Comparing Direct and Indirect Methods of Audio Quality Evaluation in Virtual Reality Scenes of Varying Complexity

Thomas Robotham, Olli S. Rummukainen, Miriam Kurz, Marie Eckert, and Emanuël A. P. Habets

Fig. 1. Audio-visual objects used within VR scenes to demand various amounts of interactivity from the user. From left to right; static loudspeaker, animated train-set, interactive radio, and interactive remote control drone.

Abstract—Many quality evaluation methods are used to assess uni-modal audio or video content without considering perceptual, cognitive, and interactive aspects present in virtual reality (VR) settings. Consequently, little is known regarding the repercussions of the employed evaluation method, content, and subject behavior on the quality ratings in VR. This mixed between- and within-subjects study uses four subjective audio quality evaluation methods (viz. multiple-stimulus with and without reference for direct scaling, and rank-order elimination and pairwise comparison for indirect scaling) to investigate the contributing factors present in multi-modal 6-DoF VR on quality ratings of real-time audio rendering. For each between-subjects employed method, two sets of conditions in five VR scenes were evaluated within-subjects. The conditions targeted relevant attributes for binaural audio reproduction using scenes with various amounts of user interactivity. Our results show all referenceless methods produce similar results using both condition sets. However, rank-order elimination proved to be the fastest method, required the least amount of repetitive motion, and yielded the highest discrimination between spatial conditions. Scene complexity was found to be a main effect within results, with behavioral and task load index results implying more complex scenes and interactive aspects of 6-DoF VR can impede quality judgments.

Index Terms—Multi-modal, virtual reality, 6-Degrees-of-freedom, audio quality, direct scaling, indirect scaling, evaluation methods

1 Introduction

Subjective quality evaluation methods for media technology are highly dependent on the context-of-use and user demographic. Consequently, many standards now exist to address these different scenarios [68]. For traditional audio quality, most recommendations consider contexts where audio is judged as an independent sensory input (uni-modal), as opposed to in conjunction with multiple sensory inputs (multi-modal). With the rising popularity of immersive and interactive systems such as virtual reality (VR), investigations within complex multi-modal environments are becoming more prevalent in the research community to understand higher-order quality covariates such as cyber-sickness and plausibility [7, 63, 75]. Nevertheless, the ability to conduct quality evaluations of audio rendering within a multi-modal setting is still advantageous for algorithmic optimization. The choice of method in psychoacoustic evaluations can directly influence the ability to identify statistically significant perceptual differences which drive many decisions in the domain of sensory sciences. To the authors' knowledge, no standard currently exists for VR audio evaluation, and little research has comparatively studied the use of standardized quality evaluation methods in multi-modal VR environments.

- *Thomas Robotham, and Emanuël A. P. Habets are with International Audio Laboratories Erlangen. E-mail: {thomas.robotham | emanuel.habets}@audiolabs-erlangen.de*
- *Olli S. Rummukainen, Miriam Kurz, and Marie Eckert are with Fraunhofer-Institut für Integrierte Schaltungen IIS. E-mail: {olli.rummukainen | kurzmm | eckertme}@iis.fraunhofer.de*

Manuscript received 6 Sept. 2021; revised 3 Dec. 2021; accepted 7 Jan. 2022. Date of publication 15 Feb. 2022; date of current version 29 Mar. 2022. Digital Object Identifier no. 10.1109/TVCG.2022.3150491

This study provides a comparison of four evaluation methods (viz. multi-stimulus with and without reference for direct scaling, and rank-order elimination and pairwise comparison for indirect scaling) to identify subjective audio quality within a multi-modal VR context. Five simple scenes are employed to target differing degrees of interaction available to users representative of typical VR content: static, animated, and interactive objects (shown in Fig. 1). In traditional signal-focused quality evaluation, the choice of the evaluation method is also informed by the estimated magnitude of quality differences of the presented conditions. However, as all audio conditions in VR are presented in combination with visual, proprioception, and vestibular cues, these differences can no longer be assumed to have the same perceived quality levels and cognitive impact in comparison to a uni-modal presentation. Consequently, by altering the available task complexity within controlled VR scenes, we also observe the effect multi-modal input imposes on condition discernability. Given a comparison of subjective results of using direct and indirect scaling evaluation methods, scene interactivity, user behavior, and factors related to subjects' cognitive load, the results provide developers and quality evaluation experts a novel contribution on how to evaluate and analyze audio quality in multi-modal VR, as well as highlighting the implications of content complexity in the absence of any explicit auditory reference.

2 Background

2.1 Audio Quality Evaluation in the Context of 6-DoF VR

In six-degrees-of-freedom (6-DoF) VR, users may orientate their field of view (FoV) and move position in all directions. Tracking data is then used to drive rendering systems (*e.g.*, audio, video) in real-time to deliver a coherent egocentric representation of an interactive virtual environment (IVE). Multi-sensory integration (MSI) is the multifaceted neurocognitive process that combines our perception of these renderings, allowing us to orientate spatially and successfully navigate a

space [23, 66]. In this study, this comprises of our visual, auditory, vestibular, and proprioceptive responses. For audio, systems that provide us with high-fidelity sensory information can contribute to a greater sense of immersion and presence within IVEs [37, 42]. Therefore, subjectively evaluating these systems is a key step in maximizing the use of limited computational budgets or bandwidth, for example. However, many test considerations that apply to traditional audio quality evaluation do not strictly fit the context of 6-DoF IVEs.

Traditional audio quality evaluation procedures often aim to minimize the effect of influencing visual factors on quality judgments. Research into auditory localization demonstrates how our auditory perception is changed given various amounts of spatio-temporally aligned visual cues [9, 24]. However, visual information is a necessity in nearly all VR contexts. Now, quality evaluation within multi-modal IVEs moves us towards MSI conflict situations, whereby two modalities may receive incoherent information regarding a particular stimulus [22]. On the one hand, MSI conflicts may lead to multiple sensory streams being combined into a multi-modal percept in a near-optimal manner [1, 8, 14]. On the other hand, processing information with two different sensory systems, which are activated synchronously by the same multi-modal stimulus, increases the likelihood that we can correctly identify diverging sensory signals [65, 74]. The latter demonstrates how quality judgments may be aided in multi-modal settings, where the expected behavior of the modality under test does not spatially, temporally, or otherwise correlate with another [34, 35].

For interacting with audio content in traditional evaluation settings, the only dimension of variability is via a loop control allowing subjects to set start- and end-points for loop segments along the presented waveform. In 6-DoF VR, the amount of interaction offered within a VR scene leads to completely non-linear exploration ranging from purely 6-DoF movement to manipulating and interacting with individual audio-visual objects. This non-linearity means any subject may experience the content completely different from all others [69]. For object interaction, allowing users to actively 'grab' an audio source with a controller means the auditory position of the sound source then undergoes MSI with both visual and proprioception cues. Auditory localization studies indicate proprioceptive feedback can improve localization accuracy [27], and even aid our auditory system to procedurally calibrate and adapt over a given exposure period to maintain localization performance [49]. How this is achieved at a neurocognitive level is not yet fully understood [44, 62]. However, increasing levels of interactivity may place a higher cognitive and physical demand on subjects leading to several undesirable effects such as increased test time, mental fatigue, loss of motivation, and frustration. Studies involving dual-task evaluations [41] in MSI research demonstrate that an increase in cognitive processes and manual tasks, which occupy our working memory, leave us susceptible to false perceptions of our other senses, thereby influencing task performance [26, 59, 60]. Considering all the available factors contributing to mental workload, such as: VR controls, menu interaction, interactive elements within the scene, the quality judgment task, and processing of sensory information, understanding any challenges that impede (or mask) the quality judgment process is of particular use for future research.

Lastly, depending on the evaluation conditions and quality criteria, naïve or expert listeners may be used [61]. However, an additional component within IVEs is the subject's experience with VR systems. Works regarding intuitive interaction design demonstrate that intuition is not the innate simplicity of a system, but rather a cognitive process transforming knowledge or prior experience to a current application [11, 43]. Following this principle, participants who are experienced in similar systems as VR have already developed the cognitive schema associated with particular human-computer interfaces and will consequently find certain aspects less challenging, allowing them to focus more on the quality judgment task.

2.2 Audio Evaluation Methods

2.2.1 Direct Scaling

In sensory evaluation, direct scaling allows subjects to directly assign a value that represents their estimated magnitude of a perceived sensation or attribute to a stimulus [45]. As this value is prescribed directly from the subject, it is seemingly unambiguous and may often be easily compared across conditions in terms of mean opinion scores (MOS) and confidence intervals (CIs). For audio evaluation, two of the most commonly used direct scaling methods are BS.1116 [28], targeting small quality differences and BS.1534-3 [30] **Mu**ltiple **S**timulus with **H**idden **R**eference and **A**nchor (MUSHRA) tests, for larger deviations compared to an explicit reference [4]. In MUSHRA tests, multiple conditions (typically ≥ 5), including a hidden reference and a low-quality anchor, are presented in parallel to the subject. A rating may then be attributed to a condition via a slider representing a continuous quality scale ranging from 0 - 100, annotated with five descriptive labels: *Bad*, *Poor*, *Fair*, *Good*, and *Excellent*. One of the main advantages of MUSHRA testing is the ability to compare multiple conditions during one test scene, making the test efficient whilst yielding high-quality data [80]. Whilst not mandatory, subjects are also encouraged to switch between conditions and reference as often as possible [47]. Subjects also inherently provide a form of relative ranking through paired comparison between the presented conditions [64], meaning the quality of presented conditions also affects the interpretation of the scale. To control the scale usage, a low-quality anchor is used along with the explicit open reference. How subjects are instructed to interpret and rate anchor signals is likely to change the observed statistical differences [61]. Consequently, great care and consideration should be taken in the design and guidance of any anchor conditions. No standard currently exists detailing what constitutes an appropriate anchor for real-time immersive audio evaluations.

Both BS.1116 and BS.1534-3 direct scaling methods are common within signal-related audio quality evaluation largely due to the use of a reference condition. However, IVEs are often computer-simulated, and any real-time audio rendering pipeline for 6-DoF VR will require signal processing to reflect user movements and acoustic auralization [46]. Therefore, a 'ground-truth' reference model for audio quality is seldom available. Instead, direct scaling methods that do not utilize an open reference condition could be used in VR. Tests such as Absolute Category Ratings (ACR) [32] present only a single condition to subjects which, in comparison to MUSHRA testing of audio codec qualities, has proven to be efficient and repeatable for low and intermediate degradations, and less stable for high-quality conditions [21]. Recent experiments and standardization activities also employ multi-stimulus experiments, whereby direct scaling is still employed but in the absence of a reference and hidden reference, making a subject's quality judgments relative to one another and scale labels [31, 56, 80]. However, issues related to response mapping of the scale and label interpretations [83] may be more prevalent in the absence of a reference condition and thus further susceptible to inter-subject differences.

