JASA ARTICLE

Perceptual implications of different Ambisonics-based methods for binaural reverberation

Isaac Engel,[1,a)] Craig Henry,[1] Sebastià V. Amengual Garí,[2] Philip W. Robinson,[2] and Lorenzo Picinali[1]
[1]*Dyson School of Design Engineering, Imperial College London, London SW7 2DB, United Kingdom*
[2]*Facebook Reality Labs, Redmond, Washington 98052, USA*

ABSTRACT:
Reverberation is essential for the realistic auralisation of enclosed spaces. However, it can be computationally expensive to render with high fidelity and, in practice, simplified models are typically used to lower costs while preserving perceived quality. Ambisonics-based methods may be employed to this purpose as they allow us to render a reverberant sound field more efficiently by limiting its spatial resolution. The present study explores the perceptual impact of two simplifications of Ambisonics-based binaural reverberation that aim to improve efficiency. First, a "hybrid Ambisonics" approach is proposed in which the direct sound path is generated by convolution with a spatially dense head related impulse response set, separately from reverberation. Second, the reverberant virtual loudspeaker method (RVL) is presented as a computationally efficient approach to dynamically render binaural reverberation for multiple sources with the potential limitation of inaccurately simulating listener's head rotations. Numerical and perceptual evaluations suggest that the perceived quality of hybrid Ambisonics auralisations of two measured rooms ceased to improve beyond the third order, which is a lower threshold than what was found by previous studies in which the direct sound path was not processed separately. Additionally, RVL is shown to produce auralisations with comparable perceived quality to Ambisonics renderings.
© 2021 Author(s). All article content, except where otherwise noted, is licensed under a Creative Commons Attribution (CC BY) license (http://creativecommons.org/licenses/by/4.0/). https://doi.org/10.1121/10.0003437
(Received 16 June 2020; revised 21 December 2020; accepted 11 January 2021; published online 4 February 2021)
[Editor: Jonas Braasch] Pages: 895–910

I. INTRODUCTION

Digital reverberation (reverb) was first conceived by Schroeder and Logan (1961) and has undergone continuous evolution ever since (Välimäki *et al.*, 2016). For the most part, research in this area has been driven by the music and acoustic architecture industries in which efficiency often comes second to fidelity when producing room auralisations. More recently, however, the emergence of virtual and augmented reality has increased the demand for highly realistic interactive audiovisual experiences. The real-time requirements of such applications mean that acoustic modelling must often be simplified in favour of efficiency, even more so if audio-dedicated computational resources are limited, which is likely the case in portable devices.

A common approach to simplify the computation of a reverberant sound field is to encode it in the spherical harmonics (SH) domain, popularly known as Ambisonics in the context of audio production (Gerzon, 1985; Zotter and Frank, 2019). In this encoding process, a specific SH order (hereafter, spatial or Ambisonics order) may be chosen, which dictates the spatial resolution of the reproduced sound field as well as the computational load and memory requirements (Avni *et al.*, 2013). This can be useful for applications which demand real-time dynamic room auralisations. For instance, Schissler *et al.* (2017) describe a practical implementation which simulates sound propagation through physical and geometrical models and processes the resulting sound field in the Ambisonics domain at different spatial orders, depending on its directivity at each time instant.

Arguably, the minimum spatial order required for a binaural Ambisonics rendering is mostly dictated by the direct sound path from the source to the listener, rather than the reverb, as the former is generally more directive than the latter and, therefore, needs finer spatial resolution to be simulated accurately (Engel *et al.*, 2019; Lübeck *et al.*, 2020; Schissler *et al.*, 2017). In this study, a "hybrid Ambisonics" method is proposed in which the direct path is rendered through convolution with head related impulse responses (HRIRs) sampled in a dense spatial grid, whereas the reverb is processed in the SH domain. This contrasts with the more straightforward traditional method, here referred to as "standard Ambisonics," in which the sound field is rendered as a whole in the SH domain (Ahrens and Andersson, 2019; Zotter and Frank, 2019). Even though the goal of this study is not to perform a direct comparison between hybrid and standard Ambisonics, it is expected that the former will require a lower spatial order than the latter to produce renderings of similar perceived quality as suggested in a preliminary study by the present authors (Engel *et al.*, 2019). The evaluation of the proposed method will give insight on the minimum spatial resolution required to render Ambisonics-based reverb, assuming that the direct sound path is simulated

[a)]Electronic mail: isaac.engel@imperial.ac.uk, ORCID: 0000-0001-7355-0829.

separately and with enough accuracy. This could, in turn, lead to the development of more efficient rendering methods.

At the same time, a practical limitation of Ambisonics rendering has to do with the number of sound sources that can be processed dynamically, i.e., in an interactive environment where sources or listener change their position with time. Typically, when multiple sources are dynamically rendered through either standard or hybrid Ambisonics, it is necessary to perform separate convolutions with room impulse responses (RIRs) for each source. These RIRs must be either precomputed, which can become memory-intensive if each source-listener position pair is considered, or calculated in real time, which quickly becomes costly as the number of sources or the RIR length increases (Schissler *et al.*, 2014). In practice, RIRs need not be modified to simulate listener head rotations, e.g., as shown by Noisternig *et al.* (2003), but they must still be recomputed for translational movements of sources or listener. In this study, the reverberant virtual loudspeaker method (RVL) is presented as a way to binaurally render an arbitrary number of sources in a reverberant space with a relatively low computational cost while allowing for the dynamic addition and translational movement of the sources. This is achieved by making several assumptions such as the listener having low sensitivity to the directionality of reverb (Lindau *et al.*, 2012). The main drawback of RVL is that listener's head rotations are approximated by having the room "locked" to the head as explained in more detail in Sec. II C. This allows the algorithm to be highly efficient at simulating a large number of sources, but whether such simplifications negatively affect the realism of the rendering is something that remains to be investigated.

The general goal of this study is to explore the perception of Ambisonics-based binaural reverb with aims to make recommendations for efficient rendering techniques. More concretely, two main research questions are tackled through numerical analyses and perceptual evaluation:

(1) What is the perceptual impact of decreasing the spatial order of hybrid Ambisonics binaural reverberation? (experiment 1), and
(2) how does RVL compare to a more accurate method in terms of subjective preference, given its approximate simulation of head rotations? (experiment 2).

The rest of this paper is structured as follows. Section II provides a literature review; Sec. III describes the methods, including the measurements and binaural rendering procedure; Sec. IV presents numerical analyses of the methods under comparison; Sec. V describes the listening tests performed to perceptually evaluate the methods; Sec. VI discusses the results and potential future work, and Sec. VII summarises the findings and concludes the paper.

II. BACKGROUND AND MOTIVATIONS

A. Reverb perception: A summary

Reverberation comes as a result of pairing an acoustic source with an environment. As a sound wave propagates from the source, it interacts with its surroundings, leading to reflection and diffraction. Consequently, filtered replicas of the original wavefront arrive at a receiver through different paths at distinct times. As time passes, the echo density increases as the wave continues to interact with the room, eventually resulting in a diffuse reverberant sound field. This process highly depends on the geometry of the room and the acoustic properties of the materials therein.

The effects of room acoustics on auditory perception have long been an active research topic. The precedence effect establishes that the direct sound allows the listener to localise the source, whereas later reflections are generally not perceived as separate auditory events (Brown *et al.*, 2015; Litovsky *et al.*, 1999; Wallach *et al.*, 1949). However, strong specular early reflections can shift the perceived position of a source, broaden its apparent width (Olive and Toole, 1989), and modify its spectrum due to phase cancellations and subsequent comb-filtering (Bech, 1996). This can affect the perception of the actual space. For instance, Barron and Marshall (1981) state that the timing, direction, and spectrum of early lateral reflections contribute to the room envelopment. Similarly, the time delay between the direct sound and the first perceptually distinct early reflection has been shown to affect the perception of presence and environment dimensions in small rooms (Kaplanis *et al.*, 2014) and the intimacy of concert halls (Beranek, 2008). As the temporal density of the reflections increases, perception is governed less by temporal characteristics and more by statistical properties of the reverberant tail. Research done by Yadav *et al.* (2013) suggests that reverberation time (RT) contributes to the perception of size most significantly in large rooms, whereas early reflections are of greater importance in small rooms. With respect to binaural rendering, it has been shown that reverb improves the externalisation of sound sources in the binaural domain, even if only early reflections are used (Begault *et al.*, 2001). Also, it has been found that the congruence between presented virtual sounds and the acoustic properties of the actual listening space contribute to the level of externalisation (Werner *et al.*, 2016).

Based on the aforementioned research, various reverb-rendering methods which try to achieve high fidelity at reasonable costs have been proposed throughout the years. According to a comprehensive review by Valimaki *et al.* (2012), these methods can be generally classified in three categories: delay networks, such as feedback delay networks (Jot, 1997; Jot and Chaigne, 1991) or Schroeder reverberators (Schroeder and Logan, 1961); convolution algorithms in which a dry input signal is convolved with an omnidirectional or Ambisonics RIR; and computational acoustics, which encompass geometry-based simulations, such as the image source method (Allen and Berkley, 1979) and wave-based methods, similar to the finite-difference time-domain method (Botteldooren, 1995). In practice, these categories overlap; for instance, an RIR used for convolution may be generated through computational acoustics (Pelzer *et al.*, 2014; Schissler *et al.*, 2017).

The present work focuses on convolution methods based on Ambisonics RIRs, measured with a spherical

microphone array, and binaural room impulse responses (BRIRs), measured on a head and torso simulator. The measurement process and rendering methods will be explained in more detail in Sec. III.

B. Spatial order perception and hybrid Ambisonics reverb

Standard Ambisonics binaural rendering typically involves convolving a dry audio signal with an Ambisonics RIR (either measured or simulated) for each rendered sound source. Then, all the resulting Ambisonics signals may be accumulated into a single sound field, which is decoded to a binaural signal by means of a free-field head related transfer function (HRTF) and a method of choice, e.g., virtual loudspeakers (Bernschütz *et al.*, 2014; McKeag and McGrath, 1996) or by convolution of the sound field and the HRTF in the SH domain (Schörkhuber *et al.*, 2018; Zaunschirm *et al.*, 2018). A straightforward way of reducing the cost of this process is by decreasing the spatial order of the Ambisonics signals, but this can alter the perception of the resulting auralisations as previous studies have shown. First, Avni *et al.* (2013) performed listening tests with simulated room renderings of varying order, showing that listeners mainly relied on perception of spaciousness and timbre to discriminate them, with higher orders producing spatially sharper and brighter sounds. Later, Bernschütz (2016, Sec. 5.6.1) observed that renderings became generally indistinguishable from each other for spatial orders of 11 and above, obtaining "excellent" results for orders as low as 5 in noncritical scenarios. He also reported that perceptual differences between spatial orders were more accentuated for direct sound and early reflections than they were for diffuse reverb. More recently, Ahrens and Andersson (2019) reported that order 8 was sufficient for lateral sources when compared to auralisations based on measured BRIRs, but slight spectral differences were detected up to 29th order for frontal sources in a discrimination task.

The studies above suggest that realistic standard Ambisonics auralisations may be achievable if a sufficiently high spatial order is employed. However, these may be computationally costly and their feasibility in practice is limited as commercially available microphone arrays are generally of order four and lower. A promising alternative is to render the direct sound (and, possibly, some early reflections) by convolution with a spatially dense HRIR dataset while computing the rest of the RIR in the Ambisonics domain. The rationale is that if the direct sound path is rendered accurately, sources should still be well localised because of the precedent effect (Brown *et al.*, 2015; Wallach *et al.*, 1949), minimising perceptual degradation caused by spatial order reduction. This approach is referred to here as hybrid Ambisonics and has been previously employed by Picinali *et al.* (2017) and Engel *et al.* (2019). Due to the reasons stated above, it is hypothesised that hybrid Ambisonics could potentially achieve comparable spatial and overall quality to standard Ambisonics at lower orders, reducing computational requirements and the need for costly high-order microphone arrays.

