Object-based reverberation encoding from first-order Ambisonic RIRs

Philip Coleman¹, Andreas Franck², Dylan Menzies², and Philip J. B. Jackson¹

¹Centre for Vision, Speech and Signal Processing, University of Surrey, Guildford, Surrey, UK
²Institute of Sound and Vibration Research, University of Southampton, Southampton, Hampshire, UK

July 12, 2017

Abstract

Recent work on a reverberant spatial audio object (RSAO) encoded spatial room impulse responses (RIRs) as object-based metadata which can be synthesized in an object-based renderer. Encoding reverberation into metadata presents new opportunities for end users to interact with and personalize reverberant content. The RSAO models an RIR as a set of early reflections together with a late reverberation filter. Previous work to encode the RSAO parameters was based on recordings made with a dense array of omnidirectional microphones. This paper describes RSAO parameterization from first-order Ambisonic (B-Format) RIRs, making the RSAO compatible with existing spatial reverb libraries. The object-based implementation achieves reverberation time, early decay time, clarity and interaural cross-correlation similar to direct Ambisonic rendering of 13 test RIRs.
1 Introduction

Object-based audio representations give new opportunities to provide consumers with newly personalized and immersive audio experiences [1]. In object-based audio, audio assets are transmitted alongside metadata which describe how they should be experienced by the listener. A renderer, part of the end user’s reproduction equipment, then interprets the object-based scene and derives loudspeaker or headphone feeds. Compared to traditional channel-based or scene-based approaches [2], this means that the user may personalize the scene to their own taste, while the renderer is able to optimize the (personalized) reproduction over the target setup.

Reverberation is a key element of an immersive audio experience; however, current practice in object-based production is to encode the reverberant elements of a sound scene in a traditional channel-based or scene-based format [3]. For instance, MPEG-H [4] provides containers for channel-based and scene-based content to be transmitted in parallel to objects. Aside of any upmixing or downmixing issues leading to diminished immersion, this particularly limits the opportunities for personalization. On the other hand, object-based encoding of reverberant content allows extensive personalization, for instance: reverberant dialog could be substituted for a different language using only the dry speech source; or the reverberation time could be adjusted to facilitate improved speech intelligibility. Furthermore, the listening room can potentially be compensated by adjusting the reverberation parameters [5].

Building on work by Remaggi et al. [6], Coleman et al. [7] proposed the reverberant spatial audio object (RSAO), a metadata scheme for object-based reverberation, and demonstrated that the metadata could be edited to alter listeners’ perception of room size, source distance and envelopment. The RSAO parameters are intuitively editable and correspond closely to the room geometry. The relative merits of the RSAO compared to other state of the art parametric reverberation techniques (e.g., the spatial decomposition method [8], directional audio coding [9], and feedback delay network-based techniques [10]) are discussed in [7]. Other recent work [11] has considered possible options for encoding reverberation in an object stream described by the audio definition model (ADM) [12].

Regardless of the parameters eventually encoded in an object-based audio stream, there is a need to be able to capture these parameters from room acoustic measurements. Current methods to encode the RSAO parameters are based around signal processing applied to room impulse responses (RIRs) recorded with an array of omnidirectional microphones [6, 13], however a large number of RIRs exist in B-Format (used throughout to refer to 1st order Ambisonics). This presents an opportunity to encode a varied library of rooms in the RSAO format, for future object-based content production and subjective evaluation.

In this paper, we describe the implementation of RSAO parameterization from B-Format RIRs.
We evaluate the performance of the parameterization by the RSAO rendering with a direct Ambisonic rendering of the input test RIRs, for a number of rooms. The background to the RSAO is given in Sec. 2, and the parameterization is described in Sec. 3. The evaluation methodology is explained in Sec. 4, and the results are presented in Sec. 5. Finally, the results are discussion in Sec. 6, including impressions from informal listening, and we summarize in Sec. 7.

2 Background

The RSAO is modelled as a combination of a sparse set of specular early reflections, together with a filter describing the late reverberation. The assumed RIR model underlying the RSAO is illustrated in Fig. 1. The early reflections are encoded into parameters describing their time of arrival (TOA) and attenuation with respect to the direct sound, direction of arrival (DOA), and frequency response. The late reverberation parameters describe a temporal envelope, which is constructed into a filter during rendering. The (broadband) mixing time is estimated and encoded, and additional parameters are defined in octave subbands. For each subband, the rate of exponential decay is encoded together with the level in the neighbourhood of the mixing time. The parameters are motivated and described fully in [7].

