Influence of Multi-Modal Interactive Formats on Subjective Audio Quality and Exploration Behavior

Thomas Robotham
thomas.robotham@audiolabs-erlangen.de
International Audio Laboratories Erlangen†
Erlangen, Germany

Ashutosh Singla
Audiovisual Technology Group, TU-Ilmenau
Ilmenau, Germany

Alexander Raake
Audiovisual Technology Group, TU-Ilmenau
Ilmenau, Germany
alexander.raake@tu-ilmenau.de

Olli S. Rummukainen
International Audio Laboratories Erlangen†
Elrangen, Germany

Emanuël A. P. Habets
International Audio Laboratories Erlangen†
Elrangen, Germany

ABSTRACT

This study uses a mixed between- and within-subjects test design to evaluate the influence of interactive formats on the quality of binaurally rendered 360° spatial audio content. Focusing on ecological validity using real-world recordings of 60 s duration, three independent groups of subjects ($\frac{N}{3} = 18$) were exposed to three formats: audio only (A), audio with 2D visuals (A2DV), and audio with head-mounted display (AHMD) visuals. Within each interactive format, two sessions were conducted to evaluate degraded audio conditions: bit-rate and Ambisonics order. Our results show a statistically significant effect ($p < .05$) of format only on spatial audio quality ratings for Ambisonics order. Exploration data analysis shows that format A yields little variability in exploration, while formats A2DV and AHMD yield broader viewing distribution of 360° content. The results imply audio quality factors can be optimized depending on the interactive format.

CCS CONCEPTS

• **Human-centered computing** → **Auditory feedback**; **Virtual reality**.

KEYWORDS

Spatial Audio, Quality, Method, Behavior, Exploration, Multi-modal

ACM Reference Format:
Thomas Robotham, Ashutosh Singla, Alexander Raake, Olli S. Rummukainen, and Emanuël A. P. Habets. 2023. Influence of Multi-Modal Interactive Formats on Subjective Audio Quality and Exploration Behavior. In *ACM International Conference on Interactive Media Experiences (IMX '23), June 12–15, 2023, Nantes, France.* ACM, New York, NY, USA, 14 pages. https://doi.org/10.1145/3573381.3596155

† A joint institution of the Friedrich-Alexander-Universität Erlangen-Nürnberg (FAU) and Fraunhofer IIS, Germany.

IMX '23, June 12–15, 2023, Nantes, France

ACM ISBN 979-8-4007-0028-6/23/06.
https://doi.org/10.1145/3573381.3596155

1 INTRODUCTION

Interactive media experiences allow users to consume the same content in multiple different ways upon each viewing. For 360° video content combined with spatial audio, users can choose to look in any direction at any time and hear audio that reflects their head/camera movements. In many cases, this interactive content can be consumed in multiple ways, either simply via headphones (e.g., spatial audio podcasts, music, live performance broadcasts), over a display (desktop or portable devices), or inside virtual reality (VR) using head-mounted displays (HMDs). When consuming content across these interactive formats, various levels of sensory integration can occur between our auditory and visual systems and vestibular stimulation (i.e., gravity- and motion-receptors). In turn, cross-modality effects occur, which may impact our perception and cognition of sensory input streams [5, 54, 73, 85]. Consequently, our experience and formations of judgments (quality, spatial, or otherwise) may change.

Extensive research has been done regarding quality criteria and methods for uni-modal audio or video evaluations [13, 50, 56, 70, 81]. With the interest in VR systems increasing, the number of studies on multi-modal quality and cross-modal influences has been growing over the previous decade, leading to more insight into aspects such as presence and plausibility [9, 17, 25, 33, 36, 53]. However, many studies directed toward audiovisual cross-modality impairments do not consider the interactive format (e.g., HMDs or 2D displays) and the additional sensory information these mediums bring as an independent variable. Moreover, few studies are available that evaluate audio quality in multi-modal settings using ecologically valid stimuli or experimental conditions. Many of the methods and behavioral constraints are typical for a laboratory context which aims to minimize confounding factors or assist in discerning quality differences. However, evaluating test stimuli such as pink noise or anechoic sources in this setting, with the ability to compare between quality conditions or restricted movements, for example, will detract from a more naturalistic immersive experience. Indeed, a large appeal of spatial audio is to assist in creating immersion, encourage exploration within interactive experiences [48, 60], guide visual attention, and facilitate cognitive scene analysis. Consequently, audio quality evaluations with free exploration, with stimuli more

representative of real-world environments, using methods more focused on a single presentation of conditions, may yield results more applicable to natural experiences.

In this study, we investigate subjective audio quality within three multi-modal interactive formats. By employing binaurally rendered head-tracked spatial audio with: no video, interactive video via 2D display, and interactive video via HMD, influences from cross-modality effects and interaction types on subjective audio quality judgments and exploration behavior can be identified. The auditory conditions used are induced by audio bit-rate compression and Ambisonics spatial resolution, both highly relevant for spatial audio delivery in immersive experiences [24]. The method of evaluation is the single stimulus absolute category rating (ACR) [32]. To focus this study, the following hypothesis statements are drawn:

- $\mathcal{H}_1$: There is a significant interaction effect of interactive format and bit-rate on overall audio quality.
- $\mathcal{H}_2$: There is a significant interaction effect of interactive format and Ambisonics order on spatial audio quality.
- $\mathcal{H}_3$: Ecologically valid content can be used to determine quality differences in presented conditions.
- $\mathcal{H}_4$: Interactive format has a significant effect on subjects' exploration behavior for both degradation types.

The remainder of this article is structured as follows. Section 2 visits relevant literature regarding cross-modality interaction, focusing on audio-visual integration for multimedia content. Additionally, a basic description of immersive audio is given, including appropriate quality studies related to Ambisonics, bit-rate, and localization. Section 3 describes the evaluation paradigm used, method, and independent variables, followed by an outline of the evaluation procedure in Section 4. Section 5 then presents the results of the subjective evaluation and exploration data. These results are discussed in Section 6, with a focus on the presented hypotheses. Finally, the article ends with a conclusion in Section 7.

2 BACKGROUND

2.1 Cross-Modality Effects

2.1.1 Sensory Integration. Sensory integration is the process that combines our independent sensory streams into a single unified percept, distinct from the cognition of the individual channels [76]. The integration of spatial-temporally coherent audio and visual input can enhance neurological response to the stimuli. However, sufficiently synchronously distinct audio and visual input will not be optimally integrated with one another and thus not meet the neurological specifications for enhancement [75]. In the latter case, attention may be directed or dominated towards the more potent stimulus depending on factors such as context, task, memory, emotional affect, or experience [2, 74]. Sensory integration can also give rise to phenomena where the dominance of a particular sensory input compensates for an otherwise spatially, temporally, or semantically incongruent secondary modality. Common audio-visual examples are the ventriloquist effect [1] or McGurk effect [73]. For interactive audio-visual media experiences, there is a great interest in researching cross-modality effects to better understand perceptual thresholds and quality criteria.

2.1.2 Evaluation of Cross-Modality in Multimedia. In 1999, Rimell and Hollier published a novel architecture towards a multi-sensory perceptual model [64]. Using a 5-point ITU scale [32] they evaluated audio with a range of bit-rates, and video quality degraded via white noise and 'edge busyness'. Their results demonstrated that quality ratings for single modalities are influenced by the perceived quality of another, and the degree to which this is present is dependent on the content type. This is supported by Storms and Zyda [77], who investigated the cross-modal interactions of pixel resolution and Gaussian white noise level in the visual domain, against sampling frequency and Gaussian white noise level in the auditory domain. Regarding future work, the authors highlight the need to investigate other quality parameters and to augment the experiment to VR where new perceptual phenomena may be observed. Given its importance within multimedia experiences, bit-rate has been involved as an audio quality parameter in many more cross-modality evaluations [22, 42, 43, 49, 59, 87]. Many of these evaluations aim to determine the impact of audio-visual degradations on overall audiovisual quality using subjective and/or objective methods. Further audio degradations in cross-modal studies include *background noise*, *clipping*, *echo*, *packet-loss*, and *chopping* in transmission settings [20, 23, 44, 46]. However, there is little research addressing the medium of interaction with immersive spatial audio content as a third factor. For interactive multimedia experiences, addressing the mode of interaction and their inherent modalities (e.g., audio, visual, and vestibular systems) may yield further insight into cross-modality effects and their perceptual impact when evaluating aspects of spatial audio quality. Moreover, when spatial audio is employed in quality evaluation cross-modality research [73], it is seldom in an interactive setting, or to evaluate the impact of the interaction type. Given that spatial audio is a supporting pillar in creating immersion for interactive content, it seems pertinent to address the impact vestibular stimulation and sensory integration may have on quality judgments of various audio degradations.

2.2 Audio for Interactive Virtual Environments

2.2.1 Authoring Immersive Audio. The three main workflows available for authoring audio for immersive experiences are channel-based, object-based, and scene-based [31, 58, 69, 88]. In object-based audio, sources are treated as individual entities that are rendered at a specific location, agnostic to the reproduction format. This is opposed to channel-based content that is designed and authored to be reproduced over a specific reproduction setup [26]. For VR, object-based audio can include multiple stages within a source-receiver rendering pipeline to realize properties such as early reflections, occlusion, diffraction, reverberation, directivity, etc [63, 79, 82]. Such implementation provides access to many parameters within the acoustic auralization during run-time. Consequently, the object-based approach lends itself to highly non-linear content [78] and a complementary counterpart to computer-generated imagery. Scene-based audio combines elements of both channel- and object-based, where the number of audio signals is fixed (as with channel-based content), but has no direct relationship to the reproduction format (as with object-based audio) [57]. Ambisonics audio, a scene-based approach, can be used to describe the 3D sound field. In doing so, all aspects, such as reverberation, reflections, directivity, etc., are

inherently recorded. When paired with 360° video, Ambisonics is an attractive solution for providing spatial audio for streaming and video platforms such as Facebook, YouTube, and even web-browser-based solutions [61]. The content can be experienced over multiple devices such as phones, monitors, or head-mounted displays where the Ambisonics signals can be binaurally rendered to headphones and rotated to reflect subjects' head/camera movements [72].

2.2.2 Ambisonics Foundation. Ambisonics is a description of the spherical sound field pressure as a function of time using a set of three-dimensional orthogonal basis functions, referred to as *spherical harmonics* [28, 90]. Figure 1 depicts the spherical harmonic directivity patterns. For audio-specific implementations, the *AmbiX* format was proposed to standardize aspects such as normalization and channel ordering [51], and is commonly used in Ambisonics databases (e.g., [16, 66]). As shown in Figure 1, with an increasing Ambisonics order n, the number of spherical harmonic directivity patterns used to decompose the sound field will increase, resulting in a greater spatial resolution [21]. The addition of this spatial information comes at the cost of increased requirements for storage and transmission [52]. For an Ambisonics order of $n = 0$, the scene will have a single omnidirectional signal containing zero spatial information. For $n = 1$, referred to as $1st$-order Ambisonics (FOA), the Ambisonics will result in four channels that are essentially comprised of the initial omnidirectional signal plus three orthogonal figure-8 polar patterns. Ambisonics signals of order $n \geq 2$ are referred to as higher-order Ambisonics (HOA). Although the de-facto opinion of FOA is that it is generally insufficient to accurately localize direct sounds [19], it is not clear as to what order yields a sufficiently high perceptual spatial quality, particularly in the context of any multi-modal and/or interactive scenario.