Promising results have recently been reported by Lübeck *et al.* (2020), who showed through perceptual tests that the minimum required spatial order for early reflections and late reverb was significantly lower than it was for the direct sound path, for auralisations based on sparse BRIR grids. An important difference between that study and the present study is that they generated their sparse BRIR set by means of spatial subsampling (Bernschütz *et al.*, 2014), which introduces both aliasing and truncation error in the signals (Ben-Hur *et al.*, 2019), whereas in this work, Ambisonics RIRs were directly truncated in the SH domain. Whether the findings of Lübeck *et al.* (2020) can be extended to the rendering of order-truncated (rather than spatially subsampled) Ambisonics sound fields is a question that the present work aims to answer.

C. Multiple-source rendering and RVL

In both standard and hybrid Ambisonics renderings, the cost of the convolution stage increases (at least) linearly with the number of simulated sources as the dry audio signal of each source must typically be convolved with a separate Ambisonics RIR (Schissler *et al.*, 2017). In a low-cost scenario, this can limit the number of sources that can be rendered dynamically, e.g., allowing the addition of a new source or the changing of a source's position in real time. RVL—previously used by Picinali *et al.* (2017) and Engel *et al.* (2019) and natively implemented by the 3D Tune-In Toolkit (Cuevas-Rodríguez *et al.*, 2019)—is proposed here as an alternative computationally efficient approach to dynamically render multiple sources. Its main feature is that the number of convolutions needed to produce a reverberant sound field is independent of the number of rendered sources.

RVL is inspired by the classic virtual Ambisonics approach first outlined by McKeag and McGrath (1996) and later used by Noisternig *et al.* (2003). In the original method, one or more anechoic sound sources are encoded in an Ambisonics sound field, which is then decoded to a virtual loudspeaker grid distributed around the listener, and the resulting signals are finally convolved with the corresponding HRIRs to produce the binaural output. To implement reverb, Noisternig *et al.* (2003) suggested computing early reflections as additional sources and late reverb through a delay network. In RVL, the procedure is analogous to anechoic virtual Ambisonics except that BRIRs are used in place of HRIRs, effectively integrating the room acoustics in the binaural rendering. Also, the direct sound path is rendered separately from the reverb through convolution with discrete HRIRs (like in hybrid Ambisonics) as will be explained in Sec. III.

Because the convolutions with the BRIRs happen at the Ambisonics decoding step, once all sources have been blended into a single sound field, the number of required real-time convolutions is always $2(N+1)^2$, where N is the

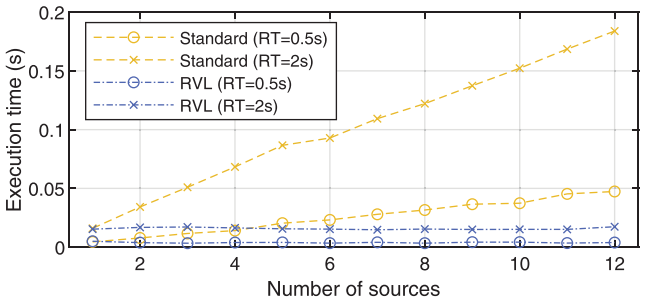


FIG. 1. (Color online) Comparison between the average execution time of the convolution stage in standard Ambisonics and a RVL binaural rendering as a function of the number of rendered sources for two different RTs. A random input signal with a length of 1024 samples was used as input. Simulations were performed in MATLAB (The MathWorks, Natick, MA) using the overlap-add method (Oppenheim *et al.*, 2001), running on a quad-core processor at 2.8 GHz.

Ambisonics order, independent of the number of sources. This feature makes RVL highly efficient at dynamically rendering multiple sources as shown in Fig. 1. Also, it allows for the simulation of virtual sound sources at any position in a sphere around the listener from a reduced set of measured BRIRs, e.g., six BRIRs for first order. Therefore, it requires fewer measurements and memory usage than traditional convolution-based methods such as standard Ambisonics, which need separate RIRs or BRIRs for every possible source-receiver pair location. The main limitation of RVL is that although the relative position of the sound sources can be changed in the Ambisonics domain, the room is head-locked due to the set of BRIRs being fixed. This means that a rotation of the listener's head is simulated by translating all sound sources in the opposite direction, which may produce inaccurate reflections as depicted in Fig. 2. The assumption of RVL is that such approximations will not be noticeable by listeners, e.g., directionality of late reverb may not generally be perceived by listeners, according to Lindau *et al.* (2012) or, at least, will not lead to implausible renderings (Lindau and Weinzierl, 2012).

A previous study by Picinali *et al.* (2017) evaluated the perceived quality of RVL auralisations of different spatial orders while the direct sound path was rendered identically for all conditions. The results suggested that first-order RVL was able to produce room auralisations which were indistinguishable from higher order simulations. This study was web-based and had some limitations, namely, that it was carried out with uncontrolled hardware in an uncontrolled environment and did not implement head tracking, thus, the perceptual effect of approximated head rotations could not be evaluated. Furthermore, renderings were simulation based instead of measurement based and it lacked a comparison with a benchmark method, all of which limited the scope of the findings.

D. Contributions

The goal of this study is to explore the perceptual effect of applying practical simplifications to binaural Ambisonics reverb—first, by investigating the perceptual impact of varying spatial order on the proposed hybrid Ambisonics approach, which is expected to be smaller than the impact on standard Ambisonics that has been reported in previous studies (Ahrens and Andersson, 2019; Avni *et al.*, 2013; Bernschütz, 2016) and, second, by comparing the proposed

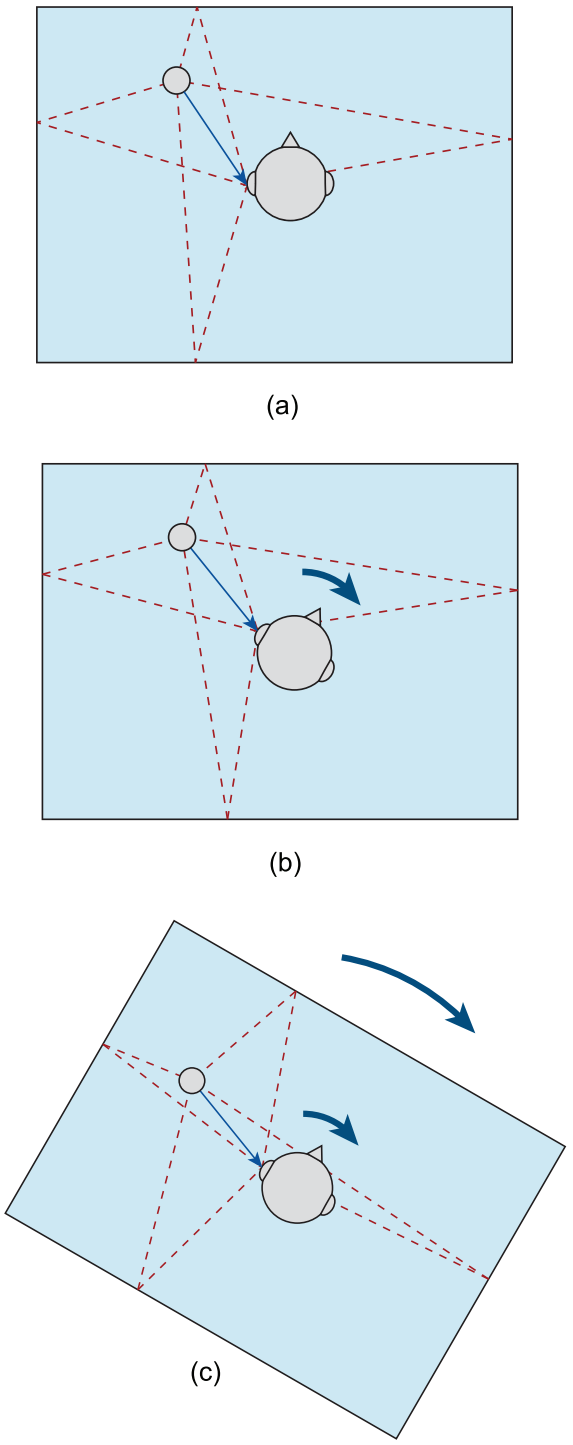
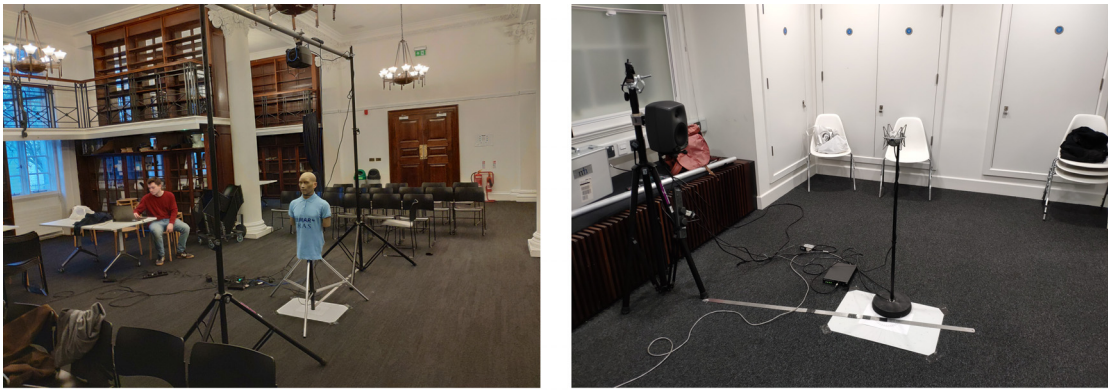


FIG. 2. (Color online) Direct sound path and first-order early reflections as they reach the left ear of a listener in the following three scenarios: (a) before any head rotation, (b) canonical rendering after a head rotation of 30 deg clockwise, and (c) RVL rendering after the same head rotation. Note that in (c), the direct sound path is accurate, whereas the room is head-locked, affecting the incoming direction of reflections.

https://doi.org/10.1121/10.0003437



(a) (b)

FIG. 3. (Color online) The two measured rooms. (a) The library is shown during a BRIR measurement. (b) The trapezoid is shown during a RIR measurement.

computationally efficient RVL to a more accurate approach in terms of subjective preference. Two different rooms were measured, and dynamic binaural renderings were produced through hybrid Ambisonics up to order four and through first-order RVL. The renderings were compared through numerical analyses and perceptual evaluations.

III. METHODS

This section describes how binaural signals were produced with both hybrid Ambisonics and RVL. First, the procedure to generate RIRs for either method is detailed. This is followed by a description of the binaural rendering process for either method. Finally, a description of the audio material employed in the experiments is provided. It is important to emphasise that the direct sound path was identically rendered across all hybrid Ambisonics and RVL conditions, hence, the output binaural signals differed only in the reverb. Non-individualised HRIRs and BRIRs from a KEMAR head and torso simulator (GRAS, Holte, Denmark) were employed throughout.

A. Measurements and RIR generation

Two rooms were measured as shown in Fig. 3. The first room (*library*) was a large, open space with a carpeted floor, high ceilings, and furniture, including chairs, desks, and bookshelves. The second space (*trapezoid*) was a small meeting room also with a carpeted floor, four slightly asymmetrical walls (two of them made of glass), and with no furniture except for some chairs. Acoustic measurements of the rooms are reported in Table I.

