R. F. Fela, N. Zacharov, and S. Forchhammer, "Assessor Selection Process for Perceptual Quality Evaluation of 360 Audiovisual Content"
J. Audio Eng. Soc., vol. 70, no. 10, pp. 824–842, (2022 October).
DOI: https://doi.org/10.17743/jaes.2022.0037

Assessor Selection Process for Perceptual Quality Evaluation of 360 Audiovisual Content

RANDY FRANS FELA,[1, 3] *AES Student Member*,

NICK ZACHAROV,[2, †] *AES Fellow* **AND SØREN FORCHHAMMER**[3, *]

[1]*SenseLab, FORCE Technology, Hørsholm, Denmark*
[2]*Meta Reality Labs, Paris, France*
[3]*Deptartment of Electrical and Photonics Engineering, Technical University of Denmark, Kgs. Lyngby, Denmark*

For accurate and detailed perceptual evaluation of compressed omnidirectional multimedia content, it is imperative for assessor panels to be qualified to obtain consistent and high-quality data. This work extends existing procedures for assessor selection in terms of scope (360° videos with high-order ambisonic), time efficiency, and analytical approach, as described in detail. The main selection procedures consisted of a basic audiovisual screening and three successive discrimination experiments for audio (listening), video (viewing), and audiovisual using a triangle test. Additionally, four factors influencing quality of experience, including the simulator sickness questionnaire, were evaluated and are discussed. After the selection process, a confirmatory study was conducted using three experiments (audio, video, and audiovisual) and based on a rating scale methodology to compare performance between rejected and selected assessors. The studies showed that (i) perceptual discriminations are influenced by the samples, the encoding parameters, and some quality of experience factors; (ii) the probability of symptom occurrence is considerably low, indicating that the proposed procedure is feasible; and (iii) the selected assessors performed better in discrimination than the rejected assessors, indicating the effectiveness of the proposed procedure.

0 INTRODUCTION

Omnidirectional media have gained popularity because they provide users with a more exploratory experience with at least three degrees of freedom. Media categorized as omnidirectional include 360° images and videos, spatial audio, and associated timed text during presentation, e.g., movie subtitles [1]. Standardization of the storage and streaming format for omnidirectional media has been under development by the Moving Picture Experts Group since 2017 and was clearly presented by [2].

One of the research directions in the field of omnidirectional media is the quantification of user-perceived quality through the use of predictive computational metrics and perceptual evaluations in a laboratory setting. Because standardization in this area is still ongoing, the methods developed for experimentation in early studies had to be adapted from earlier standards intended for traditional media. For example, *Recommendations ITU-R BT.500* [3], *ITU-T P.910* [4], and *ITU-T P.913* [5] were adopted for studies in 360° video, and *Recommendations ITU-R BS.1116* [6] and *ITU-R BS.1534* [7] for spatial audio research. For 360° video, a newly developed standard for a subjective evaluation protocol for head-mounted displays [8] was recently published. The intrinsic method, an evaluation method based on the reference materials, is a common practice found in both audiovisual and omnidirectional media quality evaluation. The technical aspect was to apply different encoding schemes, e.g., bitrates, frequency sampling in audio, and bitrates, quantization, and resolution in video, evaluated by objective metrics or subjective evaluation with full reference. Several typical studies using intrinsic methods can be found, for example, in [9–13].

On the other hand, the number of participants and their expertise are also critical for ensuring data quality in perceptual evaluation [14]. Early research in 360° video used a number of at least 15 [3, 4] or 24 [5] naive assessors, depending on the standard followed during the experiment. Especially for the evaluation of 360° videos, the latest

[*] To whom correspondence should be addressed, e-mail: sofo@fotonik.dtu.dk.

[†] This work was performed while the author was with FORCE Technology—SenseLab, but he is now employed by Meta.

standard recommends that 28 participants should be registered to achieve sufficient statistical power [8]. Additionally, in perceptual audio evaluation, a slightly different paradigm was created with at least 20 experienced assessors that are required because of the complexity of the task [6, 7]. Moreover, it is clear that the experience level of the assessors could affect the experimental results [15]. Given this different situation between audio and video domains, the question arises for the audiovisual domain: "How many assessors and what level of expertise are appropriate for unimodal and multimodal experiments if we use the same assessor panel?" This question is straightforward, but proposing appropriate solutions is not trivial.

Perceptual evaluation of multimedia quality is a challenging problem because multiple factors (e.g., user, system, and context factors) interact to influence the perception of overall quality among these multiple influencing factors. In the case of immersive multimedia systems (e.g., 360° videos with spatial audio), these influencing factors are increased primarily by the opportunity to explore the audiovisual event attentively during the evaluation and by the user's expectations about the quality level of the given stimulus. The former has an influence on user behavior and specific attention during the evaluation, which can be observed by head and eye movements on a given viewport. The latter may be due to the fact that the highest-quality 360° video is only compressed to about 75% because of projection.[1] Without a qualified assessor, this condition will affect the quality of the data captured during the experiment. As reported by Fela et al. [12], having an assessor panel that had expertise in perceptual audio evaluation did not span the results in perceptual quality scores for spatial audio and 360° video evaluation.

The rationale for this study is that efforts reported assessing the perceived quality of immersive multimedia content are limited. Additionally, there is a lack of a method to improve the qualification of an assessor panel, which may affect the quality of subjective data in the evaluation of immersive multimedia quality. The technical aspect of "multimedia quality" that this study focuses on is due to compression with a set of coding parameters that are consistent with the common evaluation procedure for 360° videos [16] and previous work [9–12, 17, 18]. Undoubtedly, there are other parameters that influence the perceived quality and quality of the experience, such as the quality and number of speaker channels, choice of headphones, arrangement of speaker channels and video display, playback method, and position of the assessor (standing-sitting, still-exploring, etc.).

Accordingly, the contribution of this paper is threefold. First, the authors proposed a practical framework for selecting assessors based on their performance during the screening process before inviting them to perform perceptual evaluation studies in omnidirectional audiovisual multimedia applications. The process aims to transform the naive assessor into a selected assessor as classified in [19], potentially increasing the reliability and robustness of the data for a future test. Second, the audiovisual material used in this study was made available for future work. The audiovisual scenes were time-synchronized and selected based on their temporal and spatial characteristics. Third, the data obtained from the selection process were analyzed primarily to answer the research questions (RQs) addressed in this study:

- RQ1: What are the effects of audio and video encoding parameters on perceptual discrimination results?
- RQ2: What are the effects of audio, video, and audiovisual content on perceptual discrimination results?
- RQ3: How is the assessor's performance on the perceptual discrimination test of audio, video, and audiovisual content?
- RQ4: How is it concluded that the proposed selection procedure is effective?

To answer questions RQ1−RQ3, three discrimination tests were conducted for audio, video, and audiovisual content. A frequency analysis and statistical methods were used to test the hypotheses made in relation to the RQ(s). Additionally, factors affecting the quality of the experience and simulator sickness were also examined to observe the effects of test samples on level of interest, difficulty of judgment, dizziness, and presence. Furthermore, a confirmatory study consisting of audio, video, and audiovisual experiments was conducted to answer RQ4. The evaluations were made using a rating scale method to compare the performance of small groups of failed and successful assessors.

The remainder of this paper is organized as follows: SEC. 1 reviews the state of the art in the selection of assessors developed in the field of sensory science and their application in multimedia perceptual studies. SEC. 2 describes the proposed procedure for forming a group of selected assessors for immersive audiovisual studies. Experimental results are discussed in SEC. 3; a confirmatory study along with the results and analysis are described in SEC. 4; and finally, SECS. 5 and 6 draw the conclusion and outlook of the study.

1 RELATED WORKS

Panel selection methods have been widely used in the sensory science for decades and are mainly used in the food and beverage industry. There are two ISO standards for the selection, training, and monitoring of panel assessors for the training [20] and expertise [21] of sensory assessors, which have recently been unified in *Standard 8586:2012* [19], which also describes the terminology of classifications of sensory assessors in relation to their expertise.

In the field of audiovisual quality evaluation, an early attempt to study panel selection dates back to the work of Hansen [22] and Toole [23], which was later followed by Bech [24, 25], Olive [26], Mattila et al. [27], Isherwood et al. [28], Florian et al. [29], Legarth et al. [30], Sontacchi et

[1] The highest quality 360° video recorded at 8K is perceptually 2K. Normally, other compression schemes are applied before delivery to users.

al. [31], and Kuusinen et al. [32]. In general, the framework defined in previous studies can be divided into three aspects, namely motivation and application, proposed procedure, and technical aspects. The formation of a panel of assessors is primarily motivated by the long-term use of the same group of assessors for a given task, with the expectation of consistent repeatability of the subject's ratings [24, 25]. Having considered the selected assessors as such, one could also investigate a number of different tasks to improve their skills in performing a broader range of subjective tests [24, 26].

Procedural aspects typically include pre-selection using a questionnaire, auditory and/or visual screening test, and series of subjective experiments. In the pre-selection phase, the questionnaire is used to understand the background of the candidates and their range of interests in terms of their potential as members of the assessor panel. The candidates recruited for the pre-selection questionnaire should be at least four times the target panel size, as suggested in ISO 8586-1 [20]. Hansen [22] and Mattila et al. [27] designed some examples of questionnaires for listening test applications, and Legarth et al. [30] for multimodal test applications. One of the most popular selection procedures was formulated in Generalized Listener Selection (GLS) [27], which has been widely used in perceptual audio evaluation with some modifications regarding augmentation [28], test tuning [28, 29], extension of reliability analysis [31], and multisensory approach [30]. It should be noted that some items in the GLS questionnaire could benefit from updating to remain relevant as immersive technology evolves.

Basic auditory and/or visual screening has been shown to be essential for assessor selection. A study conducted by Toole showed that listeners with a near-normal hearing threshold had the highest agreement and least individual variation [23]. It was also recommended that the selected assessors must be otologically normal subjects and that the hearing threshold level should not exceed 15 dB at any audiometric frequency [24, 27]. Different studies have reported different hearing threshold level criteria, ranging from <20 dB [28, 29] to any frequency or allowing only one frequency per ear exceeding 20 dB [30, 32]. As with image quality, screening is also considered necessary for visual acuity, because it is strongly correlated with perceived image quality, as reported by Ravikumar et al. [33]. Additionally, color blindness and stereopsis tests are other visual screening tests suitable for multisensory applications [30].

Depending on the purpose of the panel, a number of perceptual tests may be performed in the final stage of the selection process, including but not limited to a test of loudness, speech quality, and audio quality [27, 30], stereo width [29, 31], timbre quality [24, 31], image compression and brightness [30], and verbal fluency skills [30, 29, 32]. Ghani et al. [34] took a different approach, using a battery of psychoacoustic discrimination tests, e.g., intensity and frequency detection, masking level difference, interaural level/time difference, and gap detection, to predict panel members' abilities to judge sound quality. However, because of the low predictive accuracy with respect to assessor performance on the listening tests, this approach does not provide significant benefit.

Technical aspects of the experiment include the equipment used for audio and image/video display, test material/sample, and evaluation methods. Loudspeaker playback in screening procedures has been used by several predecessors, for example, in the work of Bech [24, 25], Isherwood et al. [28], Florian et al. [29], and Kuusinen et al. [32] to evaluate timbral quality, spatial discrimination, stereo width discrimination, and acoustic properties of concert halls. Meanwhile, sound reproduction over headphones has been widely used in typical evaluations for loudness, speech quality, audio quality, and stereo width [27, 30, 31, 34]. Similarly, various playback devices can be used for 360° video, such as a 2D monitor, mobile-based and standalone head mounted displays (HMDs), and CAVE-like displays [35]. The results of previous studies suggested that the presentation of 360° video in virtual reality (VR) mode via HMD is highly preferable and could have a positive impact on the user's spatial awareness and enjoyment [9, 35].

The effects of program material have also been studied in audio, for example, in the work of Hansen [22], how differences in recording techniques for the same pieces of music can profoundly affect the perception of quality when listening with particular pairs of loudspeakers. In evaluating video quality, Mirkovic et al. [36] found that the category of content can influence content-specific characteristics such as the user's familiarity with and expectations of the content (cognitive component), elicited emotions (affective component), and intention to repeat and recommend the content (conative component). User interest was cited as a factor contributing to the overall viewing experience after technical aspects [37]. A similar result was found by Jumisko et al., whereby test participants tended to rate familiar content lower than unfamiliar content and rate interesting content higher than that considered uninteresting [38].