2.2.3 Ambisonics Evaluation. Several studies have been conducted regarding localization accuracy, compression, and quality of Ambisonics audio. Narbutt et al. [52] evaluated the perceptual quality of compressed FOA and $3rd$-order Ambisonics signals at a variety of bit-rates using the OPUS 1.2 codec [84]. The evaluation method chosen was a multiple-stimulus with hidden reference and anchor (MUSHRA) [30] which utilizes an open reference of the known highest quality, with selected short (7 - 15 s) critical listening items. The quality criteria were *listening quality*, i.e., the perceived quality of a condition compared to the reference; and *localization accuracy* i.e., the localization of sound sources compared to the reference. Their results highlight the increased localization accuracy of HAO over FOA conditions, in addition to the impact of encoding scheme used for channel compression on listening quality. Moreover, quality ratings also differ with regard to content type, a known phenomenon within perceptual quality evaluations [37]. Results obtained by Rudzki et al. [67] show that bit-rate had minimal impact on localization accuracy which was dictated largely by Ambisonics order. However, bit-rate did have a significant impact on *timbral quality* which, as timbre significantly influences the overall audio quality, would support findings from [52]. Strictly focusing on bit-rate, Sen et al. [71] and Peters et al. [57] highlight the improved encoding quality of HOA using the MPEG-H codec compared to other approaches. To combat high bandwidth consumption due to the channel size of HOA content, the MPEG-H codec [26] was developed through competitive standardization efforts to include an efficient compression scheme for HOA signals. Instead of compressing individual channels, the Ambisonics signal is spatially decomposed into salient and background signals, which are then compressed in several transport channels. Consequently, the salient sources retain a higher proportion of bits than less spatially relevant background signals (see [29] for a detailed description). Perceptual evaluations using the MUSHRA method [30] and a mixture of $3rd$, $4th$, and $6th$ order Ambisonics show that even at lower bit-rates, the MPEG-H encoder-decoder pipeline still performs in the "good" range on a MUSHRA scale.

Regarding the Ambisonics order, numerous studies have been conducted using specific localization tests either via the method of adjustment [11, 67, 80], or acoustic/physical pointing [6, 10, 28, 62]. All these studies show improved localization accuracy of HOA over FOA. Thresh et al. [80] compared not only the localization of FOA and HOA ($3rd$ and $5th$ order) conditions but also between a real loudspeaker array and virtually rendered binaural sources over headphones. Their results show that for both groups the largest improvement of localization accuracy was from $1st$ to $3rd$ order Ambisonics, with marginal improvement from $3rd$ to $5th$ order. However, such differences were only observable once outliers had been removed for source positions which are more challenging to localize. During the listening test, subjects were instructed to maintain a forward-facing position, and thus, no head movements could be used to help determine source locations. As even small head movements can be used to significantly reduce localization confusion [3], it is unclear if the same results would be observed given free exploration. The same observations of reduced localization error between $1st$, $3rd$ and $5th$ order Ambisonics were also reached by Rudzki et al. [67]. Considering more potential cross-modality and situational effects discussed in Section 2.1, Huisman et al. [28] evaluated Ambisonics source localization with differing visual information. Comparing the results of Ambisonics orders from blindfolded subjects, the most pronounced localization error was found using FOA, particularly at increased azimuth angles, consistent with results from [80]. Furthermore, their results aggregated over HOA conditions indicate that localization error is *decreased* with the addition of real or virtual visuals.

Lastly, many of the studies above utilize noise or non-reverberant single instrument recordings for localization evaluations [6, 10, 28, 62, 67, 80]. However, there is a growing interest in utilizing stimuli that departs from clinical laboratory settings. This ecologically valid content provides an opportunity to understand quality aspects in more acoustically complex environments more representative of real-life-like settings [86]. In the context of this study, understanding if ecologically valid content can be used to evaluate certain audio quality aspects (i.e., bit-rate and dimensions of spatial quality) will lend further support to experiments that are targeted at evaluating technology for interactive, immersive experiences.

3 METHOD

To evaluate the proposed hypotheses $\mathcal{H}_1$ and $\mathcal{H}_2$, a mixed between- and within-subject test design is used. To provide applicability to real-life interactive content experiences and to understand if quality differences can be discerned in using less clinical program material, a focus on using ecologically valid stimuli is employed to evaluate

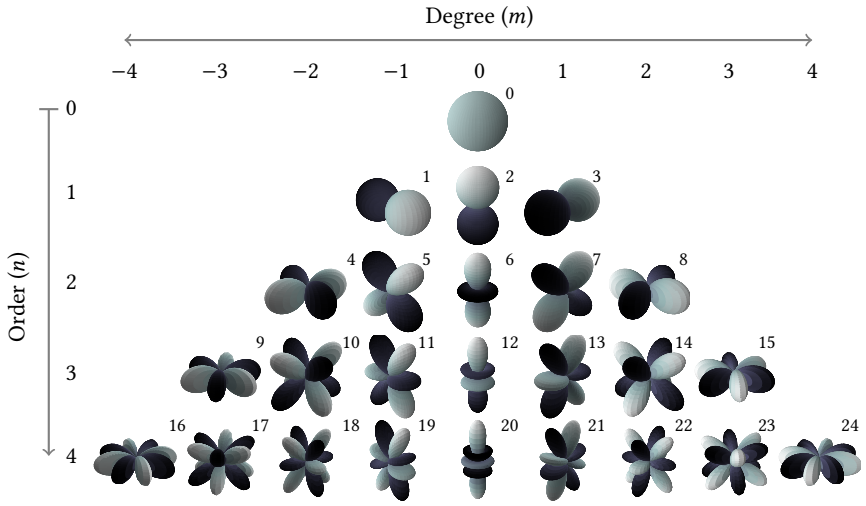


Figure 1: First 25 spherical harmonic directivity patterns in AmbiX channel ordering, of orders $n \leq 4$, and index m denoting the degree for each order limited by $-n \leq m \leq n$.

$\mathcal{H}_3$ (described in Section 3.1). The between-subject aspect presents different interactive formats to three different subject groups (described in Section 3.2). Within each of the three subject groups, two test sessions were conducted that targeted two aspects related to the perception of immersive audio for VR, using Ambisonics (described in Section 3.3). Both test sessions were conducted using a single stimulus absolute category rating (ACR) method [32]. While using this method, subjects sequentially rate stimuli based on their own merit, meaning that all judgments are made in isolation against the subject's internal reference within the restraints of the evaluation criteria. Previous research has demonstrated that the ACR method can be used for audio codec evaluations in comparison to the more commonly employed MUSHRA method [15]. The main motivation here is to yield results comparable to real-life content consumption without the ability to discriminate between all quality conditions at once. Finally, to evaluate $\mathcal{H}_4$, real-time exploration data of subjects were recorded for comparison across the interactive formats.

To conduct the experiments, a test framework has been designed to allow the automation, control, interaction, and real-time playback of Ambisonics audio and visual media. The framework is built with a server-client-based architecture using MaxMSP[1] and the Unity3D[2] game engine. As the server, MaxMSP handles the test configuration (including method, questionnaire, results management, and media files) in addition to hosting virtual studio technology (VSTs) for real-time audio rendering. As the client, the Unity3D game engine is programmed to handle all VR mechanics, user interfaces, and video playback, as well as parsing all interactive real-time information to the server. Communication between the server and client is done via Open Sound Control packets over a User Datagram Protocol connection. Real-time information relevant to audio rendering is triggered via the client's physics engine at an update rate of 50 Hz (20 ms). For this study, both the client and server operated on the same local machine. Test configuration was done by loading a JSON file into the server with all audio VSTs, rendering parameters, relative paths to all audio and video test stimuli, and auxiliary test constraints. The server would then calculate all unique test items based on the audio-visual content and ACR method, which would then be sequentially sent to the client for evaluation at runtime. Test sequences are fully randomized across both the scene recording and the degradation spatial audio conditions. Once the server has recognized a connected client, all subsequent actions can be automated and triggered via the subjects inside the virtual test environment, meaning no additional involvement is required by the test administrator.

3.1 Stimuli

After an internal review of potential stimuli, five audiovisual scenes were selected from a public database [66] and employed for both quality evaluation sessions. The database contains 360° video recordings at 8K resolution and 4th-order Ambisonics audio content in AmbiX format for reproduction with a high degree of audiovisual fidelity. While other audiovisual databases are available (e.g., [12, 55, 68]) few are publicly available that offer high-resolution spatial audio and visual fidelity, recorded in real-life recordings. Studies using non-ecologically valid stimuli may not provide realistic results of real environments [2, 86]. Consequently, scenes from the selected audiovisual database offer an opportunity to yield results more representative of content that is likely to be consumed, addressing our Hypothesis $\mathcal{H}_3$. Moreover, complementary aspects of the 'Immersive Methodolology' [45, 59] are adopted in this study to include stimuli sequences $\approx$ 60 s in duration, which is typically longer than most auditory evaluation stimuli available in other databases. With the selected database, subjects are given time to digest the context of the content and provides a more realistic media experience. Specifcally, the five recordings selected were *BuskingUnderpass*, *ForestWalk*, *ParkFountains*, *Skateboarding*, and

[1]www.cycling74.com
[2]www.unity3d.com

Train. The selected scenes were based on the criteria of providing a variety of sources for both audio and visual domains, in addition to audio-visually coupled sources. These sources then include static and dynamic movements, possess both continuous and transient audio signals, range in auditory source extent, and vary in content type (e.g., urban, speech, music, ambient, and nature sounds).

3.2 Interactive Formats

To evaluate the influence of various modalities and interaction types, three interactive formats are employed, illustrated in Figure 2. Format ***A*** is binaurally rendered audio. The audio is rendered to reflect the subject's head movements in real-time using an HTC Vive tracker mounted to the headphones. As such, this format ***A*** is a combination of vestibular and auditory sensory cues. Format ***A2DV*** refers to binaurally rendered audio with the addition of video information provided via a 2D display. Control of the viewing angle (camera) is done via a mouse with a click+hold+drag gesture, similar to many 360° video players. Consequently, the binaural audio is disconnected from a subject's real vestibular cues. Upon releasing the mouse button, the camera stops moving in the 360° scene, does not continue to turn, and does not react to any mouse movements. The monitor used for the video presentation was a 27-inch EIZO EV2795-BK with 4K resolution. A second monitor was provided with an interface for controlling the test and providing quality ratings. Finally, the ***AHMD*** format provides the 360° video content via an HMD in addition to the binaurally rendered audio over headphones. In doing so, this combines both vestibular, auditory, and visual modalities. The HMD used for this study was the Valve Index, allowing a 120° field of view and refresh rate of 90 Hz and a combined resolution of both eyes of 2880×1600. To provide subjects with a means of giving a quality rating and operating the test, a small virtual interface was programmed into the host video player. The interface could be shown/hidden at any time by subjects so as not to hinder their view, as well as being placed anywhere inside the scene at a distance of 1.5 m. When showing the interface, a laser pointer would appear that allowed subjects to interact with buttons (*Play*/*Next*) and the rating slider.