In each room, three RIRs were measured using the sine sweep technique (Farina, 2007) with the receiver placed at the centre of the room and sources at relative azimuths $\varphi = [-30°, 0°, 30°]$, an elevation of $\theta = 0°$ and a distance of $r = 1.2$ m (trapezoid) or $r = 1.5$ m (library). A 32-capsule fourth-order spherical microphone array (Eigenmike, mh acoustics, Summit, NJ, United States) acted as a receiver and a Genelec 8030 loudspeaker (Iisalmi, Finland) acted as the source. From the 32-channel RIRs, zeroth- to fourth-order Ambisonics RIRs were generated using the Eigenstudio software package (mh acoustics). RIRs of orders 0–3 were obtained by truncating the fourth-order signals. According to the manufacturer specifications, equalisation was applied such that all Ambisonics channels had a nominally flat magnitude response up to the spatial Nyquist frequency (approximately 8 kHz) and down to the lowest operating frequency of each Ambisonics channel, namely, 30 Hz for orders zero and one, 400 Hz for order two, 1 kHz for order three, and 1.8 kHz for order four (mh acoustics, 2016).

For hybrid Ambisonics renderings, all RIRs had the direct sound path removed by replacing the first 4.32 ms (trapezoid) or 3.88 ms (library) after the onset with silence and applying a Hanning window for the RIR fade-in. This time was calculated analytically by subtracting the propagation time of the direct sound path from that of the first reflection minus a safety window of 30 samples (0.68 ms). Finally, RIRs were windowed at the corresponding RT and applied a de-noising procedure to remove the noise floor (Cabrera *et al.*, 2011).

For RVL renderings, BRIRs were measured with a KEMAR head and torso simulator from six directions (front, back, left, right, up, down) using the same loudspeaker at the same distance as in the RIR measurements and applied identical post-processing, i.e., removing direct sound, windowing, and de-noising. Additionally, frontal BRIRs were used as a reference to equalise Ambisonics RIRs with a

TABLE I. Acoustic parameters of the two measured rooms, including reverberation time (RT) per octave band, early decay time (EDT) per octave band and broadband direct to reverberant ratio (DRR), calculated according to Zahorik (2002).

	f (Hz)	250	500	1000	2000	4000	8000
Library	RT (s)	1.47	1.35	1.16	0.98	0.73	0.52
	EDT (s)	1.21	1.11	1.08	0.57	0.37	0.22
	DRR (dB)	10.09	—	—	—	—	—
Trapezoid	RT (s)	0.78	0.63	0.53	0.48	0.49	0.46
	EDT (s)	0.70	0.49	0.43	0.36	0.34	0.28
	DRR (dB)	4.36	—	—	—	—	—

series of second-order filters, similar to Ahrens and Andersson (2019), with the goal of minimising spectral error due to the spatial order limitation of the signal (Avni *et al.*, 2013).

B. Binaural rendering

For hybrid Ambisonics, head-tracked binaural renderings were generated in real time as follows. The direct sound path was rendered by convolving each source's dry audio signal with an HRIR generated through barycentric interpolation from the three closest available directions, selected from a set of 8802 HRIRs measured on a KEMAR head and torso simulator (Armstrong *et al.*, 2018). HRIRs were aligned prior to interpolation, and interaural time differences were restored assuming a nominal head radius of 8.8 cm. Reverb was rendered in the Ambisonics domain at spatial orders 0–4 using the virtual loudspeaker approach (McKeag and McGrath, 1996), which is equivalent to applying spatial subsampling to the HRIR dataset (Bernschütz *et al.*, 2014). To do so, the measured Ambisonics RIRs (with the direct sound removed) were convolved offline with the dry audio signals and then decoded to loudspeaker signals using a sampling decoder (Zotter and Frank, 2019, Sec. 4.9.1). Depending on the spatial order N, an appropriate number of virtual loudspeakers $M_N \geq (N+1)^2$ was used and placed at the vertices of a platonic solid (regular and convex polyhedron) when possible: octahedron ($M_{0,1} = 6$), icosahedron ($M_2 = 12$), and dodecahedron ($M_3 = 20$). Because no platonic solid exists with 25 or more vertices, a quasi-regular pentakis-dodecahedral layout ($M_4 = 32$), which is the same one used for the capsule placement on the Eigenmike microphone array, was used for $N = 4$. Finally, the virtual loudspeaker signals were convolved with the corresponding interpolated HRIRs.

For RVL, head-tracked binaural renderings were also generated in real time as detailed in Sec. II C. The direct sound path was rendered identically to hybrid Ambisonics. The reverb was generated by encoding all sources' dry audio signals in a single first-order Ambisonics sound field, which was then decoded to six virtual loudspeaker signals using a sampling decoder and a octahedral grid (same as first-order hybrid Ambisonics), and these were finally convolved with the six measured KEMAR BRIRs (also without the direct sound).

The gain of the reverberant sound fields was adjusted offline so that the binaural renderings' direct to reverberant ratio (DRR; ratio between direct and reverberant sound energy) matched that of the frontal KEMAR recordings. The 3D Tune-In Toolkit (Cuevas-Rodríguez *et al.*, 2019) was used as a spatial audio engine, taking care of HRIR interpolation and real-time convolutions. Also, it enabled head tracking by means of an IMU-based tracker (EdTracker Pro Wireless, Wokingham, United Kingdom). Informal tests showed that the end-to-end tracking latency was low enough to not be noticeable during the perceptual evaluation.

C. Audio material

Two different types of audio material were used in the perceptual evaluation, each being an auditory scene comprising one or more spatialised sound sources. Source positions were chosen after a pilot study in which they were verified to provide good separation and externalisation. All sources were presented at a relative elevation of $\theta = 0^\circ$ and a distance of 1.5 m (library) or 1.2 m (trapezoid):

(1) *Music*: a performance of "Take Five" by Paul Desmond, consisting of dry recordings of piano, drum kit, and saxophone, spatialised as three different sound sources at azimuths $\varphi = [-30^\circ, 0^\circ, 30^\circ]$, respectively. The audio tracks were recorded separately in near-anechoic conditions and had a length of 47 s.
(2) *Speech*: dry recording of a single female speaker (Hansen and Munch, 1991) at azimuth $\varphi = -30^\circ$. The audio track had a length of 47 s.

IV. NUMERICAL ANALYSES

This section contains numerical analyses of the signals used for the perceptual evaluation. First, a descriptive analysis of the Ambisonics RIRs is performed. Then, the effect of changing the spatial order is evaluated on the synthesised BRIRs through different metrics. Finally, the effects of rendering reverb statically are explored.

A. Descriptive analysis of Ambisonics RIRs

Figure 4 illustrates the differences in spatial structure of Ambisonics RIRs of both rooms when rendered at different spatial orders. The time axis is split in three segments that will be simply referred to as *direct sound* or HRIR ($0 < t < \tau_{\text{dir}}$), *early reflections* ($\tau_{\text{dir}} < t < \tau_{\text{mix}}$), and *late reverberation* ($t > \tau_{\text{mix}}$), following typical room acoustics nomenclature. For simplicity, early reflections and late reverb as a whole may also be referred to simply as "reverb." The time instant τ_{dir} separates the direct sound and reverb and was calculated analytically as 3.88 ms for the library and 4.32 ms for the trapezoid as mentioned in Sec. III. Note that the direct sound is represented by an approximate spatial delta in Fig. 4, indicating that it was processed through convolution with an HRIR instead of along with the Ambisonics RIR. The mixing time (separation between early reflections and late reverb) was defined at $\tau_{\text{mix}} = 40$ ms, according to Olive and Toole (1989) and considering the RT of the rooms. This time approximately coincides with the precedence effect threshold for speech and music (Moore, 2012; Wallach *et al.*, 1949), meaning that reflections arriving before then are likely not to be perceived as separate auditory events but as events to shift the perception of the leading stimulus in terms of the colouration and source width (Bech, 1996; Olive and Toole, 1989). Note that τ_{mix} does not intend to follow the more rigorous definition of perceptual mixing time from Lindau *et al.* (2010), but this was not critical to the experiment as it was just used for visualisation purposes.

https://doi.org/10.1121/10.0003437

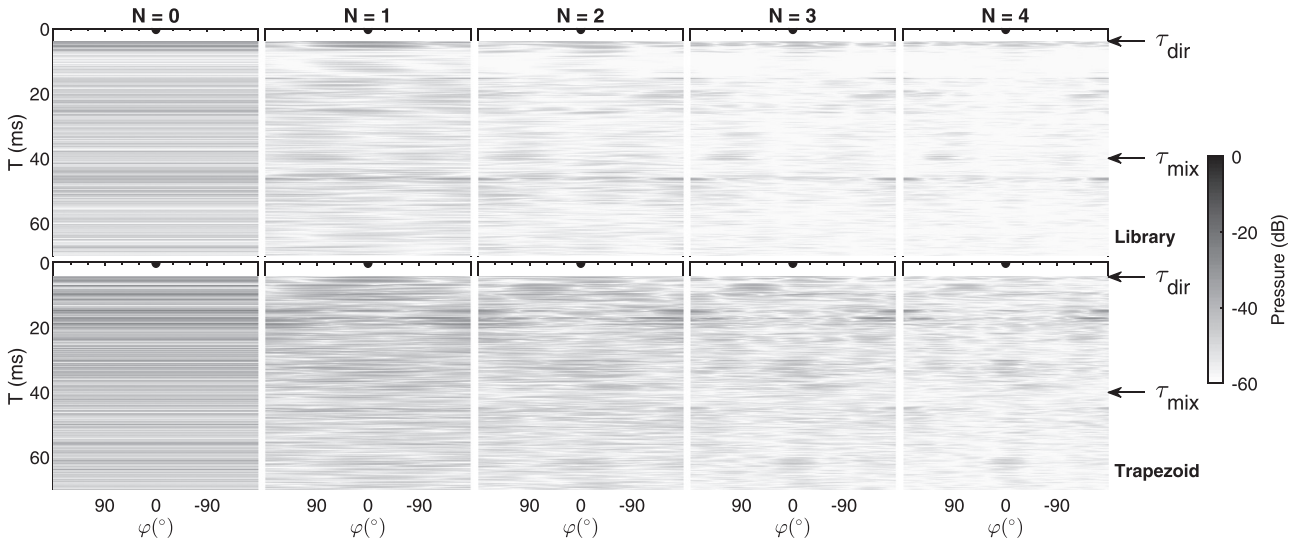


FIG. 4. Spatial RIRs of zeroth to fourth spatial orders (N) for a source placed in front of a listener in the library (top) and trapezoid (bottom) as a function of time and azimuth. The division between direct sound and reverb (τ_{dir}) and the perceptual mixing time (τ_{mix}) are indicated. The direct sound ($t < \tau_{dir}$) is represented by a discrete spatial delta at $(t = 0\,\text{s}\,, \varphi = 0^\circ)$ to indicate that it was excluded from the Ambisonics rendering and generated through convolution with an HRIR.

Given that the direct sound was unchanged throughout the different conditions, the early reflections constitute the segment of the RIR where spatial resolution is most critical. By observing Fig. 4, it is evident that reflections become more diffuse in the azimuth axis, i.e., less directional, at lower spatial orders, with $N = 0$ being the extreme case in which the signal becomes isotropic. It can also be seen how individual reflections are generally less salient in the library than in the trapezoid, which may lead to a lower requirement in terms of spatial order.