3 Parameterization Procedure

In this section, the proposed parameterization procedure to acquire RSAO parameters from B-Format is described. The approach is outlined in Fig. 2, where the upper processing path corresponds to the specular components and the lower (shaded) paths correspond to the diffuse late reverberation components. The processing blocks in these paths are described below.

3.1 Specular components

Following the upper path in Fig. 2, four parameters are estimated for each specular component: delay (TOA), DOA, level, and the spectrum (frequency response). The estimation of each of these parameters from B-Format is described below.
Figure 2: Block diagram showing the parameterization of a B-Format RIR. The upper path is used to extract specular reflection parameters, and the lower paths for the late reverberation parameters. Output parameters are indicated in bold.

**Time of Arrival.** Following [6], the TOA peaks were estimated by using the dynamic programming projected phase-slope algorithm (DYPSA) [14] to detect peaks along the omnidirectional (W) component of the B-Format signal. The DYPSA algorithm returned peaks along the first 0.5 s of the RIR under test. An initial estimate of the energy corresponding to each peak was obtained by taking the root-mean-square (RMS) value of samples in the 1 ms neighbourhood of the peak. This estimate was used to identify the direct sound, the peak immediately following it, and the most prominent peaks up to the specified number. Here, we encode 20 peaks, in order to capture both the very first reflections, which generally carry most energy and affect the timbre and perceived source width, and subsequent strong echoes which add to the sense of room size.

The samples in the neighbourhood of the selected peaks were used for subsequent processing. For the direct sound, a 4 ms window centered on the peak was used, in order to include the energy present in the first 2 ms. For the subsequent reflections, a smaller window of 64 samples (1.3 ms @ 48 kHz) was used, in order to fully capture the target reflection without also encoding neighbouring reflections. A Hamming window was applied in each segment to weight the samples towards the central peak. The level values eventually encoded in the object stream were encoded based on the directional responses, described below.

**Direction of Arrival.** The DOAs of the segmented peaks were estimated by steering a virtual cardioid microphone. The output of the steered cardioid acting on a B-Format signal \([h_W(t), h_X(t), h_Y(t), h_Z(t)]\) can be written as

\[
S(\vec{r}, t) = \frac{1}{2} [h_W(t) + r_x h_X(t) + r_y h_Y(t) + r_z h_Z(t)],
\]

where \(r_x = d \cos(\theta) \cos(\phi), \ r_y = d \sin(\theta) \cos(\phi), \ r_z = d \sin(\phi), \) and \(d = 1, \) for \(0 < \theta < 2\pi \) and \(-\pi/2 < \phi < \pi/2,\) with one degree resolution. The DOA was then estimated as the steered peak containing the most energy, i.e., \(\max_{\vec{r}} \sum_t S(\vec{r}, t)^2,\) for the \(t\)th sample of the steered peak. The cardioid-steered RIR in the maximal direction was used to estimate the level parameters and the frequency response. Eq. 1 could
straightforwardly be extended for higher-order Ambisonic (HOA) RIRs.

**Level.** The level parameters encoded into the object stream were estimated from the noise gain of the RIR steered towards the maximal DOA, i.e., \( \sqrt{\sum_t S(\vec{r}_{max}, t)^2} \).

**Spectrum.** The frequency response of the early reflections was encoded in a two-step process. First, following [6], 8 linear prediction coefficients were used to model the frequency response of the steered RIR. Then, these coefficients were converted to second-order sets of infinite impulse response (IIR) coefficients. The IIR coefficients were normalized to have unity noise gain. The IIR coefficients are currently transmitted directly in the prototype object stream.

### 3.2 Late reverberation

The late reverberation parameters are estimated following the lower (shaded) paths in Fig. 3. The first goal is to estimate the late delay, or mixing time. Here, as our goal was to parameterize a number of B-Format RIRs without prior knowledge about the recording environments, we adopted a data-based approach. Namely, we employed the normalized echo density, which was proposed in [15] as a means to estimate the mixing time. The echo density profile of a RIR typically rises slowly during the first early reflections, then rises more steeply as the diffuse sound energy builds inside the room, before stabilising around a value of 1 (i.e., when the RIR temporally approximates Gaussian noise). The mixing time was here defined as when the echo density became greater than 1, using the implementation of [16].