To determine an assessor's ability to perceive differences between audiovisual samples, various discrimination test methods have often been used. These methods include the pairwise comparison [22, 27, 28, 31], Three Alternative Forced Choice (3AFC) [34, 29], and the triangle test [30, 32]. The elementary nature of these methods makes them relatively easy for assessors with little or no training. Pairwise comparison (PC) is commonly used because it requires minimal understanding and training on the part of the assessor. At PC, multiple repetitions are required to measure subject reliability over time, which can later be analyzed using intra-rater reliability and inter-rater agreement [27, 31]. 3AFC and triangle methods are useful for increasing the objectivity of test results. In 3AFC, the assessors observe the highest or lowest intensity, whereas in the triangle test, they compare the sensory distance between stimuli. Interested readers should refer to [39] for a comparative analysis between these two methods.

The triangle test [40] was chosen for the screening procedure because it is easy for naive assessors to understand, and there is a correct answer for each trial, so it addresses the problems associated with PC highlighted in [27]. In the

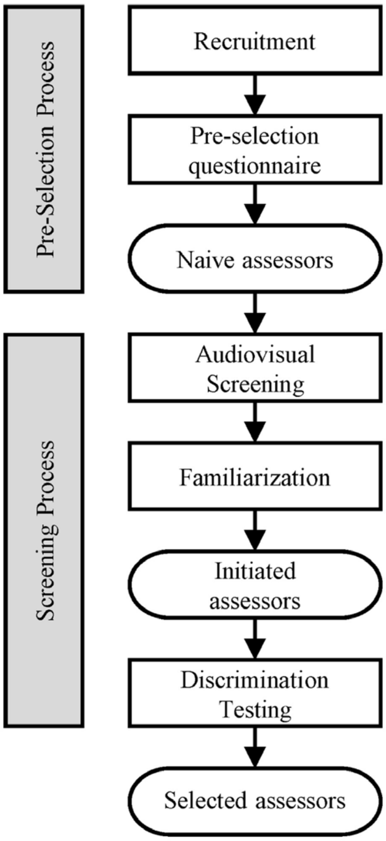


Fig. 1. Flowchart of a proposed assessor selection process.

triangle test, a triad of stimuli is presented, two of which are identical. The task is to find the odd sample from each group of three. It is at the discretion of the investigator whether a full combination of conditions or a subset is presented to the assessors. An analysis of the triangle test can also be expressed as a percentage of correct answers to provide some degree of absoluteness.

2 SELECTION PROCEDURE

This section describes the framework of the assessor selection process proposed in this study, as shown in Fig. 1. The framework consists of two stages, including 1) pre-selection and 2) audiovisual screening. The pre-selection stage aims to stimulate interest in studying the perceptual quality of immersive multimedia content at the level of the consumer or the so-called naive assessor. The audiovisual screening stage then consists of basic audiovisual screening, familiarization, and discrimination tests.

Stage 2 has two goals, namely 1) to measure the basic physiological abilities of the assessor through tests for things such as audiometry, color blindness, visual acuity, visual stereoscopy, basic binary choice, and familiarization to the test procedure and stimuli and 2) to assess the intrinsic discrimination ability of the assessor through perceptual quality tests. By fulfilling the first objective, the naive assessor can be considered an "initiated assessor," because they already know a specific task and the test procedure [19]. The discrimination test is used to measure their interest and improve their ability to recognize perceptual quality. Those who pass the discrimination test can be considered a "selected assessor," who are then expected to be more sensitive to the quality of omnidirectional multimedia content. The two-stage selection process within the proposed framework is described in the following subsections.

2.1 Stage 1: Pre-Selection

An online questionnaire was formulated and published in advance of the pre-selection. A modified version of the GLS questionnaire formulated in [30] was proposed to be more appropriate for current technological developments and the focus of the study. The questionnaire contained several types of questions and consisted of several sections, including personal data, health status, previous experience with audio and/or visual experiments, interest in audio and visual products, experience with VR/360° video, and assessor availability. The formulated questionnaire used in this study is shown in Table 1.

In this case, the target size of the panel was 20−25 assessors, so the number recruited should be at least four times the target size according to *Standard 8586-1* [20]. A total of 106 volunteers (57 males and 49 females) responded to a pre-selection questionnaire with an age range between 18 and 66 years (mean = 31.5; SD = 8.9) and represented different professions and nationalities. Applications were filtered based on these criteria:

- Age between 18−50 years old,
- No reported hearing damage,
- No reported visual damage,
- No color-blindness,
- Commitment to complete the study unless there was a reasonable condition for withdrawal, and
- Availability to participate in the tests during or outside working hours.

2.2 Stage 2: Basic Audiovisual Screening

Eighty-nine volunteers (47 males and 42 females) aged 18–50 years (mean = 30.1; SD = 6.3) met the screening criteria, of whom 44 volunteers were excluded for Stage 2 because they withdrew, had health problems, or did not respond to the invitation. Finally, 45 volunteers were invited as participants in Stage 2 and completed the audiovisual screening test. The test participants were 23 males and 22 females, aged 18−43 years (mean = 29; SD = 4.7), with different nationalities.[2] Stage 2 included basic audio and visual screening tests, as described below.

2.2.1 Audiometry

A pure-tone audiometry test was performed in a low-noise (Noise Rating 10) listening room using an Interacoustics AD 229e audiometer and calibrated Sennheiser HDA 200 headphones. The test procedure was based on the

[2] Please note that certain assessor panels may require a group of native speakers, as described in [30], but this was not a requirement for this panel.

Table 1. List of questionnaire items for pre-selection stage.

Category	ID	Question	Type
Personal data		First name	Short text
		Last name	Short text
		Address (not required)	Short text
		Post nr. + City	Short text
		E-mail	Short text
		Phone (not required)	Short text
		Native language (Beginner, Intermediate, Advanced)	Short text
		English proficiency	Multiple choice
		Age	Short text
		Gender (Male, Female)	Multiple choice
		Years of education (including elementary school)	Short text
		Current profession (if student, please specify the field)	Short text
Health	1	Do you have a known history of hearing damage?	Yes/No
	2	Do you have a visual impairment that can't be corrected by your glasses?	Yes/No
	3	Do you wear glasses in daily basis?	Yes/No
	4	Are you colorblind? (Yes, Partially, No)	Multiple choice
Experience	5	Have you previously participated in a listening test?	Yes/No
	6	If yes, how many times within the last 2 years?	Short text
	7	Please give a short description of these tests, if known.	Long text
	8	Have you previously participated in a viewing test? (watching video to rate the quality, preference, etc.)	Yes/No
	9	If yes, how many times within the last 2 years?	Short text
	10	Please give a short description of these tests, if known.	Long text
	11	Have you previously participated in VR test? (watching video to rate the quality, experience, preference, immersion, game quality, etc.)	Yes/No
	12	If yes, how many times within the last 2 years?	Short text
	13	Please give a short description of these tests, if known.	Long text
Sound	14	Do you listen to music, or podcast, or audiobook?	Yes/No
	15	Do you attend music concerts, operas, ballets, theater?	Yes/No
	16	Do you play a musical instrument or sing?	Yes/No
	17	Do you consider yourself a critical listener?	Yes/No
	18	Do you notice sounds in your environment or from products?	Yes/No
	19	Do you own a hi-fi system?	Yes/No
	20	Do you own a surround sound system?	Yes/No
	21	Do you know 3D / spatial audio?	Yes/No
	22	Are you professionally or academically involved in audio or acoustics?	Yes/No
Video	23	Do you have a TV?	Yes/No
	24	Do you take photos with a handy camera (mirrorless, DSLR, SLR)?	Yes/No
	25	Do you edit your pictures?	Yes/No
	26	Do you watch DVD movies at home?	Yes/No
	27	Do you watch from IP streaming provider (Netflix, Prime, HBO, etc.)?	Yes/No
	28	How often do you go to the cinema? (Weekly, Monthly, Seasonal, Yearly, Never)	Multiple choice
	29	Do you consider yourself a critical viewer?	Yes/No
VR/360 video (experience)	30	Do you know 360 video?	Yes/No
	31	Do you own a VR Glasses / head-mounted display (HMD)?	Yes/No
	32	If yes, estimate the value of the HMD	Short text
	33	Do you own a spherical / omnidirectional / 360 camera?	Yes/No
	34	If yes, estimate the value of the camera	Short text
	35	Have you watched 360 video from offline media player (VLC, WMP, etc.)?	Yes/No
	36	Have you watched 360 video from online media player (YouTube, Vimeo, etc.)?	Yes/No
	37	Have you watched a cinematic Virtual Reality?	Yes/No
	38	Do you use HMD to watch 360 video?	Yes/No
Availability	39	Are you working every weekday during working hours∗?	Yes/No
	40	Are you working on the weekend during working hours?	Yes/No
	41	Would you be available for the test during working hours?	Yes/No
	42	Would you be available for the test after working hours?	Yes/No
	43	Would you be available for the weekend test if necessary?	Yes/No
	44	Would you be committed to the SenseLab audiovisual test between August – September for 4 – 6 sessions within this period?	Yes/No
Consent		I have read the description of the study and how my personal data is used by SenseLab, I accept these terms.	Accept and Submit

Working hours is Monday – Friday at 9.00 – 17.00.

Hughson-Westlake method [41] (ascending method) and complied with *Standard 8253-1* [42], with a threshold determination procedure at 10 dB for the frequencies 125 Hz, 250 Hz, 500 Hz, 1 kHz, 2 kHz, 4 kHz, and 8 kHz.

2.2.2 Visual Tests

Visual ability was assessed with three basic vision tests, e.g., the color blindness, visual acuity, and stereopsis tests. Additionally, a short basic video quality test with binary choice (pairwise comparison) was added to provide participants with the basic idea of a discrimination test. The description of each test is explained as follows.

- **Color blindness**: The Ishihara test [43] was performed to detect color blindness based on red-green color deficiencies.
- **Depth perception**: The stereopsis test was performed using the RANDOT[3] stereopsis test method [44] to assess a subject's depth perception. Each subject was seated and asked to wear polarized glasses to read the RANDOT stereopsis book, which was placed on the table with natural light illuminance. Individuals with a vision prescription were asked to continue wearing their glasses/contact lenses with the polarized glasses.
- **Visual acuity**: Visual acuity testing was performed using Freiburg Vision Test (FrACT) 3.10.5 software developed by Michael Bach.[4] FrACT uses psychometric methods in combination with anti-aliasing and dithering to provide an automated, self-paced assessment of visual acuity. [45]. FrACT measures the visual angle of the smallest perceived structure, or the so-called minimum angle of resolution (MAR), which is measured on a logarithmic scale. MAR can be referred to as visual loss, and the number can be converted to Visual Acuity decimals (VA_{dec}). This test was performed in the listening room using a Thinkpad X230 laptop with a screen size of 12.5 in and an HD300 nit display. The screen brightness was set to 150 cd/m^2 and the backlight was set to 100 lux. The screen resolution was calibrated based on an observer screen distance of 104 cm described in [45].

2.2.3 Basic Video Quality Test

A basic test was performed to discriminate video quality by pairwise comparison. The source material (SRC), named "Students Looming Across Street," abbreviated as (*ss*), was downloaded from a public dataset, the LIVE Mobile Video Quality Dataset [46]. Stimuli were generated by compressing SRC into various constant rate factors 1−51. In the test, assessors were asked to state their name and click on the video that was perceived as having the lowest quality among a pair of videos. There were 30 trials in which the system had an adaptive difficulty level depending on the previous response. The user interface of this test is shown in Fig. 2.

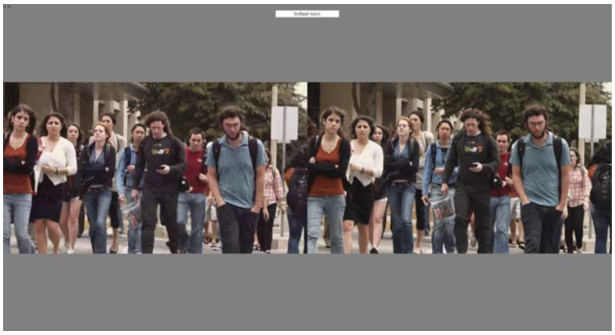


Fig. 2. Illustration of a pair of visual stimuli used in basic video quality with pairwise comparison (left panel is degraded).