For all three interactive formats, the real-time binaural Ambisonics audio was rendered using the SPARTA AmbiBin VST [47] and fed head-tracking data at a rate of 50 Hz (20 ms). A generic head-related transfer function was used to binaurally render the audio for all subjects at a sampling rate of 48 kHz. The audio presentation was done using Beyerdynamic DT 770 Pro closed headphones connected via an RME Babyface audio interface.

3.3 Audio Conditions

To evaluate audio quality for each interactive format, we selected two audio degradations; bit-rate and Ambisonics order. As discussed in Section 2.2, many previous studies focusing on cross-modality influences include bit-rate as an audio degradation. Artifacts induced by bit-rate reduction are a known quality factor, and bit-rate itself is an important resource in many media experiences. Ambisonics order is chosen due to its prevalence in multiple research studies and is the main factor contributing to the spatial quality. The results of both degradation sets can provide insights into real applicable issues for 360° audiovisual content. To create the Ambisonics signals of various bit-rate, the MPEG-H encoder-decoder pipeline was used.

For a detailed explanation see [29]. Previous studies investigating bit-rate reduction of HOA signals have used Ambisonics codes such as OPUS to compress individual channels of the Ambisonics signal. In this study, the MPEG-H codec was selected due to its deployment and adoption within the industry and its status as an international standard within the broadcast sector. To the author's knowledge, this is the first study to use the MPEG-H HOA codec in an interactive evaluation with immersive content using an ACR method. To select the appropriate bit-rate conditions, several internal preliminary listening sessions were conducted by listening experts. The bit-rates selection should possess a range of conditions from easily perceptible audio compression artifacts to near-reference like quality. In doing so, we can observe any interaction effect present via the interactive format on the differing bit-rate conditions. The five bit-rate conditions subsequently chosen for the main test were **1156** (uncompressed), **36**, **30**, and **24** kb/s, with an anchor signal compressed at 24 kb/s *and* low-pass filtered at **3.5 kHz**. Such anchor signals are often low-pass filtered in audio quality evaluations involving codec-induced artifacts. The only deviation made in this study is that the low-pass filter was conducted on an already compressed signal. A final round of internal checks was done by selected expert listeners to ensure the degraded anchor signal would not skew the scale usage due to ceiling effects, i.e., the anchor being so bad that all other conditions are compressed at the upper end of the rating scale [89].

For the conditions varying in spatial quality, Ambisonics audio was employed that varied in the order n (see Figure 1). Five conditions were selected that included HOA $4th$, $3rd$, and $2nd$-order Ambisonics, FOA $1st$-order Ambisonics, and an additional omnidirectional anchor signal ($n = 0$). Figure 3 shows the spatial resolution across $1st$ to $4th$-order Ambisonics and the resulting increased spatial accuracy. Moreover, decreased spatial resolutions not only provide less localization accuracy in the perceptual domain but also induce less prominent auditory changes with respect to head movements.

4 EVALUATION PROCEDURE

The evaluation procedure was comprised of three phases; administrative, training, and evaluation phase. For each subject, two test sessions were conducted on separate days for the (within-subject) bit-rate and Ambisonics order evaluations. All subjects who registered to participate were pseudo-randomized in one of the three (between-subject) interactive format groups. Additionally, the degradation type for subjects' first and second sessions was also pseudo-randomized.

For the administrative phase, all subjects completed a data protection form, a payment information form, and a short demographic survey. For the demographics; subject ID, gender (*male*, *female*, and *non-binary*), age, and listening experience were noted. The listening experience was divided into *naïve* and *expert*. After completing all forms, subjects were provided written information regarding the test, the test instructions (specific to each interactive format), and a non-exhaustive list of quality attributes. Consistent with previous

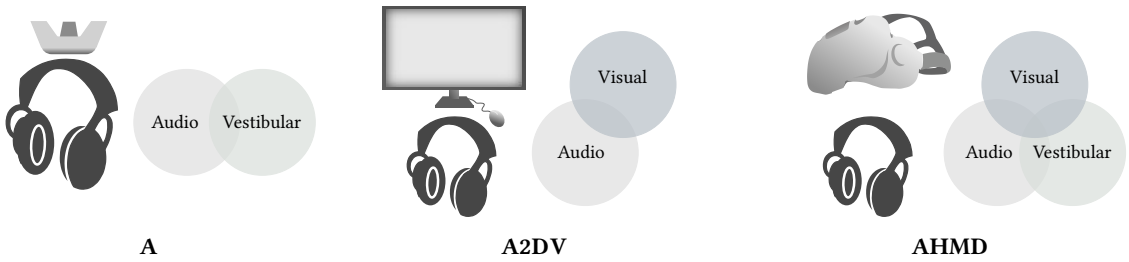


Figure 2: Interactive formats A, A2DV, and AHMD for between-subject groups depicting the stimuli-dependent sensory streams and modality integration.

research [52], two different evaluation 'questions' were given to the subjects depending on the degradation. For the bit-rate degradation test, subjects were instructed to rate *overall audio quality*. For the Ambisonics order degradation test, subjects were instructed to judge *spatial audio quality*. The quality attributes were taken from the SAQI database [39] and provided to help subjects formulate a single quality score based on potential artifacts. These included selected attributes under: *Timbre, Geometry, Time-behavior, Dynamics, Artifacts,* and *General*. At this stage, all written instructions were verbally clarified by the test administrator, along with any questions. Importantly, as an ACR single stimulus method uses no explicit reference, it was emphasized to all subjects that quality judgments should be based on comparison to their own internal reference, expectations, and prior experiences. Furthermore, subjects were informed that they may explore/interact with the content in any way that helps reach their goal of coming to a quality judgment (i.e., no behavioral restrictions were imposed).

Subjects then moved on to the training phase, where they conducted a small round of all degradations for a single scene. The purpose of the training phase is to ensure that all subjects understand on what basis they are providing quality judgments, how to operate the test, to set expectations for the range of qualities to be evaluated, and to check they can hear a difference in quality for (at minimum) the intended anchor signal. Different training scenes were used for the different degradation sessions. For the bit-rate degradation, an additional *Cheerleading* scene, and for the Ambisonics order degradation, a *Badminton* scene (both taken from the selected database [66]). None of the training scenes were used in the main evaluation. Subjects conducted the training as if it were the main test; evaluating an item, providing a rating, and moving on to the next item. For interactive groups **A** and **A2DV**, subjects were seated on a 360° swivel stool at a desk. For the **AHMD** group, subjects were sat on the same stool but centered in the middle of the lab. The same general starting direction (0±30° azimuth, 0±10° elevation) for each test item was administered for all subjects and all interactive groups. At any time, the test administrator helped with issues relating to operating the test mechanics or interfaces. All subjects reported that they could hear an audible difference in quality for (at least) one item.

Once subjects were comfortable with the test question, format, and mechanics, they moved on to the evaluation phase, where they evaluated all 25 items ($5_{scenes} \times 5_{conditions}$). The test administrator remained at an observation position at the rear of the lab and stayed to observe any health and safety issues for the **AHMD** group until the end of the test. Subjects were then thanked and escorted out of the lab. Payment for participation was 12 EUR/h. Overall, the study included 54 participants, broken into groups of 18 for the three interactive formats. For each format, the following number of *naïve* and *expert* (N:E) listeners took part, in addition to the gender distribution of *male*, *female*, and *non-binary* (M:F:NB). For format **A**, listener expertise was split $12_N{:}6_E$ over genders $12_M{:}6_F$. For format **A2DV**, listening experience was $10_N{:}8_E$, comprised of $12_M{:}6_F$. For **AHMD**, listening experience was $10_N{:}8_E$, with a gender split of $13_M{:}5_F$. The average age across participants for the complete study was 37.2 (SD = 13.824). No subjects reported motion sickness during the study, and all subjects reported normal hearing and had normal or corrected-to-normal vision.

5 RESULTS

5.1 Quality Ratings

The results for both evaluation sessions are shown in Figure 4 using mean values and 95% bootstrapped confidence intervals (CIs). To analyze the results, we performed a three-way mixed repeated measures analysis of variance (ANOVA) on the dependent variable **Rating** on independent variables **Scene** and **Condition** for mixed between groups of interactive **Format**. Normality distribution of residuals was inspected and found satisfactory for ratings acquired in both sessions. For the bit-rate session, the residual's standard error measured over all independent variables was 1.45, and 1.41 for the Ambisonics order session. Main effects are reported significant at values $p < .05$. Mauchly's test of sphericity was used to assess the assumption of equal variance across conditions. When violated, significance values after Greenhouse-Geisser ($\hat{\varepsilon}$) correction are reported. Greenhouse-Geisser correction is chosen over Huyhn-Feldt as values for Mauchly's (w) statistic return values below .75 [18]. Table 1 lists the results from the ANOVA and effect sizes for both bit-rate and Ambisonics order sessions. For both sessions, two significant main effects of **Scene** and **Condition** are observed, in addition to a significant interaction between **Scene** × **Condition**. Additionally, for the Ambisonics order session, an additional interaction effect was found between **Format** × **Condition**.

To evaluate significant interactions, multiple Tukey's HSD post-hoc t-tests were conducted for the different degradation tests. For the bit-rate test session, no significant interaction was present due to the independent variable **Format** on either the **Scene** and/or

Table 1: Table of ANOVA results detailing main effects, interaction effects, and generalized effect sizes for the bit-rate (left) and Ambisonics order (right) between-group test sessions. Significance is noted via asterisks notation (*) for given p-values at $p < .05$.

Evaluation Session	Bit-rate			Ambisonics Order		
Effect	F-value	p-value	Effect size	F-value	p-value	Effect size
Format	$F_{(2,51)} = .347$	$p = .071$	$\eta^2_G = .004$	$F_{(2,51)} = 2.896$	$p = .064$	$\eta^2_G = .019$
Scene	$F_{(4,204)} = 33.395$	$p < .001^*$	$\eta^2_G = .075$	$F_{(4,204)} = 5.563$	$p < .001^*$	$\eta^2_G = .018$
Condition	$F_{(4,204)} = 214.698$	$p < .001^*$	$\eta^2_G = .477$	$F_{(4,204)} = 246.915$	$p < .001^*$	$\eta^2_G = .542$
Format × Scene	$F_{(8,204)} = 1.096$	$p = .367$	$\eta^2_G = .005$	$F_{(8,204)} = 1.045$	$p = .401$	$\eta^2_G = .007$
Format × Condition	$F_{(8,204)} = 0.578$	$p = .367$	$\eta^2_G = .006$	$F_{(8,204)} = 2.442$	$p = .037^*$	$\eta^2_G = .023$
Scene × Condition	$F_{(16,816)} = 20.654$	$p < .001^*$	$\eta^2_G = .126$	$F_{(16,816)} = 2.949$	$p < .001^*$	$\eta^2_G = .024$
Format × Scene × Condition	$F_{(32,816)} = 1.249$	$p < .207$	$\eta^2_G = .017$	$F_{(32,816)} = 1.395$	$p = .105$	$\eta^2_G = .022$

Condition. Therefore, results across the three formats were pooled together to perform post-hoc comparisons of the interaction of **Condition** *within* each scene. As indicated by the results from the ANOVA, the number of significantly differently rated conditions varied across scenes (*BuskingUnderpass* = 6, *ForestWalk* = 9, *ParkFountains* = 6, *Skateboarding* = 7, *Train* = 9). Only the scenes *ForestWalk* and *Train* yielded significant differences between higher-quality conditions 36 vs. 1156 kb/s ($p = .002$ and $p = .012$, respectively). For lower-quality bit-rates, only the *Skateboarding* scene yielded a significant difference between conditions *LP* vs. 24 kb/s ($p = .033$).