Following estimation of the mixing time, the RIR under test was passed through an octave subband filter. For the \( b \)th subband, the late peak \( P_b \) was estimated as the noise gain in the neighbourhood of the mixing time \( T_m \),

\[
P_b = \sqrt{\frac{T_m + I_b}{\sum_{t=T_m-I_b}^{T_m+I_b} h_{W,b}(t)^2}},
\]

where \( h_{W,b} \) is the subband filtered \( h_W \) and \( I_b = 2f_s/f_b \) is a frequency-dependent window size, and \( f_s \) and \( f_b \) are respectively the sampling frequency and the subband centre frequency. In addition, in order to avoid over-estimation of the low frequency levels caused by the presence of early energy in the window, a maximum window length of \( I_b = f_s/100 \) samples was enforced, and a scaling value of 0.3 was applied in the 62.5 Hz subband. The scaling value was derived heuristically.

The RIRs \( h_{W,b} \) were also used to estimate the subband decay times. First, the late response was segmented (taking the samples from the mixing time to the end of the test RIR). Then, the Schroeder energy decay curve (EDC) was estimated for the segmented signal and an exponential curve was fitted to the EDC based on the decay over the first 20 dB (after the mixing time). This generally avoided reaching the noise floor of the input RIR. The time constant of the exponential curve was encoded, for each subband, as the late decay parameter. For each subband, the length of an onset ramp rising linearly...
<table>
<thead>
<tr>
<th>ID</th>
<th>Name</th>
<th>RT (s)</th>
<th>Dataset</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Emmanuel Main Church</td>
<td>1.22</td>
<td>S3A RIR</td>
</tr>
<tr>
<td>2</td>
<td>Emmanuel Old Church</td>
<td>1.31</td>
<td>S3A RIR</td>
</tr>
<tr>
<td>3</td>
<td>CVSSP Vislab</td>
<td>0.31</td>
<td>S3A RIR</td>
</tr>
<tr>
<td>4</td>
<td>St Andrew’s Church</td>
<td>1.39</td>
<td>OpenAIR</td>
</tr>
<tr>
<td>5</td>
<td>Maes Howe</td>
<td>0.51</td>
<td>OpenAIR</td>
</tr>
<tr>
<td>6</td>
<td>York Minster</td>
<td>7.38</td>
<td>OpenAIR</td>
</tr>
<tr>
<td>7</td>
<td>St Patrick’s Church</td>
<td>1.50</td>
<td>OpenAIR</td>
</tr>
<tr>
<td>8</td>
<td>Tyndall Bruce Monument</td>
<td>1.41</td>
<td>OpenAIR</td>
</tr>
<tr>
<td>9</td>
<td>UoY Sports Hall</td>
<td>6.04</td>
<td>OpenAIR</td>
</tr>
<tr>
<td>10</td>
<td>Nuclear Reactor Hall</td>
<td>4.60</td>
<td>OpenAIR</td>
</tr>
<tr>
<td>11</td>
<td>Alcuin Courtyard</td>
<td>1.90</td>
<td>OpenAIR</td>
</tr>
<tr>
<td>12</td>
<td>Railway Tunnel</td>
<td>1.78</td>
<td>OpenAIR</td>
</tr>
<tr>
<td>13</td>
<td>Troller’s Gill</td>
<td>1.01</td>
<td>OpenAIR</td>
</tr>
</tbody>
</table>

Table 1: Rooms for which the parameterization procedure was tested. RT stated was calculated from the W channel of the measured data, averaged over 500–2000 Hz octave bands.

from zero at the first early reflection to $P_b$ at the mixing time was also encoded. This allows diffuse energy to increase with time, even before the mixing time, according to the model in Fig. 1, and helps to give the perception of increasingly diffuse early reflections.

4 Evaluation Methodology

In this section, we introduce the selection of test RIRs, the rendering algorithms, and the evaluation metrics utilized to evaluate the RSAO parameterization and rendering pipeline.

4.1 B-Format Libraries

A total of 13 B-Format RIRs were parameterized for this paper. B-Format RIRs are readily available online and facilitate encoding of an RSAO library for many real-world spaces. RIRs were drawn from two datasets: the Surrey S3A RIR dataset [17] and the OpenAir RIR database [18]. The rooms tested, together with their reverberation times (RT30) averaged over the 500–2000 Hz octave subbands\(^1\), are listed in Table 1. The selected rooms have RTs in the range 0.31–7.38 s, and also include spaces where the assumed RIR model (Fig. 1) may not hold, for instance outside spaces (Rooms 11, 13) and a railway tunnel (Room 12). These kinds of spaces could be very useful in sound design and represent interesting test cases for the RSAO parameterization and rendering.