B. Objective binaural metrics

BRIRs were synthesised for zeroth- to fourth-order hybrid Ambisonics and first-order RVL. This was done for a source placed at $(\varphi = 30^\circ, \theta = 0^\circ)$ in both rooms. These BRIRs were analysed to quantify the expected perceived quality of each condition. First, the interaural cross-correlation coefficient (IACC) was calculated, which is an objective metric commonly used to predict spatial perception from binaural content (Beranek, 2008; Nowak and Klockgether, 2017; Okano *et al.*, 1998). As a rule of thumb and per the aforementioned studies, a lower IACC often translates to higher perceived spatial quality. Figure 5(a) shows how the IACC of the direct sound is similar across all conditions, which was expected given that the HRIR was not modified. For early reflections and late reverb, differences become larger with zeroth-order renderings showing the highest IACC. This was also expected given that they contained essentially isotropic reverb which produced highly correlated binaural signals. For the rest of the spatial orders, differences seem to increase above 1 kHz with higher orders generally showing lower IACCs. Unexpectedly, RVL reverb mostly obtained lower values than did the hybrid Ambisonics conditions, including the higher order conditions.

From the IACC, other more easily interpretable metrics may be derived. One which is typically used in room acoustics studies to estimate room spatial quality is the binaural quality index (BQI; Beranek, 2008). According to Nowak and Klockgether (2017), it may be calculated as

$$\text{BQI} = 1 - \frac{\text{IACC}_{500} + \text{IACC}_{1000} + \text{IACC}_{2000}}{3}, \qquad (1)$$

where IACC_c represents the IACC for the octave band centred at $f = c$. It is evident from Fig. 5(b) that the early reflections obtained lower overall BQI values than did the late reverb and also showed more variance across conditions, supporting the idea that they are the more perceptually critical part of the RIR. These results are consistent with previous studies, which showed a higher BQI for late reverb with values close to 0.8 for the best performing conditions (Nowak and Klockgether, 2017). Consistently with the IACC analysis, the zeroth-order BRIR obtained the lowest BQI values, therefore, predicting a low perceived spatial quality. As expected, higher orders produced higher BQI values, although slight differences were observed between the rooms. Whereas in the trapezoid the trend was preserved until order 4, the BQI in the library seems to plateau between orders one and two. When comparing to previous studies, the range of early BQI values for $N \geq 1$ seems to be lower here (0.22) than the range reported by Nowak and Klockgether (2017; 0.5), which may be explained by the fact that the direct sound path was removed here. Finally, RVL obtained the highest BQI values overall for both rooms, predicting a higher spatial quality that, again, was unexpected given the lower complexity of the method when compared to higher order hybrid Ambisonics.

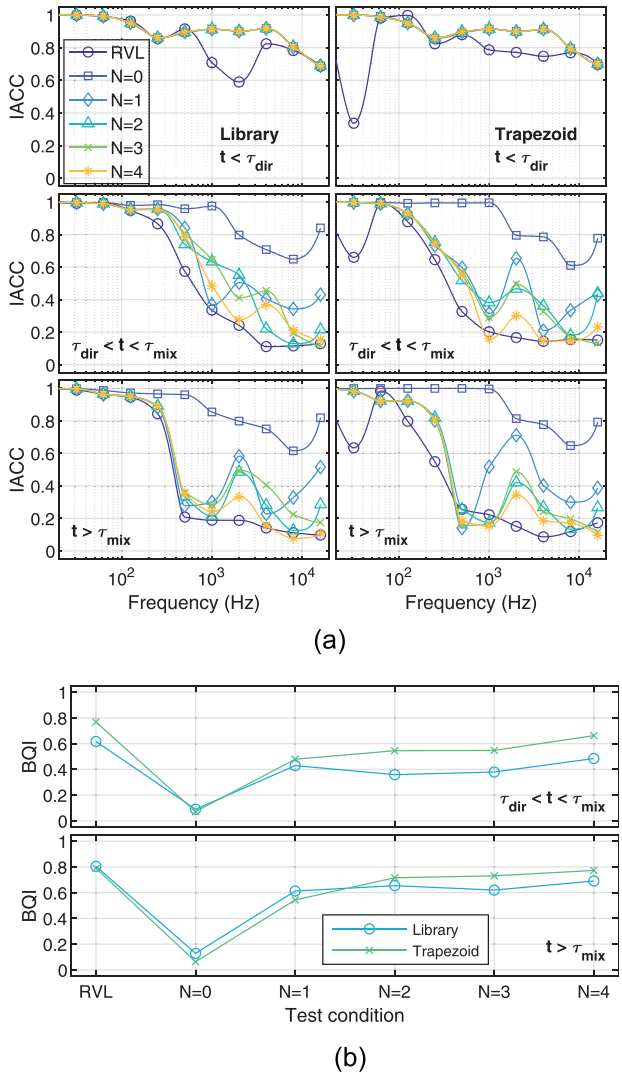


FIG. 5. (Color online) Objective binaural metrics of left-ear BRIRs at $(\varphi = 30°, \theta = 0°)$ for different rooms and test conditions. (a) The IACC per octave band, separated per BRIR segment. (b) The binaural quality index (BQI) for early reflections ($\tau_{dir} < t < \tau_{mix}$) and late reverb ($t > \tau_{mix}$).

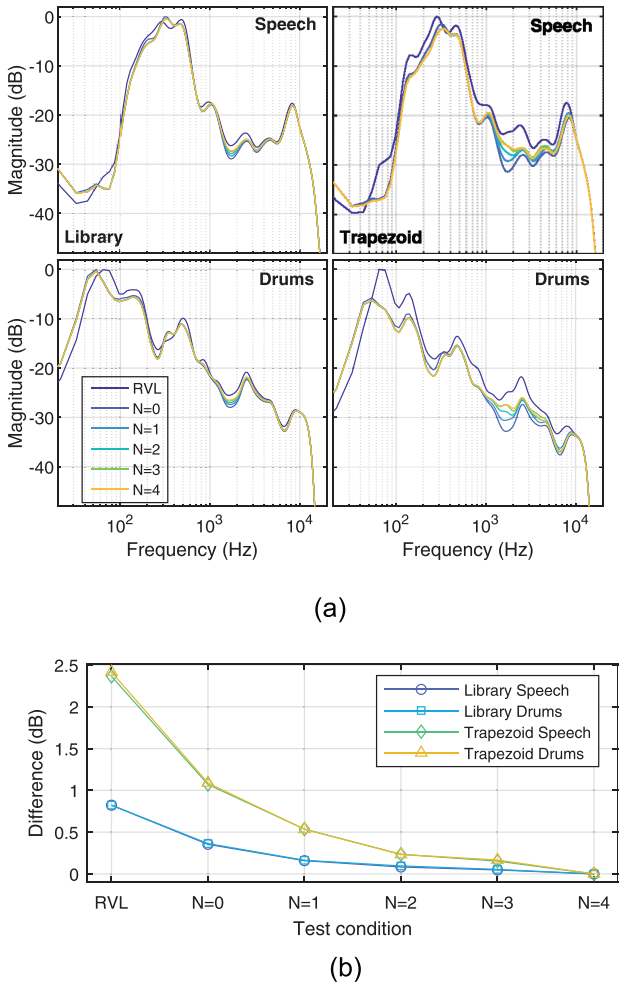


FIG. 6. (Color online) (a) Long-term average spectra of two test signals (speech and drums) convolved with left-ear BRIRs, corresponding to the different test conditions, for a source at $(\varphi = 30°, \theta = 0°)$. (b) The absolute difference between each spectrum and the reference ($N = 4$), averaged across 42 equivalent rectangular bandwidths (ERBs).

C. Spectral analysis

Next, spectral differences across the synthesised BRIRs were explored by convolving them with test signals (speech audio material and drum kit track from music audio material) and analysing the long-term averaged spectra of the results. Spectra were calculated as the average power spectral density obtained from a series of overlapping 4096-sample discrete Fourier transforms (DFTs) after applying 1/3-octave Gaussian smoothing and are shown in Fig. 6(a). The absolute deviation of each condition from a reference averaged across 42 equivalent rectangular bandwidths (ERBs) is shown in Fig. 6(b). The results were similar for the left and right channels of the BRIRs, therefore, only the former are presented for brevity. The condition $N = 4$ was chosen as the reference because it has the highest available spatial order. For the hybrid Ambisonics conditions, it can be seen how the differences are largest for $N = 0$ and decrease for higher orders as expected. In the case of RVL, deviations are clearly larger in the trapezoid than in the library, which may be related to its limitations in accuracy when rendering early reflections. Overall, the range of spectral differences was observed to be larger for the trapezoid (up to 2.4 dB) than it was for the library (up to 0.8 dB), which might be explained by the lower DRR of the trapezoid. The type of audio material did not seem to have a noticeable effect on the results.

D. Loudness stability

As mentioned in Sec. II, one of the limitations of RVL is the way it approximates head rotations by having the room rotate with the listener's head, which may have perceptual implications. On the one hand, if a strong reflection is perceived as coming from the wrong direction, it may lead to decreased externalisation, which is generally not desired. On the other hand, the loudness of the auditory scene may change more smoothly across head orientations, which might be perceived as preferable. This is particularly relevant for low-order Ambisonics renderings, which suffer

https://doi.org/10.1121/10.0003437

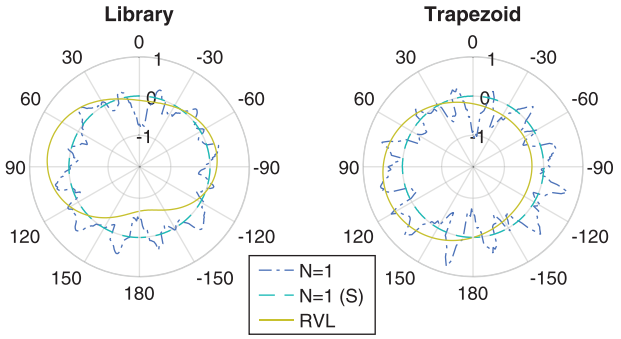


FIG. 7. (Color online) Loudness, K-weighted, relative to full scale (in dB) of the reverberant sound field, generated by convolving pulsated pink noise with a BRIR that had the direct sound removed, for different listener head orientations over the horizontal plane. Each line represents a different condition of experiment 2: first-order hybrid Ambisonics, first-order hybrid Ambisonics (static) and RVL.

from poor loudness stability across head orientations (Ben-Hur *et al.*, 2019). Figure 7 shows the predicted loudness (ITU-R, 2015b) of the reverberant sound field in first-order hybrid Ambisonics and RVL renderings. Additionally, a "static" version of the first-order hybrid Ambisonics rendering for which reverb was not updated with head rotations was also evaluated. Loudness was estimated after convolving pulsated pink noise with BRIRs of different directions across the horizontal plane. It can be seen how the static condition (indicated with an "S") shows constant loudness across the different azimuth angles, whereas the dynamic condition is less smooth. RVL, which renders reverb in a semi-static way in which the room is head-locked, falls somewhere in the middle of the other two as it is smoother than the dynamic rendering but not constant like the static rendering.