2.2.2 Indirect Scaling

Contrary to direct scaling, indirect scaling methods use data given by subjects to derive the magnitude of a stimulus' quality. Data is collected through a comparative process where conditions are prescribed a status in relation to others via a choice-based paradigm. A scale can then be constructed to show the probability that any particular condition is favored over others with relative distances. Pairwise comparison methods are some of the most powerful indirect scaling sensory tests. This may be given as a binary forced-choice paradigm [76, 78], or a bi-polar labelled scale [29]. The discriminatory power is due to subjects being presented only two stimuli, reducing the subject's task to simple pairwise judgments [77]. If no scale is employed, the possibility of subjective interpretation of the scale or labels is eliminated [16] making the task even simpler. Bradley-Terry model [13] or Thurstone's model [71] are often used for the statistical analysis to determine a quality (or preference) scale based on logistic or Gaussian distributions [81]. Although commonly described as accurate, the main disadvantage of this method is the time taken to complete all comparisons. Consequently, incomplete block designs are used to decrease the required number of paired comparisons. However, this design choice drastically increases the number of subjects required to participate in testing.

To reduce task time, rank-order methods have been used through-

out sensory evaluation to rapidly compare multiple conditions [67]. These methods require that subjects rank the presented stimuli in a specified direction. This subjective ranking can then be used to develop a relationship between all conditions within each test scene. As with pairwise comparisons, the absence of any scale removes any potential bias due to the interpretation of labels and also requires very little training for the subjects. The drawback of this method is that the rating cannot be described in terms of absolute quality [50, 79]. An adaptation of a rank-order procedure, proposed by Wickelmaier *et al.* [77], takes the method a step further. The goal is to sequentially eliminate all conditions on the premise that preceding conditions all possessed aspects that the current conditions do not [73]. By continuously diminishing the number of comparisons available within a test scene, test time can be reduced whilst still providing results comparable to other indirect scaling methods. Statistical inference can be conducted via Plackett-Luce model [51], which is comparatively similar to the Thurstone model [39].

For 6-DoF VR, ranking by elimination has proven to be an effective method for determining relative levels of audio quality [57]. While overcoming many hurdles presented by direct scaling, indirect scaling still possesses overt limitations. Evaluating a collection of '*poor*' quality audio renderers may yield a ranking, but the '*best*' condition may still be substandard to our expectations. However, in the absence of a reference, direct scaling multi-stimulus experiments may also only provide results in terms of relative quality. Another potential issue becomes apparent if all subjects make consistent rankings thus resulting in zero inter-individual error [45]. Relative distances then derived from choice probability will result in conditions that are infinitely separated along any scale. Therefore, extracting a finite distance between conditions requires more detailed statistical models than those used for direct scaling analysis.

3 Study Paradigm

This study employs a mixed between- and within-subjects test design with a sample size $N = 68$. The between element of the design can be broken down into four experiments that comprise of two separate test sessions. These four experiments refer to the four different methods employed for quality evaluation (see Sect. 3.1) shown in Fig. 2. Consequently, $\frac{N}{4} = 17$ subjects are randomly assigned to each experiment with two sessions ending in a total of 136 unique results across experiments and sessions. Within each experiment, the two test sessions employ two different sets of five conditions varying along two dimensions of audio quality degradation (see Sect. 3.4).

3.1 Evaluation Methods and Environment

Four subjective evaluation methods are compared within this study: two direct scaling methods and two indirect scaling methods. For direct scaling methods, an adaptation of BS.1534-3 [30] and a multi-stimulus comparison rating are used. Both methods are hereafter referred to as multi-stimulus with hidden reference (**MSHR**) and multi-stimulus (**MS**), respectively. As discussed in Sect. 2.2.1, the reference condition is seldom available for 6-DoF VR, particularly in scenarios evaluating different rendering algorithms. However, the MSHR test is included here to serve as a comparison against other methods. The only departure from the recommendation [30] is the anchor, as no 3.5 kHz low-pass filtered version is included. This was withheld due to the uncertain impact such a heavily degraded condition would have on scaling when used in other methods (Sect. 2.2.1). The MS method includes no open reference making all judgments relative to one another [31, 56].

For indirect scaling methods, rank-order elimination-by-aspects [77] and pairwise comparisons are employed, hereafter referred to as **EBA** and **PC**, respectively. The PC method is presented as a continuous 120-point bi-polar scale allowing subjects to represent their opinion where smaller or greater differences are perceived. In actuality, the scores given are converted to a binary format (described in Sect. 5.2). Both bi-polar scale and statistical analysis have been selected as a quality evaluation method in current standardization activities [54]. Consequently, we adopt the design choice here to be consistent with related activities. To keep the PC method at a feasible size, a complete

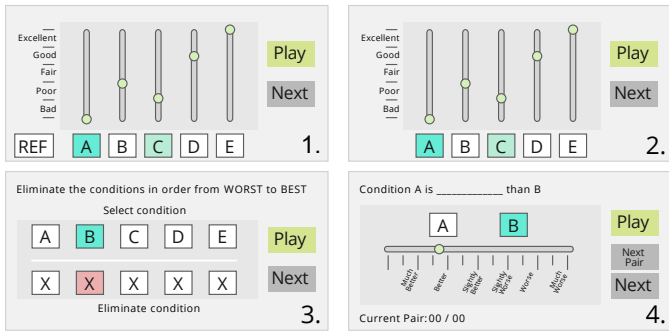

Fig. 2. VR test method interfaces. 1. Multiple-stimulus with hidden reference interface (MSHR) . 2. Multi-stimulus interface (MS). 3. Rank-order elimination interface (EBA). 4. Pairwise comparison interface (PC).

block design with no self-comparison or pairwise repetitions is used, resulting in 50 comparisons per subject. All comparison pairs are randomly presented per scene. Randomization of condition pairs across all scenes was not included to reduce waiting (loading) times when repeatedly switching scenes.

To conduct the experiments, software was built using the Unity game engine and Max/MSP inspired by Robotham *et al.* [55] and Dodds *et al.* [19]. The Unity game engine is programmed to host the VR devices, interactivity framework, graphical components, and the test method user interfaces. Max/MSP is used as a real-time audio engine to host all audio renderers, test paradigms, and results logging. Communication between the software is done via the open sound control protocol through a user datagram protocol (UDP) transport layer. For this study, the HTC Vive Pro is used as the VR head-mounted display in a SteamVR playable area measuring 2.4 m $\times$ 2 m. The headphones used are Beyerdynamic DT-990 Pro. Two controllers are used with the Vive Pro to interact with the VR environment and test interfaces. Maximum motion-to-sound latency was measured at approximately 12 ms.

For each method, a separate test panel is automatically imported into each scene. The test interfaces for all methods can be seen in Fig. 2. To show and hide the test panel, the subjects can touch-tap the north sector of the controller touchpad. The test panel will always appear in the subjects' FoV. When the test panel appears, a laser pointer is also activated to allow the users to change conditions and provide ratings using buttons and sliders respectively. Teleportation was included to allow exploration beyond the immediate playable area. Touching the south sector of the touchpad would show the target pointer and pressing the touchpad would activate. The mechanics of teleporting in VR raise certain audio rendering considerations such as Doppler effects or brief auditory discontinuities. How such mechanics are handled in complex auralizations may also impact the perceived quality of a continuous spatial soundscape. Grabbing interactive objects within the scene is done by moving the controller towards the object, followed by pressing and holding the grip button. If the interactive object has any functionality, it can be activated by pulling the trigger while gripping. To assist subjects in learning the controls, they first enter a VR configuration scene where instructions for all interaction mechanics are written on separate panels along with example objects to practice with. As the VR controls will contribute towards usability, it is important that all subjects (particularly those new to VR) feel comfortable with all mechanics to a point where the initial learning curve has been overcome, allowing them to better concentrate on the quality evaluation task.

3.2 Evaluation Content

For 6-DoF VR, varying levels of interaction are available to the user depending on the use case and context. For this study, five scenes have been designed which differ in interaction complexity and the behavior of audio-visual objects to observe any effects on the cognitive workload, and consequently, the quality judgment ratings. Table 1 provides a summary of the scenes in order of estimated complexity (based on practical VR experience) and Fig. 1 shows the main audio-visual objects included. Although treated as audio-visual, these scenes are authored to be minimalistic in terms of visual realism. The intention

is to provide a 6-DoF environment and relevant visual counterparts to support multi-sensory integration. All scenes are presented in a randomized order across all tests to minimize any learning effects.

The most simplistic scene is the ***Static*** scene. An audio-visual loudspeaker is positioned on a stand with the acoustic axis at roughly 1.7 m height, allowing subjects to explore the complete space around the audio-visual object. The loudspeaker plays a short music excerpt lasting ≈ 16 seconds which is continuously looped.

The ***Animated*** scene possesses a tabletop toy train-set with a small animated train that continuously runs around a circuit. The circuit varies height at three distinct positions, and the table is purposefully modeled to allow subjects to enter a small 2 meter wide cut-out, thus being enveloped by the animation. Subjects may also explore around all sides of the table to judge from all angles. While running, the train plays pulse-modulated Brownian noise imitating a train engine, and when passing the station plays a short whistle. The total duration of the train audio and animation is ≈ 20 seconds.

The ***Distraction*** scene augments the ***Animated*** scene using five more additional sound sources. Two audio-visual loudspeakers are positioned on stands on either side of the train model, which loop left and right channels of a ≈ 20 second stereo music excerpt. The sound of a pond is positioned at the water source indicated on the table. Finally, two more audio sources are included, which have no direct visual counterpart but have semantic meaning within the scene's context. Birds singing positioned at a cluster of trees and a platform bell positioned at the station. The platform bell is triggered to play only when the train arrives. The intention is to make it challenging for the subjects to judge the quality of only the toy train-set in the presence of competing audio. All audio sources are subject to HRTF degradations (see Sect. 3.4).

The ***Interaction*** scene introduces more complex interaction with an audio-visual object. An analog-style radio is situated on a table and the subjects are instructed to grab and move the radio around their person while evaluating the audio content. The radio loops a music excerpt of ≈ 34 seconds duration. By interacting with one controller and operating the evaluation interface with another controller, it is estimated that the cognitive load on subjects will be increased, making the evaluation more challenging than previous passive scenes.