4.2 Signal Flows

The signal flow between the test RIR and the output signals is shown in Fig. 3. Loudspeaker positions for a virtual 3D loudspeaker array were passed to the renderer in a configuration file. The loudspeakers comprised the ITU-standardized 9+10+3 positions [19, System H, no LFE] with additional loudspeakers

\(^1\)RT30s were calculated using the irStats function in the IoSR Matlab Toolbox https://github.com/IoSR-Surrey/MatlabToolbox
Figure 3: Block diagram showing the acquisition of B-Format signals comparing the rendered original RIRs against the rendered parameteric RIRs.

B+135 and B-135 [19]. Channel feeds were obtained using an audio object renderer developed within the S3A project, which contains both an object-based reverberation renderer (Render RSAO) and a HOA renderer (Render HOA). The former is a fully parametric object render based on a single audio channel per object, while the latter receives an Ambisonic signal together with metadata describing the Ambisonic order and level. The renderers are described in more detail below.

Test audio of a unit impulse was used for the evaluation. To acquire the object render output, the audio was passed to the audio object renderer, and for the HOA render and Reference outputs, the unit pulse was convolved with the incoming test RIR. Metadata for the object and HOA renders was sent over UDP\(^2\), and audio was routed via the Jack audio connection kit\(^3\). Once the loudspeaker feeds had been generated, they were ‘recorded’ back to B-Format using a virtual B-Format microphone (i.e., by summing the loudspeaker feeds with angle-dependent weightings).

4.2.1 Object rendering

To obtain the object render, the extracted RSAO parameters were sent to the object renderer along with the test unit impulse. The direct sound was treated as a point source and rendered using vector-base amplitude panning (VBAP, [20]). The early reflections were each delayed, scaled, filtered, and spatialized (using VBAP) based on the encoded parameters. In the current implementation, it is not possible to filter the direct path audio without influencing the rest of the reverberation. Therefore, the direct path spectral filters proposed in [6] were not applied.

To render the late reverberation, the audio object is convolved with a finite impulse response (FIR) filter. The filter is constructed by applying time envelopes to filtered white noise (in each octave subband). The envelope is set to zero for samples before the first early reflection, rises linearly to \(P_b\) (Eq. 2) at the mixing time, and subsequently decays exponentially based on the encoded time constant. The convolved object audio is rendered to all loudspeakers by decorrelating with a set of random-phase allpass FIR

\(^3\)http://jackaudio.org
4.2.2 HOA rendering

The HOA renderer is based on the All-round Ambisonic Decoding (AllRAD) approach [21]. AllRAD operates in two stages: first, the HOA signal is decoded onto a regular virtual loudspeaker array; second, the signals corresponding to the virtual loudspeakers are panned to the real loudspeaker setup using VBAP. This enables robust HOA rendering to arbitrary, including irregular, loudspeaker setups. The HOA-rendered RIRs represent a direct application of the test signals to the specified loudspeakers. Here, all HOA-rendered scenes were first-order.

4.3 Evaluation Metrics

A number of evaluation metrics were adopted to compare the performance of the object render to that of the HOA render (and the reference RIR). To analyse the late reverberation, the reverberation time RT30 was calculated from the EDCs (i.e., estimating the 60 dB decay by curve fitting over the range -5 to -35 dB) in octave subbands from 250–8000 Hz. The RT30 is linked to the overall perception of room size [22]. For real-world stimuli where the signal continues during the decay, such as music, the early decay time (EDT) is thought to be a better indicator of perceived reverberance [23, 24, 25]. The EDT is also calculated from the EDC, but uses the range 0 to -10 dB to estimate the 60 dB decay. Where stated as a single figure, the RT30 and EDT values are averaged over the 500–2000 Hz octave bands.

In order to give an overall measure of the rendered early reflection energy, the clarity index, estimated over the first 50 ms following the direct sound (C50), was also used. The C50 is the ratio of the energy in the first 50 ms to that in the remainder of the RIR [25]. The C50 is here calculated as a broadband measure. The IoSR Matlab Toolbox was used for the RT30, EDT and C50 calculations (see Footnote 1), based on the W component of the virtual B-Format recording of the loudspeaker channels.