For the Ambisonics order test, no significant interaction was present between **Format** × **Scene** × **Condition**. Consequently, the same post-hoc analysis was conducted to observe the effect of **Scene** on the number of significantly different conditions. As indicated by the smaller generalized effect size η^2_G in Table 1, the interaction effect of **Scene** × **Condition** is not as strong as with bit-rate, resulting in a more consistent number of significant differences across scenes (*BuskingUnderpass* = 3, *ForestWalk* = 7, *ParkFountains* = 7, *Skateboarding* = 7, and *Train* = 7). In contrast to other scenes, the *BuskingUnderpass* scene yielded no significant differences in comparisons between 1*OA* vs. 2*OA*, 3*OA*, and 4*OA* (i.e., first-order vs. all higher-order Ambisonics).

Unlike the bit-rate evaluation, a significant effect was indicated by the ANOVA analysis for **Format** × **Condition** for the Ambisonics order degradation. Therefore, three further separate post-hoc analyses were conducted on results pooled over all scenes for the individual formats. For format **A**, post-hoc analysis shows 7/10 comparisons to be significantly different between conditions 0*OA* vs. all Ambisonics conditions, and 1*OA* (first-order Ambisonics) vs. 2*OA*, 3*OA*, and 4*OA* (higher-order Ambisonics). For format **A2DV**, post-hoc analysis shows the same significantly different pairs as format **A**. For the **AHMD** format, post-hoc analysis reveals 8/10 significantly different comparisons of conditions, the same seven as with previous formats with the addition of 4*OA* vs. 2*OA* ($p < .001$).

5.2 Behavioral Data

To observe any effect the interactive format may have on viewing behavior, we analyze subjects' head rotations (gained from the recorded tracking data of respective equipment) in addition to the time taken to complete the test session. Figure 5 shows the normalized density distribution of rotational yaw movements over the potential 360° viewing/listening azimuth angle of the content. Figure 5a (top) shows results for the bit-rate degradation session, and 5b (bottom) for the Ambisonics order degradation session. For both degradation sessions, the distribution of viewing angle over time differs between both the interactive format and scene content resulting in various uni-/bi-/multi-modal shapes. The distribution data is characterized by mean (M) and standard deviations (SD) in Table 2 over all conditions. Additionally, Figure 5 presents absolute values for cumulative yaw rotation and total test time. For the bit-rate session, the mean time taken for subjects to complete each scene for the three formats was; **A** = 46.1 s, **A2DV** = 48.5 s, and **AHMD** = 47.9 s. For the Ambisonics order session, average test times were; **A** = 47.8 s, **A2DV** = 50.1 s, and **AHMD** = 46.1 s.

6 DISCUSSION

6.1 Quality Ratings and Interactive Format

Based on the analysis in Section 5, the interactive **Format** had no significant effect on quality ratings for bit-rate **Conditions**. As seen in Figure 4, mean quality ratings are highly consistent for all three interactive formats. This suggests that even with the inclusion of visual information and head movements, subjects evaluate artifacts induced by bit-rate reduction in a similar manner. The cause for this is likely due to the decoupling of the visual input and vestibular cues with the inherent artifacts induced by bit-rate reduction. These compression artifacts are frequency band limitations and temporal smearing [14, 40, 41] that are not strictly sensory-coupled with any visual or vestibular input. Hence, subjects may be able to disconnect their auditory processing without any potential biases. As the three formats also differ in cross-modal interactions, it is reasonable to suggest no cross-modal effect was present (or significant enough) to mask or influence quality judgments. Consequently, we reject our hypothesis $\mathcal{H}_1$.

On the other hand, a significant interaction was observed between the quality of Ambisonics order **Conditions** and interactive **Format**. Therefore, we do not reject Hypothesis $\mathcal{H}_2$. In this case, the auditory degradation is directly coupled with visual information and mouse-induced camera movements. For example, the visual appearance of a fountain has audiovisual attributes such as position and distance (i.e., spatial location), in addition to attributes such as physical size and auditory blur (see Figure 3). Therefore, depending

Table 2: Mean (M) and standard deviation (SD) in degrees of viewing angle over time, pooled over conditions.

Evaluation Session	Bit-rate						Ambisonics Order					
Format	A		A2DV		HMD		A		A2DV		HMD	
Scene	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*
BuskingUnderpass	20.13	28.26	27.02	64.53	-7.82	77.52	14.49	57.2	25.96	73.75	8.0	91.95
ForestWalk	25.24	26.53	-39.03	87.9	-62.33	87.15	20.9	54.39	-35.83	94.75	-45.35	96.93
ParkFountains	19.08	31.3	18.45	89.6	-11.45	95.93	15.45	60.05	22.59	92.72	5.37	98.35
Skateboarding	25.57	30.15	-36.04	93.44	-50.8	90.15	22.5	53.38	-36.64	101.78	-48.98	94.89
Train	23.98	25.59	-25.86	96.6	-43.93	90.83	19.41	52.87	-25.49	103.14	-40.64	89.89

on the interaction type or visual stimulus, various degrees of perceptual incongruency may be present that influences our sensitivity to spatial differences, thus impacting quality judgments. Post-hoc analysis performed in Section 5 describes a significant difference between conditions $2OA$ vs. $4OA$ for the **AHMD** group, in comparison to interactive formats **A** and **A2DV**. Looking at Figure 4, the quality ratings aggregated over scenes also suggest that formats **A** and **A2DV** yield similar responses, whereas the quality ratings for decreasing Ambisonics order presented in group **AHMD** appear to fall more consistently across the scenes. Considering format **A** is binaural audio with vestibular stimulation and no visuals, and format **A2DV** is binaural audio with visuals but no direct vestibular stimulation coherent with visual or auditory streams, it seems only the combination and integration of all three sensory inputs that are completely synchronized with the evaluation content in format **AHMD** had a significant effect on quality judgments. Previous research describes that adding additional sensory channels aids in detecting artifacts in sensory-coupled stimuli [27]. However, the results here indicate that the integration of audio and visual modalities, when driven by subjects' own body movements, culminate in more critical judgments of spatial audio quality for lower Ambisonics orders. Complementary to this, cognitive aspects related to formats **A** and **A2DV** may also decrease sensitivity to spatial quality due to the lack of a visual reference (**A**) or additional non-immersive visuals, unrelated vestibular stimulation, and mouse controls (**A2DV**).

6.2 Quality Ratings with Ecologically Valid Stimuli

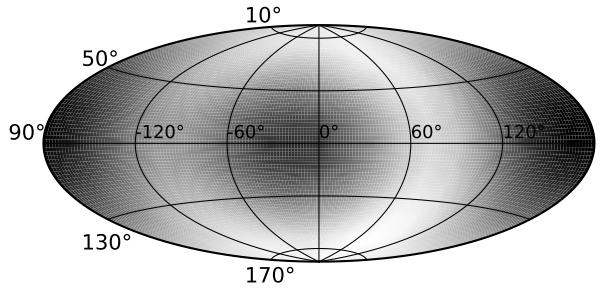
Another component of this study was the evaluation of bit-rate and Ambisonics orders using stimuli with a higher ecological validity to determine if quality differences could be perceived. Results from the ANOVA (Table 1) reveal a significant interaction effect between **Scene** × **Condition**. Ratings within each scene show overall trends of declining quality with respect to bit-rate but provide either different levels of absolute quality or a finer granularity of quality ratings between high and low-quality conditions depending on the scene content. For example, the *BuskingUnderpass* scene, which features a saxophone player in a highly reverberant space, spans ≈ 35 points between mean scale values 35 and 60, yielding six significantly different conditions. In contrast, the *ForestWalk* scene, which features noise-like water, speech, and background music, spans ≈ 70 points between mean scale values 15 and 85, yielding nine significantly different pairs. In general, the overall scores shown in Figure 4 (aggregated over **Scene**) show a distortion curve of higher-quality conditions comparable to previous studies using the MPEG-H codec and MUSHRA method [71]. However, due to the lack of an explicit reference in the ACR method, the curve is shifted lower on the quality scale, as demonstrated in [15].

For the Ambisonics order evaluation, significant interaction was also shown between **Scene** × **Condition**, although with a reduced generalized effect size than that seen in the bit-rate evaluation. The ratings within each scene in Figure 4 imply that this reduced effect size is due to the conditions being rated more consistently within scenes for all interactive formats, which is also supported by post-hoc analysis, yielding a more consistent number of significant differences.

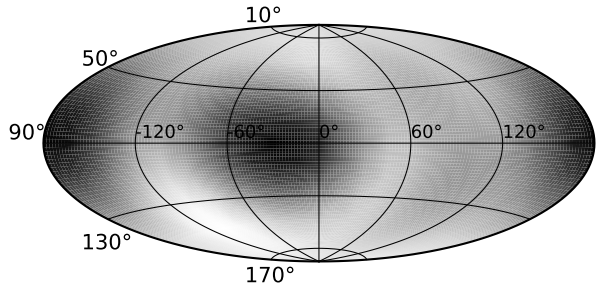
To decipher any relationship between individual acoustic scene properties and quality ratings for both the bit-rate and Ambisonics order evaluations is beyond the scope of this study. However, for the bit-rate evaluation, it is reasonable to suspect that differences are induced by the spectral and temporal structure (i.e., transient or steady-state signals) within the content. It is also interesting to note that this influence on content type has a stronger effect on bit-rate quality than Ambisonics order spatial quality. Transient signal properties and frequency content also play a role in localization [7], and one might therefore suspect equal amounts of variation across scenes compared to the bit-rate evaluation, which was not the case. Moreover, the specific questions of *overall audio quality* and *spatial audio quality* will (intentionally) alter subjects' internal criteria of assessment, thus altering the priority of certain quality features within the stimuli. Nevertheless, despite the cause, the results and discussion highlight that the employed content can be used to determine differences in bit-rate and Ambisonics order. The significant interaction between **Condition** × **Format** for the Ambisonics order also suggests that for such real-life-like content, a higher spatial quality may be needed, which is not specifically dependent on the differences between scene content. Therefore, we see sufficient support not to reject our hypothesis $\mathcal{H}_3$. This finding supports the use of ecologically valid stimuli for subjective studies in immersive experiences while still being able to evaluate specific quality attributes critically.