Finally, the ***Task*** scene is intended to be the most cognitively demanding. Set in a forest environment, the task involves flying a drone which subjects must control using a virtual remote controller. Upon starting the scene, the drone starts at a particular height and then proceeds to descend slowly. Subjects can move the forwards/sideways position of the drone to any location indicated in space via a yellow sphere by pulling the controller trigger. To increase the height of the drone, the subjects must pull and hold the trigger (> 0.2 seconds). If subjects let the drone sink to the ground, all audio playback from the drone will stop, forcing subjects to continuously monitor the drone height and use the virtual controller if they are to continue the evaluation. Two audio files are attached to the drone. One plays a continuous fan-motor, the other plays a revved fan-motor. The revved fan-motor is only triggered when the user pulls the trigger and thus moves the drone position. Four additional audio sources are positioned in the scene: two ambient forest sounds, an owl source, and a woodpecker source. These have no direct visual counterpart but support the contextual setting. Overall, the mental workload of monitoring the drone position, controlling the drone with one hand, and the test panel with another hand is estimated to be the most demanding.

3.3 Questionnaire Scene

When evaluating methods, previous research [76, 77] has shown additional surrounding metrics such as task time, cognitive demand, and emotional responses (burden, stress, frustration [10]) to be valuable in understanding subjective responses that may contribute to the overall reliability and ease-of-use [33]. Consequently, we include an additional questionnaire scene presented after each VR scene to collect more information regarding the perceived difficulty of the evaluation content. This is programmed into the VR software eliminating the need for subjects to remove the VR equipment to provide manual pen and paper responses. The intention being that the experience of the previous VR scene will be more present in their memory. The NASA Task Load Index (TLX) questionnaire [25] is employed which asks for six responses along the following dimensions: *mental demand, physical demand, temporal demand, performance, effort,* and *frustration*. Time spent within the questionnaire scene is not analyzed; therefore, subjects can also use the scene to take a break if desired.

3.4 Condition Pool

To compare the evaluation methods, we devised two degradation categories that highlight important aspects of real-time binaural audio rendering using head-related transfer functions (HRTFs). One test session employs conditions that are spatially degraded and hereafter referred to as $\mathbf{HRTF}_{Spat}$. The other test session employs conditions that posses frequency differences, hereafter referred to as $\mathbf{HRTF}_{Freq}$. As a starting point and high-quality non-degraded open reference for the MSHR method, we used the generic HRTF set measured from the KEMAR head and torso simulator retrieved from the spatial audio for domestic interactive entertainment (SADIE) database [3]. The measurements comprise of 8802 source positions measured at 1.2 m distance with a 1° azimuth and 15° elevation resolution, resulting in 35737 KB of data. The final HRTF set is diffuse-field equalized and the filter length is 256 samples. Due to practical limitations of measuring 68 subjects' individualized HRTFs, the methodological focus of this study, and that we study here the perceptual impact of degradation's originating from a singular HRTF dataset, generic HRTFs are used. All audio was rendered using convolution of the nearest HRTF angular filter and direct sound with the addition of distance attenuation according to the inverse-square law. No interpolation was performed to focus the perceptual quality on the HRTF dataset, as opposed to further signal processing optimization. A gain limiter is active at 0.5 m from the center of the head to avoid amplitude clipping at close distances. No additional acoustic auralization was included to focus on the quality effects of HRTF fidelity degradations. Block size for real-time audio playback was 256 at a sampling rate of 44.1 kHz. All conditions within both sessions were fully randomized to minimize learning effects.

For the $\mathbf{HRTF}_{Spat}$ session, the spatial azimuth resolution of the HRTF grid was reduced to produce HRTFs with azimuth grid resolutions of 5°, 10°, 20°, and 30° (illustrated in Fig. 3 - Top). The lower resolution HRTF files were constructed by subsampling the original HRTF at all measurement points that equaled the desired azimuth angles. The resulting number of source positions for the respective degrees are 1586, 794, 398, and 226. While a high-resolution HRTF set may produce higher spatial fidelity, the motivation to include a lower number of source positions drastically reduces the amount of data (6400, 3214, 1620, and 1089 KB) and allows a faster capture time.

For the $\mathbf{HRTF}_{Freq}$ session, the frequency resolution of the HRTF set was altered by reconstructing the original HRTF using principal component analysis (PCA) [36] on the HRTF magnitiude response. The process entails identifying the reoccurring data patterns within a data set, which can then be used to reduce the data such that its defining components may still be expressed given a combination of basis vectors. The number of principal components (bases) used defines the order of dimensions to characterize our original data. In this study, the frequency resolution for each HRTF source position is re-synthesized using PCA with 2, 8, 16, 64, and 128 bases (Fig. 3 - Bottom), where 128 bases result in a reconstruction of the original HRTF (code available at [72]). By altering the frequency response of our original HRTF, we simulate possible degraded localization cues of height due to altering the finer detail of higher frequency content [12], and overall impression regarding coloration and timbre differences. The influence of timbre is known to play a large role in the perception of overall audio quality [58]. However, such evaluations are often in comparison to a high-quality open reference condition. In this study, the conditions in $\mathbf{HRTF}_{Freq}$ grant us more insight when comparing results between methods that do not employ an open reference.

4 Evaluation Procedure

At the start of each test, all subjects were asked to fill out consent and data protection forms and a short demographic form asking for

Table 1. Evaluation content employed. **Object** refers to the audio-visual object within the scene, where N/A means no visual or audio counterpart is present. For **Audio**, certain audio files are 'triggered', meaning the audio file only plays back at specific points during the scene.

Scene ID	Area ($x \times z$)	Source No#	Object	Audio	Starting Position	Interactivity
Static	12×12	1	Loudspeaker	Music excerpt (mono)	[1.9 1.7 0.0]	Static
Animated	24×24	1	Toy train	Pulsed Brownian noise & whistle	[-0.1 1.0 -5.6]	Dynamic Trajectory
Distraction	24×24	1	Toy train	Pulsed Brownian noise & whistle	[-0.1 1.0 -5.6]	Dynamic Trajectory
		2	Loudspeaker (left)	Music excerpt (left)	[7.0 1.7 -7.0]	Static
		3	Loudspeaker (right)	Music excerpt (right)	[-7.0 1.7 -7.0]	Static
		4	N/A	Birds chirping	[-2.4 1.15 -6.0]	Static
		5	Table pond	Slow water splashing / pond noise	[1.6 1.0 -6.2]	Static
		6	N/A	Platform bell (triggered)	[-0.8 1.2 -5.8]	Static
Interaction	12×12	1	Analogue radio	Music excerpt (mono)	[0.9 0.4 0.4]	Interactive
Task	20×17	1	Remote drone	Drone fans	[1.0 4.0 4.0]	Interactive
		2	Remote drone	Drone fans revving (triggered)	[1.0 4.0 4.0]	Interactive
		-	Drone controller	N/A	On person	Interactive
		3	N/A	Forest ambience / crickets	[-25.0 4.0 0.0]	Static
		4	N/A	Forest ambience / crickets	[25.0 4.0 0.0]	Static
		5	N/A	Tawney owl (sporadic)	[18.1 8.4 -20.5]	Static
		6	N/A	Woodpecker (sporadic)	[-9.3 11.3 12.0]	Static

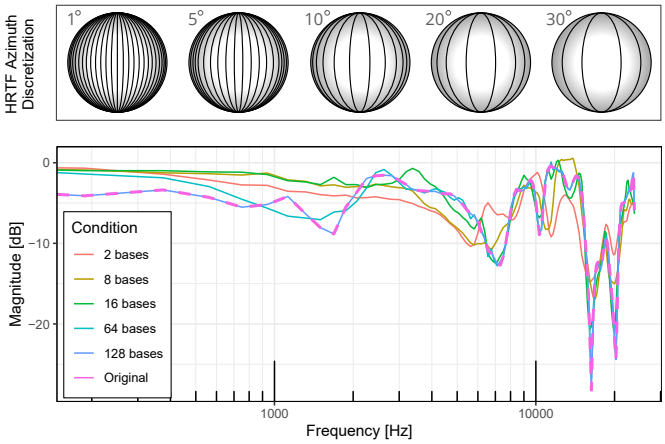

Fig. 3. Top: Illustration of azimuth discretization of the HRTF set for the HRTF$_{Spat}$ test session. Bottom: Magnitude response at position 0° azimuth and 0° elevation of the reconstructed HRTF using 2, 8, 16, 64, 128 basis functions, used as conditions for the HRTF$_{Freq}$ test session.

name, age, gender (*male, female,* and *non-binary*), VR experience, and listening experience. Subjects could freely choose where they saw themselves given written explanations of the various choices. VR experience was broken down into five categories; *naïve*, *occasional*, *semi-regular*, *frequent*, and *expert* user. Listening experience was broken into three categories: *naïve*, *experienced*, and *expert* listener. All subjects recruited for the test reported normal or corrected-to-normal vision and hearing and were all compensated via monetary payment. In total, 24 female and 44 male subjects participated with an average age of 31 (SD = 11.7), 23 subjects were *expert* listeners, and 22 subjects were *naïve* regarding VR experience. At least five expert listeners participated in each separate evaluation method.

A written description of the test was given to the subjects and verbally explained by the administrator detailing, the method they will be using, and how to provide their quality ratings. The criteria for how subjects base their ratings is method-dependent. As MSHR is the only method that provides a reference, the *overall audio quality* of the conditions should be rated against the open reference. For the other methods MS, EBA, and PC, the *overall audio quality* judgments should be based on comparisons against other sensory cues (i.e., visual and proprioceptive), and subjects' inner reference and expectation. Instructions were provided in each VR scene (Table 1) to remind subjects of the evaluation task. For the ***Distraction*** and ***Task*** scene, subjects were informed to evaluate the overall audio quality of the toy train and remote drone audio-visual objects, respectively. To aid the more naïve subjects, some major attribute categories of audio quality were provided and described: *localization quality, timbral quality, time behavior, and dynamics*. The chosen attributes were selected from established standards of traditional audio quality evaluation and research considering more dynamic aspects of IVEs [30, 38]. It was stressed to subjects that this list is not exhaustive, and any other differences they perceived may also be used to help them provide quality judgments.

Once the subjects confirmed they had understood the test, they were given the VR equipment and entered into a virtual familiarization scene. Here, subjects have the opportunity to practice all the controls and interactions they will encounter in the test including showing / hiding the test interface, teleporting, basic interaction with the radio, and advanced interaction with the drone. No time restriction was placed during the familiarization phase, and once subjects felt comfortable with all controls, they could begin the evaluation.

