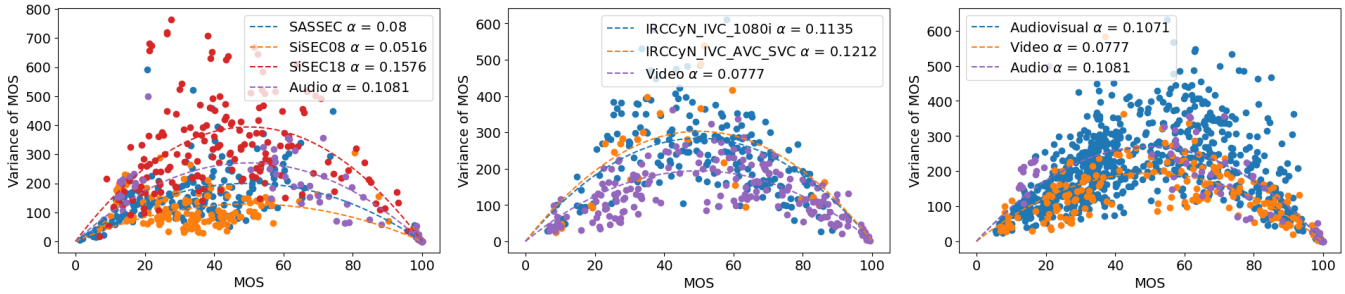

Fig. 4: SOS analysis for audio (left), video (middle) and audiovisual (right) experiments against existing datasets of the literature.

participate in three consecutive subjective experiments carried out at FORCE Technology SenseLab including I: listening/audio, II: viewing/video, and III: AV test. The experiments were performed in a standardized listening room that meets the acoustical requirements of EBU 3276 [23] and ITU-R BS.1116-3 [24], compliants for listening and VR experiment with head-mounted display. The experimental setup for AV experiment is depicted in Fig 1.

To avoid bias occurring between auditory and visual memory, the subjective test ordering was audio, then video, and finally AV experiment. We used SenseLabOnline 4.2 [25] as the user interface and ran double blinded-randomized trials. The participant was seated on a rotating chair, located in the acoustically sweet spot, and used a pad controller to perform the test. The user interface was displayed on a projection screen for experiment I and virtually in the Head Mounted Display (HMD) for experiments 2 and 3. Multiple stimulus with hidden reference without anchor (modified MUSHRA) was used to generalize the common evaluation methodology found in SAMVIQ [26] for video and MUSHRA [27] for intermediate audio quality. Each assessor was asked to rate their overall perceived quality of the audio, video, and audiovisual to a continuous rating scale between 0 and 100 categorized into 5 labels (Bad, Poor, Fair, Good, Excellent). We limited the number of stimuli on each trial set to 7 according to Miller's Law [28].

In experiment I, the participants rated the audio quality of sound stimuli reproduced over a 26-channel setup of 8040A Genelec loudspeakers. The sound level of stimuli were subjectively calibrated to 65 – 73 dB for most comfortable loudness, depending on the samples. Experiment II was a viewing test, where the task was to assess the perceived quality of 360 videos. Visual stimuli were displayed in VR using a Samsung Odyssey+ HMD which has a display resolution of 1440×1600 per eye, 110° horizontal field of view and 90Hz refresh rate. Finally, experiment III was an AV test where the participants rated the integrated AV quality in overall experience. One subject was withdrawn in the middle of experiment II due to a comfort issue, therefore the total number of participants was 20 subjects (12 males, 8 females; mean = 28.1, std = 4.0) for the last two experiments.

2.4. Evaluation of objective quality metrics

We evaluated 360 video quality metrics: S-PSNR [29], WS-PSNR [30], and CPP-PSNR [31]. In addition, we calculated VMAF [32, 33] and its components, VIF and DLM [34, 35] as shown in previous studies [36, 37]. Lastly, we calculated 2D image quality metrics: PSNR, SSIM[38], and MS-SSIM[39]. For audio quality metrics, we evaluated PEAQ [40, 41], ViSQOL [42, 43], and AMBIQUAL [44] for W-channel component of ambisonic as it represents both ambient and direct signals, and center channel of NHK format. Only listening quality feature of AMBIQUAL was computed. Three frequency bands were calculated for ViSQOL: denoted as $\text{ViSQOL}_{\text{nb}}$ (150 – 3400Hz), $\text{ViSQOL}_{\text{wb}}$ (50 – 8000Hz), and $\text{ViSQOL}_{\text{aswb}}$ (50 – 16000Hz).