6.3 Immersive Experience Expectations

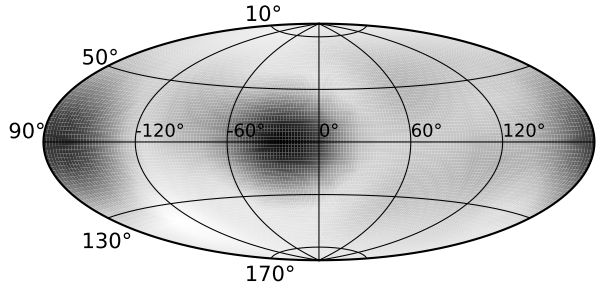
Although there is support for our hypothesis $\mathcal{H}_2$, the reason why the interactive format yields significant interaction with Ambisonics order requires further investigation. Instead of the varying spatial resolutions, an alternative justification for the significance may be



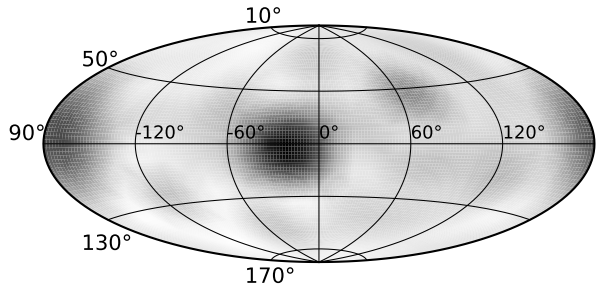
(a) 1*st* order



(b) 2*nd* order



(c) 3*rd* order



(d) 4*th* order

Figure 3: Stereographic projection of directional integrated sound power level of *FountainPark* item, showing the increasing spatial resolution and reduced blur of Ambisonics orders from n =1 to 4. Dark shaded areas indicate higher sound power level.

that the interactive format **AHMD** altered subjects' inner reference and expectation as to what "Excellent" and "Poor" spatial audio quality is, as presented on the ACR scale, compared to formats **A** and **A2DV**. Our inner reference and perception of the world are built on expectations gained from our previous experiences [8, 34, 35]. It is not unreasonable to hear audiovisual sources around us without seeing them, as much as it is to experience spatial audio without directly being driven by our own head movements. However, our most trained reference model is from everyday life with audiovisual sensory input driven by our body movements. Consequently, a distinction could be made between subjects' inner reference model being shared across interactive formats and the sensitivity to different spatial resolutions changing, or sensitivity of spatial resolution persisting and the scale being stretched due to changes in expectation, optimal modality integration, or other contextual factors [83]. The two possibilities are also not mutually exclusive, and both an ability to detect lower spatial quality and an altered inner reference may occur.

For every test, subjects undertook a training phase where the range of quality was presented to calibrate subjects' interpretation and expectations of higher and lower quality. However, in accordance with [32], subjects were not explicitly told that the worst and best quality conditions necessarily correspond to the lowest and highest ends of the scale. Visual inspection of Figure 4 suggests that a saturation point is reached at around "Good", for the highest quality conditions in both degradation tests and for all interactive formats. Indeed, the 1156 kb/s and 4OA items exactly represent the same stimuli for both tests and provide remarkably similar subjective ratings. Previous studies have highlighted the difference between ACR and MUSHRA methods, particularly regarding the higher end of the scale [4, 15], where higher-quality ACR ratings also reached a saturation point of ≈ 80-points. However, having a consistent saturation point across all formats suggests their interpretation of the highest-quality stimuli remains similar. The ACR evaluation method (as with MUSHRA) also exists under a broader category of evaluation methods known as *direct scaling*. Here, subjects are asked to *directly* prescribe a number that reflects their perceived magnitude of a sensation or attribute. Such scaling methods can be prone to several biases [89], particularly when presented without an explicit high-quality reference condition for comparison [65].

The ACR method differs from most direct scaling methods due to the sequential single stimulus presentation. Consequently, subjects' opinions on higher and lower-quality conditions may change as they are progressively exposed to more content. For example, subjects may have reservations about using the upper end of the scale for test items initially presented to leave headroom for items of a potentially higher quality later in the test, even if the presented item is of the highest quality. Therefore, quality saturation may occur simply due to a proportion of high-quality items being rated lower because they appear too early, and higher later on because subjects have a better grasp of high quality. To check for leading stimulus effects, ratings for the first 10 items were collected and inspected to see the ratio of ratings above and below 70-points. Using 10 items allows the possibility of subjects to have heard at least two of the different conditions, and a benchmark of 70-points appears to be the quality saturation level. For instances where the highest-quality condition 4OA and 1156 kb/s were presented within the first 10

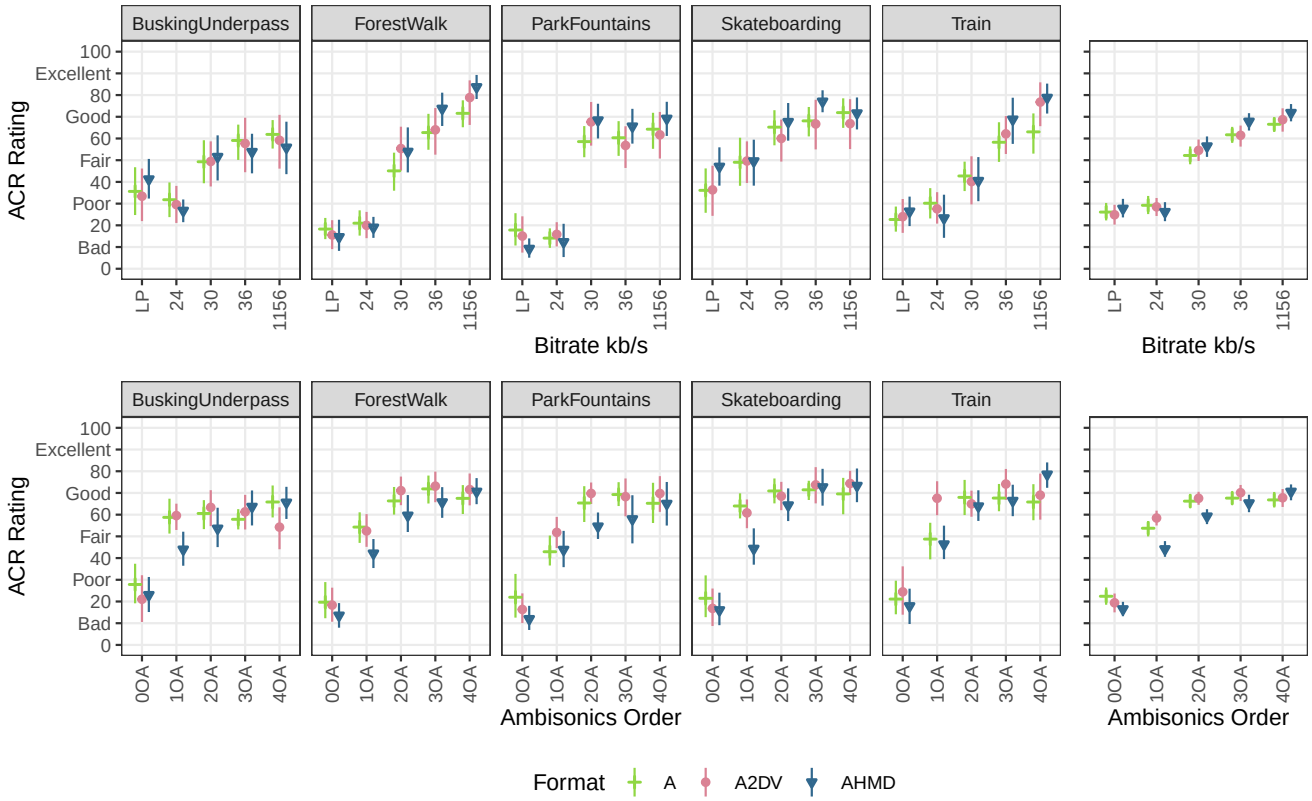


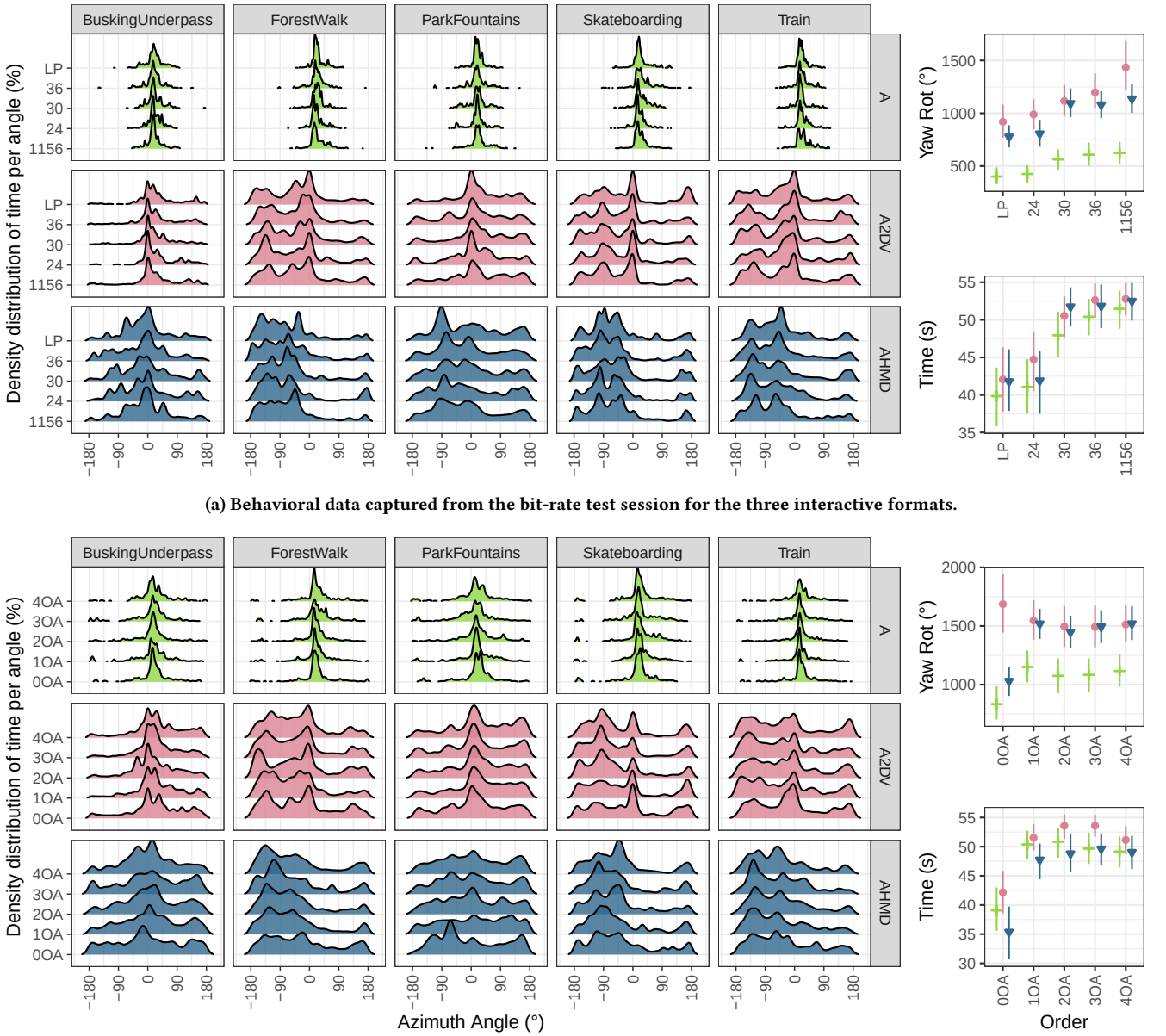
Figure 4: For each session (top = bit-rate, bottom = Ambisonics order), MOS and bootstrapped 95% CIs for each scene are shown (from left to right) across the first five plots. The last plot shows the mean MOS values and 95% bootstrapped CI values for data culminated across all scenes. Ratings for each condition are presented on the x-axis, and are grouped in three for each interactive format (A, A2DV, and AHMD).

items, 52% and 59% were above 70-points respectively. As such, we can infer that subjects had no reservations about using the scale's upper quartile for the highest-quality condition when presented early in the test, and not simply in the latter half of the test after exposure to more stimuli. Therefore, a reasonable argument may be made that hypothesis $\mathcal{H}_2$ cannot be rejected, and the sensitivity to spatial resolutions is impacted via the interactive format, and not scale usage or inner reference expectations.