5 Results

5.1 Direct Scaling Results

For both MSHR and MS methods, a two-way repeated-measures analysis of variance (ANOVA) was conducted separately for both HRTF$_{Spat}$ and HRTF$_{Freq}$ test sessions for independent variables **Scene** and **Condition** on dependent variable subjective **Rating**. In all instances, significant main effects were obtained when present at $p < 0.05$. Mauchly's test of sphericity was performed for all data. In cases where sphericity was not met, significant main effects after Huynh-Feldt correction are reported. Cohen's operational definitions for effect sizes are reported for interpreting the magnitude of correlation coefficients in accordance with [18]. Normality of residuals for all ANOVAs was found satisfactory with a standard deviation of residuals ranging $16.74 \leq s \leq 22.14$, suitable for the available response scale between 0 - 100.

5.1.1 Multiple Stimulus with Hidden Reference

For the MSHR method, Fig. 4 shows the mean opinion scores and 95 % bootstrapped CIs. For the HRTF$_{Spat}$ session, a large main effect of condition ($F_{(4,64)} = 35.23$, $p < .001$, $\eta_G^2 = .33$) and a small effect of scene ($F_{(4,64)} = 3.42$, $p = .03$, $\eta_G^2 = .03$) were found. For the HRTF$_{Freq}$ session, a large effect of condition ($F_{(4,64)} = 38.5$, $p < .001$, $\eta_G^2 = .38$) was found.

Post-hoc Tukey pairwise comparison t-tests of conditions within each scene for the HRTF$_{Spat}$ session revealed the following main conclusions. No significant difference was found between conditions *1°* and *5°* across all scenes. Conditions *1°* and *5°* consistently yielded significant differences against conditions *20°* and *30°* for all scenes.

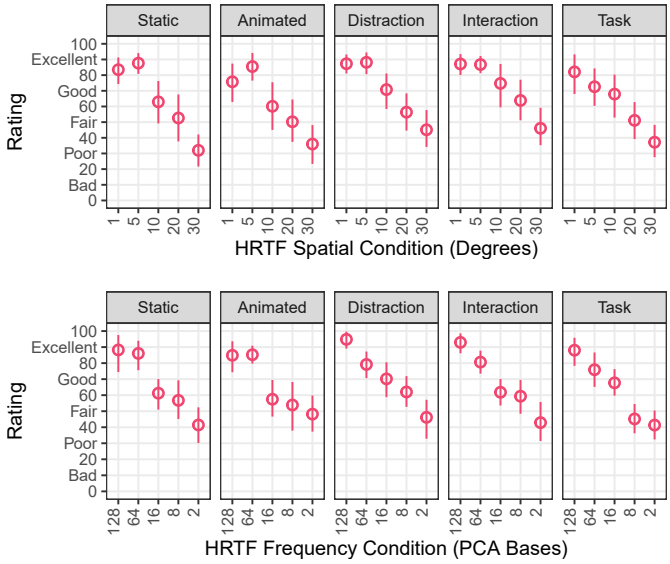

Fig. 4. MOS and bootstrapped 95 % CIs of HRTF$_{Spat}$ (top) and HRTF$_{Freq}$ (bottom) test sessions using **MSHR** method. $N_{Spat} = 17$ and $N_{Freq} = 17$.

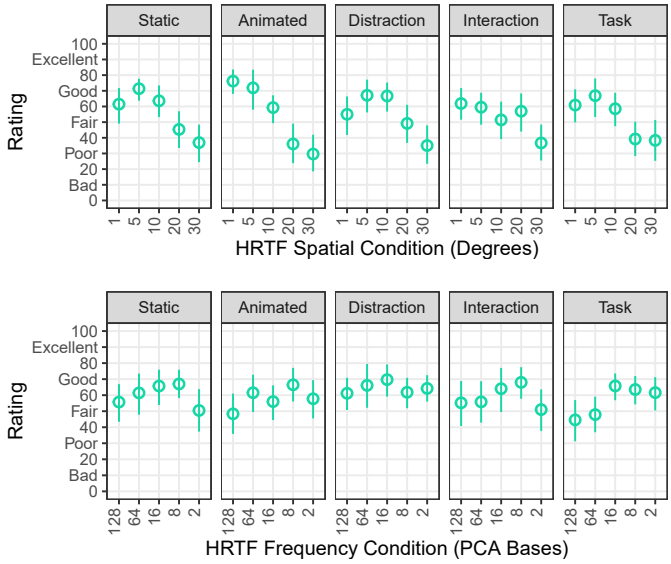

Fig. 5. MOS and bootstrapped 95 % CIs of HRTF$_{Spat}$ (top) and HRTF$_{Freq}$ (bottom) test sessions using **MS** method. $N_{Spat} = 17$ and $N_{Freq} = 17$.

Overall, the number of significant differences found between all conditions for each scene was Static (8) > Animated (6) > Distraction (5) = Interaction (5) > Task (4).

Post-hoc analysis for the HRTF$_{Freq}$ session revealed similar results to the spatial conditions. No significant difference was found between *128* and *64* bases in any scene, the closest to significance being the ***Distraction*** scene ($p = 0.059$). For all scenes, significant differences were found between high bases count conditions *128* and *64* against lower bases count conditions *8* and *2*. The number of significant differences found between all conditions for each scene was Static (7) = Interaction (7) = Task (7) > Animated (6) = Distraction (6).

5.1.2 Multi-Stimulus

For the MS method, Fig. 5 shows the mean opinion scores and 95 % bootstrapped CIs. Analysis of the HRTF$_{Spat}$ session revealed a large main effect of the condition ($F_{(4,64)} = 29.55$, $p < .001$, $\eta_G^2 = .2$) and small effect of scene × condition ($F_{(16,256)} = 1.82$, $p < .031$, $\eta_G^2 = .05$). The HRTF$_{Freq}$ session possessed a small effect from the condition ($F_{(4,64)} = 5.31$, $p = .001$, $\eta_G^2 = .04$).

Post-hoc Tukey pairwise comparison t-tests of conditions within each scene for the HRTF$_{Spat}$ session revealed the following observations. Across all scenes, no significant difference was found between any combination of *1°*, *5°*, *and*, *10°*. In scenes ***Static Distraction***, and ***Interaction***, condition *1°* had no significant difference vs. *20°* and in one scene, no difference vs. *30°*. The number of significant differences from pairwise t-tests observed across scenes ranged from two in the ***Interaction*** scene between *1°* vs. *30°* ($p = .01$), and *5°* vs. *30°* ($p = .03$), up to six in the ***Animated*** scene. Overall, number of significant differences per scene was Animated (6) > Static (5) > Task (4) > Distraction (2) > Interaction (2). Post-hoc analysis for the HRTF$_{Freq}$ session revealed the only scene with a significant difference was the ***Task*** scene between bases count conditions *128* vs. *16* ($p = .019$).

5.2 Indirect Scaling Results

To analyze the EBA and PC methods, we employed the scale modeling technique described by Pérez-Ortiz and Mantiuk [53], where Thurstone's Case V model is used to calculate maximum likelihood estimation of quality differences. The results were then converted to a just-objectionable-difference (JOD) scale, with confidence intervals calculated via bootstrap resampling [40]. The statistical analysis is the same as in current standardization efforts [54]. Just-noticeable-difference (JND) scores imply discernability along a single dimension, where two conditions exceed a difference limen regarding a certain psychophysiological sensation [48], conventionally exceeding a 75 % probability of one condition being chosen over another. However, in [53] the authors argue JOD better represents differences along multiple dimensions. Although we manipulate only a single parameter for separate test sessions, the multi-modal nature of 6-DoF VR means subjects are offered multiple dimensions they can utilize to give their rating. Consequently, it is not guaranteed that subjects will employ the same difference criteria to come to their conclusions.

To use the data within the model, results from both methods required a pre-processing step. For EBA, the rankings for each scene were converted into a count matrix [15]. For the PC data, the subjective ratings for all sequential condition pairs A and B were converted into a binary form. For a rating r, the binary value was chosen given $(-60 \geq r < -5) \rightarrow A = 0$, $B = 1$, and $(5 < r \leq 60) \rightarrow A = 1$, $B = 0$, where 0 indicates the selected condition. A 10-point margin was given to allow for slider inaccuracies when indicating no difference between conditions A and B. All ratings between ±5, were collected on a scene-condition-pair basis and randomized to have an equal distribution between 0 and 1. This essentially simulates the same outcome of a 2-alternative forced-choice paradigm where 50 % of subjects would randomly select A as being better and the other 50 % B. After conversion, the same count matrix as with the rank-order data was formulated. For both methods two-tailed tests were ran to inspect significant differences ($\alpha = 0.05$ $p < 0.025$ for two-ended distribution) between JOD values.

5.2.1 Rank Order Elimination-by-Aspects

Fig. 6 shows JOD scores and 95 % bootstrapped CIs for the EBA method. Results from the HRTF$_{Spat}$ session, two-tailed significance tests yielded the following differences. For the ***Static*** scene, six significantly different JODs were found for all combinations of *1°*, *5°*, and *10°* vs. *20°* and *30°*. Within the ***Animated*** scene, nine significant differences were found. The only pair with no significant difference was between *1°* vs. *5°*. For the ***Distraction*** scene, four JODs were significantly different between combinations of *1°* and *5°* vs. *20°* and *30°*. For the ***Interaction*** scene, six significant differences were found between combinations of *1°*, *5°*, and *10°* vs. *20°* and *30°* (same as the Static scene). For the ***Task*** scene, six significant differences were found between varying higher and lower azimuth degree conditions, with no significant difference found between *1°* vs. *5°* vs. *10°*. Overall the number of significant differences within each scene was Animated (9) > Static (6) = Interaction (6) = Task (6) > Distraction (4).

For the HRTF$_{Freq}$ session two-tailed tests within the ***Interaction***

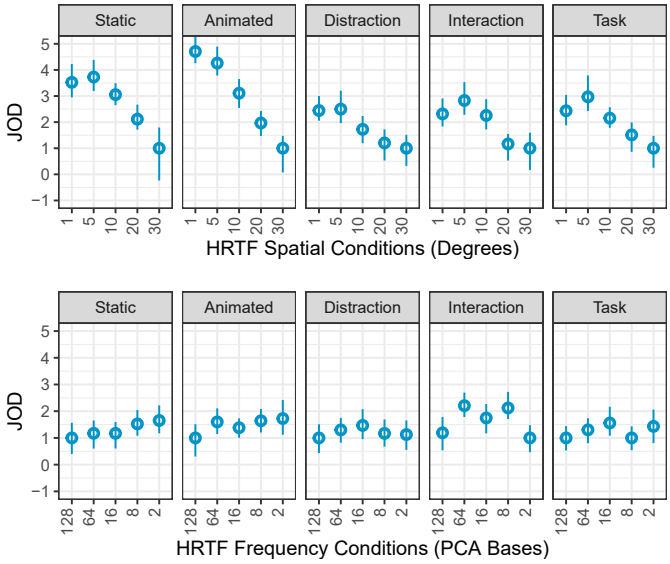

Fig. 6. Mean JODs and bootstrapped 95 % CIs for HRTF$_{Spat}$ (top) and HRTF$_{Freq}$ (bottom) test sessions using **EBA** method. $N_{Spat} = 17$ and $N_{Freq} = 17$.