2.3 Stimuli

The video SRC used in this study was captured with a professional VR camera Insta360 Pro2,[5] a spherical 360° camera consisting of six lenses that capture multiple angles of a scene at once. Audio was recorded using an em32 Eigenmike, a spherical microphone array that allows recording of acoustic signals from 32 array microphones. During recording, the camera was mounted below the HOA microphone on the customized mounting rig. An attempt was made to match the height of the rig so that the height of the camera and microphone were appropriate to simulate a first point of view. The final raw video format had 8K (7,680 × 3,840) resolution, 30 fps, 8-bit color depth, and YUV 4:2:2 chroma subsampling.

The output of audio recording was in a raw 32-channel ambisonic A-format, which was then converted to a fourth-order Ambisonic B-format AmbiX (25 channels) with Ambisonic Channel Number and SN3D normalization. All audio files were in Pulse-Code Modulation (PCM) format with 24 bits and 48 kHz sampling rate. Regarding the spatial characteristics of the microphone, a previous study reported that the em32 has the highest directional accuracy compared with all other high-order sound field microphones [47, 48]. The SRCs used in this study, as shown in Fig. 3, are publicly available by request from the Higher-Order Ambisonic Sound Scene Repository (HOA-SSR) Database project page [17].[6]

2.4 Encoding

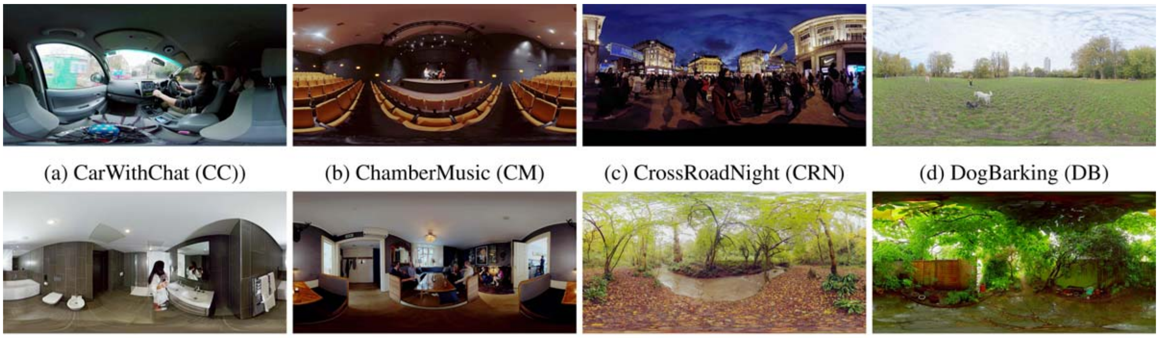
Stimuli were created using the audio-video encoding process. Before video encoding, all video SRCs were converted to raw YUV422 format and downgraded to playable YUV420 format. Video stimuli were created with libx265

[3] https://www.stereooptical.com/products/stereotests-color-tests/randot/.

[4] https://michaelbach.de/fract/.

[5] Kandao Qoocam 8K camera was used for contents recorded in small spaces, e.g., clip "CarWithChat (CC)."

[6] https://bit.ly/HOA-SSR-Dataset.



(a) CarWithChat (CC)) (b) ChamberMusic (CM) (c) CrossRoadNight (CRN) (d) DogBarking (DB)

(e) HairDrying (HD) (f) LoungeBar (LB) (g) BurblingStream (BS) (h) NatureThunder (NTR)

Fig. 3. Equirectangular preview of audiovisual samples used in the study. (a) CarWithChat (CC), (b) ChamberMusic (CM), (c) CrossRoadNight (CRN), (d) DogBarking (DB), (e) HairDrying (HD), (f) LoungeBar (LB), (g) BurblingStream (BS), and (h) NatureThunder (NTR).

(H.265/High Efficiency Video Coding) in FFmpeg using three quantization parameters (QPs; 22, 27, and 37) and three video resolutions (1,920 × 1,080; 3,840 × 1,920; and 6,144 × 3,072).

For the audio files, the first 16 channels were extracted from the original files, representing ambisonic audio files third order. Audio encoding was performed in FFmpeg using Advanced Audio Coding–Low Complexity (AAC-LC) in five bitrates/channels (16, 24, 64, and 128 kbps and in the original PCM format). All audio files were decoded into a 26-channel speaker system using the All-Round Ambisonic Decoding algorithm proposed in [49], following the standard layout in [50]. The All-Round Ambisonic Decoding algorithm is designed to be robust even with irregular loudspeaker configurations (such as the one used in the study), providing on average good energy conservation over all directions and good localization acuity. The level of audio playback for each encoded file was measured at a listening position and subjectively calibrated by experts to be between 47.9 and 66.7 dB for the most comfortable listening levels [51], depending on the samples.

2.5 Stage 2: Discrimination Tests

Three consecutive discrimination experiments were conducted for the listening, viewing, and audiovisual tests as a combination of the first two. As for the triangle test, six balanced triads (AAB, ABA, ABB, BAB, BAA, and BBA) should be presented to the assessors according to *Standard 4120* [40] when conducting the test to determine the stimuli A and B. The theoretical basis of this method has been well discussed, e.g., in [52–54]. A random triad was selected to be presented for each triad of stimulus. In general, the number of trials N_{trials} performed in the experiment can be calculated by

$$N_{trials} = \frac{n_{system} - 1}{2} \times n_{system} \times n_{sample} \times n_{repl}, \qquad (1)$$

where n_{system} refers to the encoding parameters, n_{sample} is a number of audiovisual samples, and n_{repl} is the number of repetitions of the same triad. The number of triads and their representations used for each discrimination experiment are described as follows.

2.5.1 Experiment 1: Audio Quality Discrimination

From the HOA-SSR database, two audio excerpts [DogBarking (DB) and NatureThunder (NTR)] were used for the familiarization tasks, and four excerpts [HairDrying (HD), BurblingStream (BS), CarWithChat (CC), and CrossRoadNight (CRN)] were used for the samples for all experiments. With the motivation to enrich the samples with musical character, two audio excerpts [Organ (ORG) and Acappella (ACP)] were added from the 3D Microphone Array Comparison recording dataset [55]. The selected recording files were recorded at St. Paul's Concert Hall in Huddersfield using EigenMike em32. The dataset was in raw ambisonic A-format, allowing signal processing identical to the HOA-SSR dataset. The encoding process resulted in 10 triads for each excerpt, for a total of 60 triads [Eq. (1)].

2.5.2 Experiment 2: Video Quality Discrimination

There were nine combinations of encoding parameters in the video (three resolutions and three QPs). After calculating with Eq. (1), 21 of 36 possible triads were selected for this experiment in terms of approximate difficulty level according to the estimated perceptual distance, resulting in a total number of 84 triads.

2.5.3 Experiment 3: Audiovisual Quality Discrimination

The samples and configurations used in the audiovisual part were identical to those used in the video quality part. Three audio bitrates (24, 64, and 128 kbps) and three video resolutions (1,920 × 1,080; 3,840 × 1,920; and 6,144 × 3,072) were paired to produce different audiovisual levels. Twenty-one out of 36 possible triads were selected for testing, resulting in a total of 84 triads. A pilot experiment was conducted to determine the character of the stimulus selected and the level of difficulty for each triad in the audio, video, and audiovisual discrimination test. A pair of systems for testing the triads and estimated difficulty levels are described in Tables 2−4.

All discrimination experiments were conducted in the SenseLab listening room at FORCE Technology, which is compliant with EBU Technical Report 3276 [56] and

Table 2. Pairs for each stimulus triad in the triangle test for audio quality discrimination test.

Pair	System A	System B	Level
A01	16 kbps	24 kbps	*
A02	16 kbps	64 kbps	*
A03	16 kbps	128 kbps	*
A04	16 kbps	PCM	*
A05	24 kbps	128 kbps	*
A06	24 kbps	PCM	*
A07	24 kbps	64 kbps	**
A08	64 kbps	PCM	**
A09	64 kbps	128 kbps	***
A10	128 kbps	PCM	****

PCM = Pulse-Code Modulation.

Table 3. Pairs for each stimulus triad in the triangle test for video quality discrimination test.

Pair	System A	System B	Level
V01	qp22_4k	qp37_4k	*
V02	qp22_6k	qp37_6k	*
V03	qp27_2k	qp27_4k	*
V04	qp27_2k	qp27_6k	*
V05	qp27_4k	qp37_4k	*
V06	qp27_6k	qp37_6k	*
V07	qp22_6k	qp27_4k	*
V08	qp27_4k	qp37_6k	*
V09	qp22_4k	qp37_6k	*
V10	qp22_6k	qp27_2k	**
V11	qp27_6k	qp37_4k	**
V12	qp22_6k	qp37_4k	**
V13	qp27_2k	qp37_4k	**
V14	qp27_2k	qp37_6k	**
V15	qp22_4k	qp27_2k	**
V16	qp37_4k	qp37_6k	**
V17	qp27_4k	qp27_6k	***
V18	qp22_4k	qp27_6k	***
V19	qp22_4k	qp22_6k	***
V20	qp22_4k	qp27_4k	***
V21	qp22_6k	qp27_6k	***

Recommendation ITU-R BS.1116-3 [6]. To avoid the occurrence of bias between auditory and visual memory, the test order was audio, then video, and finally audiovisual. SenseLabOnline 4.2 [57] was used to conduct the randomized double-blind trials and was used as the user interface during testing. For all tests, the subject sat on a swivel chair located in the acoustic sweet spot and was given the pad controller to perform the test. The user interface was projected onto the acoustically transparent screen for Experiment 1 and virtually projected into the HMD for Experiments 2 and 3.

2.6 Selection Criteria

For the selection process, assessors with normal hearing and vision who could pass the video quality test with over 50% correct responses were selected. Additionally, intrinsic criteria should also be considered, such as personality and personal motivation or enthusiasm [30]. The selection criteria for hearing and vision are defined as follows.

Table 4. Pairs for each stimulus triad in the triangle test for audiovisual quality discrimination test.

Pair	System A	System B	Level
AV01	2k_128kbps	4k_128kbps	*
AV02	2k_128kbps	6k_24kbps	*
AV03	2k_128kbps	6k_128kbps	*
AV04	2k_128kbps	4k_64kbps	*
AV05	4k_24kbps	6k_64kbps	*
AV06	4k_128kbps	6k_24kbps	*
AV07	4k_64kbps	6k_24kbps	*
AV08	2k_128kbps	4k_24kbps	*
AV09	2k_128kbps	6k_64kbps	*
AV10	4k_24kbps	6k_128kbps	**
AV11	4k_24kbps	4k_128kbps	**
AV12	4k_24kbps	4k_64kbps	**
AV13	6k_24kbps	6k_64kbps	**
AV14	6k_24kbps	6k_128kbps	**
AV15	4k_64kbps	6k_64kbps	**
AV16	4k_128kbps	6k_128kbps	**
AV17	4k_128kbps	6k_64kbps	***
AV18	4k_64kbps	6k_128kbps	***
AV19	4k_24kbps	6k_24kbps	***
AV20	6k_64kbps	6k_128kbps	***
AV21	4k_64kbps	4k_128kbps	***

2.6.1 Hearing

A normal level of hearing was expected. Criteria defined by Legarth and Zacharov [30], where a person's hearing level (HL) should be $\leq$15 dB HL for all frequencies, was adopted. However, a deviation of 20-dB HL for one frequency per ear was considered acceptable.

2.6.2 Vision

- No deficiency was detected in the test of color vision.
- Stereopsis was better than 250 s of arc and preferably better than 50 s of arc.
- A visual acuity of 1.0 $\mathrm{VA}_{\mathrm{dec}}$ or higher was preferable, which is equivalent to the 0.00 LogMAR or 20/20 Snellen test (ft). However, $\mathrm{VA}_{\mathrm{dec}} \geq 0.8$ or equal to Snellen 20/25 was also considered normal visual range according to [58] and is therefore still valid.

2.6.3 Basic Video Quality Test

A basic test of video quality was conducted to familiarize participants with a specific video quality task. Although it was a simple test with a pairwise comparison, it had a steep compression distance and the presentation order was highly adaptable depending on the previous response. This condition can be very challenging for first-time users, and therefore high expectations were not set. Participants had to answer at least 50% of the responses correctly to pass.