Finally, the interaural cross-correlation (IACC) was calculated. The IACC gives an indication of the spatial properties of the rendered RIRs: a high IACC indicates that the sound is mostly localized; a low IACC indicates that there is significant diffuse energy. The IACC is calculated as the maximum value of the interaural cross-correlation function, IACC= max |\rho_{lr}(\tau)|, for -1 ms \leq \tau \leq +1 ms [26], where

\begin{equation}
\rho_{lr}(\tau) = \frac{\int_{t_1}^{t_2} p_l(t)p_r(t+\tau)dt}{\sqrt{\int_{t_1}^{t_2} p_l^2(t)dt \int_{t_1}^{t_2} p_r^2(t)dt}}
\end{equation}

where \(p_l\) and \(p_r\) are the sound pressure signals recorded at the left and right ears, and \(t_1\) and \(t_2\) define the range of samples over which the IACC is calculated. The binaural signals \(p_l\) and \(p_r\) were obtained by convolving the loudspeaker channel feeds with a dense set of anechoic head-related transfer functions.
Figure 4: Results comparing the HOA Render and Object Render performance for each room under the metrics (a) reverberation time RT30, (b) early decay time EDT, (c) clarity C50, (d) IACC Early (first 80 ms), (e) IACC Late (80 ms–). The dashed lines correspond to just-noticable differences. Selected rooms are marked.

recorded with a Neumann KU100 dummy head [27], choosing the closest measured direction to the loudspeaker position, and summing the contribution of each loudspeaker to each ear. The IACC is calculated for two time periods: $IACC_{Early}$ is calculated with $t_1 = 0$, $t_2 = 80$ ms; $IACC_{Late}$ is calculated with $t_1 = 80$ ms, $t_2 = \infty$. Therefore, $IACC_{Early}$ gives an indication of the accuracy of spatial rendering covering the early reflections, and $IACC_{Late}$ measures how diffuse the rendered rooms are during the late reverberation.

5 Results

The 13 rooms listed in Table 1 were parameterized and rendered, and the results are discussed in this section. The results shown compare the object render and HOA render cases (cf. Fig. 3); it was separately verified that the HOA render corresponded very closely to the reference measurements under all metrics adopted.

5.1 Overview

An overview of the results can be seen in Fig. 4, which compares the HOA render with the object render under the metrics of RT30, EDT, C50, and IACC. On each plot, information about the just-noticable difference (JND) is also plotted with dashed lines. On Figs. 4(a, b), the red dashed lines correspond to 5% variation, which is classically thought of as the JND for these metrics [28, 25]. However, recent work [29] suggested that for non-interrupted stimuli, the JND for RT30 is above 20%. On Fig. 4(a), the 20% tolerance is therefore also shown with blue lines. Only two rooms fall outside the latter JND: Room 3 (an acoustically treated laboratory) and Room 12 (railway tunnel). These cases are discussed in Sec. 5.2.

In general, we can conclude that the reverberation times are perceptually well modelled by the object rendering. The EDT values in Fig. 4(b) are also generally close between the HOA render and object render. Room 8 is further outside of the 5% JND range, and is mentioned further in Sec. 5.2.

---

4The file HRIR_FULL2DEG.sofa was used.
The C50 results plotted in Fig. 4(c) show that the object render and HOA render generate quite similar values. The JND lines are plotted at ±1 dB (which strictly corresponds to the clarity JND over the first 80 ms [25]). As above, Room 3 falls further away from the JND than the other rooms, for which the object render generally gives C50 of the same order as the HOA render.

Finally, for the IACC results in Fig. 4(d, e), the JND values plotted are ±0.075 [25]. In each case, just over half of the rooms (7/13) fall within the JND, meaning that the overall spatial perception of the object render can be said to be perceptually similar to the HOA reference. Of the six rooms falling outside the JND (which are different between the early and late evaluation of IACC), those below the curve were, overall, rendered with too much diffuse energy, while those above the curve did not have sufficient diffuse energy in the object render. Rooms 1 and 7, which fell outside the JND range for both IACC early (Fig. 4(d)) and IACC late (Fig. 4(e)), are analysed further below.