3. RESULTS AND ANALYSIS

3.1. Perceived audio, video, and audiovisual quality

Figure 3 depicts the results obtained from audio, video, and AV subjective experiments to analyze perceptual difference between encoding parameters. The data is presented as Mean Opinion Score (MOS) over all stimuli with 95% confidence interval (CI). In listening test (Figure 3(a)), there is a large perceptual gap between 16 and 32 kbps, and between 64kbps and PCM, where PCM is the reference. Although the difference is statistically significant, mean score difference is perceptually small between 32 and 64kbps. Video quality test presented in Figure 3(b) shows by nature of perceived video quality over encoding parameters, that MOS score is, as expected, higher in lower QP and higher resolution, and vice versa. It can also be seen that there is a significant difference between QPs at each resolution. Finally, Figure 3(c-d) show the results from the audiovisual experiment averaged over resolutions and QPs in relation to audio bitrates, respectively. It is shown that perceived audiovisual quality difference in all bitrates is statistically different if video quality is higher (QP $\leq$ 28, resolution $\geq$ 4K).

3.2. SOS analysis

When investigating perceptual quality scores, one may gain the information from the collected data by observing user rating diversity. 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 [45] for Absolute Category Rating (ACR) use case 1 – 5 to our continuous rating scale 0 – 100,

$$\sigma^2(MOS) = \alpha(MOS - 0)(100 - MOS) \quad (1)$$

Several existing audio and video datasets were benchmarked with ours to investigate rating diversity in typical studies. We selected three datasets available in the SEBASS-DB database named SASSEC, SiSEC08 and SiSEC18 datasets originally collected for evaluation of audio source separation algorithms [46], and IRCCyN datasets for video quality evaluation case [47]. We focused on the datasets which use multiple stimulus methodologies, MUSHRA for audio and SAMVIQ for video. However, no available AV dataset exists with this rating paradigm.

Figure 4 summarizes SOS analysis for audio and video experiments benchmarked with existing datasets, and Figure 4(right) represents SOS analysis for our three experiments. Our observation is on the SOS parameter α. In audio experiments, α is 0.051 for

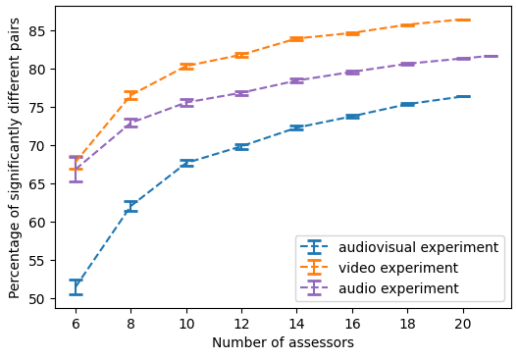

Fig. 5: The evolution of the percentage of significantly different pairs (y-axis) with an increasing number of assessors (x-axis) for the 3 experiments: audio, video, and AV. Over 100 simulations, the curves represent mean percentages and the error bars represent 95% confidence intervals.

SiSEC08 and 0.157 for SiSEC18. In comparison, our audio dataset has an α of 0.1081, which is in the range observed for audio studies with MUSHRA. The SOS score and the α value indicate that the dataset has consistent rating scores with low diversity.

For video datasets, α is in the same value range as both IRCCyN datasets, since α for our dataset is 0.077. Compared to our dataset, the diversity of user judgments in benchmark datasets is higher for scores between 20 and 60. By plotting our three datasets in Figure 4(right), even if the number of systems tested and signals rated are different, it is shown that the α values are small within the range of 0.077 and 0.108. These small values of α indicate the consistency of user experience quality, as we argue also as the benefit of using trained assessors [48, p. 105].

Overall, the difference for all tested datasets in this study is considerably low compared to previous audio studies, which ranges between 0.269 and 0.590 [46]. A wide range of benchmark studies also showed that the range of α in image and video QoE is 0.037 – 0.590. Nevertheless, the primary purpose of the SOS analysis in this study is to support comparisons and reliability checks between subjective studies. A large and in-depth benchmarking study is required in order to categorize the level of diversity based on α and SOS score, particularly in multiple stimulus rating methodologies.