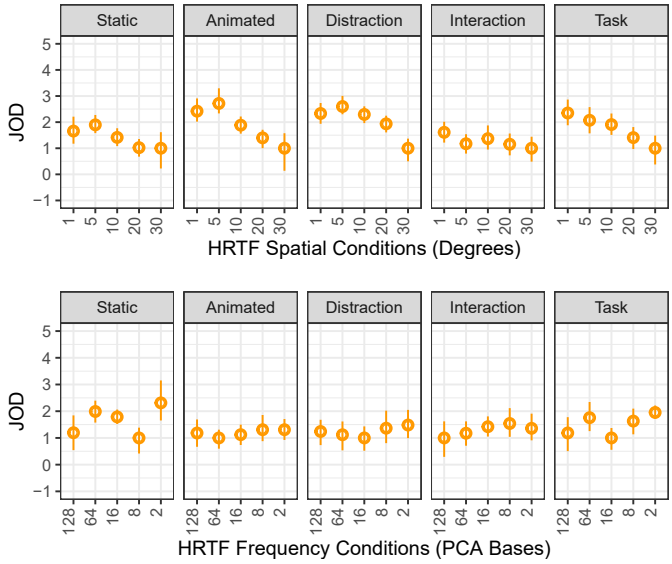

Fig. 7. Mean JODs and bootstrapped 95 % CIs HRTF$_{Spat}$ (top) and HRTF$_{Freq}$ (bottom) test sessions using **PC** method. $N_{Spat} = 17$ and $N_{Freq} = 17$.

scene revealed four significantly different JODs for pairs: *128* vs. *64* bases ($z = 2.19$, $p = .014$), *128* vs. *8* bases ($z = 2.11$, $p = .018$), *64* vs. *2* bases ($z = 2.5$, $p = .006$), and *8* vs. *2* bases ($z = 2.92$, $p = .002$).

5.2.2 Pairwise Comparison

Fig. 7 shows the JOD values and 95 % bootstrapped CIs for the PC method. Results from the HRTF$_{Spat}$ two-tailed tests revealed the following differences. For the ***Static*** scene, 5° vs. 20° ($z = 2.81$, $p = .003$). For the ***Animated*** scene six JODs were significantly different resulting from tests between the higher azimuth resolutions 1° and 5° vs. lower resolutions 10°, 20°, and 30°. For the ***Distraction*** scene five significant differences were found between 30° and all other conditions, and between 5° vs. 20°. No significant differences were found within the ***Interaction*** scene and four significant differences within the ***Test*** scene between 1° vs. 20° ($z = 2.02$, $p = .022$), for 1° vs. 30° ($z = 2.23$, $p = .013$), for 5° vs. 30° ($z = 2.11$, $p = .012$), and for 10° vs. 30° ($z = 2.21$, $p = .014$). Overall, the number of significant differences between all conditions per scene was Animated (5) > Distraction (4) > Task > (3) > Static (1) > Interaction (0). For the HRTF$_{Freq}$ session, two-tailed tests revealed only a single difference between *8* vs. *2* bases in the Static scene.

5.3 NASA-TLX and Tracking Results

Mean opinion scores and 95 % bootstrapped CIs for the TLX questionnaire are plotted in Fig. 8. To analyze the data we ran a one-way between-groups ANOVA on each question for independent variable **Scene** on dependent variable NASA-TLX **Rating** between the **Method** groups. The results shown in Table 2 revealed a main effect of the **Scene** was found for all questions. A main effect between the different groups **Method** was found for *Mental Demand*, *Effort*, and *Performance*, and a weak interaction was found between **Scene** and **Method** for Temporal demand.

For the between group **Methods**, Tukey pairwise comparison t-tests revealed significant differences for: *Mental Demand* between MSHR vs. EBA ($p = .007$), *Effort* between MSHR vs. EBA ($p = .04$) and MSHR vs. PC ($p = .045$), and *Performance* between MSHR vs. MS ($p = .017$). For *Physical Demand*, which yielded the strongest main effect of scene ($\eta_G^2 = .242$), post-hoc t-tests shows scenes ***Task*** and ***Interaction*** were significantly different from each other ($p = .001$), and significantly different from all other scenes. Scenes ***Static***, ***Animated***, and ***Distraction*** yielded no significant difference for *physical demand*.

Finally, the results for the tracking data are shown in Fig. 9 for time (minutes), virtual distance traveled (meters), and head rotations (degrees) across all different scenes for each method. The total average duration for each method was: MSHR $\approx$ 24 minutes, MS $\approx$ 28 minutes, EBA $\approx$ 22 minutes, and PC $\approx$ 45 minutes. Within all methods, the ***Interaction*** scene generally provided the lowest average distance traveled and head movements.

Table 2. Results of individual one-way between-groups analysis of variance for all questions.

Question	Effect	F-Value	p-value	Effect Size
Mental	Method	$F_{(3,132)} = 3.77$	$p = .012$	$\eta_G^2 = .051$
	Scene	$F_{(4,528)} = 18.45$	$p < .001$	$\eta_G^2 = .05$
Physical	Scene	$F_{(4,528)} = 77.07$	$p < .001$	$\eta_G^2 = .242$
Temporal	Scene	$F_{(4,528)} = 21.69$	$p < .001$	$\eta_G^2 = .064$
	Scene:Method	$F_{(12,528)} = 2.02$	$p = .029$	$\eta_G^2 = .019$
Effort	Method	$F_{(3,132)} = 3.19$	$p = .026$	$\eta_G^2 = .037$
	Scene	$F_{(4,528)} = 18.55$	$p < .001$	$\eta_G^2 = .062$
Perform	Method	$F_{(3,132)} = 3.13$	$p < .028$	$\eta_G^2 = .042$
	Scene	$F_{(4,528)} = 7.94$	$p < .001$	$\eta_G^2 = .022$
Frust	Scene	$F_{(4,528)} = 20.94$	$p < .001$	$\eta_G^2 = .063$

6 Discussion

6.1 Methods

For the MSHR method (Fig. 4), results revealed a reasonable number of significant differences between conditions for both test sessions, with a larger variation for the HRTF$_{Spat}$ session depending on the scene (8 to 4) than HRTF$_{Freq}$ (7 to 6). The larger variation in number of significant differences using HRTF$_{Spat}$ conditions implies some effect of scene complexity. This observation is supported through the main effect of the scene shown in the ANOVA analysis, in addition to the difference range for MOS across HRTF$_{Spat}$ conditions ($\approx \Delta 15$-points) between scenes ***Interaction*** and ***Static***. The HRTF$_{Spat}$ session showed results unsurprising given the degradations of the condition set and open reference. The absence of a significant difference between conditions 1° vs. 5° in all scenes implies a perceptual threshold of 5° azimuth resolution for HRTF azimuth discretization consistent with most data-sets and literature [2, 82]. Results for the HRTF$_{Freq}$ session also demonstrate subjects can discriminate between frequency resolutions using an au-

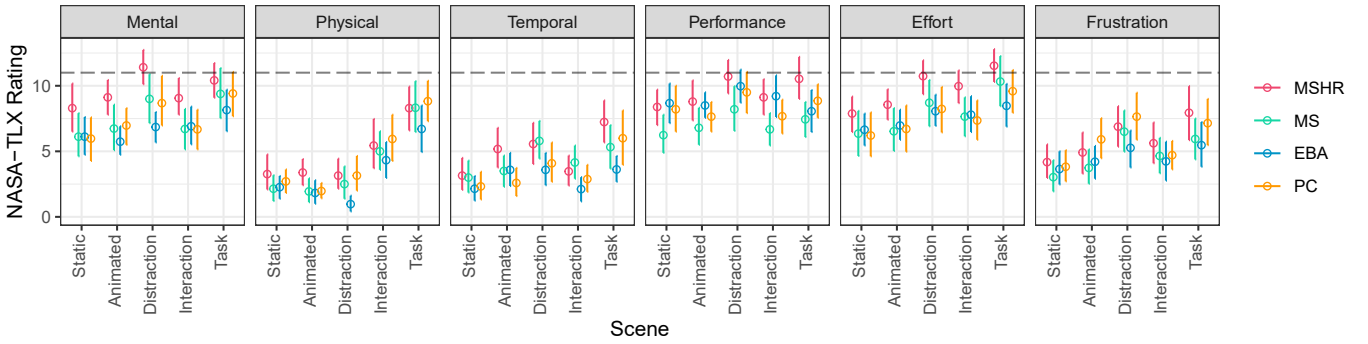

Fig. 8. Mean opinion scores and bootstrapped 95 % CIs of NASA-TLX questionnaire results for both HRTF$_{Spat}$ and HRTF$_{Freq}$ test sessions across all scenes, for each method. The dashed line indicates the midway point 11 along a 21-point scale. For all questions apart from *performance*, scale value 0 = low, and 21 = high. For *performance*, 0 = good, 21 = poor. $N = 68$.

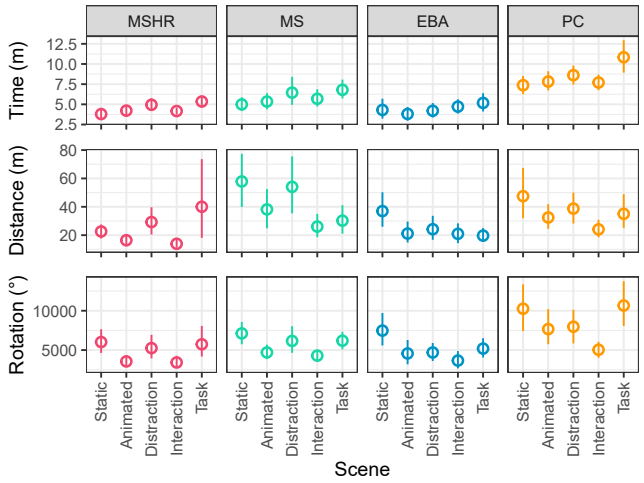

Fig. 9. Mean and bootstrapped 95 % CIs for; Top: Evaluation time (minutes), Mid: Virtual distance travelled (meters) and Bot: Total head rotations (degrees). $N = 68$.

dible reference. The MSHR results validate the use of the employed conditions, showing that differences can be perceived and rated accordingly representative of the impaired spatial and frequency resolutions. Given this, we proceed further with the discussion of other methods using MSHR as a suitable comparison.