E. Summary of numerical analyses

An objective evaluation of the reverb-rendering methods being tested has been presented. First, an overview of the RIR characteristics for each room was given through descriptive analysis. Then, different metrics of the synthesised BRIRs were analysed to try and predict their perceived quality. It was observed that the BQI seemed to saturate at an earlier spatial order in the library (it did not vary much for $N \geq 1$) than it did in the trapezoid (it increased monotonically up to $N = 4$). Similarly, spectral differences with respect to the reference ($N = 4$) decreased as the spatial order increased as expected but were smaller for the library (all conditions under 1 dB) than they were for the trapezoid (under 1 dB only for $N \geq 1$) and not much affected by the type of audio material. These results suggest that the room characteristics will influence the minimum spatial order needed to achieve a certain subjective quality. Once paired with a perceptual evaluation, these observations may help to identify which objective metrics are more useful to predict the perceived quality of the binaural reverb.

RVL was found to be more challenging to evaluate objectively against the other conditions because of its different nature, i.e., its renderings were based on measured BRIRs, whereas the hybrid Ambisonics renderings were built from Eigenmike measurements—which led to larger spectral deviations from the reference than other conditions. According to the BQI data, BRIRs generated with this method were predicted to have a higher spatial quality than for even the highest order hybrid Ambisonics conditions discussed above, which was an unexpected result. Also, it was observed that one of its potential limitations, namely, the way in which head rotations are implemented, caused the rendered sound scene to have smoother loudness variations across different head orientations than did first-order hybrid Ambisonics. However, it is still not clear how this will impact the perceived quality, which should, therefore, be assessed through a perceptual evaluation.

V. PERCEPTUAL EVALUATION

The methods under study were perceptually evaluated through two separate experiments in which a total of 32 listeners participated voluntarily. The mean listener age was 32 years old [standard deviation (SD) = 9.6 yr]. Of the 32 listeners, 24 declared to have previous experience in similar listening tests, 31 declared to have no hearing impairments (the remaining one was excluded from both experiments in post-screening), 30 declared to have prior knowledge of highly realistic or binaural audio reproduction, and 18 declared to possess advanced musical knowledge or have received formal musical education. Listeners were split in two groups: the first one (21 listeners) performed the experiment in an acoustically dead laboratory environment (lab), whereas the second one (11 listeners) did so in situ in the actual measured rooms (library and trapezoid). The reason for this split was to evaluate the effect of "room divergence" or how the listener's exposure to the actual room acoustics affects their perception of a virtual rendering of the same room (Werner *et al.*, 2016).

A. Experiment 1: Paradigm

In the first experiment, listeners were asked to rate the quality of sound scenes rendered through hybrid Ambisonics at different spatial orders ($0 \leq N \leq 4$) with the direct sound being rendered through HRIR convolution identically for all conditions as explained in Sec. III. A double-blind listening test paradigm was used, based on the MUSHRA (multiple stimulus test with hidden reference and anchor) format (ITU-R, 2015a). The highest order rendering ($N = 4$) was used as the reference and a dry rendering (without reverb) was used as a low-quality anchor. Listeners were asked to rate the similarity of each stimulus to the reference on a scale from 0 to 100, where the latter meant "identical to the reference." The user interface was implemented in Max 7 (Cycling '74, Walnut, CA, United States). Listeners were encouraged to use head movements to explore the scene. Each listener completed one trial per combination of room (library or trapezoid) and type of audio material for a total of four trials. Post-screening was applied to exclude ratings of unreliable listeners from the data analysis, i.e., those who

rated the hidden reference lower than 90 points or the anchor higher than 50 points for more than 25% of the trials.

B. Experiment 1: Results

Results of the first experiment are shown in Fig. 8. Data for 11 listeners (5 *in situ*, 6 laboratory) were excluded in post-screening. Listeners took an average time of 186.44 s (SD = 141.48 s) to complete each MUSHRA trial. The displayed data were normalised so that the highest rating of every trial is set to 100 and the lowest rating is set to 0. Note that this was done for the sake of visualisation and all inferential analysis was performed on non-normalised data. Descriptive analysis shows that the ratings were generally higher for higher spatial orders. Dry renderings consistently obtained the lowest ratings, followed by zeroth- and first-order renderings. Mean and median ratings seem to be similar across the higher order conditions, placed close to the top of the rating scale.

Inferential analysis was performed through a repeated measures analysis of variance (RM-ANOVA). The main dependent variable was the reverb spatial order, but its interactions with other variables, such as room, type of audio material, and test location (*in situ* vs laboratory), were investigated as well. Following the MUSHRA recommendation (ITU-R, 2015a), the Huynh-Feldt correction was applied to reduce type I errors as the data did not pass the Mauchly sphericity test ($p < 0.001$) and showed a Greenhouse-Geisser epsilon higher than 0.75 ($\varepsilon = 0.81$). A significance value of $\alpha = 0.05$ was used.

(1) *Effect of spatial order*: The RM-ANOVA found a significant effect of spatial order on listeners' ratings [$F(5, 380) = 747.17$, $p < 0.001$]. *Post hoc* multiple dependent sample *t*-tests were run using a corrected significance level of $\alpha' = 0.0033$. Significant differences were found between all pairs of conditions [$t(83) \leq -3.65$, $p < 0.001$] except between the third- and fourth-order conditions [$t(83) = -1.67$, $p = 0.098$].

(2) *Effect of room*: A significant effect was found for the interaction between room and spatial order [$F(5, 380) = 2.92$, $p = 0.019$]. *Post hoc* multiple dependent samples *t*-tests were run on data separated by rooms using a corrected significance level of $\alpha' = 0.0033$. For the library, differences between the second- and fourth-order [$t(41) = -2.80$, $p = 0.008$] conditions and between the third- and fourth-order [$t(41) = 0.20$, $p = 0.846$] conditions were not significant, whereas significant differences were found for all other pairs of conditions [$t(41) \leq -3.34$, $p \leq 0.002$]. The fact that a significant difference was found between the second- and third-order but not between the second- and fourth-order may seem surprising at first. However, it is worth mentioning that the latter was very close to being significant ($0.008 \not\gg \alpha'$). Also, this result can be explained by the fact that the RM-ANOVA is a parametric analysis which relies on comparisons between means and the fourth-order data presented some outliers which slightly lowered the mean rating, bringing it closer to that of the second-order (cf. Fig. 8, top-middle plot). For the trapezoid, on the other hand, differences between the second- and third-order [$t(41) = -1.67$, $p = 0.103$] conditions and between the third- and fourth-order ($t(41) = -2.39$, $p = 0.021$) conditions were not significant, whereas all other pairs of conditions showed significant differences [$t(41) \leq -4.15$, $p < 0.001$].

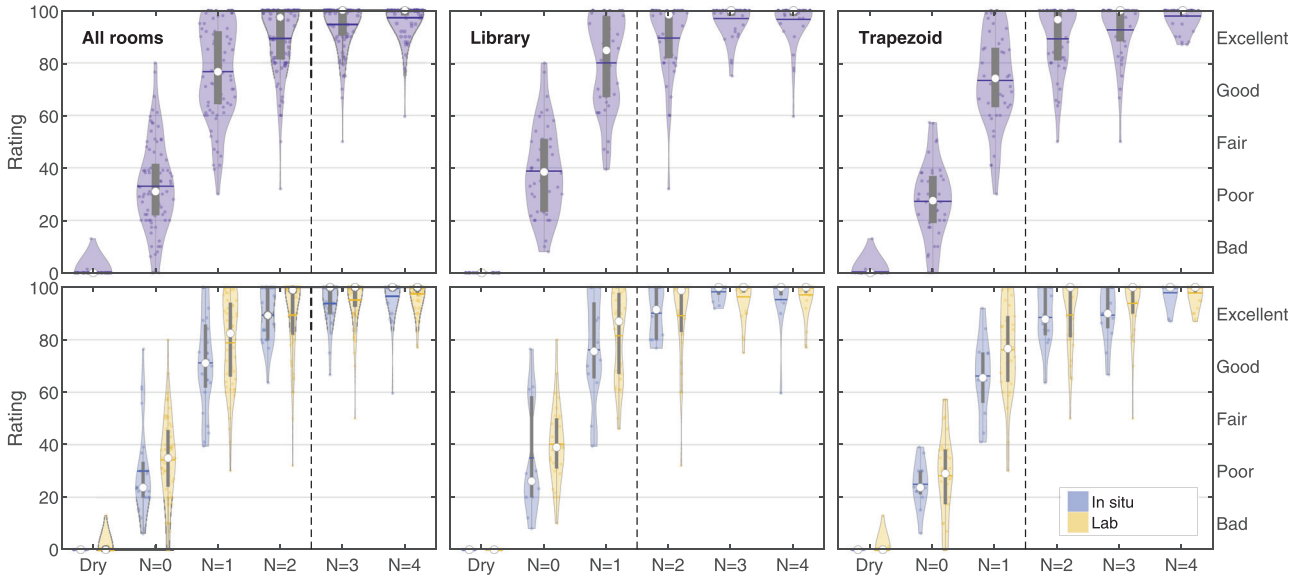


FIG. 8. (Color online) Results from experiment 1 represented by violin plots (Hintze and Nelson, 1998), which show the probability density of the data, median (circle), interquartile range (box), and mean (horizontal line). (Top) Both test locations pooled together. (Bottom) Separated per test location (from left to right, "*in situ*" and "lab"). From left to right, both rooms pooled together, library, and trapezoid are shown. For this visualisation, data were scaled on a per-trial basis by setting the lowest rating of every MUSHRA (multiple stimulus test with hidden reference and anchor) trial to 0 and setting the highest rating of every MUSHRA trial to 100. The vertical dotted lines indicate that the groups on the left are significantly different ($p < 0.05$) from the groups on the right (before normalisation).

https://doi.org/10.1121/10.0003437

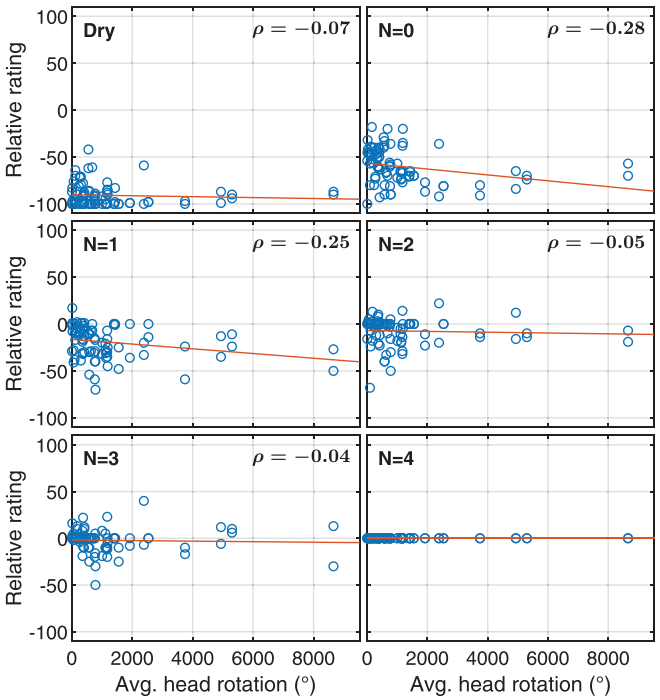


FIG. 9. (Color online) Scatterplots showing all relative ratings (absolute ratings minus reference ratings) from experiment 1 and the amount of head rotation (azimuth) that listeners performed for each rating. A linear fit of the data is displayed in red, and the Pearson correlation coefficient is indicated at the top. Data for both rooms were pooled together.

(3) *Other interactions*: No significant interactions were found between the spatial order and test location [$F(5, 380) = 1.28$, $p = 0.275$], type of audio material [$F(5, 380) = 1.94$, $p = 0.099$], or any of the three-way interactions ($p > 0.6$).