2.6.4 Discrimination Tests

In the discrimination experiments with the triangle test, difficulty is related to perceptual distance, resulting in the encoding parameters of a stimulus pair shown in Tables 2−4. There were four levels in Experiment 1 and three levels in Experiments 2 and 3, where levels 1–4 can be expressed as 1: easy, 2: moderate, 3: difficult, and 4: very

Table 5. Questionnaires related to quality of experience (QoE) factors.*

Q	Questions and rating scale
Q1	How interesting is the sound you listen to from the content being tested? How is your visual interest in the content being tested? (1 = Boring, 2 = Uninteresting, 3 = Neutral, 4 = Interesting, 5 = Intriguing)
Q2	In your opinion, how easy was it to give an answer to the selected audio file? In your opinion, how easy was it to give an answer to the selected video file? (1 = Very easy, 2 = Easy, 3 = Medium, 4 = Difficult, 5 = Very difficult)
Q3	How is the level of dizziness or nausea for each content during the VR viewing experiment? (1 = Very dizzy, 2 = Dizzy, 3 = Slightly dizzy, 4 = No dizzy, 5 = Absolutely no dizzy)
Q4	Please rate how you feel the "Presence" was for each content. Presence can be interpreted as a sensation of being in the video environment that you watched. (1 = Not at all, 5 = High presence)

difficult. Participants were not expected to give 100% correct answers on all tests, but there was a minimum score that they had to achieve. The total number of correct responses was expected to be at least as high as the total number of pairs for an easy and moderate level for Experiment 1, and an additional 50% of difficult levels for Experiments 2 and 3. Thus, the minimum threshold was 80%, 85%, and 85% for Experiments 1, 2, and 3, respectively.

2.7 Factors Affecting Quality of Experience

Factors affecting quality of experience (QoE) when rating immersive multimedia were examined, e.g., personal interest in each sample, difficulty in giving a rating, level of dizziness and presence. In addition, the impact of the entire selection process on cybersickness was investigated using the simulator sickness questionnaire (SSQ).

2.7.1 Human–Content Related Aspects

The questions were adopted from [9] and presented to participants to rate each sample using a five-point categorical rating scale, as shown in Table 5. Questions were asked after Experiments 1 and 2 for Q1 and Q2, after Experiments 2 and 3 for Q3, and after Experiment 3 for Q4 only.[7] The mean opinion score (MOS) with a 95% confidence interval (CI) was calculated as the sum of R_n, the individual score for a given stimulus of subject n, divided by the total number of subjects N.

$$MOS = \frac{\sum_{n=1}^{N} R_n}{N}. \quad (2)$$

2.7.2 Simulator Sickness Questionnaire

Due to the fact that the experiments were lengthy and involved exposure to multiple tests/modalities with omnidirectional media, a simulator sickness questionnaire (SSQ) was also evaluated using the SSQ in [59], which consists of 19 associated symptoms rated on a 0–3 rating scale (0 = None, 1 = Slight, 2 = Moderate, and 3 = Severe). The detailed use of the SSQ in VR research was described in [60].

3 RESULTS OF THE ASSESSOR SELECTION PROCEDURE EXPERIMENT

This section presents the results of the screening process in Stage 2, including audiometry, visual screening, and discrimination testing. The discrimination test data were statistically analyzed, and the method and results are also discussed. To conclude the assessor selection procedure proposed in the study, the selected assessors are indicated by their SubjectID (S01−S45). An analysis of the factors affecting the QoE is also presented.

3.1 Audiometry

Using the hearing threshold defined in Sec. 2.6.1, 39 and 42 participants passed the audiometry test for the left and right ears, respectively. Thirty-nine participants (18 males and 21 females), aged 18–43 years, passed the test for both ears (mean = 28.0; SD = 4.4). Figs. 4(a) and 4(b) show the HL curve for each frequency test for each ear, and Fig. 4(c) shows the mean curves for both ears of selected participants with a 95% CI.

3.2 Visual Tests and the Basic Video Quality Test

The results of the Ishihara color blindness test showed that none of the participants were color-blind. In the visual acuity test using FrACT, the results of VA_{dec} varied in the range of 0.75−1.34 and a LogMAR between 0.13 and −0.13. Note that a low number of VA_{dec} or high number of LogMAR indicates a loss of visual acuity. With a minimum threshold of 0.9 VA_{dec}, 42 participants (23 males and 19 females) with an age range of 18−43 years (mean = 28.7; SD = 4.8) passed the visual acuity test. Finally, 42 participants (21 males and 21 females) with an age range of 22−43 years (mean = 28.8; SD = 4.6) passed the video quality test. Of the 42 participants, eight participants had a score of ≥25, and 23 participants had a score between 20 and 24. The highest score was 27 and was achieved by participant S45.

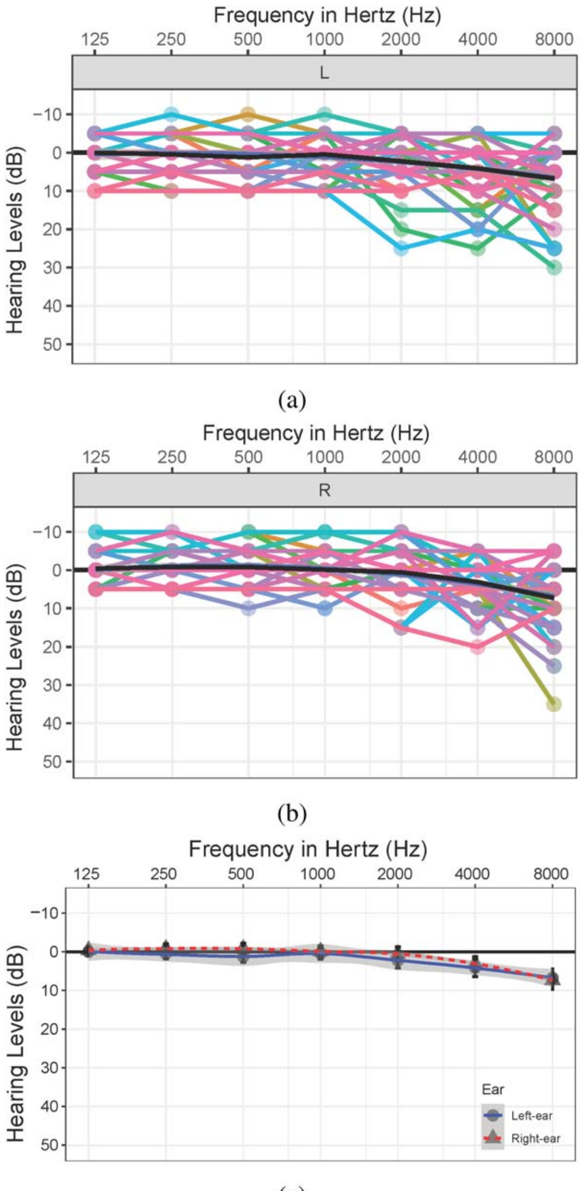
3.3 Discrimination Tests

In connection with the answers to questions RQ1−RQ3, two additional questions were posed related to the analysis of the discrimination test:

- What is the percentage score distribution as a function of sample?
- Do individual participants rate each triangle triad similarly?

In general, the triangle discrimination test can be analyzed with the percentage of correct answers by dividing the number of correct answers of the pair X given by asses-

[7] Pre-test: before Experiment 1; mid-session: after Experiment 2; end-session/post-test: after Experiment 3.



(a)

(b)

(c)

Fig. 4. Audiogram hearing level (HL) in decibels for (a) left ear and (b) right ear and (c) mean HL [with 95% confidence interval (CI)] in both ears of selected assessors.

sor (n) by the total number of assessors (N), as in Eq. (3).

$$\%_{correct} = \frac{\sum_{n=1}^{N} X_n}{N} \times 100\%. \tag{3}$$

The number of pairs X was 60 in Experiment 1 and 84 in Experiments 2 and 3. The use of absolute results, as shown in Eq. (3), was demonstrated by Legarth and Zacharov [30] and Kuusinen et al. [32] for each test performed in the study. Additionally, a statistical analysis was performed in [32] using a cumulative binomial probability and a Z-test.

The analysis based on the binomial test only, however, can be refuted because of a Type I error, which is due to the variance between trials. Therefore, the β-binomial model, which may provide greater benefit in dealing with overdispersion if it exists, was employed. The β-binomial model compounds β and binomial distributions and assumes that probability of a correct answer, p_c, in the binomial distribution follows a β distribution with parameters α and β. The probability function of the β-binomial distribution is expressed as

$$Pr(x) = \binom{n}{x} \frac{\Gamma(\alpha + x)\Gamma(\beta + n - x)\Gamma(\alpha + \beta)}{\Gamma(\alpha + \beta + n)\Gamma(\alpha)\Gamma(\beta)}, \tag{4}$$

where $\Gamma()$ denotes the gamma function, x denotes the number of correct answers, and n denotes the number of trials on given observations, $\alpha > 0$, $\beta > 0$, and $x = 0, 1, ..., n$. The parameters α and β can be re-parameterized into a mean μ of the binomial parameter p_c, $\mu = \alpha/(\alpha + \beta)$, and a scale parameter γ measuring the variation of p_c, $\gamma = 1/(\alpha + \beta + 1)$. In order to estimate the parameters μ and γ, maximum likelihood estimation can be used. Interested readers should follow the computational steps in [61].

From here, the statistical difference test of the two systems (of each encoding pair), i.e., a test of the null hypothesis H_0: $\mu = \mu_0$ against the alternative hypothesis H_A: $\mu \neq \mu_0$, can be performed. Hypothesis tests were computed using the **sensR** package with R. According to [62], in this case, the parameters are $\mu = p_c$ and $\mu_0 = p_{c_0} = 1/3$, where the hypothesis tests are:

$\boldsymbol{H_0 : p_c \leq p_{c_0}}$, there is no difference between two systems (encoding parameters).

$\boldsymbol{H_A : p_c \geq p_{c_0}}$, there is a significant difference between two systems (encoding parameters).

Fig. 5 shows the results of the responses to RQ1 and RQ2 and the questions regarding the percentage of correct answers and perceived discrimination of the participants to the given stimuli during the experiment. Data were calculated for each sample and as a combination of all samples. The p value was calculated using the β-binomial difference test based on the null and alternative hypotheses. The p values were divided into six levels to determine the strength of discrimination, where $p \leq 0.0001$ represents high discrimination or that the system pair is relatively easy to discriminate (reject $\boldsymbol{H_0}$, accept $\boldsymbol{H_A}$), and conversely, a p value close to or equal to 1 means that the task of finding the difference was relatively difficult (accept $\boldsymbol{H_0}$).

As for the percentage of correct answers in Fig. 5, in Experiment 1 (audio), it is clear from the participants' point of view that the pair A01−A07 is distinguishable, as shown by the results of >75% for the individual and total samples. There is also agreement for the pair A10, for which only about 30% of the participants could find the unique stimulus. It is argued that pair A10 is very similar in terms of quality, as evidenced by both the percentage level and $p = 1$. A10 includes 128 kbps AAC-LC and PCM, where it is very difficult to tell the difference between these two bitrates. Note that the difference between 128 kbps and PCM depends on the content of the sample and the criticality of the sample. In a previous study reported by [63], it was found that the objective difference grade calculated by the PEAQ algorithm for the AAC codec is −0.15 (the objective difference grade range is from −4 to 0, where 0 represents perceptually lossless quality) [64]. Also, a subjective listening test presented in [65] showed that there was