5.2 Late reverberation analysis

Generally, the RT30 was well estimated and reproduced in the object-based framework. To illustrate the performance in more depth, the power spectrograms for Room 1 reproduced by object rendering and HOA rendering are shown in Fig. 5. It can be seen that there is a good match in terms of the time-frequency distribution of energy between the two cases. The roughness visible in the late response for the object render is caused by the decorrelation filters, and they are also partially responsible for the slight low frequency boost. Moreover, it can be seen that the prominent reflections at around 100 ms are modeled well by the object-based system.

The estimated RT30 and EDT for Room 1 are shown over frequency in Fig. 6. For the RT30, it can be seen that the general trend of the HOA rendered reference is matched by the object render, although the object render gives a slightly lower RT30 than expected in the 250 Hz and 500 Hz octave bands. The EDT values are generally lower for the object render than the HOA render, although they follow the
In Fig. 4, Rooms 3, 8 and 12 were seen to be less well estimated than most of the other rooms in terms of the RT30 (Rooms 3 & 12) and EDT (Room 8). The EDC for each of these rooms is plotted in Fig. 7, showing the 500 Hz subband. The EDC for Room 3 shows that the object rendered energy decays at the correct rate, but slightly later than the HOA render. This might imply that the mixing time was over-estimated, especially given that in this specific room the acoustic treatment led the late energy to decay quite quickly. This result also leads to the slightly underestimated C50 for Room 3. Nevertheless, the absolute error in RT30 estimate between the HOA render and object render was only 24 ms (averaged over 250–8000 Hz octave bands).

Considering the EDCs for Room 8, it is evident that, overall, insufficient energy was encoded in the early reflections for the object render. This leads to the underestimation of EDR, which is based only on the first 10 dB of decay. However, the overall RT30 was estimated within the JND for this room.

Finally, referring to the EDCs in Fig. 7 for Room 12, it can be seen that the late decay over the first 25 dB is very well estimated. However, the reference room then ceases to decay exponentially. In other words, the complex acoustics of the railway tunnel do not conform to the RIR model in Fig. 1.
Figure 8: Normalized absolute amplitude and energy decay for Room 1, showing the first 20 ms for the HOA render (solid) and object render (dashed).

Figure 9: Direction of arrival over first 20 ms of Room 1, showing the HOA render and object render.

5.3 Early reflection analysis

Further insight can be gained by considering some examples of object rendering of early reflections. The normalized absolute amplitude and EDC, comparing the HOA render and object render for Room 1 over the first 20 ms of the RIR, are shown in Fig. 8. From the normalized amplitude, it can first be verified that the early reflection peaks in the object render are well matched to those in the HOA render, i.e., the TOA estimation in the parameterization stage worked well. The early reflection peak amplitudes also contain the energy from the neighbouring samples (for instance, the first reflection (around 3 ms) also contains the energy for the following peak in the HOA render (around 4 ms)). In this case, EDC over the first 3 ms shows that the object render does not decay as quickly as the HOA render. However, thereafter the early reflections and building late tail match the decay of the HOA render very well (this trend continues beyond the first 20 ms shown in Fig. 8). The extra energy in the object render helps to explain why the IACC early (Fig. 4(d)) reported over-diffuse rendering for this room. Similarly, the IACC late (Fig. 4(e)) reported a slight under-diffuse rendering for Room 1; this effect is due to there being more energy in strong echoes (at around 100 ms, cf. Fig. 5) in the object render than the HOA rendered reference.
Nevertheless, one of the main objectives to encoding distinct early reflections was to provide a format-agnostic representation suitable for reproduction over 3D audio systems. For this use case, the DOA of the rendered early reflections is also important. Figure 9 shows a representation of the recordings at the virtual B-Format microphones for the HOA render and object render, showing the spatial response in azimuth with respect to time, over the first 20 ms, also for Room 1. The spatial responses were obtained by sample-wise steering a virtual cardioid microphone (Eq. 1). The object render in Fig. 9 essentially represents a sparse version of the HOA render: the energy peaks have similar TOA and DOA. As the object-encoded peaks include the energy in neighbouring samples, the peak levels in the object render plot are generally higher than the HOA render. The corresponding EDCs (Fig. 8) show this approach to be effective in terms of the overall energy decay.