3.3. Subjective scores discriminability analysis

As suggested in [49], we can examine the evolution of discriminability for subjective scores with an increasing number of assessors. A two-sample Wilcoxon test is performed on all the possibles pairs of stimuli and a p_{value} of 0.05 is used to compute the percentage of pairs significantly different. The number of possible pairs for audio, video, and AV experiments is 2016 pairs, 18336 pairs, and 294528 pairs, respectively. The result is presented in Figure 5 with 95% confidence interval over 100 simulations.

Regarding the overall trend of the curves, video and AV experiments show the highest and lowest discriminability, respectively. Audio and video tests started with similar discriminability, where audio CI is larger than video. However, by increasing the number of assessors, the rate of discriminability of video is higher than audio, which is shown by the gap between the two curves. This could be that the stimuli and step sizes between video stimuli are larger and easier for assessor compared to audio. For AV task, two modalities were used, making a larger cognitive load, thus the task of evaluating critically both AV was harder, hence the results in Figure 5. As previously investigated by [50], perceptual evaluation of single modality is less complex compared to multiple modalities.

Table 1: Performances of Full-Reference metrics to relation to the DMOS scores of the audio dataset. Bold best performance score and underlined scores are not significantly different from the best score.

Metric	PLCC	SRCC	KRCC	$C_0\%$	AUC_{BW}	AUC_{DS}
Ambiqual	0.878	0.905	0.746	0.933	0.989	0.916
PEAQ	0.753	0.753	0.552	0.815	0.944	0.851
$ViSQOL_{nb}$	0.864	0.884	0.720	0.914	0.970	0.851
$ViSQOL_{wb}$	0.897	0.912	0.755	0.938	0.981	0.885
$ViSQOL_{aswb}$	**0.924**	**0.938**	**0.800**	**0.958**	**0.995**	**0.921**

Table 2: Performances of Full-Reference metrics to relation to the DMOS scores of the video dataset. Bold best performance score and underlined scores are not significantly different from the best score.

Metric	PLCC	SRCC	KRCC	$C_0\%$	AUC_{BW}	AUC_{DS}
VMAF HD	0.859	0.927	0.767	0.924	0.973	0.760
VMAF 4K	**0.919**	**0.957**	**0.822**	**0.954**	0.983	0.828
VMAF B	0.915	0.955	0.816	0.952	0.988	0.824
VMAF Neg	0.917	0.957	0.819	0.954	**0.989**	**0.830**
DLM	0.893	0.938	0.787	0.935	0.980	0.788
VIF_{scale0}	0.770	0.765	0.586	0.826	0.910	0.691
SSIM	0.693	0.823	0.645	0.856	0.922	0.680
MS-SSIM	0.671	0.843	0.662	0.867	0.921	0.648
PSNR	0.616	0.719	0.538	0.800	0.891	0.678
S-PSNR	0.628	0.743	0.559	0.811	0.902	0.691
WS-PSNR	0.617	0.720	0.538	0.800	0.892	0.682
CPP-PSNR	0.622	0.731	0.547	0.805	0.897	0.686

3.4. Objective quality metrics

The correlation between objective metrics and subjective data is presented in Table 1 and 2, respectively, for audio and video quality metrics. We computed Pearson, Spearman and Kendall correlation coefficients. In addition, we ran statistical pairwise analysis on the performances of the different metrics, using the indicators presented in [51]: percentage of correct classification, $C_0\%$, from pairs with statistical significance differences, and AUCs from ROC analysis (AUC_{BW} and AUC_{DS}). In the tables, underlined metrics are not significantly different compared to the best performing metric reported in bold.

From table of audio quality metrics, the best performing metric on the audio dataset is $ViSQOL_{aswb}$, where we use the largest frequency bands in the calculation. PEAQ has the lowest score followed by AMBIQUAL which competing with $ViSQOL_{wb}$. It is well known that AMBIQUAL was specifically designed to compute listening quality and localization accuracy of ambisonic signal as the performance proved in [44] for low bitrate codec in 1st and 3rd order ambisonic. However, the performance of AMBIQUAL compared to other metrics, especially its predecessor (ViSQOL) remains unexplored.