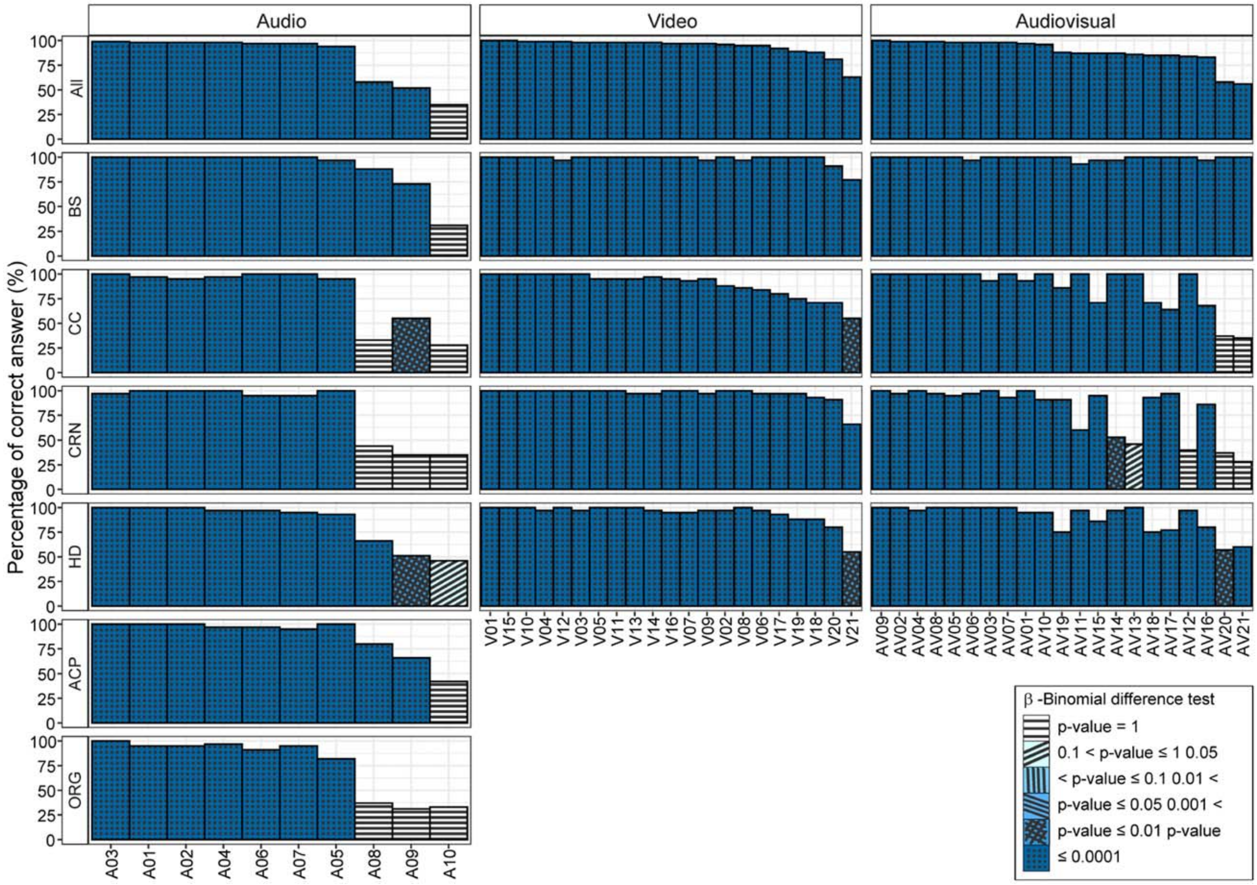


Fig. 5. Bar plot of the percentage of correct answers (%) in relation to *p* value from β-binomial difference test for each pair of system.

no statistical difference between the 160 kbps AAC-LC and uncompressed audio files.

On the other hand, the pair A08 and A09 has a lower percentage compared to A01−A07, whose percentages vary depending on the sample. The stimuli with pure music (ORG), ambient noise (CRN), and low-frequency ambient noise with speaker (CC) were more difficult to distinguish. The audio sample has a contribution to the difficulty level corresponding to 1) the type of sample (music, ambient sound, speech, white noise, singing, and nature sounds) and 2) the frequency distribution [wide-range (HD), specific range (speech/CC), low-frequency dominance (BS and CRN)].

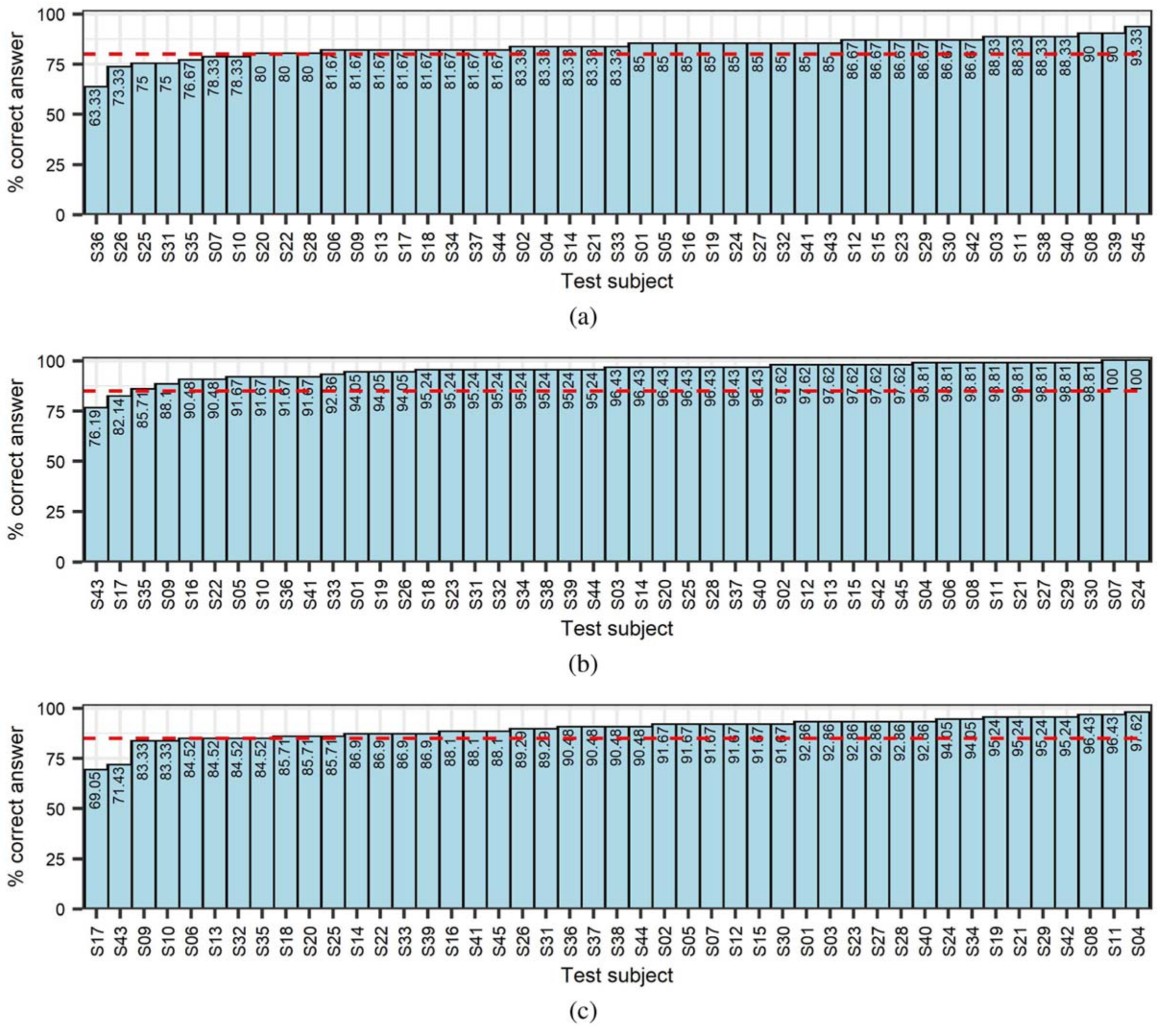
In Experiment 2 (360° video), the percentage decreases with the number of paired systems and varies with the sample. The pair V17−V20 has a lower percentage compared to the pair <V17, except for the clip BS, where the percentage decreases from V20. It is noticeable that V21 has the lowest percentage in all samples and has $0.001 < p \leq 0.01$ in the clips CC and HD, which means that the significant difference between the two systems was still noticeable. According to Table 3, it can be said that the pair containing 4K, 6K, QP22, and QP27 as encoding parameters in the system has a lower discrimination level or is less different.

Finally, the results in Experiment 3 (audiovisual) showed a similar trend where the percentage decreases across the system triad. However, the sample analysis shows a large variation, for example, in the clip BS, whose results are relatively high for all pairs (∼90%). Clips CC and HD show a similar distribution when trending downward, except for CC AV18 and AV19. In comparison, clip CRN shows a different distribution because of the lower values in AV11−AV14.

See Table 4; AV11 and AV12 both have 4K video resolution, but they differ in audio bitrate (AV11: 24 vs. 128 kbps; AV12: 24 vs. 64 kbps). They have a similar pair as AV14 and AV13, except that the video resolution is 6K for AV13 and AV14. It appears that clip CRN is less discriminating than other clips, which will be discussed later in QoE factor analysis. This is because clip CRN consists of an outdoor scene with multiple objects and sound sources. A fact that A11>A14 and A12<A13 can be argued that assessors paid more attention when evaluating lower quality stimuli. This can be even more difficult when the distinction between audio is smaller (e.g., the small distance between audio bitrates), as in AV20 and AV21, which compare 64 kbps with 128 kbps at the same video resolution (6K and 4K for AV20 and AV21, respectively). In terms of *p* value, the authors argue that clip CRN has lower discriminative power and lower percentage in many pairs with $p = 1$ in AV11, AV20, and AV21 and $0.1 < p \leq 1$ in AV13.

3.4 Selected Panels

Because one of the main goals of this study was to propose a framework for selecting assessors for immersive



(a)

(b)

(c)

Fig. 6. Percentage of correct answers for each assessor on the discrimination tests of (a) audio, (b) video, and (c) audiovisual.

audiovisual quality experiments, here, assessors' performance during their participation in all three experiments (RQ3) is evaluated. The performance of each participant in rating the *i*th pair *N* for the *j*th *M* sample can be calculated as follows:

$$\%_{correct} = \frac{\sum_{i=1}^{M} \sum_{j=1}^{N} X_{i,j}}{M \times N} \times 100\%. \quad (5)$$

To answer RQ3, Fig. 6 shows the results from Eq. (5) in three experiments, plotted by subjects (*y* axis) and percentage of correct answers (*x* axis). The bar graph was arranged in descending order to facilitate identification of unsuitable candidates. As mentioned earlier, the minimum scores of ≥80%, ≥85%, and ≥85% were set for Experiments 1, 2, and 3, respectively, and are drawn with a dashed line.

In Experiment 1, the percentage range is between 63% and 93%. There were seven participants (S36, S26, S25, S31, S35, S07, and S10) who failed the test with a percentage range of 63%−78%, and three participants (S20, S22, and S28) had a score within the threshold of 80%. In Experiment 2, on the other hand, the percentage range was between 76% and 100%. Two participants (S24 and S07) were able to perform perfectly, and only two participants (S43 and S17) failed. It can be assumed that the perceptual discrimination of the video was relatively high, because participants were able to discriminate almost all given triads. This result confirms the previously described content analysis in conjunction with Fig. 5.

Finally, for Experiment 3, the range was from 69.0% to 97.6%, with eight participants (S10, S43, S09, S10, S06, S13, S32, and S35) failing the test with a percentage score below the threshold of 85%. S17 and S43 had scores below 72%, consistently failed both the video and audiovisual experiments, and were therefore rejected. The remaining six subjects with scores of ≥80% in the video and audiovisual experiments can still be considered a selected assessor for a new group because the scores are close to the threshold. The minimum target can probably be met by additional training sessions and reassessment.

Using the results of Stage 2, the selection criteria were strictly applied to all 45 participants to form a group of selected assessors. Twenty-six participants (14 males and 12 females) with an age range of 22−36 years (mean = 27.2; SD = 3.2) fully met all selection criteria. Thus, a selected assessor in the group has a high potential to receive further

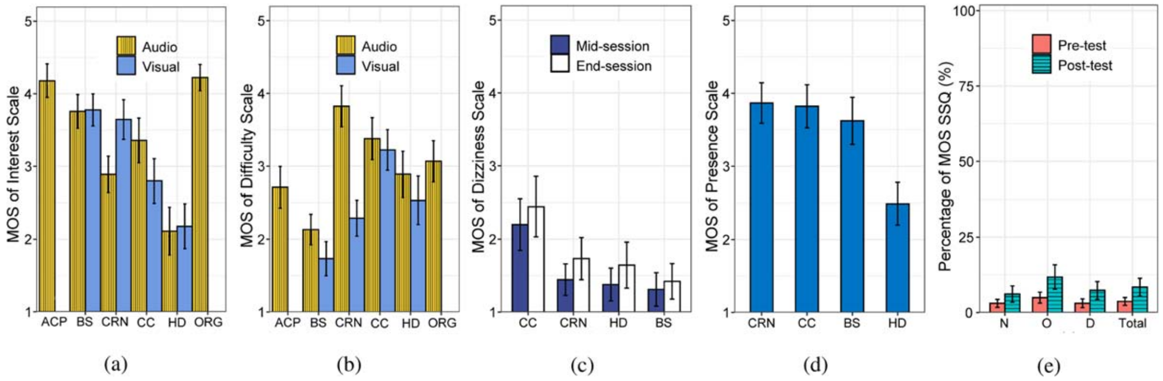


Fig. 7. Mean opinion scores for each factor influencing quality of experience (QoE) including (a) interest level, (b) difficulty level, (c) dizziness level, (d) presence level, and (e) percentage of simulator sickness questionnaire (SSQ) score. MOS = mean opinion score.

training and become a member of an expert group. However, it cannot be guaranteed that all assessors will perform sufficiently to achieve expert status. Therefore, an oversized alternative group of selected assessors is needed as a contingency. Each person in the alternative group passed all basic discrimination tests (audio, video, and audiovisual) but failed either the audiometric or visual screening tests. This alternative group consisted of six assessors (five males and one female) with an age range of 18−38 years (mean = 30.5; SD = 7.0).