The results for the MS method (Fig. 5) HRTF$_{Spat}$ session revealed a similar trend to the MSHR method with notable differences. The range of average MOS for the MS method is compressed, suggesting two mutually compatible possibilities. One, the best high-quality conditions seen in the MSHR results are still not what subjects would consider *'excellent'* in MS. The freedom to account for MSI conflicts beyond the audible differences observed against a fixed reference may have augmented subjects' impression of overall audio quality. For example, the unrealistic timbre or spatial extent of an audio source given its visual physical dimensions. Two, the different interpretation of labels and scale usage across the subjects may result in any number of biases, initially described by Poulton [52]. Biases such as range equalizing bias, centering bias, and contraction bias, are all potential causes of error for direct scaling that are even more prevalent in the absence of an open reference. Research regarding evaluations with and without an open reference also show similar effects [5, 6, 83]. Irrespective of scale positioning, the relative differences also vary. The statistical similarity between conditions *1°*, *5°*, and *10°* across all scenes would suggest a higher perceptual threshold of azimuth resolution than that observed in MSHR results. Lastly, the range of MOS across conditions had a difference of $\approx \Delta 30$-points between scenes, double that of MSHR. The difference suggests that content and scene complexity influence subjective judgments to a greater degree in the absence of a reference, an observation also supported by the varying number of significantly different condition pairs between scenes.

For the HRTF$_{Freq}$ all MS results are completely different compared to MSHR. One might conclude from the MSHR results that higher frequency resolution conditions produce 'excellent' quality, while the MS results show the overall audio quality is 'fair' - 'good'. This emphasizes the potential misinterpretation when using a reference that is presupposed to be the best quality. However, without an explicit open reference, the tendency of quality to decrease with HRTFs of lower bases derived from the same HRTF set is not observed. This implies the differences in frequency resolution of a generic HRTF set yield no decrease in overall quality, and that preference or the fit of the HRTF$_{Freq}$ conditions against subjects' personal HRTF may alone be justification for equal ratings. Further study would be needed to validate if reduction of bases on an individualized HRTF yields the same result, given the more bespoke modeling of the high-frequency peaks and notches.

Results for the EBA method (Fig. 6) are represented on a JOD scale which provides a more comparable means of interpretation than probability against MOS ratings [81]. In the HRTF$_{Spat}$ session, no significant differences were consistently found between conditions *1°* and *5°* for all scenes, similar to the observation from the MSHR and MS method. However, the notably higher number of significant differences found between the conditions for EBA is more consistent with MSHR compared to MS. The only difference in tracking data between EBA and MS is a higher average distance traveled in favor of MS, with time and head rotations remaining comparatively similar for both methods. This would suggest no advantage for the EBA results can be seen in subjects' behavior and therefore, a higher discriminatory power inherent in the EBA method compared to MS in the context of 6-DoF VR. Discernability between conditions also appears to be less susceptible to scene complexity for the EBA method, where ***Interaction*** and ***Task*** scenes yielded a high number of significant differences. While the argument may be proposed that no absolute quality level for conditions can be drawn as with direct scaling, the potential scaling biases in MS make it hard to quantify what any scale labels actually represent. For indirect scaling, differences in a latent multi-dimensional space can be modeled post-evaluation by observed responses rather than attempting to be directly estimated. As such, the JOD scale representation does not assume any maximum quality and acknowledges latent variables impacting perceived audio quality inclusive of MSI. For the HRTF$_{Freq}$ session, the same trend from the MS is present where few significant differences were found across all scenes, thus leading to the same hypothesis regarding preference or personal fit to the HRTFs.

The results for the PC method (Fig. 7) showed a compressed range of JOD values in comparison to the EBA method but with a smaller range for CIs. The result exhibits a similar trend in mean quality levels compared to EBA and MS. In many contexts, the PC method is often credited as having a high discriminatory power due to a psychological tendency of humans to be better at discrimination than judging absolutes [67]. However, the results here indicate no advantage is given in providing more consistent or accurate ratings over EBA or MS methods. Results for the HRTF$_{Spat}$ still suggest a statistical similarity between

higher azimuth resolutions, with a number of significant differences comparable to MS. The range of significant differences across scenes (5 - 0) also indicates a stronger influence of scene complexity similar to MS than EBA and MSHR. Results for the HRTF_{Freq} session lead to the same observations of no perceptual difference in overall audio quality as with MS and EBA.

An apparent difference between indirect methods EBA and PC is the time taken and tracking data (Fig. 9). For PC, subjects must repeat actions for every condition pair resulting in a higher average distance traveled and head movements than in the EBA method (Fig. 9). If movements are not repeated, the subjects are more likely to neglect interactions that induce audible differences. The PC test time is also roughly double compared to the EBA method, similar to previous comparison studies [77]. Both observations are likely due to the repetitive nature of PC plus the proactive actions required by subjects inside 6-DoF VR. The distance traveled for the MS method also shows a comparatively higher distance traveled than EBA. Following the same logic of repetitive motion, even when presented in parallel, subjects have to carry out an extra step of prescribing magnitude estimations. Consequently, additional movements are required to listen, assign, relisten, and adjust for all conditions. Overall, the EBA method appears to combine benefits of parallel presentation of all conditions seen in direct scaling MS and MSHR methods together with the unambiguous response format of indirect scaling. The sequential elimination of parallel conditions in EBA means subjects do not have to repeat actions resulting in a more efficient use of subjects' movements.

6.2 Scene Complexity and Task Load Index

Statistical analysis of the TLX questionnaire revealed independent variable **Scene** to be a main effect for all questions. This finding coincides with main effects in the direct scaling ANOVA HRTF_{Spat} results, along with the varying range of mean JODs for indirect scaling methods. For all methods, HRTF_{Spat} ratings for the ***Animated*** scene show a large range in MOS and JOD values, in addition to the highest number of significant differences for MS, EBA, and PC methods. In all TLX questions, the ***Animated*** scene is also one of the least demanding. Both results are likely due to the scene producing the most prominent auditory cues. Considering that the HRTF_{Spat} conditions were altered only in azimuth resolution combined with the animated train's mostly-lateral trajectory, subjects are more likely to be exposed to binaural localization cues such as inter-aural time and level differences with more passive behavior. In contrast, the ***Distraction*** scene (which included the same animation) proved to: 1) be one of the most demanding scenes for all TLX questions excluding *Physical Demand*, 2) require more movements and head rotations from subjects, and 3) result in quality ratings with decreased differentiation between conditions for all methods. Research regarding complex scenes with multiple specialized audio sources shows directed attention towards a specific object helps detect localization changes [20]. However, our results suggest an additional cognitive overhead may be required in auditory scene analysis and stream segregation [17, 70] that can lower our sensitivity in detecting changes between the employed spatial differences.

The TLX dimension with the most variation and strongest effect size from the ANOVA analysis (Table 2) across scenes was *Physical Demand*, with a clear pattern of mean values showing ***Static*** = ***Animated*** = ***Distraction*** < ***Interaction*** < ***Task***, for all methods. Interestingly, for all referenceless methods, the scenes ***Static***, ***Animated***, and ***Distraction*** generally result in more subject movements than the ***Interaction*** scene. This scene is designed to include more complex interactions, allowing subjects to interact with a radio which subjects mostly moved around their head. Consequently, it is reasonable to suggest the observed physical demand is a consequence of movements linked to more 'coordinated' interactions rather than 6-DoF translational motion. Moreover, for MS and PC methods the ***Interaction*** scene resulted in a low number of significant differences between conditions for the HRTF_{Spat} session. As controller and menu operations remain consistent across all scenes, the significantly higher physical demand for interactive scenes suggest more complex interactions to be a contributing factor towards the lower discrimination between audio quality conditions.

Finally, for *Mental Demand*, *Performance* and *Effort* MSHR was rated worse than referenceless methods, with the largest significant difference being between MSHR and EBA for *mental demand*. Although an open reference provides a clearer target of overall audio quality, it also seems to increase the cognitive load on subjects, in contrast to EBA where subjects can rather quickly remove conditions without further consideration towards relative rating positions on a scale. Inspection of Fig. 9 also suggests the inclusion of an open reference in MSHR seems to reduce 6-DoF exploration in comparison to the other direct scaling method MS with no reference. The implication here is that an open reference causes a higher concentration (and thus *mental demand*) and in turn possibly deters subjects from more active exploration.

6.3 Outlook and Limitations

The results of this study imply the type of interactivity within a VR scene can influence subjects' ability to discern between conditions in the absence of an open reference. More research into the perceptual and cognitive effects of complex interactions on quality judgments through controller tracking data is needed to support this observation. Moreover, all scenes were made intentionally simple for evaluation of psychoacoustic thresholds using auditory rendering parameters of HRTFs given multi-modal input. Contrary to what one might see with realistic virtual scenes, clear visual focal points meant that subjects' movements had little exploratory variation. Research using more real-life-like scenes with high-resolution textures, environmental settings, and background objects may yield a different behavioral response potentially reducing discernability between conditions. In the same vein, the condition sets employed vary only a single parameter of the HRTFs. When testing various algorithmic approaches for audio rendering, the differences in the latent multi-dimensional quality space may be much more difficult for subjects to judge. A continuation of this study using multi-dimensional condition degradations and more realistic acoustic auralization would provide further confidence in method stability.

7 Conclusion

This study compared four quality evaluation methods: multiple-stimulus with and without a reference (MSHR and MS), rank-order elimination-by-aspects (EBA), and pairwise comparison (PC) using two conditions sets targeting quality differences important for real-time binaural audio rendering. The conditions sets were constructed altering spatial (HRTF_{Spat}) and frequency (HRTF_{Freq}) resolutions of a high-quality HRTF file. The evaluation content employed virtual reality scenes of varying complexity. For all methods, subjects were tasked with rating differences in overall audio quality.

Our results show that all methods produce a similar trend in average overall audio quality ratings for HRTF_{Spat} session but differ regarding the scale ranges and the number of significantly different conditions. For the HRTF_{Freq} session, all referenceless methods produce comparable results implying perceived differences in frequency characteristics of HRTFs may not impact overall quality in the absence of an open reference. Overall, the EBA method yields: 1) the fastest evaluation times, 2) less repetitive 6-DoF movement from subjects due to sequential elimination of conditions, 3) results with a high discernability between HRTF_{Spat} conditions for all scenes, and 4) significantly less *mental demand* and *effort* from subjects compared to MSHR. Hence, our results demonstrate that EBA is a pragmatic and sensitive method that should be considered in tackling broader research and standardization challenges when comparatively evaluating spatial audio rendering for multi-modal VR where no high-quality reference is available. Finally, scene complexity is shown to affect quality judgments for HRTF_{Spat} conditions, with scenes that possess more hand interactions (linked to *physical demand*) or a higher number of audio sources showing a tendency to reduce discernability between conditions.