For video quality metrics, VMAF 4K, and the bootstrapped version (VMAF B) outperforms other metrics. It is also interesting to see the performance of VMAF Neg on part with the previous one. This finding is supported by other studies that demonstrated the superiority of VMAF 4K in terms of perceptual correlation in 360 video compared to other video quality metrics [36, 37]. Metrics based on PSNR and directly optimized for 360 contents (S-PSNR, WS-PSNR, CPP-PSNR) are not providing any gain compared to their 2D counterparts computed on the ERP.

4. CONCLUSION

In this paper we present an audiovisual dataset comprising 360 video and ambisonic spatial audio with associated subjective scores. The work focused on the exploratory analysis of subjective data

to understand 1) the overall perceptual difference between each encoding parameter as perceived by assessors, 2) the span of subjective scores, and 3) the improvement of subjective scores accuracy as function of assessor number. Furthermore, we showed the performance of the dataset in relation to a set of objective quality metrics for audio and video.

The findings provided in this research confirm that there are perceptual differences for different encoding parameters. The SOS analysis confirms that our dataset has a low α value and variance for stimuli in the middle of the rating scale, proving the quality of the proposed dataset. We also found that the α value for audio part of the dataset is comparable to other works that used MUSHRA methodology. However, the threshold remain unclear for α value in this application and further benchmark analysis is required. In subjective scores discriminability analysis, video experiment was placed the highest, followed by audio and AV experiment. All curves have a low CI, with a steady trend after 12 number of assessors, confirming our choice of 20 assessors on each task. Lastly, objective metrics analysis concludes that $\text{ViSQOL}_{\text{aswb}}$ and VMAF 4K outperform other audio and video quality metrics, respectively.

The proposed dataset and findings in this research open new possibilities for future studies on primarily, but not limited to, AV quality evaluation in 360 videos with ambisonic spatial audio. Furthermore, the dataset from experiments can be used to advance existing objective quality metrics as well as propose a new ones by employing ML/DL based models. Our future work is to extend subjective and objective analysis together with the development of an AV perceptual quality model for 360 content.

5. ACKNOWLEDGEMENTS

We convey the acknowledgment to FORCE Technology, Bang & Olufsen, Demant, GN Store Nord, Sonova, WSA and industrial partners who created the 360 audio-visual datasets under the HOA-SSR joint project, and to XRHub Team for the great help in dealing with technicalities, filming and field recording. We also thank VRTonung and Vtopia360 for providing additional high-quality AV datasets for testing. This research was cofounded by Danish ministry of science and tech and the Marie Skłodowska-Curie grant agreement No.765911 RealVision.

6. REFERENCES

[1] B Choi, YK Wang, MM Hannuksela, Y Lim, and A Murtaza, "Information technology–coded representation of immersive media (MPEG-I)–part 2: Omnidirectional media format," *ISO/IEC*, pp. 23090–2, 2017.

[2] Huiyu Duan, Guangtao Zhai, Xiaokang Yang, Duo Li, and Wenhan Zhu, "IVQAD 2017: An immersive video quality assessment database," in *2017 International Conference on Systems, Signals and Image Processing (IWSSIP)*. IEEE, 2017, pp. 1–5.

[3] Mai Xu, Chen Li, Zhenzhong Chen, Zulin Wang, and Zhenyu Guan, "Assessing visual quality of omnidirectional videos," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 29, no. 12, pp. 3516–3530, 2018.

[4] Chen Li, Mai Xu, Xinzhe Du, and Zulin Wang, "Bridge the gap between VQA and human behavior on omnidirectional video: A large-scale dataset and a deep learning model," in *Proceedings of the 26th ACM international conference on Multimedia*, 2018, pp. 932–940.

[5] Hyunkook Lee and Dale Johnson, "An open-access database of 3D microphone array recordings," in *Audio Engineering Society Convention 147*. Audio Engineering Society, 2019.

[6] Marc Ciufo Green and Damian Murphy, "EigenScape: A database of spatial acoustic scene recordings," *Applied Sciences*, vol. 7, no. 11, pp. 1204, 2017.

[7] Adam Weisser, Jörg M Buchholz, Chris Oreinos, Javier Badajoz-Davila, James Galloway, Timothy Beechey, and Gitte Keidser, "The ambisonic recordings of typical environments (ARTE) database," *Acta Acustica United With Acustica*, vol. 105, no. 4, pp. 695–713, 2019.