3.5 Factors Affecting QoE

Five content-related factors affecting QoE were analyzed and presented in Figs. 7(a)–7(e), including interest in content, difficulty of content, dizziness, presence, and SSQ. Regarding the assessment of interest in Fig. 7(a), ACP and ORG have relatively similar auditory interest scores, followed by BS, CC, CRN, and HD for Experiment 1. BS and CRN have slightly similar visual interest, followed by CC and HD. It is also evident that BS, as with a nature scene, can elicit both high auditory and visual interest. In contrast, HD has the lowest score for both interests. CRN has high visual interest but low auditory interest, whereas the opposite is true for CC. It is argued that CC has speech content that arouses people's interest in informative sound, whereas CRN has a city setting with a lot of visual information to explore during the experience.

In terms of difficulty scores in Fig. 7(b), it can be seen that although BS can arouse interest, BS is less difficult for both modalities, implying that discrimination is relatively high. In contrast, HD is less interesting but relatively difficult to judge. This condition is similar to CRN's for visual perspective, which proved to be the most difficult scene of the group to judge. It was also found that the quality of speech-related content (CC) was relatively difficult to judge but kept people interested. For musical content, ACP tended to be slightly more distinguishable than ORG, as indicated by the difficulty MOS.

For Experiments 2 (in the middle of the session) and 3 (at the end of the session), dizziness was measured after each experiment. According to the dizziness score in Fig. 7(c), the scores at the end of the session were generally higher than at the middle of the session for all observed scenes. However, MOS is still somewhat low in both cases (MOS < 2.5). The dynamic scene, CC, caused more dizziness than the static scenes, especially for the nature scenes. However, there is an indirect relationship between the difficulty level of a dynamic scene and dizziness, as in CC, but high interest with a low difficulty level can lead to lower dizziness, as in BS. However, CRN and HD have almost the same score in both cases.

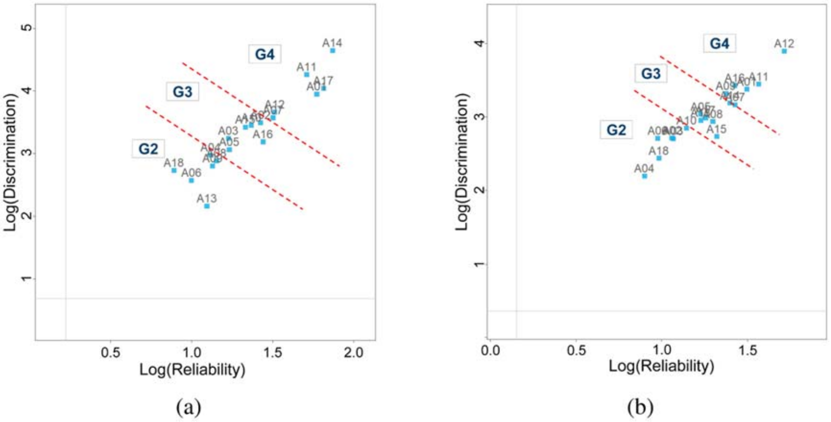


Fig. 8. eGauge [66, 67] plots for perceptual evaluation of (a) audio quality and (b) video quality.

The presence score in Fig. 7(d) shows that CRN could provide the highest presence, followed by CC and BS (MOS > 3.5), and HD had the lowest score (MOS $= 2.5$) compared with the others. It can be concluded that this is because of the fact that HD is considered a low-interest content, as shown in Fig. 7(a). Finally, the SSQ scores are drawn in Fig. 7(e) and analyzed for nausea (N), oculomotor (O), disorientation (D), and total score. The score is the ratio of the total score for each symptom to the maximum score multiplied by the weights defined in [59] and is expressed as a percentage. It can be seen that the post-test scores (after Experiment 3) are always higher than the pre-test scores (before Experiment 2). However, all symptoms and total scores are considerably low $<$12.5% , with symptom O having the highest score compared with the others. In general, no sickness issue was found in the discrimination experiments in Stage 2, indicating that the experiment is feasible.

4 CONFIRMATORY EXPERIMENT

To draw a conclusion about the effectiveness of the proposed procedure addressed in RQ4, a perceptual quality experiment was conducted to investigate whether

- A group of selected assessors could also obtain reliable results using another type of test, such as the rating scale method, and
- The performance of a group of selected assessors is statistically more accurate than that of failed assessors.

Eighteen selected assessors (A01−A18) were invited to perform audio, video, and audiovisual tests. Six samples, namely CC, CM, DB, HD, LB, and NTR (see Fig. 3), were used for the experiment. Bitrate encoding was implemented for the audio samples in 16, 32, and 64 kbps and PCM per channel using AAC-LC. Video samples were encoded using libx265 in FFmpeg at three resolutions (6,144 × 3,072; 3,840 × 1,920; and 1,920 × 1,080) and four QPs (0, 22, 28, and 34). The experiment was in full factorial design and was run over multiple stimuli with a hidden reference without an anchor. The playback of the system and user interface used in the experiments were the same as in the assessor selection experiments.

To confirm the first condition, a post-screening analysis was performed by computing the eGauge [66, 67] of the experiments. Replication was also considered by analyzing how each assessor performed in the test based on their reliability and discrimination scores, as depicted in Figs. 8(a) and 8(b). It can be seen that all the selected assessors were in the upper threshold range, indicating statistically good performance in general. Four assessors (A01, A11, A14, and A17) and one assessor (A12) had rather higher performances compared with the others in audio and video quality test, respectively. According to the variability, the assessors were put into three groups (G2, G3, and G4), each consisting of six assessors, based on the region of their position in the reliability-discrimination plot. To satisfy the second investigation, a group of six assessors (G1) who had not passed the selection procedure were also invited to participate in the same experiments. In the analysis, two additional groups were defined as baselines: G5, which represents all selected assessors (G2−G4, $n = 18$), and G6, which represents all assessors (G1−G4, $n = 24$).

4.1 Result 1: Listening Test

Results are shown in Figs. 9 and 10 for the audio, video, and audiovisual experiments. For the listening test in Fig. 9(a), the range of scores assessed by G1 is smaller than that of G2−G4; the range increases in proportion to expertise between G2−G4. Similarly, precision increases from G1 to G4, as indicated by the CI. G1 shows the largest CI, whereas G4 has the smallest CI, for all audio bitrates. A small distinction can be seen for the difficult part between 32 and 64 kbps, where G1 shows a barely discernible difference between the two conditions. G2 performs slightly better results with respect to MOS but continues to overlap in CI. In contrast, G3 and G4 could show a significant difference between these conditions.

In summary, all assessors in G2−G4 improved in terms of CI and mean score, regardless of assessor type. This is not a surprise, because all assessors have high eGauge discrimination and reliability scores, indicating that they are contributing reliable and discriminatory data to the dataset. G5's MOS generally fall between G2 and G4. When G1 is added to G5, the MOS at the lowest and highest bitrates are similar to G2 but with a lower CI. At 32 and 64 kbps, the results are only slightly different from G5. The results suggest that adding 25% of incompetent assessors could improve the CI while maintaining discrimination.

4.2 Result 2: Video Quality Test

The result of the video quality test is presented in Fig. 9(b). In general, the difference in performance between the groups is relatively insignificant, although the CI values for all encoding levels are relatively low. This common match can be interpreted as human sensitivity to a visual stimulus, which could elicit a common match more quickly than reception with an auditory stimulus. However, G1 generally fails to discriminate between QP22 and QP28 at 2K resolution, as shown by the scarce result at the mean CI. The CI is relatively low at the lowest (2K-QP34) and highest video quality (6K-QP0), whereas it varies for all encoding parameters. These encoding levels, i.e., 4K-28 and 6K-34, result in an overlap of perceived quality and show an insignificant difference between the two conditions as the CI increases.

4.3 Result 3: Audiovisual Quality Test

Fig. 10(a) shows the effect of video resolutions over audio bitrates on the mean opinion rating G1−G6. In general, the contribution of improving audio quality compared to increasing video resolution is relatively smaller than the difference in perception. A significant difference between moderate audio quality ($\geq$32 kbps) is only observed for high-resolution video ($\geq$4K) and a large number of asses-

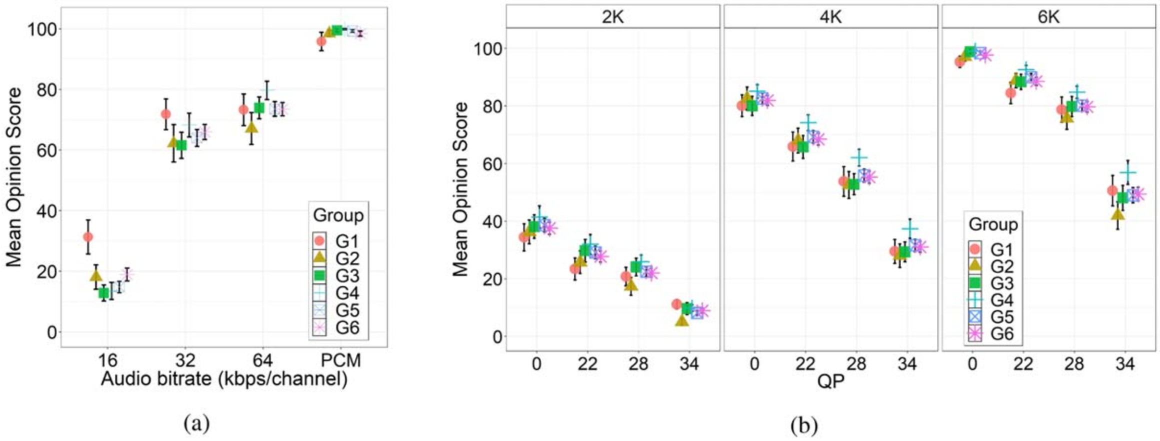


Fig. 9. Comparison of Mean: 95% confidence interval (CI) between six groups for (a) listening quality and (b) video quality. PCM = Pulse-Code Modulation.

sors (G5−G6). In groups with few assessors, the quality becomes transparent between 64 kbps and PCM for the same video resolution. When audio quality is very poor (16 kbps), improving video resolution from 4K to 6K results in an insignificant difference in perception.

The difference between each group lies in the ability to distinguish different audio qualities at the same quality level of a video. As far as this discrimination ability is concerned, the result increases from G1 to G6. At 2K and 4K resolutions, it can be seen that G1 is not able to distinguish audio quality from 32 kbps to PCM. Additionally, a relatively similar result was obtained by G1 at 6K resolution with 64 kbps and PCM. G2 is able to improve the perceptual difference by up to 64 kbps but could not distinguish between 64 kbps and PCM, especially at 2K and 4K resolutions.

Fig. 10(b) shows the effect of video QPs and audio bitrate on the average opinion rating by G1−G6. In general, increasing audio quality above 64 kbps does not improve perceptual quality in QP28 and QP34. Similarly, an improvement in video quality from QP22 to QP0 does not lead to a significant improvement in perceived quality at audio bitrates of 16 and 32 kbps. As for group performance, G1 fails to discriminate audio quality at low and medium video quality (QP 22−QP34), whereas G2 fails at low video quality (QP28−QP34).