Listeners performed an average absolute head rotation (in azimuth) of 1202° (SD = 1672°) or 6.68 half-circle rotations per MUSHRA trial. Due to the high variance in the amount of head movement across subjects, the potential effect of head rotations on MUSHRA ratings was explored. Figure 9 shows the MUSHRA relative ratings, i.e., deviations from the reference's rating, plotted against the average head rotation per trial, calculated from the total head movement during the full test for each listener. For each test condition, the Pearson correlation coefficient (ρ) between the average head rotation and the relative ratings was calculated. By inspecting Fig. 9, it is evident that correlation was small ($|\rho| < 0.1$) for all conditions except the zeroth- and first-order conditions, which show negative correlation values, indicating that they were rated lower by listeners who employed more head movements. The low correlation on the dry condition indicates that it produced low ratings regardless of the amount of head movement, which was expected as it is the anchor condition. Meanwhile, the low correlations for orders greater than one suggest that listeners did not perceive them as more similar or different to the reference by performing additional head movements. Note that similar trends were observed when separating data per room (library/trapezoid) and test location (*in situ*/laboratory), but the analysis is not reported here due to space constraints.

C. Experiment 2: Paradigm

The second experiment aimed to compare the proposed computationally efficient RVL to a more accurate rendering approach in terms of subjective preference. Because the main limitation of RVL renderings is the way head rotations are implemented as the room "rotates" with the listener's head, this experiment focused on that aspect to evaluate RVL's performance in an adverse scenario. Thus, three different rendering methods were evaluated.

(1) $N = 1$: head-tracked first-order hybrid Ambisonics, i.e., identical to condition $N = 1$ from experiment 1,
(2) $N = 1(S)$: hybrid Ambisonics where the direct sound path was head-tracked but the reverb was not, i.e., the reverb did not change according to head movements, and
(3) *RVL*: head-tracked first-order RVL.

Therefore, RVL was compared to an approach which implemented head rotations properly. Additionally, a static method $N = 1$ (S) was introduced as an anchor condition where head movements did not influence the incoming direction of reverb and, therefore, simulated head rotations with a lower accuracy than for RVL. As in the previous experiment, the direct sound path was rendered through HRIR convolution and was head-tracked for all conditions.

Because the methods under comparison were generated from different measurements (Eigenmike RIRs for hybrid Ambisonics and KEMAR BRIRs for RVL), there existed significant spectral differences between the renderings, as shown in Sec. IV C, that were not trivial to compensate through equalisation. Preliminary tests with discriminability and MUSHRA tasks showed that listeners focused mainly on these spectral differences rather than on other attributes such as reverb directionality. Therefore, a preference task was performed instead where no reference was provided and listeners evaluated whether the renderings met their internal expectations after seeing a picture of the room or from their own experience, i.e., for those who conducted the test *in situ*.

A double-blind pairwise comparison listening test paradigm was used. In each trial, listeners were shown a picture of the rendered room (library or trapezoid) and a diagram of the sound scene and were presented two stimuli (A and B). These stimuli were two binaural renderings of the same room, generated with two different methods out of the possible three (except in null pairs, where A and B were identical). Listeners were asked the question, "*Considering the given room, which example is more appropriate?*". To answer, they would use a continuous rating scale from −2 to +2 (with one decimal place) from *definitely A* to *definitely B*. Listeners could freely switch between the synchronised stimuli during a trial, and head movements were encouraged to explore the scene. As in experiment 1, the user interface was implemented in Max 7. For each room and type of

audio material, listeners evaluated all possible pairs of conditions plus two null pairs, where *A* and *B* were identical (randomly chosen), totalling 16 paired comparisons. Because no reference was provided, there were no correct or wrong answers except for the null pairs, which a reliable listener should rate as zero as no audible differences existed in those. Post-screening was applied to exclude listeners who rated the null pairs with an average absolute value higher than 0.25.

Note that the $N = 1$ (S) and RVL conditions were not included in the MUSHRA test (experiment 1) to keep the unidimensionality across test conditions, i.e., spatial order. The reason for using first-order hybrid Ambisonics was to have a fair comparison in the sense that all conditions used the same number of virtual loudspeakers (six).

D. Experiment 2: Results

Results of the second experiment are shown in Fig. 10. Data for eight listeners (four *in situ*, four laboratory) were excluded in post-screening, and six of those were also screened out from experiment 1. Listeners took an average time of 41.33 s (SD = 18.55 s) to complete each paired comparison. Descriptive analysis shows that the mean rating was close to zero for the null pairs, the preference between RVL and hybrid Ambisonics reverb changed depending on the room and type of audio material, and static first-order hybrid Ambisonics renderings were perceived as very similar to the dynamic renderings with a slight trend toward favouring the former. The fact that RVL was not systematically rated lower than hybrid Ambisonics suggested that the simplifications in the RVL rendering do not significantly impair subjective preference.

The inferential analysis tried to find whether listeners had a significant preference on each paired comparison and whether this was affected by factors such as room, type of audio material, and test location. To that end, a RM-ANOVA was first run on the data of each paired comparison to study the effect of the different variables and their interactions. Then, data were grouped accordingly and *t*-tests were run to evaluate whether the samples deviated from a normal distribution with a mean equal to zero. The Mauchly sphericity test was passed by the $N = 1$/RVL and $N = 1(S)$/RVL pairs ($p > 0.05$) but not by the $N = 1/N = 1(S)$ or the null pair for which the Greenhouse-Geisser correction was applied ($\varepsilon < 0.75$). As in the previous experiment, a significance value of $\alpha = 0.05$ was used.

(1) *Null pair*: No significant effect of any variable was found, thus, data were grouped for all types of audio material, rooms, and test locations. A one-sample *t*-test showed that the mean rating was not significantly different from zero [$t(95) = 1.69, p = 0.094$].
(2) *N = 1/N = 1(S)*: A significant effect of the room was found [$F(1, 22) = 4, 38, p = 0.048$], therefore, the data were separated by rooms. *t*-tests showed that the trapezoid mean rating was significantly different from zero [$t(47) = 2.57, p = 0.013$] with the $N = 1(S)$ renderings being preferred. In the case of the library, the mean rating was not significantly different from zero [$t(47) = -1.15, p = 0.254$].
(3) *N = 1/RVL*: Significant effects of the room [$F(1, 22) = 39.18, p < 0.001$] and the interaction of room and location [$F(1, 22) = 11.95, p = 0.002$] were found, thus, data were separated by room and location (*in situ*/laboratory). *t*-tests showed that conditions library–*in situ* [$t(13) = -5.30, p < 0.001$], library– laboratory

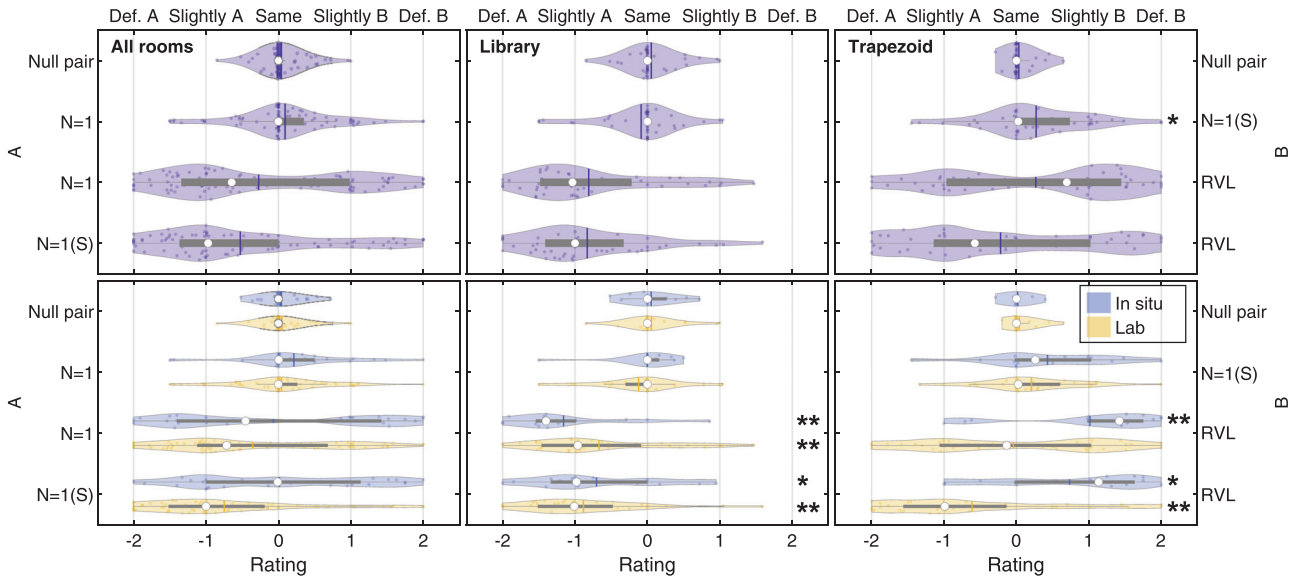


FIG. 10. (Color online) Results from experiment 2 represented by violin plots (Hintze and Nelson, 1998), which show the probability density of the data, median (circle), interquartile range (box), and mean (vertical line). (Top) Both test locations pooled together. (Bottom) Separated per test location (from top to bottom, "*in situ*" and "lab"). From left to right, both rooms pooled together, library, and trapezoid. Ratings range from "definitely prefer A" (−2) to "definitely prefer B" (2). Asterisks indicate whether the mean is significantly different from zero: * for $p < 0.05$ and ** for $p \leq 0.005$.

https://doi.org/10.1121/10.0003437

[$t(33) = -4.26$, $p < 0.001$], and trapezoid–*in situ* [$t(13) = 3.69$, $p = 0.003$] all had mean ratings significantly different from zero. $N = 1$ renderings were preferred for the two library conditions, whereas RVL was preferred for the trapezoid–*in situ* condition.

(4) $N = 1(S)/RVL$: Significant effects of the room [$F(1, 22) = 10.62$, $p = 0.004$], location [$F(1, 22) = 7.93$, $p = 0.010$], and room/location interaction [$F(1, 22) = 4.93$, $p = 0.037$] were found, therefore, data were again separated by room and location. *t*-tests showed that all room–location combinations had mean ratings significantly different from zero: library–*in situ* [$t(13) = -2.89$, $p = 0.013$], library–laboratory [$t(33) = -6.21$, $p < 0.001$], trapezoid–*in situ* [$t(13) = 2.61$, $p = 0.022$], and trapezoid–laboratory [$t(33) = -2.98$, $p = 0.005$]. $N = 1$(S) renderings were preferred for both of the library conditions and trapezoid–laboratory, whereas RVL was preferred for the trapezoid–*in situ* condition.

Listeners performed an average absolute head rotation (in azimuth) of 449° (SD = 373°) or 2.49 half-circle rotations per paired comparison. This indicates that, as instructed, they employed head movements to inform their ratings. However, given the relative high variance of the head tracking data across subjects, the potential effect of head rotations on the paired comparison ratings was investigated. Figure 11 shows the relation between the ratings and the average head rotation per trial, calculated from the total head movement during the full test for each listener. The Pearson correlation coefficient (ρ) between the average head rotation and the ratings was also calculated and indicated in Fig. 11. Inspection of these data shows that head rotations had a near-zero correlation with the ratings for the $N = 1/N = 1(S)$ pair. For the other two pairs, larger correlations between head rotations and ratings were reported in the library than were reported in the trapezoid, particularly for the $N = 1$/RVL pair in which a correlation of $\rho = -0.44$ was observed. Correlation values were generally small ($|\rho| < 0.1$) for all trapezoid data, which might suggest that listeners' ratings in this room did not importantly change when additional head movements were performed. Note that similar trends were observed when separating data per test location (*in situ*/laboratory), but the analysis is not reported here due to space constraints.