6.4 Modality Driven Behavior

The results presented in Table 2 and Figure 5 show that the interactive format has a notable effect on subjects' behavior and exploration patterns. The density distributions depicted in Figures 5a and 5b reveal that for interactive format **A**, subjects' Yaw movements are highly focused towards the original seating position with very little exploration, resulting in a uni-modal distribution. For the bit-rate evaluation, this is supported by a consistent mean angle of $\approx 20^\circ$ and standard deviation of $\approx 27^\circ$ across all presented scenes. For the Ambisonics order session, the mean values are slightly more varied and occupy a larger standard deviation of $\approx 56^\circ$, implying that greater attempts to explore the spatial audio were made under the context of this evaluation question. For the interactive formats **A2DV** and **AHMD** where visual information is presented, exploration behavior becomes far more scene specific. Through visual inspection of the behavior patterns for both degradation types in Figure 5, it can be concluded that subjects' exploration is not only more spread but also has dominant viewing angles, which differs from the original viewing direction. While previous research demonstrates the role of auditory cues in guiding viewing attention [12, 38], our results here also demonstrate the changes in exploration behavior given the introduction of visuals in immersive content.

Comparing the exploration of format **A2DV** across the bit-rate and Ambisonics order sessions shows that mean viewing angles are remarkably similar. This is supported through the comparable density distributions depicted in both tests (pink A2DV distribution in Figures 5a and 5b.) As with format **A**, standard deviation values per scene are slightly larger again for all scenes in the spatial audio evaluation than in the bit-rate evaluation. However, there is some evidence suggesting the inclusion of self-driven head motions in addition to visuals in the **AHMD** format has a further influence on exploration behavior. Considering both **A2DV** and **AHMD** formats include visuals and spatial audio, it might be reasonable to assume they share the same viewing patterns. However, variations



(a) Behavioral data captured from the bit-rate test session for the three interactive formats.

(b) Behavioral data captured from the Ambisonics order test session for the three interactive formats.

Figure 5: Behavioral data captured from the bit-rate (top) and Ambisonics order (bottom) test sessions. Data across all plots are split into three interactive formats, which are represented via three colors. For both test sessions, three sets of data are provided; a *scene* × *format* × *condition* matrix showing density distributions of subjects' yaw (y-axis) head rotation data, absolute cumulative yaw rotation aggregated over all scenes for each condition, and time taken spent aggregated over all scenes for each condition.

in mean viewing angle have noticeable differences. For example, the *BuskingUnderpass* scene, **A2DV** vs. **AHMD** $\approx \Delta 34°$ in the bit-rate session, and **A2DV** vs. **AHMD** $\approx \Delta 17°$ in the Ambisonics order. Although quite similar in standard deviation, the distributions of the two different formats (pink vs. blue) within different evaluation sessions also show some different multi-modal shapes. The absolute cumulative Yaw rotation and evaluation time shown on the right in Figures 5a and 5b highlight that for most conditions, both the **A2DV**

and **AHMD** interactive formats yield roughly the same amount of total movement and test time. Therefore, the differences described in density distributions can only be due to differences in exploration, and not simply total movement or test time. When considering hypothesis $\mathcal{H}_4$, we find support for the statement based on the clear distinction between interactive formats **A** vs. **A2DV** and **AHMD**. However, given the discussion above, it is also reasonable not to reject the hypothesis based on the differences between formats **A2DV** vs. **AHMD**.

7 CONCLUSION

This study evaluated two spatial audio quality aspects relevant to interactive immersive experiences using Ambisonics technology in a mixed between- and within-subjects experimental design. To observe any cross-modality influences on quality judgments, the evaluation was conducted using three different interactive formats: 360° head-tracked spatial audio only (**A**), mouse/camera controlled spatial audio with 360° 2D visuals (**A2DV**), and head-tracked spatial audio with 360° visuals over a head-mounted display (**AHMD**). Quality assessment was split into two sessions, *overall audio quality* with conditions degraded through Ambisonics bit-rate using the MPEG-H codec, and *spatial audio quality* with conditions degraded via Ambisonics order. An emphasis on ecological validity is established by employing real-life-like stimuli of a longer duration ($\approx$ 60 s) and evaluated using the single stimulus ACR method.

Firstly, our results show that the interactive format had a significant effect on quality ratings for Ambisonics order but not for bit-rate. Specifically, the ratings for **AHMD** format imply a perceptual benefit of $4th$-order Ambisonics over $2nd$-order, in contrast to formats **A** and **A2DV**, where no increase in quality was perceived beyond $2nd$-order. This suggests that the optimal sensory integration of audio, visual, and vestibular modalities in the **AHMD** format provided subjects with a higher degree of sensitivity to differences between higher-order Ambisonics conditions; and that different cognitive factors as a result of non-stimuli dependent information in formats **A** and **AHMD** may reduce our sensitivity to detect lower spatial resolutions. Secondly, our results demonstrate that ecologically valid stimuli can be used to yield significant differences in audio quality conditions within interactive experiences. Moreover, further evidence is provided to show that such stimuli can be used in conjunction with the ACR single stimulus method and unrestricted head movements. Both observations support the position that *overall* and *spatial* audio quality can be evaluated with stimuli, methods, and behavioral conditions more comparable to real-life. Thirdly, our behavioral analysis shows that the interactive format has a considerable impact on exploration behavior. While the introduction of the visual modality has the greatest impact on exploration, differences between formats **A2DV** and **AHMD** can also be observed, suggesting viewing behavior can also be influenced if audiovisual sensory streams are coupled with vestibular stimulation. Overall, our results demonstrate that trade-offs between certain parameters and perceptual quality can be dependent on the interactive format and that to maintain quality levels, format-dependent audio rendering may be needed. Furthermore, the changes in behavior across interactive formats suggest the potential for format-specific optimization based on exploration patterns likely to be observed.

ACKNOWLEDGMENTS

The authors would like to thank all IIS employees and external subjects who participated in the subjective experiments. In particular, Pavan Kantharaju for his input and expertise regarding MPEG-H and Adrian Herzog for assisting in Ambisonics analysis. The authors would also like to thank the anonymous referees for their valuable comments and helpful suggestions, and William Menz for additional proof reading. This research was funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Project No.: 444832250.

REFERENCES

[1] David Alais and David Burr. 2004. The Ventriloquist Effect Results from Near-Optimal Bimodal Integration. *Current Biology* 14, 3 (2004), 257–262. https://doi.org/10.1016/j.cub.2004.01.029

[2] Hudson D. Bailey, Aidan B. Mullaney, Kyla D. Gibney, and Leslie D. Kwakye. 2018. Audiovisual Integration Varies With Target and Environment Richness in Immersive Virtual Reality. *Multisensory Research* 31, 7 (2018), 689–713. https://doi.org/10.1163/22134808-20181301

[3] Durand R. Begault and Elizabeth M. Wenzel. 2001. Direct Comparison of the Impact of Head Tracking, Reverberation, and Individualized Head-Related Transfer Functions on the Spatial Perception of a Virtual Speech Source. *Journal of the Audio Engineering Society* 49, 10 (2001), 904–916.

[4] Kathryn Beresford, Natanya Ford, Francis Rumsey, and Slawomir K. Zielinski. 2006. Contextual Effects on Sound Quality Judgements: Listening Room and Automotive Environments. In *Audio Engineering Society 120th Convention*. Paris, France, 1–13. http://www.aes.org/e-lib/browse.cfm?elib=13452

[5] Paul Bertelson. 1999. Ventriloquism: A Case of Crossmodal Perceptual Grouping. *Advances in Psychology* 129 (1999), 347–362. https://doi.org/10.1016/s0166-4115(99)80034-x

[6] Stéphanie Bertet, Jérôme Daniel, Etienne Parizet, Laëtitia Gros, and Oliver Warusfel. 2007. Investigation of the Perceived Spatial Resolution of Higher Order Ambisonic Sound Fields: A Subjective Evaluation Involving Virtual and Real 3D Microphones. In *30th AES International Conference on Intelligent Audio Environments*. Saariselkä, Finland, 1–9.

[7] Jens Blauert. 1996. *Spatial Hearing: The Psychophysics of Human Sound Localization* (revised edition ed.). MIT Press, Cambridge, MA, USA. https://doi.org/10.7551/mitpress/6391.001.0001

[8] Jens Blauert and Ute Jekosch. 2012. A Layer Model of Sound Quality. *Journal of the Audio Engineering Society* 60, 1/2 (2012), 4–12.

[9] Karsten Bormann. 2005. Presence and the Utility of Audio Spatialization. *Presence* 14 (2005), 278–297. https://doi.org/10.1162/105474605323384645

[10] Sebastian Braun and Matthias Frank. 2011. Localization of 3D Ambisonic Recordings and Ambisonic Virtual Sources. In *1st International Conference on Spatial Audio*. Detmold, Germany, 1–6.

[11] Ji-Ho Chang and Wan-Ho Cho. 2018. Impairments of Binaural Sound Based on Ambisonics for Virtual Reality Audio. In *10th Sensor Array and Multichannel Signal Processing Workshop (SAM)*. Sheffield, UK, 341–345. https://doi.org/10.1109/sam.2018.8448493

[12] Fang-Yi Chao, Cagri Ozcinar, Chen Wang, Emin Zerman, Lu Zhang, Wassim Hamidouche, Olivier Deforges, and Aljosa Smolic. 2020. Audio-Visual Perception of Omnidirectional Video for Virtual Reality Applications. *2020 IEEE International Conference on Multimedia & Expo Workshops (ICMEW)* 00 (2020), 1–6. https://doi.org/10.1109/icmew46912.2020.9105956

[13] Shyamprasad Chikkerur, Vijay Sundaram, and Martin Reisslein. 2011. Objective Video Quality Assessment Methods: A Classification, Review, and Performance Comparison. *IEEE Transactions on Broadcasting* 57, 2 (2011), 165–182. https://doi.org/10.1109/tbc.2011.2104671

[14] Markus Erne. 2001. Perceptual Audio Coders "What to Listen For". In *AES 111th Convention*. New York, NY, USA, 1–10.