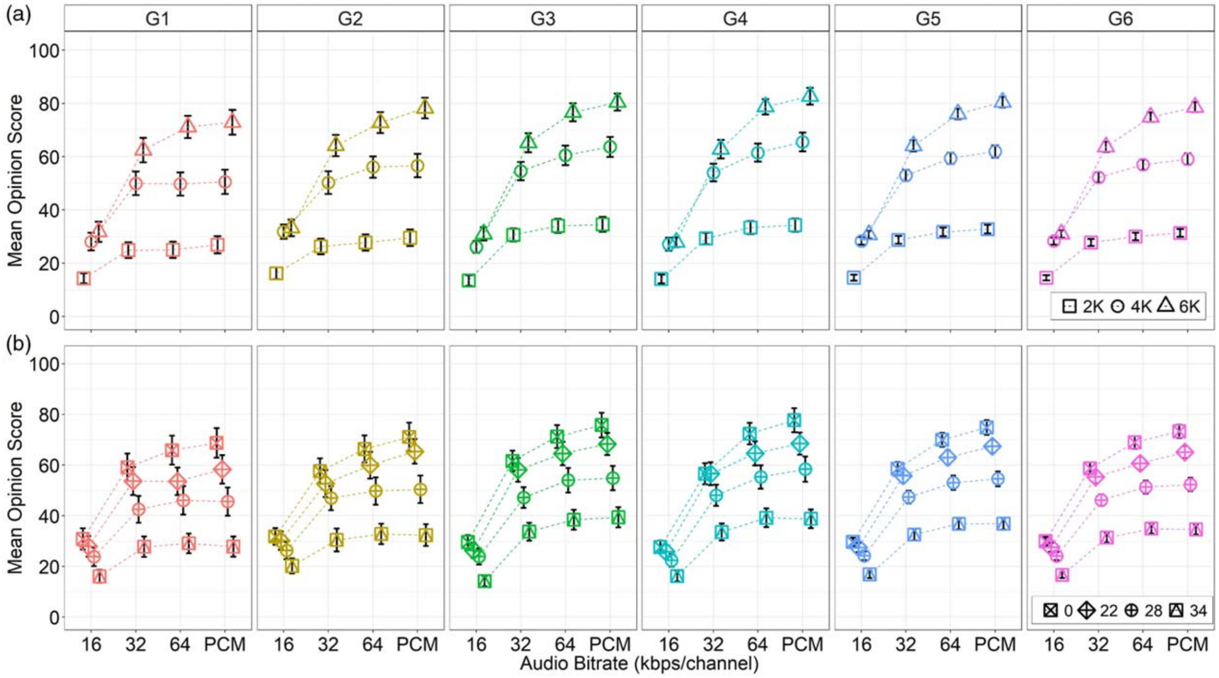


Fig. 10. Comparison of Mean: 95% confidence interval (CI) between six groups in relation to audio bitrate with (a) video resolution and (b) video quantization parameter (QP) in audiovisual quality assessment. PCM = Pulse-Code Modulation.

5 CONCLUSION

The present study presented and validated a procedure for selecting assessors to evaluate perceptual quality in immersive multimedia applications. Forty-five participants were selected in the pre-selection stage (Stage 1). Finally, 26 of the 45 participants passed Stage 2 and were included as selected assessors.

Basic audiovisual screening revealed that participants had audiometric HLs of 35 dB to −10 dB, no color blindness was detected, and participants had normal visual depth ($\text{VA}_{dec} = 0.75-1.34$). Perceptual discrimination with a triangle test and β-binomial analysis revealed that the percentage of correct responses depended on a sample and parameter encoding of the stimulus triads. Low discrimination ($p \geq 0.05$) was found under the following conditions:

- Pair A10 (128 kbps−PCM) in Experiment 1;
- Identical video resolution with one-step QP, i.e., pair V20−V21 in Experiment 2; and
- Identical video resolution combined with audio bitrate ≥24 kbps, i.e., AV11−AV13 and AV20−AV21 in Experiment 3.

When assessing factors affecting QoE, interest scores tended to be inversely proportional to difficulty scores but proportional to presence scores. Dizziness and SSQ showed a similar trend, increasing from the first to the second evaluations, but both had relatively low scores. These results indicate that the sickness caused by the experiments is very low, so the proposed procedure is considered feasible.

Additionally, modest confirmation experiments were conducted and analyzed to compare the failed assessor groups with the selected assessors who had different levels of discrimination and reliability categorized by the eGauge. It can be seen that G1 tends to perform less reliably on all tests, as evidenced by a narrow perceptual range, a wide CI, and low discrimination ability. The performance of each group improves with increasing reliability. Adding a number of assessors does not differ in MOS but improves CI accuracy. In the audiovisual experiment, the contribution of audio quality was relatively smaller than that of visual quality.

Although the confirmatory study was based on a small sample of participants, the results suggest that the selected assessors were more reliable in the rating scale method and showed statistically more accurate performance in the audio, video, and audiovisual quality experiments. This suggests that the proposed method may provide an advantage for future selection procedures.

6 LIMITATION AND FUTURE WORK

The study presented in this paper and its results are application agnostic. The proposed procedure should not be overlooked as a gold standard for various immersive multimedia applications. Instead, it provides an overall picture for the selection process that has been proven to work in this case, in which quality has been defined generally in terms of compression schemes. The authors argue that this proposal is still applicable in the case in which the definition of quality is similar. However, when completely different systems or artifacts are included, e.g., for parameters related to the temporal aspect of video (e.g., frame rates and buffering), audio spatialization (number of playback channels and binaural rendering method), audiovisual asynchrony, attribute evaluation, etc., at least two strategies can be suggested.

First, if a panel of selected assessors is available, the next step is to continue the framework shown in Fig. 1 by training the panelist with different evaluation schemes depending on the case, as recommended in [19]. The example of this augmentation process is presented in [28] from the original GLS procedure [27]. Second, if the panelist is not available because there is no prior selection process, and the naive assessor is invited, it is recommended that the candidates be trained to the “initiated assessor” level and continue the training with the study of interest [30, 32].

It should be emphasized that the assessor selection process may have direct utility for practitioners and researchers who routinely conduct perceptual evaluation and seek relatively consistent results with a smaller number of assessors. Although this procedure offers some benefit for the long-term use of qualified assessors, it should be recognized that using this procedure means adding a full stage at the beginning of the development phase. Depending on the purpose of the study and time and cost constraints, it may make more sense to use naive assessors and maintain statistical power with a higher number of assessors.

Future work should aim to investigate the application of this selection procedure to a wide range of multimedia applications, ranging from simple audiovisual cues to high-fidelity multisensory environments (3D light field video, VR, augmented reality). In terms of improvements, it would be interesting to develop a threshold mechanism for assessor selection that can be used universally for audio, video, and audiovisual tests. It might be useful to examine a range of audiovisual content categorized by different characteristics to understand the relationships between content and perceptual responses. The development of objective measures to represent perceptual distance and the introduction of a more robust approach to the evaluation process are also strongly recommended. Finally, a panel of assessors with high discriminative ability and reliability can improve the quality of the data and be very useful for more accurate prediction in machine learning.

7 ACKNOWLEDGMENT

This research was supported by the European Union’s Horizon 2020 research and innovation program under the Marie Skłodowska-Curie grant agreement number 765911 RealVision. The authors acknowledge FORCE Technology, Bang & Olufsen, Demant, GN Store Nord, Sonova, WS Audiology, and industrial partners who created the 360 audiovisual datasets under the HOA-SSR joint project funded by Danish Ministry of Science and Technology, and they acknowledge XRHub Team for the great help in dealing with technicalities, filming, and field recording. Finally, the authors appreciate all anonymous subjects who participated in the audiovisual tests described in this paper.

8 REFERENCES

[1] ISO/IEC, “Information Technology — Coded Representation of Immersive Media (MPEG-I) — Part 2: Omnidirectional Media Format,” *Standard 23090-2* (2017 Jan.).

[2] M. M. Hannuksela and Y.-K. Wang, “An Overview of Omnidirectional MediA Format (OMAF),” *Proc. IEEE*, vol. 109, no. 9, pp. 1590–1606 (2021 Sep.). https://doi.org/10.1109/JPROC.2021.3063544.

[3] ITU-R, “Methodology for the Subjective Assessment of the Quality of Television Pictures,” *Recommendation ITU-R BT.500-13* (2012 Jan.).

[4] ITU-T, “Subjective Video Quality Assessment Methods for Multimedia Applications,” *Recommendation ITU-T P.910* (2008 Apr.).

[5] ITU-T, “Methods for the Subjective Assessment of Video Quality, Audio Quality and Audiovisual Quality of Internet Video and Distribution Quality Television in Any Environment,” *Recommendation ITU-T P.913* (2016 Mar.).

[6] ITU-R, “Methods for the Subjective Assessment of Small Impairments in Audio Systems,” *Recommendation ITU-R BS.1116-3* (2015 Feb.).

[7] ITU-R, “Method for the Subjective Assessment of Intermediate Quality Levels of Audio Systems,” *Recommendation ITU-R BS.1534-3* (2015 Oct.).

[8] ITU-T, “Subjective Test Methodologies for 360° Video on Head-Mounted Displays,” *Recommendation ITU-T P.919* (2020 Oct.).

[9] H. T. T. Tran, N. P. Ngoc, C. T. Pham, Y. J. Jung, and T. C. Thang, “A Subjective Study on User Perception Aspects in Virtual Reality,” *Appl. Sci.*, vol. 9, no. 16, paper 3384 (2019 Aug.). https://doi.org/10.3390/app9163384.

[10] M. S. Anwar, J. Wang, W. Khan, et al., â€œSubjective QoE of 360-Degree Virtual Reality Videos and Machine Learning Predictions,” *IEEE Access*, vol. 8, pp. 148084–148099 (2020 Aug.). https://doi.org/10.1109/ACCESS.2020.3015556.

[11] A. Singla, W. Robitza, and A. Raake, “Comparison of Subjective Quality Test Methods for Omnidirectional Video Quality Evaluation,” in *Proceedings of the IEEE 21st International Workshop on Multimedia Signal Processing (MMSP)*, pp. 1–6 (Kuala Lumpur, Malaysia) (2019 Sep.). https://doi.org/10.1109/MMSP.2019.8901719.

[12] R. F. Fela, N. Zacharov, and S. Forchhammer, “Towards a Perceived Audiovisual Quality Model for Immersive Content,” in *Proceedings of the 12th International Conference on Quality of Multimedia Experience (QoMEX)*, pp. 1–6 (Athlone, Ireland) (2020 May.). https://doi.org/10.1109/QoMEX48832.2020.9123134.

[13] R. F. Fela, N. Zacharov, and S. Forchhammer, “Perceptual Evaluation of 360 Audiovisual Quality and Machine Learning Predictions,” in *Proceedings of the IEEE 23rd International Workshop on Multimedia Signal Processing (MMSP)*, pp. 1–6 (Tampere, Finland) (2021 Oct.). https://doi.org/10.1109/MMSP53017.2021.9733677.

[14] N. Zacharov and G. Lorho, “What Are the Requirements of a Listening Panel for Evaluating Spatial Audio Quality?” in *Proceedings of the International Workshop on Spatial Audio and Sensory Evaluation Techniques* (Guilford, UK) (2006 Apr.).

[15] M. Elwardy, H.-J. Zepernick, V. Sundstedt, and Y. Hu, “Impact of Participants’ Experiences With Immersive Multimedia on 360° Video Quality Assessment,” in *Proceedings of the 13th International Conference on Signal Processing and Communication Systems (ICSPCS)*, pp. 1–10 (Gold Coast, Australia) (2019 Feb.). https://doi.org/10.1109/ICSPCS47537.2019.9008739.

[16] J. Boyce, E. Alshina, A. Abbas, and Y. Ye, “JVET Common Test Conditions and Evaluation Procedures for 360 Video,” in *Proceedings of the 7th Meeting: Joint Video Exploration Team (JVET) of ITU-T SG 16 WP 3 and ISO/IEC JTC 1/SC 29/WG 11*, JVET-G1030, pp. 13–21 (Torino, Italy) (2017 Jul.).

[17] R. F. Fela, A. Pastor, P. L. Callet, et al., “Perceptual Evaluation on Audio-Visual Dataset of 360 Content,” in *Proceedings of IEEE International Conference on Multimedia and Expo (ICME)* (Taipei, Taiwan) (2022 Jul.).

[18] T. Rudzki, I. Gomez-Lanzaco, J. Stubbs, et al., “Auditory Localization in Low-Bitrate Compressed Ambisonic Scenes,” *Appl. Sci.*, vol. 9, no. 13, paper 2618 (2019 Jun.). https://doi.org/10.3390/app9132618.

[19] ISO, “Sensory Analysis—General Guidelines for the Selection, Training and Monitoring of Selected Assessors and Expert Sensory Assessors,” *Standard 8586:2012* (2012 Dec.).

[20] ISO, “Sensory Analysis—General Guidance for the Selection, Training, and Monitoring of Assessors—Part 1: Selected Assessors,” *Standard 8586-1:1993* (1993 Mar.).

[21] ISO, “Sensory Analysis—General Guidance for the Selection, Training, and Monitoring of Assessors—Part 2: Experts,” *Standard 8586-2:1994* (1994 Jun.).

[22] V. Hansen, “Establishing a Panel of Listeners at Bang & Olufsen: A Report,” in *Proceedings of the Symposium on Perception of Reproduced Sound*, pp. 89–98 (Gammel Avernæs, Denmark) (1987 Aug.).