Acknowledgments

This research was partially funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) project number 444832250 - SPP 2236

REFERENCES

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

[2] A. Andreopoulou, D. R. Begault, and B. F. G. Katz. Inter-Laboratory Round Robin HRTF Measurement Comparison. *IEEE Journal of Selected Topics in Signal Processing*, 9(5):895–906, Aug, 2015. doi: 10.1109/jstsp.2015.2400417

[3] C. Armstrong, L. Thresh, D. Murphy, and G. Kearney. A Perceptual Evaluation of Individual and Non-Individual HRTFs: A Case Study of the SADIE II Database. *Applied Sciences*, 8(11):1–21, Oct, 2018. doi: 10.3390/app8112029

[4] S. Bech and N. Zacharov. *Perceptual Audio Evaluation - Theory, Method and Application*. John Wiley & Sons, Chichester UK, 2006. doi: 10.1002/9780470869253

[5] K. Beresford, N. Ford, F. Rumsey, and S. K. Zielinski. Contextual Effects on Sound Quality Judgements: Listening Room and Automotive Environments. In Proc. *Audio Engineering Society 120th Convention*, pp. 1–13. Paris, France, 2006.

[6] K. Beresford, N. Ford, F. Rumsey, and S. K. Zielinski. Contextual Effects on Sound Quality Judgements: Part II – Multi-stimulus vs. Single Stimulus Method. In Proc. *Audio Engineering Society 121st Convention*, pp. 1–19. San Francisco, CA, USA, 2006.

[7] I. Bergstrom, S. Azevedo, P. Papiotis, N. Saldanha, and M. Slater. The Plausibility of a String Quartet Performance in Virtual Reality. *IEEE Transactions on Visualization and Computer Graphics*, 23(4):1352–1359, Jan, 2017. doi: 10.1109/tvcg.2017.2657138

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

[9] P. Bertelson and G. Aschersleben. Automatic Visual Bias of Perceived Auditory Location. *Psychonomic Bulletin & Review*, 5(3):482–489, Sept, 1998. doi: 10.3758/bf03208826

[10] N. Bevan, J. Carter, J. Earthy, T. Geis, and S. Harker. New ISO Standards for Usability, Usability Reports and Usability Measures. In Proc. *Human-Computer Interaction. Theory, Design, Development and Practice*, pp. 268–278. Springer International Publishing, Cham, Switzerland, 2016. doi: 10.1007/978-3-319-39510-4_25

[11] A. Blackler, V. Popovic, and D. Mahar. The Nature of Intuitive Use of Products: An Experimental Approach. *Design Studies*, 24(6):491–506, Nov, 2003. doi: 10.1016/s0142-694x(03)00038-3

[12] J. Blauert. *Spatial Hearing: The Psychophysics of Human Sound Localization*. MIT Press, Cambridge, MA, USA, revised edition ed., 1996. doi: 10.7551/mitpress/6391.001.0001

[13] R. A. Bradley and M. R. Terry. Rank Analysis of Incomplete Block Designs: I. The Method of Paired Comparisons. *Biometrika*, 39(3/4):324–345, Dec, 1952. doi: 10.2307/2334029

[14] J.-P. Bresciani, M. O. Ernst, K. Drewing, G. Bouyer, V. Maury, and A. Kheddar. Feeling What You Hear: Auditory Signals Can Modulate Tactile Tap Perception. *Experimental Brain Research*, 162(2):172–180, Dec, 2005. doi: 10.1007/s00221-004-2128-2

[15] A. Brown and A. Maydeu-Olivares. Item Response Modeling of Forced-Choice Questionnaires. *Educational and Psychological Measurement*, 71(3):460–502, May, 2011. doi: 10.1177/0013164410375112

[16] A. V. Cardello. Commentary: Direct Versus Indirect Scaling: The Gnashing of Psychophysical Worldviews. *Journal of Sensory Studies*, 20(4):373–379, Aug, 2005. doi: 10.1111/j.1745-459x.2005.00032.x

[17] R. P. Carlyon. How the Brain Separates Sounds. *Trends in Cognitive Sciences*, 8(10):465–471, Oct, 2004. doi: 10.1016/j.tics.2004.08.008

[18] J. Cohen. Statistical Power Analysis for the Behavioral Sciences. 1988. doi: 10.4324/9780203771587

[19] P. Dodds, S. V. A. Garí, O. W. Brimijoin, and P. W. Robinson. Auralization Systems for Simulation of Augmented Reality Experiences in Virtual Environments. In Proc. *5th International Conference on Spatial Audio*, pp. 1–6. Ilmenau, Germany, 2019.

[20] R. Eramudugolla, D. R. Irvine, K. I. McAnally, R. L. Martin, and J. B. Mattingley. Directed Attention Eliminates 'Change Deafness' in Complex Auditory Scenes. *Current Biology*, 15(12):1108–1113, June, 2005. doi: 10.1016/j.cub.2005.05.051

[21] B. Feiten, A. Raake, M.-N. Garcia, U. Wüstenhagen, and J. Kroll. Subjective Quality Evaluation of Audio Streaming Applications on Absolute and Paired Rating Scales. In Proc. *Audio Engineering Society 126th Convention*, pp. 1 – 9. Munich, Germany, 2009.

[22] B. d. Gelder and P. Bertelson. Multisensory Integration, Perception and Ecological Validity. *Trends in Cognitive Sciences*, 8(1):460–467, Jan, 2004. doi: 10.1016/j.tics.2003.08.014

[23] S. Girard, M. Pelland, F. Lepore, and O. Collignon. Impact of the Spatial Congruence of Redundant Targets on Within-Modal and Cross-Modal Integration. *Experimental Brain Research*, 224(2):275–285, Nov, 2012. doi: 10.1007/s00221-012-3308-0

[24] W. D. Hairston, M. T. Wallace, J. W. Vaughan, B. E. Stein, J. L. Norris, and J. A. Schirillo. Visual Localization Ability Influences Cross-Modal Bias. *Journal of Cognitive Neuroscience*, 15(1):20–29, Jan, 2003. doi: 10.1162/089892903321107792

[25] S. G. Hart and L. E. Staveland. Development of NASA-TLX (Task Load Index): Results of Empirical and Theoretical Research. *Advances in Psychology*, 52:193–183, Jan, 1988. doi: 10.1016/s0166-4115(08)62386-9

[26] R. M. A. v. d. Heiden, C. P. Janssen, S. F. Donker, and J. L. Kenemans. The Influence of Cognitive Load on Susceptibility to Audio. *Acta Psychologica*, 205:103058, April, 2020. doi: 10.1016/j.actpsy.2020.103058

[27] A. Honda, H. Shibata, S. Hidaka, J. Gyoba, Y. Iwaya, and Y. Suzuki. Effects of head movement and proprioceptive feedback in training of sound localization. *i-Perception*, 4(4):253–264, 2013. doi: 10.1068/i0522

[28] International Telecommunications Union: Radiocommunication Sector. BS.1116-3: Methods for the Subjective Assessment of Small Impairments in Audio Systems. Feb, 2015.

[29] International Telecommunications Union: Radiocommunication Sector. BS.1284-2: General methods for the Subjective Assessment of Sound Quality. Jan, 2019.

[30] International Telecommunications Union: Radiocommunication Sector. BS.1534-3: Method for the Subjective Assessment of Intermediate Quality Level of Audio Systems. Oct, 2015.

[31] International Telecommunications Union: Radiocommunication Sector. BS.2132-0: Method for the Subjective Quality Assessment of Audible Differences of Sound Systems Using Multiple Stimuli Without a Given Reference. Oct, 2019.

[32] International Telecommunications Union: Telecommunications Sector. P.800: Methods for Subjective Determination of Transmission Quality. Sept, 1996.

[33] S. R. Jaeger and A. V. Cardello. Direct and Indirect Hedonic Scaling Methods: A Comparison of the Labeled Affective Magnitude (LAM) Scale and Best–Worst Scaling. *Food Quality and Preference*, 20(3):249–258, April, 2009. doi: 10.1016/j.foodqual.2008.10.005

[34] A. C. Kern and W. Ellermeier. Audio in VR: Effects of a Soundscape and Movement-Triggered Step Sounds on Presence. *Frontiers in Robotics and AI*, 7:20, 2020. doi: 10.3389/frobt.2020.00020

[35] H. Kim, L. Remaggi, P. J. B. Jackson, and A. Hilton. Real VR – Immersive Digital Reality, How to Import the Real World into Head-Mounted Immersive Displays. *Lecture Notes in Computer Science*, pp. 293–318, 2020. doi: 10.1007/978-3-030-41816-8_13

[36] D. J. Kistler and F. L. Wightman. A Model Of Head-Related Transfer Functions Based On Principal Components Analysis And Minimum-Phase Reconstruction. *The Journal of the Acoustic Society of America*, 91(3):1637–1647, March, 1992. doi: 10.1121/1.402444

[37] P. Larsson, D. Västfjäll, and M. Kleiner. Better Presence and Performance in Virtual Environments By Improved Binaural Sound Rendering. In Proc. *Audio Engineering Society 22nd International Conference on Virtual, Synthetic and Entertainment Audio*, pp. 1–8. Espoo, Finland, 2002.

[38] A. Lindau, V. Erbes, S. Lepa, H.-J. Maempel, F. Brinkman, and S. Weinzierl. A Spatial Audio Quality Inventory (SAQI). *Acta Acustica united with Acustica*, 100(5):984–994, Sept, 2014. doi: 10.3813/aaa.918778

[39] R. D. Luce. The Choice Axiom After Twenty Years. *Journal of Mathematical Psychology*, 15(3):215–233, June, 1977. doi: 10.1016/0022-2496(77)90032-3

[40] R. K. Mantiuk. mantiuk/pwcmp. GitHub Repository, April, 2020. `https://github.com/mantiuk/pwcmp` @ 48c9192.