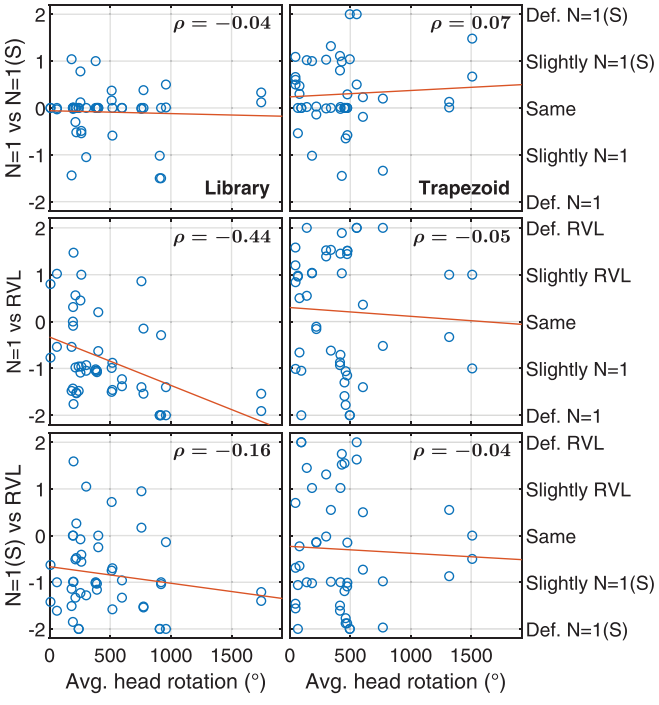


FIG. 11. (Color online) Scatterplots showing all paired comparisons' ratings (except null pairs) from experiment 2 and the amount of head rotation (azimuth) that listeners performed for each rating. A linear fit of the data is displayed in red, and the Pearson correlation coefficient is indicated at the top. Data were separated per room (library on the left and trapezoid on the right).

VI. DISCUSSION

The main goal of this study was to investigate how much spatial reverb rendering can be simplified without degrading perceived quality, assuming that the direct sound path is rendered as accurately as possible. First, the effect of reducing the spatial order of hybrid Ambisonics was investigated. Because the direct sound carries essential information for source localisation and, in some cases, contains most of the energy of the RIR, it was expected that excluding it from the spatial order reduction process would mitigate perceptual degradation. This would imply that provided the direct sound path is rendered separately and with sufficient accuracy, the perceptual impact of reducing the spatial order may be lower than reported in previous studies which used standard Ambisonics (Ahrens and Andersson, 2019; Bernschütz, 2016). Second, the effect of simplifying the implementation of head rotations in dynamic reverb rendering was investigated by comparing a first-order hybrid Ambisonics rendering against a computationally efficient alternative (RVL), which implemented head rotations in a simplified way, and against a static version where reverb was not head-tracked at all.

A. Effect of spatial order

Evaluation of objective metrics, such as the IACC and BQI, predicted a large improvement in the spatial quality when increasing the order from zero to one, but the differences became smaller as the order increased. In fact, early BQI ratings measured in this study did not vary as much across the higher spatial orders as those found by Nowak and Klockgether (2017). This may be explained by the fact that removing the most directional part of the RIR, i.e., the direct sound path, led to a more diffuse sound field, which could be rendered accurately with a lower spatial order. This would also explain why the room with less directional reverb, i.e., library, produced a lower variance in the BQI values. Similarly, it was observed that spectral differences between each BRIR and the reference (order four) increased as the spatial order decreased with the largest jump happening between orders zero and one and with said differences

being under 1 dB for all $N \geq 1$. Also, spectral differences were larger for the trapezoid, which may be explained by its lower DRR or its more salient early reflections having a larger impact on the signal spectrum, e.g., due to the comb-filtering effects (Bech, 1996).

Results of experiment 1 showed that perceived differences were large between orders zero and one and smaller for higher orders, which was in line with the numerical analysis, and a room dependence effect was observed. Data from third-order renderings are particularly representative in that aspect: although their ratings were not significantly different from the hidden reference for either room, it seems that listeners rated them consistently lower in the trapezoid than they did in the library relative to the reference. In fact, data suggest that in the latter room, third- and fourth-order renderings obtained almost identical ratings. This would agree with the numerical analyses in that the trapezoid, with its lower DRR and less diffuse reverberant sound field, displays larger differences between spatial orders than does the library. It is also worth noting that the differences in rating among orders equal or higher than two were observed to be similar across listeners regardless of the amount of head movements that they employed to explore the auditory scene.

According to Avni *et al.* (2013), spaciousness and timbre are the most relevant perceptual attributes that listeners use when evaluating sound fields of varying spatial resolution. The present results suggest that when the direct sound path is rendered accurately, the degradation in both spatial and spectral qualities becomes perceptually less relevant, particularly for more diffuse reverberant sound fields, i.e., large rooms. For the conditions tested here, it was shown that the perceived quality of binaural renderings did not improve beyond an order between two and three. This is notably lower than the eighth-order suggested by Ahrens and Andersson (2019), who included the direct sound path in the Ambisonics rendering—saving the differences in experimental paradigm, which was an *A-B* comparison with attribute scaling (timbre and spaciousness) rather than a MUSHRA test with a single global attribute. However, spatial orders higher than four, not included here due to limitations of the measurement equipment, should be evaluated to draw a more complete comparison to previous studies. Regardless of this limitation, the present results are in line with the findings of Lübeck *et al.* (2020), who showed that reverb may be rendered through BRIRs sampled on a spherical grid of a spatial order as low as three without degrading perceived quality. At the sight of this, future work could investigate the effect of spatial resolution on each RIR segment separately, i.e., early reflections and late reverb, and how this may depend on the auralised room in a similar fashion to the work by Lübeck *et al.* (2020) but applied to order-truncated Ambisonics signals instead of spatially subsampled signals. Outcomes could be used to inform perceptual models to evaluate spatial audio quality and enable efficient parametric reverb rendering, e.g., by determining the amount of early reflections needed to generate plausible virtual scenes (Brinkmann *et al.*, 2020).

B. Dynamic aspects

Results of experiment 2, which compared dynamic and static first-order hybrid Ambisonics reverb with RVL, were more challenging to interpret. The absence of a reference led to bimodal data distributions in some cases (cf. trapezoid data in Fig. 10), meaning that listeners could discriminate pairs of conditions, but neither was unanimously preferred. An unexpected outcome was that the static version of hybrid Ambisonics [$N = 1$ (S)], which was initially conceived as a low-quality anchor, was found to be preferred, in some cases, to the more accurate dynamic version ($N = 1$). In particular, this was true for the trapezoid but not for the library. *Post hoc* informal interviews suggested that this could be due to the dynamic reverb being perceived as less "stable" when head rotations were performed as pointed out in Sec. IV D. This might be explained by the fact that virtual loudspeaker decoding approaches yield angle dependent spectral distortions (Solvang, 2008), which often result in poor loudness stability at low orders (Ben-Hur *et al.*, 2019). This is supported by the fact that BRIR-predicted loudness is less smooth in the trapezoid than it is in the library as observed in Fig. 7. Analysis of head tracking data suggested that even listeners who performed a more exhaustive exploration of the scene through head movements did not rate the dynamic version significantly higher than the static version. This result suggests that, provided direct sound is rendered dynamically through convolution with an HRIR, it may actually be detrimental to render reverb dynamically if a low Ambisonics order is used and the loudness stability is not accounted for.

When comparing RVL and hybrid Ambisonics reverb, the spectral analysis showed that strong colouration differences should be expected, which may have led to polarised ratings in the perceptual evaluation. In the case of the trapezoid, RVL was clearly preferred by listeners that performed the test *in situ*, which suggests that this method captured the room characteristics more accurately than did hybrid Ambisonics. However, the opposite was true for the library, where listeners assigned lower ratings to RVL (even more so when exhaustive head movements were employed during the test), the reasons for which are yet unclear. In any case, the room divergence effect seemed to play a more important role in the second test than it did in the first test because a reference was not provided and listeners provided ratings based on their expectations, which depended on previous exposure to the rendered rooms.

VII. CONCLUSIONS

This study addressed the issue of the trade-off between computational complexity and perceived quality for binaural Ambisonics-based reverb. It introduced the concept of hybrid Ambisonics or the separation between the direct sound path and the reverb in Ambisonics-based binaural rendering, obtaining the former by convolution with a dense HRIR dataset and encoding the latter in an Ambisonics sound field. It was hypothesised that the perceived quality of

the renderings would stop improving at a lower spatial order than in previous studies where the direct sound was not processed separately (Ahrens and Andersson, 2019; Bernschütz, 2016) as the directional information of the signal would be better preserved. Results from the perceptual evaluation suggest that when the direct sound path is computed accurately, an Ambisonics order of two or three may be enough to render binaural reverb, depending on the room characteristics. For instance, rooms with lower DRR or more salient early reflections are likely to require a higher spatial order than are rooms where reverb is more diffuse. In any case, the scope of this study was limited to two measured rooms and spatial orders up to four and, therefore, further evaluations on different spaces and with higher orders are needed to generalise these results. Additionally, future work could look into rendering the most relevant early reflections (Brinkmann *et al.*, 2020) at a higher spatial order, which could lower the spatial resolution requirements for the diffuse reverb.

Additionally, RVL was introduced as a computationally efficient approach to dynamically render binaural reverb for a large number of sources. It was observed that renderings produced with this method were comparable to (and, in some cases, better than) those obtained through less flexible Ambisonics-based approaches in terms of subjective preference. Considering the advantages of RVL, namely, its efficiency and ease of implementation, this method should be worthy of consideration for convolution-based binaural reverb generation in low-cost scenarios.

ACKNOWLEDGMENTS

This work was partly supported by the PLUGGY project,[1] European Union's Horizon 2020 research and innovation programme under Grant No. 726765.

[1]See www.pluggy-project.eu (Last viewed January 28, 2021).