[8] Mohammad Goudarzi, Lingfen Sun, and Emmanuel Ifeachor, "Audiovisual quality estimation for video calls in wireless applications," in *2010 IEEE Global Telecommunications Conference GLOBECOM 2010*. IEEE, 2010, pp. 1–5.

[9] Christian Keimel, Arne Redl, and Klaus Diepold, "The TUM high definition video datasets," in *2012 Fourth international workshop on quality of multimedia experience*. IEEE, 2012, pp. 97–102.

[10] Margaret H Pinson, Lucjan Janowski, Romuald Pépion, Quan Huynh-Thu, Christian Schmidmer, Phillip Corriveau, Audrey Younkin, Patrick Le Callet, Marcus Barkowsky, and William Ingram, "The influence of subjects and environment on audiovisual subjective tests: An international study," *IEEE Journal of Selected Topics in Signal Processing*, vol. 6, no. 6, pp. 640–651, 2012.

[11] Werner Robitza, Yohann Pitrey, Matej Nezveda, Shelley Buchinger, and Helmut Hlavacs, "Made for mobile: a video database designed for mobile television," in *Sixth International Workshop on Video Processing and Quality Metrics for Consumer Electronics (VPQM)*, 2012.

[12] Toni Mäki, Dragan Kukolj, Dragana Dordević, and Martín Varela, "A reduced-reference parametric model for audiovisual quality of IPTV services," in *2013 Fifth International Workshop on Quality of Multimedia Experience (QoMEX)*. IEEE, 2013, pp. 6–11.

[13] Edip Demirbilek and Jean-Charles Grégoire, "INRS audiovisual quality dataset," in *Proceedings of the 24th ACM international conference on Multimedia*, 2016, pp. 167–171.

[14] Olli Rummukainen, Jenni Radun, Toni Virtanen, and Ville Pulkki, "Categorization of natural dynamic audiovisual scenes," *PloS one*, vol. 9, no. 5, pp. e95848, 2014.

[15] Werner Bailer, Chris Pike, Rik Bauwens, Reinhard Grandl, Mike Matton, and Marcus Thaler, "Multi-sensor concert recording dataset including professional and user-generated content," in *Proceedings of the 6th ACM multimedia systems conference*, 2015, pp. 201–206.

[16] Aakanksha Rana, Cagri Ozcinar, and Aljosa Smolic, "Towards generating ambisonics using audio-visual cue for virtual reality," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 2012–2016.

[17] F. Chao, C. Ozcinar, C. Wang, E. Zerman, L. Zhang, W. Hamidouche, O. Deforges, and A. Smolic, "Audio-visual perception of omnidirectional video for virtual reality applications," in *IEEE International Conference on Multimedia & Expo Workshops (ICMEW)*, 2020.

[18] Enda Bates, Marcin Gorzel, Luke Ferguson, Hugh O'Dwyer, and Francis M. Boland, "Comparing ambisonic microphones —- Part 1," in *2016 AES International Conference on Sound Field Control*. Audio Engineering Society, 2016.

[19] Enda Bates, Sean Dooney, Marcin Gorzel, Hugh O'Dwyer, Luke Ferguson, and Francis M Boland, "Comparing ambisonic microphones -— Part 2," in *Audio Engineering Society Convention 142*. Audio Engineering Society, 2017.

[20] Catherine Guastavino, Véronique Larcher, Guillaume Catusseau, and Patrick Boussard, "Spatial audio quality evaluation: comparing transaural, ambisonics and stereo," Georgia Institute of Technology, 2007.

[21] Franz Zotter and Matthias Frank, "All-round ambisonic panning and decoding," *Journal of the Audio Engineering Society*, vol. 60, no. 10, pp. 807–820, 2012.

[22] ITU-R Rec. BS.2159-7, "Multichannel sound technology in home and broadcasting applications," 2015.

[23] EBU. Tech. 3276 - 2nd edition, "Listening conditions for the assessment of sound programme material: Monophonic and two–channel stereophonic," 1998.

[24] ITU-R Rec. BS.1116-3, "Methods for the subjective assessment of small impairments in audio systems including multichannel sound systems," 2015.

[25] Guillaume Le Ray and Junaid Khalid, "SenseLabOnline: Combining agile database administration with strong data analysis," in *The R User Conference, useR!*, 2013, vol. 10, p. 38.