Another room of interest based on the IACC results was Room 7, which was estimated to be too diffuse both for the early and late time windows. The normalized absolute amplitude and EDCs for this room are plotted in Fig. 10, showing the first 80 ms of the response (over which the IACC early, Fig. 4(d), was calculated). It is immediately apparent that there is significant peak energy in the HOA render, and there are many more prominent peaks than the 20 available in the object-based encoding. Furthermore, the energy only decays by 4.4 dB over the first 80 ms (for Room 1, this level of energy decay occurred after just the direct sound). In fact, the source-receiver distance is large for this measurement (9 m), and there is lots of clutter (wooden pews, pillars, etc.) on the paths between the source and receiver. The reflection density is therefore very high, even surrounding the direct sound, and energy from many reflection paths arrives in the first 80 ms. Moreover, the estimated mixing time for this room was 39 ms, where inspection of the echo density profile reveals that the normalized echo density already reaches 0.7 by 18 ms. Therefore, the object model might seek to better match the signal EDC by increasing the energy in the late reverberation (although this could not be expected to increase the early IACC). Increasing the number of early reflections could potentially alleviate this problem. Overall, Room 7 does not seem to conform to the signal model in Fig. 1, and is challenging to encode. This is a topic for future research.

6 Discussion

In practice, although objective metrics are important to quantify the performance of the object-based framework, subjective plausibility of the object-rendered reverberation is the overall goal. Informal listening to the (HOA and object) rendered omnidirectional RIRs (i.e., with a unit impulse as the programme material), over headphones, revealed that the overall perception of the room (e.g., size and source distance) was retained after object-based encoding and rendering. However, some timbral detail was lost. In the late tail this was partially due to the decorrelation filtering process, which made the
Figure 10: Normalized absolute magnitude and energy decay for Room 7, showing the first 80 ms for the HOA (solid) and object (dashed) renders.

decay less smooth over frequency (cf. Fig. 5). Using shorter decorrelation filters helps to alleviate this, but results in a significant low frequency boost. In addition, for rooms where there is still significant early reflection energy around the estimated mixing time, the late levels can be over-estimated, which can make the underlying white noise audible.

Nevertheless, listening on a 24 channel reproduction system (with loudspeakers at the same positions simulated in this paper), the spatial effects of both the early part and late part were quite convincing. By rendering the reflections in 3D to the correct location, the perception of source width was generally retained between the HOA render and the object render. In addition, the diffusion filters (and later directional echoes) created a well matched sense of envelopment for many rooms. One room where the spatial effects were especially convincing was the courtyard, where the distinct echoes were very well reproduced by the object render. In addition, for less critical programme items (such as musical stimuli), the rendered reverberation plausibly gave a sense of being in the same room as the spatial HOA rendered reference.

The construction of a pipeline to convert Ambisonic RIRs into object metadata parameters is an important asset for object-based reverberation, as it means that a significant number of measured RIRs can be converted to an object-based form. The pipeline extends straightforwardly to higher-order measurements. The object-based representation brings opportunities for listeners to personalize the reverb (for instance to balance between clarity and envelopment), and allows the renderer to optimize the reverb rendering for the target reproduction setup. Similarly, for interactive scenarios such as augmented or virtual reality, the ability to adjust parameters afforded by the object-based approach can help to create good binaural externalization and convincing content. These applications will be investigated in future work.
7 Summary

The aim of the work presented here was to create a library of rooms for the ongoing development of a RSAO, exploiting the large number of first-order Ambisonic measurements available. A pipeline from Ambisonic RIRs to object-rendered loudspeaker feeds was proposed, and the reverberation rendered in this way was compared to reverberation rendered directly with an HOA renderer. Over metrics including reverberation time, early decay time, clarity and interaural cross-correlation, the object render was shown to give similar performance to the HOA rendered reference.

Future work will seek to refine the mixing time and late level estimation, and will also investigate possible refinements to the decorrelation filter design. Beyond this, the format-agnostic aspect of the rendering will be investigated, comparing the HOA and object-based approaches using real-world stimuli.

8 Acknowledgements

This work was supported by EPSRC Programme Grant S3A: Future Spatial Audio for an Immersive Listener Experience at Home (EP/L000539/1), and the BBC as part of the Audio Research Partnership. The authors would like to thank Luca Remaggi and Chris Pike for making available the DYPSA implementation and binaural convolution engine, respectively.

References

[27] Bernschütz, B., “A spherical far field HRIR/HRTF compilation of the Neumann KU 100,” in Proc. of the 40th...