[15] Bernhard Feiten, Alexander Raake, Marie-Neige Garcia, Ulf Wüstenhagen, and Jens Kroll. 2009. Subjective Quality Evaluation of Audio Streaming Applications on Absolute and Paired Rating Scales. In *AES 126th Convention*. Munich, Germany, 1–9.

[16] Randy F. Fela, Andréas Pastor, Patrick Le Callet, Nick Zacharov, Toinon Vigier, and Søren Forchhammer. 2022. Perceptual Evaluation on Audio-visual Dataset of 360 Content. *arXiv* (2022), 1–6. arXiv:2205.08007

[17] Richard E. Ferdig, Karl W. Kosko, and Enrico Gandolfi. 2020. The Use of Ambisonic Audio to Improve Presence, Focus, and Noticing While Viewing 360 Video. *Journal of Virtual Worlds Research* 13, 2-3 (2020), 1–4.

[18] Andy Field, Jeremy Miles, and Zoe Field. 2012. *Discovering Statistics Using R*. SAGE, London, UK.

[19] Matthias Frank, Franz Zotter, and Alois Sontacchi. 2015. Producing 3D Audio in Ambisonics. In *AES 57th International Conference: The Future of Audio Entertainment Technology - Cinema, Television and the Internet*. Hollywood, CA, USA, 1–8.

[20] Marie-Neige Garcia, Robert Schleicher, and Alexander Raake. 2011. Impairment-Factor-Based Audiovisual Quality Model for IPTV: Influence of Video Resolution, Degradation Type, and Content Type. *EURASIP Journal on Image and Video Processing* 2011, 1 (2011), 1–14. https://doi.org/10.1155/2011/629284

[21] Michael A. Gerzon. 1973. Periphony: With Height Sound Reproduction. *Journal of the Audio Engineering Society* 1, 21 (1973), 2–8.

[22] David Hammerschmidt and Clemens Wöllner. 2017. Audio-Visual Quality Perception in Musical Performance Videos. *Musilpsychologie* 1, 27 (2017), 112–127.

[23] Naomi Harte, Eoin Gillen, and Andrew Hines. 2015. TCD-VoIP, A Research Database of Degraded Speech for Assessing Quality in VoIP Applications. *2015 Seventh International Workshop on Quality of Multimedia Experience (QoMEX)* (2015), 1–6. https://doi.org/10.1109/qomex.2015.7148100

[24] Erik Hellerud, Audun Solvang, and Peter Svensson. 2009. Spatial Redundancy in Higher Order Ambisonics and its Use for Low Delay Lossless Compression. In *IEEE International Conference on Acoustics, Speech and Signal Processing (IEEE International Conference on Acoustics, Speech and Signal Processing)*. 269–272. https://doi.org/10.1109/icassp.2009.4959572

[25] Claudia Hendrix and Woodrow Barfield. 1996. The Sense of Presence within Auditory Virtual Environments. *Presence* 5, 3 (1996), 290–301. https://doi.org/10.1162/pres.1996.5.3.290

[26] Jürgen Herre, Johannes Hilpert, Achim Kuntz, and Jan Plogsties. 2014. MPEG-H Audio - The New Standard for Universal Spatial 3D Audio Coding. *Journal of the Audio Engineering Society* 62, 12 (2014), 821–830.

[27] Michael P. Hollier, Andrew N. Rimell, David S. Hands, and Robin. M. Voelcker. 1999. Multi-Modal Perception. *BT Technology Journal* 17, 1 (1999), 35–46. https://doi.org/10.1023/a:1009666623193

[28] Thirsa Huisman, Axel Ahrens, and Ewen MacDonald. 2021. Ambisonics Sound Source Localization with Varying Amount of Visual Information in Virtual Reality. *Frontiers in Virtual Reality* 2, 722321 (2021), 1–11. https://doi.org/10.3389/frvir.2021.722321

[29] ISO/IEC. 2021. Information Technology - High Efficiency Coding and Media Delivery in Hetrogeneous Environments - Part 3: 3D Audio. *International Standards Organization / Internation Electrotechnical Commision* (2021), 1–822.

[30] ITU-R International Telecommunications Union - Radiocommunications Sector. 2015. *BS.1534-3: Method for the Subjective Assessment of Intermediate Quality Level of Audio Systems*. Geneva, Switzerland.

[31] ITU-R International Telecommunications Union - Radiocommunications Sector. 2019. *BS.2076-2: Audio Definition Model*. Geneva, Switzerland.

[32] ITU-T International Telecommunications Union - Telecommunications Sector. 1998. *P.911: Subjective Audiovisual Quality Assessment Methods for Multimedia Applications*. Geneva, Switzerland.

[33] Angelika C. Kern and Wolfgang Ellermeier. 2020. Audio in VR: Effects of a Soundscape and Movement-Triggered Step Sounds on Presence. *Frontiers in Robotics and AI* 7, 20 (2020), 1–13. https://doi.org/10.3389/frobt.2020.00020

[34] Peter Kok, Gijs J. Brouwer, Marcel A. .J. van Gerven, and Floris P. de Lange. 2013. Prior Expectations Bias Sensory Representations in Visual Cortex. *Journal of Neuroscience* 33, 41 (2013), 16275–16284.

[35] Floris P. de Lange, Micha Heilbron, and Peter Kok. 2018. How Do Expectations Shape Perception? *Trends on Cognitive Science* 22, 9 (2018), 1–34. https://doi.org/10.1016/j.tics.2018.06.002

[36] Pontus Larsson, Daniel Västfjäll, and Mendel Kleiner. 2002. Better Presence and Performance in Virtual Environments By Improved Binaural Sound Rendering. In *AES 22nd International Conference on Virtual, Synthetic and Entertainment Audio*. Espoo, Finland, 1–8.

[37] Julie Lassalle, Laetitia Gros, Thierry Morineau, and Gilles Coppin. 2012. Impact of the Content on Subjective Evaluation of Audiovisual Quality: What Dimensions Influence Our Perception?. In *IEEE International Symposium on Broadband Multimedia Systems and Broadcasting*. Seoul, South Korea, 1–6. https://doi.org/10.1109/bmsb.2012.6264250

[38] Jong-Seok Lee, Francesca De Simone, and Touradj Ebrahimi. 2009. Influence of audio-visual attention on Perceived Quality of Standard definition multimedia content. In *2009 International Workshop on Quality of Multimedia Experience*. San Diego, CA, USA, 1–6.

[39] Alexander Lindau, Vera Erbes, Steffen Lepa, Hans-Joachim Maempel, Fabian Brinkman, and Stefan Weinzierl. 2014. A Spatial Audio Quality Inventory (SAQI). *Acta Acustica united with Acustica* 100, 5 (2014), 984–994. https://doi.org/10.3813/aaa.918778

[40] Paulo Marins. 2011. Characterizing the Perceptual Effects Introduced by Low Bit Rate Spatial Audio Codecs. In *AES 131st Convention*. New York, NY, USA, 1–8.

[41] Paulo Marins, Francis Rumsey, and Slawomir Zielinski. 2007. The Relationship Between Basic Audio Quality and Selected Artefacts in Perceptual Audio Codecs – Part II – Validation Experiment. In *AES 122nd Convention*. Vienna, Austria, 1–11.

[42] Helard B. Martinez and Mylène C. Q. Farias. 2014. Full-Reference Audio-Visual Video Quality Metric. *Journal of Electronic Imaging* 23, 6 (2014), 1–12. https://doi.org/10.1117/1.jei.23.6.061108

[43] Helard B. Martinez and Mylène C. Q. Farias. 2018. Combining Audio and Video Metrics to Assess Audio-Visual Quality. *Multimedia Tools and Applications* 77, 18 (2018), 23993–24012. https://doi.org/10.1007/s11042-018-5656-7

[44] Helard B. Martinez and Mylène C. Q. Farias. 2019. Analyzing the Influence of Cross-Modal IP-Masked Degradations on the Perceived Audio-Visual Quality. *Electronic Imaging* 2019, 10 (2019), 324–1–324–7. https://doi.org/10.2352/issn.2470-1173.2019.10.iqsp-324

[45] Helard B. Martinez, Mylène C. Q. Farias, and Andrew Hines. 2018. Perceived Quality of Audio-Visual Stimuli Containing Streaming Audio Degradations. In *26th European Signal Processing Conference*. Rome, Italy, 2529–2533. https://doi.org/10.23919/eusipco.2018.8553541

[46] Helard B. Martinez, Andrew Hines, and Mylène C. Q. Farias. 2021. Perceptual Quality of Audio-Visual Content with Common Video and Audio Degradations. *Applied Sciences* 11, 13 (2021), 5813. https://doi.org/10.3390/app11135813

[47] Leo McCormack and Archontis Politis. 2019. SPARTA & COMPASS: Real-Time Implementations of Linear and Parametric Spatial Audio Reproduction and Processing Methods. In *AES Conference on Immersive and Interactive Audio*. York, UK, 1–12.

[48] Radha N. Meghanathan, Patrick Ruediger-Flore, Felix Hekele, Jan Spilski, Achim Ebert, and Thomas Lachmann. December 2021. Spatial Sound in a 3D Virtual Environment: All Bark and No Bite? *Big Data and Computive Computing* 5, 79 (December 2021), 1–16. https://doi.org/10.3390/bdcc5040079

[49] Xiongkuo Min, Guangtao Zhai, Jiantao Zhou, Mylène C. Q. Farias, and Alan C. Bovik. 2020. Study of Subjective and Objective Quality Assessment of Audio-Visual Signals. *IEEE Transactions on Image Processing* 29 (2020), 6054–6068. https://doi.org/10.1109/tip.2020.2988148

[50] Pia N. P. Moreta, Søren Bech, Jon Francombe, Jan Østergaard, Steven van de Par, and Neofytos Kaplanis. 2022. Sensory Evaluation of Spatially Dynamic Audiovisual Sound Scenes: A Review. In *AES 152nd Convention*. Online, 1–10.

[51] Christian Nachbar, Franz Zotter, Etienne Deleflie, and Alois Sontacchi. 2011. Ambix - A Suggested Ambisonics format. In *Ambisonics Symposium (Springer Topics in Signal Processing)*. Lexington, KY, USA, 1–11.