[26] 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.

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

[28] George A Miller, "The magical number seven, plus or minus two: Some limits on our capacity for processing information.," *Psychological review*, vol. 63, no. 2, pp. 81, 1956.

[29] Matt Yu, Haricharan Lakshman, and Bernd Girod, "A framework to evaluate omnidirectional video coding schemes," in *2015 IEEE International Symposium on Mixed and Augmented Reality*. IEEE, 2015, pp. 31–36.

[30] Yule Sun, Ang Lu, and Lu Yu, "Weighted-to-spherically-uniform quality evaluation for omnidirectional video," *IEEE Signal Process. Lett.*, vol. 24, no. 9, pp. 1408–1412, 2017.

[31] Vladyslav Zakharchenko, Kwang Pyo Choi, and Jeong Hoon Park, "Quality metric for spherical panoramic video," in *Optics and Photonics for Information Processing X*. International Society for Optics and Photonics, 2016, vol. 9970, p. 99700C.

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

[33] R. Rassool, "VMAF reproducibility: Validating a perceptual practical video quality metric," in *Proc. IEEE Int. Symp. Broadband Multimedia Syst. Broadcast. (BMSB)*. IEEE, 2017, pp. 1–2.

[34] Hamid R Sheikh and Alan C Bovik, "Image information and visual quality," *IEEE Transactions on image processing*, vol. 15, no. 2, pp. 430–444, 2006.

[35] Songnan Li, Fan Zhang, Lin Ma, and King Ngi Ngan, "Image quality assessment by separately evaluating detail losses and additive impairments," *IEEE Transactions on Multimedia*, vol. 13, no. 5, pp. 935–949, 2011.

[36] Marta Orduna, César Díaz, Lara Muñoz, Pablo Pérez, Ignacio Benito, and Narciso García, "Video Multimethod Assessment Fusion (VMAF) on 360VR contents," *IEEE Trans. Consum. Electron.*, vol. 66, no. 1, pp. 22–31, 2019.

[37] Randy Frans Fela, Nick Zacharov, and Søren Forchhammer, "Perceptual evaluation of 360 audiovisual quality and machine learning predictions," *arXiv preprint arXiv:2112.12273*, 2021.

[38] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli, "Image quality assessment: from error visibility to structural similarity," *IEEE transactions on image processing*, vol. 13, no. 4, pp. 600–612, 2004.

[39] Zhou Wang, Eero P Simoncelli, and Alan C Bovik, "Multiscale structural similarity for image quality assessment," in *The Thirty-Seventh Asilomar Conference on Signals, Systems & Computers, 2003*. IEEE, 2003, vol. 2, pp. 1398–1402.

[40] ITU-R Rec. BS.1387-1, "Method for objective measurements of perceived audio quality," 2001.

[41] Thilo Thiede, William C Treurniet, Roland Bitto, Christian Schmidmer, Thomas Sporer, John G Beerends, and Catherine Colomes, "PEAQ – The ITU standard for objective measurement of perceived audio quality," *J. Audio Eng. Soc.*, vol. 48, no. 1/2, pp. 3–29, 2000.

[42] 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.

[43] Colm Sloan, Naomi Harte, Damien Kelly, Anil C Kokaram, and Andrew Hines, "Objective assessment of perceptual audio quality using ViSQOLAudio," *IEEE Trans. Broadcast.*, vol. 63, no. 4, pp. 693–705, 2017.

[44] Miroslaw Narbutt, Jan Skoglund, Andrew Allen, Michael Chinen, Dan Barry, and Andrew Hines, "AMBIQUAL: Towards a quality metric for headphone rendered compressed ambisonic spatial audio," *Appl. Sci.*, vol. 10, no. 9, pp. 3188, 2020.

[45] 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.

[46] 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.

[47] 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.

[48] Søren Bech and Nick Zacharov, *Perceptual audio evaluation-Theory, method and application*, John Wiley & Sons, 2007.

[49] 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.

[50] Randy Frans Fela, Nick Zacharov, and Søren Forchhammer, "Towards a perceived audiovisual quality model for immersive content," in *Proceedings of the 2020 Twelfth International Conference on Quality of Multimedia Experience (QoMEX)*, 2020, pp. 2472–7814.

[51] 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)*. IEEE, 2016, pp. 1–6.