[23] F. E. Toole, “Subjective Measurements of Loudspeaker Sound Quality and Listener Performance,” *J. Audio Eng. Soc.*, vol. 33, no. 1/2, pp. 2–32 (1985 Feb.).

[24] S. Bech, “Selection and Training of Subjects for Listening Tests on Sound-Reproducing Equipment,” *J. Audio Eng. Soc.*, vol. 40, no. 7/8, pp. 590–610 (1992 Jul.).

[25] S. Bech, “Training of Subjects for Auditory Experiments,” *Acust. United Acta Acust.*, vol. 1, pp. 89–99 (1993).

[26] S. E. Olive, “A Method for Training Listeners and Selecting Program Material for Listening Tests,” presented at the *97th Convention of the Audio Engineering Society* (1994 Nov.), paper 3893.

[27] V.-V. Mattila and N. Zacharov, “Generalized Listener Selection (GLS) Procedure,” presented at the *110th Convention of the Audio Engineering Society* (2001 May), paper 5405.

[28] D. Isherwood, G. Lorho, N. Zacharov, and V.-V. Mattila, “Augmentation, Application and Verification of

the Generalized Listener Selection Procedure," presented at the *115th Convention of the Audio Engineering Society* (2003 Oct.), paper 5984.

[29] F. Wickelmaier and S. Choisel, "Selecting Participants for Listening Tests of Multichannel Reproduced Sound," presented at the *118th Convention of the Audio Engineering Society* (2005 May), paper 6483.

[30] S. V. Legarth and N. Zacharov, "Assessor Selection Process for Multisensory Applications," presented at the *126th Convention of the Audio Engineering Society* (2009 May), paper 7788.

[31] A. Sontacchi, H. Pomberger, and R. Höldrich, "Recruiting and Evaluation Process of an Expert Listening Panel," in *Proceedings of Fortschritte der Akustik (NAG/DAGA)*, pp. 1552–1555 (Rotterdam, The Netherlands) (2009 Mar.).

[32] A. Kuusinen, H. Vertanen, and T. Lokki, "Assessor Selection and Behavior in Individual Vocabulary Profiling of Concert Hall Acoustics," in *Proceedings of the AES 38th International Conference: Sound Quality Evaluation* (2010 Jun.), paper 7-1.

[33] A. Ravikumar, E. J. Sarver, and R. A. Applegate, "Change in Visual Acuity is Highly Correlated With Change in Six Image Quality Metrics Independent of Wavefront Error and/or Pupil Diameter," *J. Vision*, vol. 12, no. 10, paper 11 (2012 Sep.). https://doi.org/10.1167/12.10.11.

[34] J. Ghani, W. Ellermeier, and K. Zimmer, "A Test Battery Measuring Auditory Capabilities of Listening Panels," in *Proceedings of the Forum Acusticum Congress*, pp. 1677–1681 (Budapest, Hungary) (2005 Aug.).

[35] A. MacQuarrie and A. Steed, "Cinematic Virtual Reality: Evaluating the Effect of Display Type on the Viewing Experience for Panoramic Video," in *Proceedings of the IEEE Virtual Reality (VR)*, pp. 45–54 (Los Angeles, CA) (2017 Mar.). https://doi.org/10.1109/VR.2017.7892230.

[36] M. Mirkovic, P. Vrgovic, D. Culibrk, D. Stefanovic, and A. Anderla, "Evaluating the Role of Content in Subjective Video Quality Assessment," *Sci. World J.*, vol. 2014, paper 625219 (2014 Jan.). https://doi.org/10.1155/2014/625219.

[37] W. Li, P. Spachos, M. Chignell, et al., "Impact of Technical and Content Quality on Overall Experience of OTT Video," in *Proceedings of the 13th IEEE Annual Consumer Communications Networking Conference (CCNC)*, pp. 930–935 (Las Vegas, NV) (2016 Jan.). https://doi.org/10.1109/CCNC.2016.7444912.

[38] S. H. Jumisko, V. P. Ilvonen, and K. A. Vaananen-Vainio-Mattila, "Effect of TV Content in Subjective Assessment of Video Quality on Mobile Devices," in *Proceedings of the SPIE: Multimedia on Mobile Devices*, vol. 5684, pp. 243–254 (San Jose, CA) (2005 Mar.).

[39] A. MacRae, "Visualizing the Difference Between Triangle and 3AFC Judgements," *Food Qual. Prefer.*, vol. 6, no. 4, pp. 315–320 (1995 Apr.). https://doi.org/10.1016/0950-3293(95)00034-8.

[40] ISO, "Sensory Analysis – Methodology – Triangle Test," *Standard 4120:2021* (2021 Mar.).

[41] W. Hughson and H. Westlake, "Manual for Program Outline for Rehabilitation of Aural Casualties both Military and Civilian," *Trans. Am. Acad. Ophthalmol. Otolaryngol.*, vol. 48, pp. 1–15 (1944 Jan).

[42] ISO, "Acoustics — Audiometric Test Methods — Part 1: Basic Pure Tone Air and Bone Conduction Threshold Audiometry," *Standard 8253-1:1989* (1989 Nov.).

[43] J. Clark, "The Ishihara Test for Color Blindness," *Am. J. Physiol. Opt.*, vol. 5, pp. 269–276 (1924).

[44] B. Julesz, *Foundations of Cyclopean Perception* (University of Chicago Press, Chicago, IL, 1971).

[45] M. Bach, "The Freiburg Visual Acuity Test—Automatic Measurement of Visual Acuity," *Optom. Vis. Sci.*, vol. 73, no. 1, pp. 49–53 (1996 Jan.). https://doi.org/10.1097/00006324-199601000-00008.

[46] A. K. Moorthy, L. K. Choi, A. C. Bovik, and G. de Veciana, "Video Quality Assessment on Mobile Devices: Subjective, Behavioral and Objective Studies," *IEEE J. Sel. Top. Signal Process.*, vol. 6, no. 6, pp. 652–671 (2012 Oct.). https://doi.org/10.1109/JSTSP.2012.2212417.

[47] E. Bates, M. Gorzel, L. Ferguson, H. O'dwyer, and F. M. Boland, "Comparing Ambisonic Microphones — Part 1," in *Proceedings of the AES International Conference on Sound Field Control* (2016 Jul.), paper 6-3.

[48] E. Bates, S. Dooney, M. Gorzel, et al., "Comparing Ambisonic Microphones — Part 2," presented at the *142nd Convention of the Audio Engineering Society* (2017 May.), paper 9730.

[49] F. Zotter and M. Frank, "All-Round Ambisonic Panning and Decoding," *J. Audio Eng. Soc.*, vol. 60, no. 10, pp. 807–820 (2012 Oct.).

[50] ITU-R, "Multichannel Sound Technology in Home and Broadcasting Applications," *Report ITU-R BS.2159-7* (2015 Feb.).

[51] S. Coren, "Most Comfortable Listening Level as a Function of Age," *Ergonomics*, vol. 37, no. 7, pp. 1269–1274 (1994 Jul.). https://doi.org/10.1080/00140139408964905.

[52] M. O'Mahony, "Who Told You the Triangle Test Was Simple?" *Food Qual. Prefer.*, vol. 6, no. 4, pp. 227–238 (1995 Apr.). https://doi.org/10.1016/0950-3293(95)00022-4.

[53] J. Kunert and M. Meyners, "On the Triangle Test With Replications," *Food Qual. Prefer.*, vol. 10, no. 6, pp. 477–482 (1999 Nov.). https://doi.org/10.1016/S0950-3293(99)00047-6.

[54] J. Bi, *Sensory Discrimination Tests and Measurements: Statistical Principles, Procedures and Tables* (Wiley, Hoboken, NJ, 2007).

[55] H. Lee and D. Johnson, "An Open-Access Database of 3D Microphone Array Recordings," presented at the *147th Convention of the Audio Engineering Society* (2019 Oct.), e-Brief 543.

[56] EBU, "Listening Conditions for the Assessment of Sound Programme Material: Monophonic and Two–Channel Stereophonic," Tech. Rep. 3276 (1998 May.).

[57] G. Le Ray and J. Khalid, "SenseLabOnline: Combining Agile Data Base Administration With Strong

Data Analysis," in *Proceedings of the R User Conference (useR!)*, vol. 10, p. 38 (Albacete, Spain) (2013 Jul.).

[58] Visual Functions Committee ICO, "Visual Acuity Measurement Standard," *ICO Standard* (1984 May).

[59] R. S. Kennedy, N. E. Lane, K. S. Berbaum, and M. G. Lilienthal, "Simulator Sickness Questionnaire: An Enhanced Method for Quantifying Simulator Sickness," *Int. J. Aviat. Psychol.*, vol. 3, no. 3, pp. 203–220 (1993). https://doi.org/10.1207/s15327108ijap0303_3.

[60] P. Bimberg, T. Weissker, and A. Kulik, "On The Usage of the Simulator Sickness Questionnaire for Virtual Reality Research," in *Proceedings of the IEEE Conference on Virtual Reality and 3D User Interfaces Abstracts and Workshops (VRW)*, pp. 464–467 (Atlanta, GA) (2020 Mar.). https://doi.org/10.1109/VRW50115.2020.00098.

[61] D. M. Ennis and J. Bi, "The Beta-Binomial Model: Accounting for Inter-Trial Variation in Replicated Difference and Preference Tests," *J. Sens. Stud.*, vol. 13, no. 4, pp. 389–412 (1998 Dec.). https://doi.org/10.1111/j.1745-459X.1998.tb00097.x.

[62] R. H. B. Christensen, "Statistical Methodology for Sensory Discrimination Tests and Its Implementation in sensR," unpublished document (2015 Mar.).

[63] M. Šalovarda, I. Bolkovac, and H. Domitrović, "Comparison of Audio Codecs Using PEAQ Algorithm," in *Proceedings of the 6th CARNet Users' Conference (CUC)*, paper B1 (Zagreb, Croatia) (2004 Sep.).

[64] T. Thiede, W. C. Treurniet, R. Bitto, et al., "PEAQ – The ITU Standard for Objective Measurement of Perceived Audio Quality," *J. Audio Eng. Soc.*, vol. 48, no. 1/2, pp. 3–29 (2000 Feb.).

[65] K. Grivcova, C. Pike, and T. Nixon, "A Subjective Evaluation of High Bitrate Coding of Music," presented at the *144th Convention of the Audio Engineering Society* (2018 May), paper 10001.

[66] G. Lorho, G. Le Ray, and N. Zacharov, "eGauge—A Measure of Assessor Expertise in Audio Quality Evaluations," in *Proceedings of the AES 38th International Conference: Sound Quality Evaluation* (2010 Jun.), paper 7-2.

[67] ITU-R, "Methods for Assessor Screening," *Recommendation ITU-R BS.2300-0* (2014 Apr.).

THE AUTHORS

Randy Frans Fela

Nick Zacharov

Søren Forchhammer

Randy Frans Fela has an engineering physics background from Bandung Institute of Technology (M.Sc.), majoring in the field of acoustics and spatial audio evaluation in immersive multimedia. He is currently an Early Stage Researcher of the European Union (EU) Marie Sklodowska-Curie Actions (MSCA) Innovative Training Network (ITN) RealVision and is doing his research in FORCE Technology and Technical University of Denmark. His research interests include optimal experimental design, perceptual evaluation, and predictive quality metric with machine learning approach.

•

Nick Zacharov [D.Sc. (Tech.), M.Sc., B.Eng. (Hons.), C.Eng., Fellow of the Audio Engineering Society] is a lead technology manager of perceptual audio evaluation at Meta Reality Labs, focusing on sound quality research. With an academic background in electroacoustics, acoustics, and signal processing, Nick has broad industrial experience in the audio profession spanning from mobile phone audio to AR/VR devices and professional studio monitor design. Nick is the co-author of *Perceptual Audio Evaluation–Theory, Method and Application* and also editor/co-author of *Sensory Evaluation of Sound*. He has been an active member of the Audio Engineering Society and has more than 90 publications and patents to his name.

•

Søren Forchhammer is a Professor with the Department of Electrical and Photonics Engineering, Technical University of Denmark, where he has been since 1988. He is Head of the Coding and Visual Communication Group at DTU Fotonik. He is currently Coordinator of the European Union (EU) Marie Sklodowska-Curie Actions (MSCA) Innovative Training Network (ITN) RealVision. His research interests include source coding, image and video coding, processing of image and video, processing for image displays, quality of coded multimedia data, multi-camera and light field images and video, communication theory, 2D information theory, and visual communications.