Ahrens, J., and Andersson, C. (**2019**). "Perceptual evaluation of headphone auralization of rooms captured with spherical microphone arrays with respect to spaciousness and timbre," J. Acoust. Soc. Am. **145**(4), 2783–2794.
Allen, J. B., and Berkley, D. A. (**1979**). "Image method for efficiently simulating small-room acoustics," J. Acoust. Soc. Am. **65**(4), 943–950.
Armstrong, C., Thresh, L., Murphy, D., and Kearney, G. (**2018**). "A perceptual evaluation of individual and non-individual HRTFs: A case study of the SADIE II database," Appl. Sci. **8**(11), 2029.
Avni, A., Ahrens, J., Geier, M., Spors, S., Wierstorf, H., and Rafaely, B. (**2013**). "Spatial perception of sound fields recorded by spherical microphone arrays with varying spatial resolution," J. Acoust. Soc. Am. **133**(5), 2711–2721.
Barron, M., and Marshall, A. H. (**1981**). "Spatial impression due to early lateral reflections in concert halls: The derivation of a physical measure," J. Sound Vib. **77**(2), 211–232.
Bech, S. (**1996**). "Timbral aspects of reproduced sound in small rooms. II," J. Acoust. Soc. Am. **99**(6), 3539–3549.
Begault, D. R., Wenzel, E. M., and Anderson, M. R. (**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," J. Audio Eng. Soc. **49**(10), 904–916.
Ben-Hur, Z., Alon, D. L., Rafaely, B., and Mehra, R. (**2019**). "Loudness stability of binaural sound with spherical harmonic representation of sparse head-related transfer functions," EURASIP J. Audio, Speech, Music Process. **2019**(1), 5.
Beranek, L. L. (**2008**). "Concert hall acoustics—2008," J. Audio Eng. Soc. **56**(7/8), 532–544.
Bernschütz, B. (**2016**). "Microphone arrays and sound field decomposition for dynamic binaural recording," Doctoral thesis, Technische Universität Berlin, Berlin.
Bernschütz, B., Giner, A. V., Pörschmann, C., and Arend, J. (**2014**). "Binaural reproduction of plane waves with reduced modal order," Acta Acust. Acust. **100**(5), 972–983.
Botteldooren, D. (**1995**). "Finite-difference time-domain simulation of low-frequency room acoustic problems," J. Acoust. Soc. Am. **98**(6), 3302–3308.
Brinkmann, F., Gamper, H., and Tashev, N. R. a. I. (**2020**). "Towards encoding perceptually salient early reflections for parametric spatial audio rendering," in *Audio Engineering Society Convention 148* (Audio Engineering Society, New York), available at aes.org/e-lib/browse.cfm?elib=20797 (Last viewed January 28, 2021).
Brown, A. D., Stecker, G. C., and Tollin, D. J. (**2015**). "The precedence effect in sound localization," JARO: J. Assoc. Res. Otolaryngol. **16**(1), 1–28.
Cabrera, D., Lee, D., Yadav, M., and Martens, W. L. (**2011**). "Decay envelope manipulation of room impulse responses: Techniques for auralization and sonification," in *Acoustics 2011*, Gold Coast, Australia, p. 5, available at acoustics.asn.au/conference_proceedings/AAS2011/papers/p70.pdf (Last viewed January 28, 2021).
Cuevas-Rodríguez, M., Picinali, L., González-Toledo, D., Garre, C., de la Rubia-Cuestas, E., Molina-Tanco, L., and Reyes-Lecuona, A. (**2019**). "3D Tune-In Toolkit: An open-source library for real-time binaural spatialisation," PLoS One **14**(3), e0211899.
Engel, I., Henry, C., Amengual Garí, S. V., Robinson, P. W., Poirier-Quinot, D., and Picinali, L. (**2019**). "Perceptual comparison of ambisonics-based reverberation methods in binaural listening," in *EAA Spatial Audio Signal Processing Symposium*, Paris, France, pp. 121–126.
Farina, A. (**2007**). "Advancements in Impulse Response Measurements by Sine Sweeps," in *Audio Engineering Society Convention 122*, Audio Engineering Society, available at aes.org/e-lib/browse.cfm?elib=14106 (Last viewed January 28, 2021).
Gerzon, M. A. (**1985**). "Ambisonics in Multichannel Broadcasting and Video," J. Audio Eng. Soc. **33**(11), 859–871.
Hansen, V., and Munch, G. (**1991**). "Making Recordings for Simulation Tests in the Archimedes Project," J. Audio Eng. Soc. **39**(10), 768–774.
Hintze, J. L., and Nelson, R. D. (**1998**). "Violin Plots: A Box Plot-Density Trace Synergism," Am. Statistician **52**(2), 181–184.
ITU-R. (**2015a**). "BS.1534: Method for the subjective assessment of intermediate quality levels of coding systems," available at itu.int/rec/R-REC-BS.1534 (Last viewed January 28, 2021).
ITU-R. (**2015b**). "BS.1770: Algorithms to measure audio programme loudness and true-peak audio level," available at itu.int/rec/R-REC-BS.1770 (Last viewed January 28, 2021).
Jot, J.-M. (**1997**). "Efficient models for reverberation and distance rendering in computer music and virtual audio reality," in *ICMC: International Computer Music Conference*, Thessaloniki, Greece, pp. 236–243, available at hal.archives-ouvertes.fr/hal-01106168 (Last viewed January 28, 2021).
Jot, J.-M., and Chaigne, A. (**1991**). "Digital delay networks for designing artificial reverberators," in *Audio Engineering Society Convention 90* (Audio Engineering Society, New York), available at aes.org/e-lib/browse.cfm?elib=5663 (Last viewed January 28, 2021).
Kaplanis, N., Bech, S., Jensen, S. H., and van Waterschoot, T. (**2014**). "Perception of reverberation in small rooms: A literature study," in *Audio Engineering Society Conference: 55th International Conference: Spatial Audio, Audio Engineering Society*, available at aes.org/e-lib/browse.cfm?elib=17348 (Last viewed January 28, 2021).
Lindau, A., Kosanke, L., and Weinzierl, S. (**2010**). "Perceptual evaluation of physical predictors of the mixing time in binaural room impulse responses," in *Audio Engineering Society Convention 128* (Audio Engineering Society, New York), available at aes.org/e-lib/browse.cfm?elib=15386 (Last viewed January 28, 2021).
Lindau, A., Kosanke, L., and Weinzierl, S. (**2012**). "Perceptual evaluation of model- and signal-based predictors of the mixing Time in binaural room impulse responses," J. Audio Eng. Soc. **60**(11), 887–898.

Lindau, A., and Weinzierl, S. (**2012**). "Assessing the plausibility of virtual acoustic environments," Acta Acust. Acust. **98**(5), 804–810.
Litovsky, R. Y., Colburn, H. S., Yost, W. A., and Guzman, S. J. (**1999**). "The precedence effect," J. Acoust. Soc. Am. **106**(4), 1633–1654.
Lübeck, T., Pörschmann, C., and Arend, J. M. (**2020**). "Perception of direct sound, early reflections, andreverberation in auralizations of sparsely measured binaural room impulse responses," in *Audio Engineering Society Conference: 2020 AES International Conference on Audio for Virtual and Augmented Reality* (Audio Engineering Society, New York), available at aes.org/e-lib/browse.cfm?elib=20865 (Last viewed January 28, 2021).
McKeag, A., and McGrath, D. S. (**1996**). "Sound field format to binaural decoder with head tracking," in *Audio Engineering Society Convention 6r* (Audio Engineering Society, New York), available at aes.org/e-lib/browse.cfm?elib=7477 (Last viewed January 28, 2021).
mh acoustics (**2016**). "Eigenbeam datasheet," available at mhacoustics.com/sites/default/files/Eigenbeam\%20Datasheet_R01A.pdf (Last viewed January 28, 2021).
Moore, B. C. J. (**2012**). *An Introduction to the Psychology of Hearing* (Brill, Leiden, Netherlands).
Noisternig, M., Musil, T., Sontacchi, A., and Holdrich, R. (**2003**). "3D binaural sound reproduction using a virtual ambisonic approach," in *IEEE International Symposium on Virtual Environments, Human-Computer Interfaces and Measurement Systems, 2003. VECIMS '03*, pp. 174–178.
Nowak, J., and Klockgether, S. (**2017**). "Perception and prediction of apparent source width and listener envelopment in binaural spherical microphone array auralizations," J. Acoust. Soc. Am. **142**(3), 1634–1645.
Okano, T., Beranek, L. L., and Hidaka, T. (**1998**). "Relations among interaural cross-correlation coefficient (IACCE), lateral fraction (LFE), and apparent source width (ASW) in concert halls," J. Acoust. Soc. Am. **104**(1), 255–265.
Olive, S. E., and Toole, F. E. (**1989**). "The detection of reflections in typical rooms," J. Audio Eng. Soc. **37**(7/8), 539–553.
Oppenheim, A. V., Buck, J. R., and Schafer, R. W. (**2001**). *Discrete-Time Signal Processing* (Prentice Hall, Upper Saddle River, NJ), Vol. 2, available at repository.vnu.edu.vn/handle/VNU_123/34218 (Last viewed January 28, 2021).
Pelzer, S., Aspöck, L., Schröder, D., and Vorländer, M. (**2014**). "Interactive real-time simulation and auralization for modifiable rooms," Build. Acoust. **21**(1), 65–73.
Picinali, L., Wallin, A., Levtov, Y., and Poirier-Quinot, D. (**2017**). "Comparative perceptual evaluation between different methods for implementing reverberation in a binaural context," in *AES Convention 142*, Berlin, Germany, available at hal.archives-ouvertes.fr/hal-01790217 (Last viewed January 28, 2021).
Schissler, C., Mehra, R., and Manocha, D. (**2014**). "High-order diffraction and diffuse reflections for interactive sound propagation in large environments," ACM Trans. Graph. **33**(4), 1–39:12.
Schissler, C., Stirling, P., and Mehra, R. (**2017**). "Efficient construction of the spatial room impulse response," in *2017 IEEE Virtual Reality (VR)*, pp. 122–130.
Schörkhuber, C., Zaunschirm, M., and Höldrich, R. (**2018**). "Binaural rendering of Ambisonic signals via magnitude least squares," in *Daga 2018*, Munich, Germany, pp. 339–342, available at researchgate.net/publication/325080691_Binaural_Rendering_of_Ambisonic_Signals_via_Magnitude_Least_Squares (Last viewed January 28, 2021).
Schroeder, M., and Logan, B. (**1961**). "'Colorless' artificial reverberation," IRE Trans. Audio **AU-9**(6), 209–214.
Solvang, A. (**2008**). "Spectral impairment of two-dimensional higher order Ambisonics," J. Audio Eng. Soc. **56**(4), 267–279.
Välimäki, V., Parker, J., Savioja, L., Smith, J. O., and Abel, J. (**2016**). "More than 50 years of artificial reverberation," in *Audio Engineering Society Conference: 60th International Conference: DREAMS (Dereverberation and Reverberation of Audio, Music, and Speech)* (Audio Engineering Society, New York), available at aes.org/e-lib/browse.cfm?elib=18061 (Last viewed January 28, 2021).
Valimaki, V., Parker, J. D., Savioja, L., Smith, J. O., and Abel, J. S. (**2012**). "Fifty years of artificial reverberation," IEEE Trans. Audio, Speech, Lang. Process. **20**(5), 1421–1448.
Wallach, H., Newman, E. B., and Rosenzweig, M. R. (**1949**). "A precedence effect in sound localization," J. Acoust. Soc. Am. **21**(4), 468–468.
Werner, S., Klein, F., Mayenfels, T., and Brandenburg, K. (**2016**). "A summary on acoustic room divergence and its effect on externalization of auditory events," in *2016 Eighth International Conference on Quality of Multimedia Experience (QoMEX)*, pp. 1–6.
Yadav, M., Cabrera, D. A., Miranda, L., Martens, W. L., Lee, D., and Collins, R. (**2013**). "Investigating auditory room size perception with autophonic stimuli," in *Audio Engineering Society Convention 135* (Audio Engineering Society, New York), available at aes.org/e-lib/browse.cfm?elib=16984 (Last viewed January 28, 2021).
Zahorik, P. (**2002**). "Direct-to-reverberant energy ratio sensitivity," J. Acoust. Soc. Am. **112**(5), 2110–2117.
Zaunschirm, M., Schörkhuber, C., and Höldrich, R. (**2018**). "Binaural rendering of Ambisonic signals by head-related impulse response time alignment and a diffuseness constraint," J. Acoust. Soc. Am. **143**(6), 3616–3627.
Zotter, F., and Frank, M. (**2019**). *Ambisonics: A Practical 3D Audio Theory for Recording, Studio Production, Sound Reinforcement, and Virtual Reality* (Springer International, Cham), Vol. 19 of Springer Topics in Signal Processing, available at link.springer.com/10.1007/978-3-030-17207-7 (Last viewed January 28, 2021).