[41] R. McGarrigle, K. J. Munro, P. Dawes, A. J. Stewart, D. R. Moore, J. G. Barry, and S. Amitay. Listening Effort and Fatigue: What Exactly Are We Measuring? A British Society of Audiology Cognition in Hearing Special Interest Group 'White Paper'. *International Journal of Audiology*, 53(7):433–445, March, 2014. doi: 10.3109/14992027.2014.890296

[42] R. Mehra, L. Antani, S. Kim, and D. Manocha. Source and Listener Directivity for Interactive Wave-Based Sound Propagation. *IEEE Transactions on Visualization and Computer Graphics*, 20(4):495–503, March, 2014.

doi: 10.1109/tvcg.2014.38

[43] M. Mihajlov, E. L.-C. Law, and M. Springett. Intuitive Learnability of Touch Gestures for Technology-Naïve Older Adults. *Interacting with Computers*, 27(3):344–356, May, 2015. doi: 10.1093/iwc/iwu044

[44] B. Morillon and S. Baillet. Motor Origin of Temporal Predictions in Auditory Attention. *Proceedings of the National Academy of Sciences*, 114(42):E8913–E8921, Oct, 2017. doi: 10.1073/pnas.1705373114

[45] H. R. Moskowitz. Thoughts on Subjective Measurement, Sensory Metrics and Usefulness of Outcomes. *Journal of Sensory Studies*, 20(4):347–362, Aug, 2005. doi: 10.1111/j.1745-459x.2005.00029.x

[46] M. Naef, O. Staadt, and M. Gross. Spatialized Audio Rendering for Immersive Virtual Environments. In Proc. *ACM Symposium on Virtual Reality Software and Technology*, pp. 65–72. Hong Kong, China, 2002. doi: 10.1145/585740.585752

[47] F. Nagel, T. Sporer, and P. Sedlmeier. Towards a Statistically Well-Grounded Evaluation of Listening Tests - Avoiding Pitfalls, Misuse, and Misconceptions. In Proc. *Audio Engineering Society 128th Convention*, pp. 1–10. London, UK, 2010.

[48] E. B. Newman. The Validity of the Just Noticeable Difference as a Unit of Psychological Magnitude. *Transactions of the Kansas Academy of Science*, 36:172–175, April, 1933. doi: 10.2307/3625353

[49] G. Parseihian and B. F. G. Katz. Rapid head-related transfer function adaptation using a virtual auditory environment. *The Journal of the Acoustical Society of America*, 131(4):2948–2957, 2012. doi: 10.1121/1.3687448

[50] T. Pimentel, A. G. d. Cruz, and R. Deliza. Sensory Evaluation: Sensory Rating and Scoring Methods. *Encyclopedia of Food and Health*, pp. 744–749, Jan, 2016. doi: 10.1016/b978-0-12-384947-2.00617-6

[51] R. L. Plackett. The Analysis of Permutations. *Journal of the Royal Statistical Society*, 24(2):193–202, Jan, 1975. doi: 10.2307/2346567

[52] E. C. Poulton. Models for Biases in Judging Sensory Magnitude. *Psychological Bulletin*, 86(4):777–803, 1979. doi: 10.1037/0033-2909.86.4.777

[53] M. Pérez-Ortiz and R. K. Mantiuk. A Practical Guide and Software for Analysing Pairwise Comparison Experiments. *arXiv:1712.03686*, Dec, 2017.

[54] S. R. Quackenbush and J. Herre. MPEG Standards for Compressed Representation of Immersive Audio. *Proceedings of the IEEE*, 109(99):1578–1589, May, 2021. doi: 10.1109/jproc.2021.3075390

[55] T. Robotham, O. Rummukainen, J. Herre, and E. A. P. Habets. Evaluation of Binaural Renderers in Virtual Reality Environments: Platform and Examples. In Proc. *Audio Engineering Society 145th Convention*, pp. 1–5. New York, NY, USA, 2018.

[56] T. Robotham, O. Rummukainen, J. Herre, and E. A. P. Habets. Online vs. Offline Multiple Stimulus Audio Quality Evaluation for Virtual Reality. In Proc. *Audio Engineering Society 145th Convention*, pp. 1–10. New York, NY, USA, 2018.

[57] O. Rummukainen, T. Robotham, S. J. Schlecht, A. Plinge, J. Herre, and E. A. P. Habets. Audio Quality Evaluation in Virtual Reality: Multiple Stimulus Ranking with Behavior Tracking. In Proc. *Audio Engineering Society Conference on Audio for Virtual and Augmented Reality*, pp. 1–10. Redmond, WA, USA, 2018.

[58] F. Rumsey, S. Zieliński, R. Kassier, and S. Bech. On the Relative Importance of Spatial and Timbral Fidelities in Judgments of Degraded Multichannel Audio Quality. *The Journal of the Acoustical Society of America*, 118(2):968–976, Aug, 2005. doi: 10.1121/1.1945368

[59] A. Sarampalis, S. Kalluri, B. Edwards, and E. Hafter. Objective Measures of Listening Effort: Effects of Background Noise and Noise Reduction. *Journal of Speech, Language, and Hearing Research*, 52(5):1230–1240, Oct, 2009. doi: 10.1044/1092-4388(2009/08-0111)

[60] M. Scheer, H. H. Bülthoff, and L. L. Chuang. Steering Demands Diminish the Early-P3, Late-P3 and RON Components of the Event-Related Potential of Task-Irrelevant Environmental Sounds. *Frontiers in Human Neuroscience*, 10:73, March, 2016. doi: 10.3389/fnhum.2016.00073

[61] N. Schinkel-Bielefeld and A. K. Leschanowsky. How Much is the Use of a Rating Scale by a Listener Influenced by Anchors and by the Listener's Experience. In Proc. *Audio Engineering Society 138th Convention*, pp. 1–10. Warsaw, Poland, 2015.

[62] C. E. Schroeder, D. A. Wilson, T. Radman, H. Scharfman, and P. Lakatos. Dynamics of Active Sensing and perceptual selection. *Current Opinion in Neurobiology*, 20(2):172–176, April, 2010. doi: 10.1016/j.conb.2010.02.010

[63] J.-C. Servotte, M. Goosse, S. H. Campbell, N. Dardenne, B. Pilote, I. L. Simoneau, M. Guillaume, I. Bragard, and A. Ghuysen. Virtual Reality Experience: Immersion, Sense of Presence, and Cybersickness. *Clinical Simulation in Nursing*, 38:35–43, Jan, 2020. doi: 10.1016/j.ecns.2019.09.006

[64] T. Sporer, J. Liebetau, and S. Schneider. Statistics of MUSHRA Revisited. In Proc. *Audio Engineering Society 127th Convention*, pp. 1–9. New York, NY, USA, 2009.

[65] B. E. Stein, B. Rowland, P. Laurienti, and T. R. Stanford. *Multisensory Covergence and Integration*. Academic Press, Oxford, UK, 2009. doi: 10.1016/b978-008045046-9.01112-8

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

[67] H. Stone and J. L. Sidel. *Sensory Evaluation Practices*. Elsevier Academic Press, San Diego, CA, USA, 3rd ed., 2004.

[68] R. C. Streijl, S. Winkler, and D. S. Hands. Mean Opinion Score (MOS) Revisited: Methods and Applications, Limitations and Alternatives. *Multimedia Systems*, 22(2):213–227, Dec, 2014. doi: 10.1007/s00530-014-0446-1

[69] J. Susal, K. Krauss, N. Tsingos, and M. Altman. Immersive Audio for VR. In Proc. *2016 Audio Engineering Convention on Audio for Virtual and Augmented Reality*, pp. 1–8. Los Angeles, CA, USA, 2016.

[70] E. S. Sussman. Integration and Segregation in Auditory Scene Analysis. *The Journal of the Acoustical Society of America*, 117(3):1285–1298, March, 2005. doi: 10.1121/1.1854312

[71] L. L. Thurstone. A Law of Comparative Judgment. *Psychological Review*, 34(4):273–286, 1927. doi: 10.1037/h0070288

[72] A. Tongue. alextongue/hrtf-pca. GitHub Repository, April, 2020. `https://github.com/alextongue/hrtf-pca` @ 2e01968.

[73] A. Tversky. Elimination by Aspects: A Theory of Choice. *Psychological Review*, 79(4):281–299, 1972. doi: 10.1037/h0032955

[74] J. Udesen, T. Piechowiak, and F. Gran. Vision Affects Sound Externalization. In Proc. *Audio Engineering Society 55th International Conference on Spatial Audio*, pp. 1–4. Helsinki, Finland, 2015.

[75] S. Weech, S. Kenny, and M. Barnett-Cowan. Presence and Cybersickness in Virtual Reality Are Negatively Related: A Review. *Frontiers in Psychology*, 10:158, Feb, 2019. doi: 10.3389/fpsyg.2019.00158

[76] T. Welti, O. Khonsaripour, S. E. Olive, and D. Pye. A Comparison of Test Methodologies to Personalize Headphone Sound Quality. In Proc. *Audio Engineering Society 147th Convention*, pp. 1–10. New York, NY, USA, 2019.

[77] F. Wickelmaier, N. Umbach, K. Sering, and S. Choisel. Comparing Three Methods for Sound Quality Evaluation with Respect to Speed and Accuracy. In Proc. *Audio Engineering Society 126th Convention*, pp. 1–10. Munich, Germany, 2009.

[78] N. Zacharov. A Rapid Listening Test Environment - Helping Managers Make Better Decisions. In Proc. *Audio Engineering Society 122nd Convention*, pp. 1–13. Vienna, Austria, 2007.

[79] N. Zacharov, J. Huopaniemi, and M. Hämäläinen. Round Robin Subjective Evaluation of Virtual Home Theatre Sound Systems at the AES 16th International Conference. In Proc. *Audio Engineering Society 16th International Conference on Spatial Sound Reproduction*, pp. 544–556. New York, NY, USA, 1999.

[80] N. Zacharov, C. Pike, F. Melchior, and T. Worch. Next Generation Audio System Assessment Using the Multiple Stimulus Ideal Profile Method. In Proc. *2016 Eighth International Conference on Quality of Multimedia Experience*, pp. 1–6. Lisbon, Portugal, 2016. doi: 10.1109/qomex.2016.7498966

[81] E. Zerman, V. Hulusic, G. Valenzise, R. K. Mantiuk, and F. Dufaux. The Relation Between MOS and Pairwise Comparisons and the Importance of Cross-Content Comparisons. *Electronic Imaging*, 2018(14):1–6, Jan, 2018. doi: 10.2352/issn.2470-1173.2018.14.hvei-517

[82] X.-L. Zhong and B.-S. Xie. Maximal Azimuthal Resolution Needed in Measurements of Head-Related Transfer Functions. *The Journal of the Acoustical Society of America*, 125(4):2209–2220, Feb, 2009. doi: 10.1121/1.3087433

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