[52] Miroslaw Narbutt, Seán O'Leary, Andrew Allen, Jan Skoglund, and Andrew Hines. 2017. Streaming VR for Immersion: Quality Aspects of Compressed Spatial Audio. In *23rd International Conference on Virtual System & Multimedia (VSMM)*. Dublin, Ireland, 1–6. https://doi.org/10.1109/vsmm.2017.8346301

[53] Annika Neidhardt and Anna M. Zerlik. 2021. The Availability of a Hidden Real Reference Affects the Plausibility of Position-Dynamic Auditory AR. *Frontiers in Virtual Reality* 2, 678875 (2021), 1–17. https://doi.org/10.3389/frvir.2021.678875

[54] Brain Odegaard, David R. Wozny, and Ladan Shams. 2015. Biases in Visual, Auditory, and Audiovisual Perception of Space. *PLoS Computational Biology* 11, 12 (2015), 1–23. https://doi.org/10.1371/journal.pcbi.1004649

[55] Marta Olko, Dennis Dembeck, Yun-Han Wu, Andrea F Genovese, and Agnieszka Roginska. 2017. Identification of Perceived Sound Quality Attributes of 360° Audiovisual Recordings in VR Using a Free Verbalization Method. In *AES 143rd Convention*. New York, NY, USA, 1–10.

[56] Torben H. Pederson and Nick Zacharov. 2015. The Development of a Sound Wheel for Reproduced Sound. In *AES 138th Convention*. Warsaw, Poland, 1–13.

[57] Nils Peters, Deep Sen, Moo-Young Kim, Oliver Wuebbolt, and S. Merrill Weiss. 2016. Scene-Based Audio Implemented with Higher Order Ambisonics. *SMPTE Motion Imaging Journal* 125, 9 (2016), 16–24. https://doi.org/10.5594/jmi.2016.2623398

[58] Chris Pike, Richard Taylor, Tom Parnell, and Frank Melchior. 2016. Object-Based Spatial Audio Production for Virtual Reality Using the Audio Definition Model. In *AES Conference on Audio for Virtual and Augmented Reality*. Los Angeles, CA, USA, 1–7.

[59] Margaret H. Pinson, Lucjan Janowski, Romuald Pépion, Quan Huynh-Thu, Christian Schmidmer, Phillip Corriveau, Audrey Younkin, Patrick Le Callet, Marcus Barkowsky, and William Ingram. 2012. The Influence of Subjects and Environment on Audiovisual Subjective Tests: An International Study. *IEEE Journal of Selected Topics in Signal Processing* 6, 6 (2012), 640–651. https://doi.org/10.1109/jstsp.2012.2215306

[60] Sandra Poeschl, Konstantin Wall, and Nocola Doering. 2013. Integration of Spatial Sound in Immersive Environments: An Experimental Study on Effect of Spatial Sound on Presence. In *IEEE Virtual Reality*. Orlando, Florida, 129–130.

[61] Archontis Politis and David Poirier-Quinot. 2016. JSAmbisonics: A Web Audio Library for Interactive Spatial Sound Processing on the Web. In *Interactive Audio Systems Symposium*. York, UK, 1–8.

[62] Ville Pulkki and Toni Hirvonen. 2005. Localization of Virtual Sources in Multichannel Audio Reproduction. *IEEE Transactions on Speech and Audio Processing* 13, 1 (2005), 105–119. https://doi.org/10.1109/tsa.2004.838533

[63] Schuyler R. Quackenbush and Jurgen Herre. May, 2021. MPEG Standards for Compressed Representation of Immersive Audio. *Proc. IEEE* 109, 99 (May, 2021), 1578–1589. https://doi.org/10.1109/jproc.2021.3075390

[64] Andrew Rimell and Michael Hollier. 1999. The Significance of Cross-Modal Interaction in Audio-Visual Quality Perception. In *IEEE Third Workshop on Multimedia Signal Processing*. Copenhagen, Denmark, 509–514. https://doi.org/10.1109/mmsp.1999.794134

[65] Thomas Robotham, Olli S. Rummukainen, Miriam Kurz, Marie Eckert, and Emanuël A. P. Habets. 2022. Comparing Direct and Indirect Methods of Audio Quality Evaluation in Virtual Reality Scenes of Varying Complexity. *IEEE Transactions on Visualization and Computer Graphics* 28, 5 (2022), 2091–2101. https://doi.org/10.1109/tvcg.2022.3150491

[66] Thomas Robotham, Ashutosh Singla, Olli S. Rummukainen, Alexander Raake, and Emanuël A. P. Habets. 2022. Audiovisual Database with 360 Video and Higher-Order Ambisonics Audio for Perception, Cognition, Behavior, and QoE Evaluation Research. In *14th International Conference on Quality of Multimedia Experience*. Lippstadt, Germany, 1–4. https://doi.org/10.1109/qomex55416.2022.9900893

[67] Tomasz Rudzki, Pierce Hening, Ignacio Gomez-Lanzaco, Jessica Stubbs, Thomas McKenzie, Jan Skoglund, Damian Murphy, and Gavin Kearney. 2019. Perceptual Evaluation of Bitrate Compressed Ambisonic Scenes in Loudspeaker Based Reproduction. In *AES Conference on Immersive and Interactive Audio*. York, UK, 1–10.

[68] Olli Rummukainen, Jenni Radun, Toni Virtanen, and Ville Pulkki. 2014. Categorization of Natural Dynamic Audiovisual Scenes. *PLoS ONE* 9, 5 (2014), e95848. https://doi.org/10.1371/journal.pone.0095848

[69] Francis Rumsey and Tim McCormick. 2021. *Sound and Recording Applications and Theory*. Taylor & Francis Group, New York, USA.

[70] Michael Schoeffler, Andreas Silzle, and Jürgen Herre. 2017. Evaluation of Spatial/3D Audio: Basic Audio Quality Versus Quality of Experience. *IEEE Journal of Selected Topics in Signal Processing* 11, 1 (2017), 1–14.

[71] Deep Sen, Nils Peters, Moo Young Kim, and Martin Morrel. 2016. Efficient Compression and Transportation of Scene Based Audio for Television Broadcast. In *Audio Engineering Society Conference on Field Sound Control*. Guildford, UK, 1 – 8.

[72] Shankar Shivappa, Martin Morrell, Deep Sen, Nils Peters, and S. M. Akramus Salehin. 2016. Efficient, Compelling and Immersive VR Audio Experience Using Scene Based Audio / Higher Order Ambisonics. In *AES Audio for Virtual and Augmented Reality*. Los Angeles. CA, USA, 1–10.

[73] Abubakr Siddig, Alessandro Ragano, Hamed Z. Jahromi, and Andrew Hines. 2019. Fusion Confusion: Exploring Ambisonic Spatial Localisation for Audio-Visual Immersion Using the McGurk Effect *(MMVE '19)*. Association for Computing Machinery, New York, NY, USA, 28–33. https://doi.org/10.1145/3304113.3326112

[74] Barry E. Stein, William S. Huneycutt, and M. Alex Meredith. 1988. Neurons and Behavior: The Same Rules of Multisensory Integration Apply. *Brain Research* 448, 2 (1988), 355–358. https://doi.org/10.1016/0006-8993(88)91276-0

[75] Barry. E. Stein, Benjamin A. Rowland, Paul J. Laurienti, and Terrance. R. Stanford. 2009. Multisensory Covergence and Integration. In *Encyclopedia of Neuroscience*. Academic Press, Oxford, UK, 1119–1124. https://doi.org/10.1016/b978-008045046-9.01112-8

[76] Barry E. Stein, Terrence R. Stanford, and Benjamin A. Rowland. Aug, 2014. Development of Multisensory Integration from the Perspective of the Individual Neuron. *Nature Reviews Neuroscience* 15, 8 (Aug, 2014), 520–535. https://doi.org/10.1038/nrn3742

[77] Russel L. Storms and Zyda J. Michael. 2000. Interactions in Perceived Quality of Auditory-Visual Displays. *Presence: Teleoperators & Virtual Environments* 9, 6 (2000), 557–580. https://doi.org/10.1162/105474600300040385

[78] Joel Susal, Kurt Krauss, Nicolas Tsingos, and Marcus Altman. 2016. Immersive Audio for VR. In *2016 AES Convention on Audio for Virtual and Augmented Reality*. Los Angeles, CA, USA, 1–8.

[79] Peter U Svensson. 2002. Modelling Acoustic Spaces for Audio Virtual Reality. In *1 st IEEE Benelux Workshop on Model based Processing and Coding of Audio*. Leuven, Belgium, 109–116.

[80] Lewis Thresh, Calum Armstrong, and Gavin Kearney. 2017. A Direct Comparison of Localisation Performance When Using First, Third and Fifth Order Ambisonics For Real Loudspeaker And Virtual Loudspeaker Rendering. In *AES 143rd Convention*. New York, NY, USA, 1–9.

[81] Matteo Torcoli, Thorsten Kastner, and Jürgen Herre. 2020. Objective Measures of Perceptual Audio Quality Reviewed: An Evaluation of Their Application Domain Dependence. *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 29 (2020), 1530–1541. https://doi.org/10.1109/taslp.2021.3069302 arXiv:2110.11438

[82] Nicolas Tsingos, Emmanuel Gallo, and George Drettakis. 2004. Perceptual Audio Rendering of Complex Virtual Environments. *ACM Transactions on Graphics* 23, 3 (2004), 249–258. https://doi.org/10.1145/1015706.1015710

[83] Daniel Västfjäll. 2004. Contextual Influences on Sound Quality Evaluation. *Acta Acustica united with Acustica* 90, 1 (2004), 1029–1036.

[84] Jean-Marc Villan, Koen Vos, and Timothy B. Terriberry. 2012. RFC 6716: Definition of the OPUS Codec. *Internet Engineering Task Force (IETF)* (2012).

[85] Hans Wallach. 1940. The Role of Head Movements and Vestibular Cues in Sound Localization. *Journal of Experimental Psychology* 27, 4 (1940), 339–368.

[86] Adam Weisser, Jörg M. Buchholz, Chris Oreinos, Javier Badajoz-Davila, James Galloway, Timothy Beechey, and Gitte Keidser. May, 2019. The Ambisonic Recordings of Typical Environments (ARTE) Database. *Acta Acustica united with Acustica* 105, 4 (May, 2019), 695–713. https://doi.org/10.3813/aaa.919349

[87] Stefan Winkler and Christof Faller. 2006. Perceived Audiovisual Quality of Low-Bitrate Multimedia Content. *IEEE Transactions on Multimedia* 8, 5 (2006), 973–980. https://doi.org/10.1109/tmm.2006.879871

[88] Yulia Yagunova, Mark A. Poletti, and Paul D. Teal. 2021. Ambisonics and Sonic Simulation in Virtual Reality. In *IEEE Region 10 Conference (TENCON)*, Vol. 00. Auckland, New Zealand, 369–374. https://doi.org/10.1109/tencon54134.2021.9707299

[89] Slawomir K Zielinski and Francis Rumsey. 2008. On Some Biases Encountered in Modern Audio Quality Listening Tests: A Review. *Journal of the Audio Engineering Society* 56, 6 (2008), 427–451.

[90] Franz Zotter and Matthias Frank. 2019. *Ambisonics: A Practical 3D Audio Theory for Recording, Studio Production, Sound Reinforcement, and Virtual Reality*. Springer Nature, Cham, Switzerland.