OPEN ACCESS Freely available online
Journal of the Audio Engineering Society
Vol. 65, No. 1/2, January/February 2017 (© 2017)
DOI: https://doi.org/10.17743/jaes.2016.0059

Object-Based Reverberation for Spatial Audio*

PHILIP COLEMAN,[1] *AES Member*, **ANDREAS FRANCK,**[2] *AES Member*,
(p.d.coleman@surrey.ac.uk)

PHILIP J. B. JACKSON,[1] *AES Member*, **RICHARD J. HUGHES,**[3] *AES Associate Member*,

LUCA REMAGGI[1], **AND FRANK MELCHIOR,**[4] *AES Member*

[1]*Centre for Vision, Speech and Signal Processing, University of Surrey, Guildford, Surrey, GU2 7XH, UK*
[2]*Institute of Sound and Vibration Research, University of Southampton, Southampton, Hampshire, SO17 1BJ, UK*
[3]*Acoustics Research Centre, University of Salford, Salford, M5 4WT, UK*
[4]*BBC Research and Development, Dock House, MediaCityUK, Salford, M50 2LH, UK*

Object-based audio is gaining momentum as a means for future audio content to be more immersive, interactive, and accessible. Recent standardization developments make recommendations for object formats; however, the capture, production, and reproduction of reverberation is an open issue. In this paper parametric approaches for capturing, representing, editing, and rendering reverberation over a 3D spatial audio system are reviewed. A framework is proposed for a Reverberant Spatial Audio Object (RSAO), which synthesizes reverberation inside an audio object renderer. An implementation example of an object scheme utilizing the RSAO framework is provided, and supported with listening test results, showing that: the approach correctly retains the sense of room size compared to a convolved reference; editing RSAO parameters can alter the perceived room size and source distance; and, format-agnostic rendering can be exploited to alter listener envelopment.

1 INTRODUCTION

Creation of a room effect is a critical part of audio production, whether the intention is to convey the sense of being in a specific real room or to carry the listener into a new world imagined by an artist. Technology for capture, production, and reproduction of reverberation must support these and other applications, requiring the reverberation to be recordable, intuitively editable, efficient to transmit, and reproducible on a wide range of reproduction systems.

A spatial audio scene, or components of a scene, can be represented by channel-based, transform-based, or object-based approaches [1]. For channel-based approaches, the engineer must mix for each target reproduction system, and the actual loudspeaker feeds are transmitted. Transform-based (or scene-based [2]) approaches map the scene onto orthogonal basis functions that can be decoded at the receiver and mapped to the available loudspeakers. In these approaches, the audio elements are fixed at transmission, limiting the flexibility further along the production chain.

In object-based audio, a scene is instead composed of a number of objects, each comprising audio content and metadata. The metadata are interpreted by a renderer, which derives the audio to be sent to each loudspeaker with knowledge of the specific target reproduction system (e.g., the loudspeaker positions). Similar to the transform-based approaches, object-based audio allows audio content to be format-agnostic [3], i.e., produced once and replayed on many different kinds of devices or adapted to non-ideal reproduction systems. However, as all objects are available to the renderer, the object-based approach is very flexible and has the potential to bring powerful new ways for consumers to interact with and personalize audio content.

Existing object metadata schemes provide a basic description of the object properties, such as its position in space and level. Alongside these, the ITU standard audio definition model (ADM) [4] allows an object to be diffuse or to have a size, and MPEG-H [2,5] includes a spread parameter. The signal-processing steps necessary to render these kinds of objects might also be used to reproduce reverberant signals. Even so, contemporary object standards do not support the concept of a reverberation object. Previous

*Portions of this paper were presented in P. Coleman et al., "On Object-Based Audio with Reverberation," at *AES 60th International Conference*, Leuven, Belgium, Feb. 2016.

proposals for object schemes containing room modeling have been proposed [6, 7], but have not been widely adopted. Rather, a common approach is to use a set of (localized) objects in conjunction with channel-based or Ambisonic-encoded tracks containing the ambience or reverberation [5,8–9]. Reverberation encoded in this way is limited in three ways: the spatialization is fixed and based on an assumed ideal reproduction setup, the language of any dialog is embedded into the reverberation, and control of the reverberation level is limited to a simple wet/dry mix. The benefits of object-based audio cannot therefore be fully realized.

Conversely, a fully object-based representation of reverberation implies that the reverberation is synthesized at the renderer. This could give greater immersion by allowing the renderer to reproduce early reflections independently and precisely, taking into account the reproduction layout. Opportunities for personalization or interaction can be envisaged both for producers (e.g., editing the acoustics of captured rooms) and consumers (e.g., choosing commentary language). Ideally, object-based reverberation would allow for independent control of the room effect due to each source, allowing objects to behave intuitively if, for instance, the level is adjusted. Finally, speech intelligibility may be enhanced in some cases by allowing listeners to increase the direct to reverberant ratio (DRR).

Describing the reverberation with a set of parameters implies that the reverberant signals will be synthesized in the renderer. Digital synthesis of reverberation is a topic that has attracted much research over many decades [10–12]. In particular, part of the MPEG-4 standard [13, 14] describes parametric reverberation techniques based on high-level descriptions of the desired room. Other approaches, such as spatial impulse response rendering (SIRR) [15] and the spatial decomposition method (SDM) [16] are based on low-level analysis and synthesis of a specific recorded acoustic space. Preliminary work on a reverberant spatial audio object (RSAO) [17] proposed a representation based on specular early reflections and diffuse late reverberation, having the advantage of being capturable from a real room yet editable and with easily implemented rendering.

In this article we extend [17] and our work in [18]. In Sec. 2 we outline the assumed room model and discuss the key perceptual characteristics of the room effect. Then, in Sec. 3, following [18], we draw together and compare parametric approaches to reverberation synthesis and discuss their application to object-based content, considering the whole production chain of capture, production, editing, and rendering. We propose the RSAO framework, following [17], in Sec. 4. In Sec. 5 we give an example of calculating RSAO metadata (refined compared to [17]) from recorded room impulse responses (RIRs), and extend both former works by presenting new listening test results demonstrating the potential of the approach to be edited by a producer and rendered over different loudspeaker systems. Finally, in Sec. 6 we discuss the RSAO implementation and future prospects, and we summarize in Sec. 7.

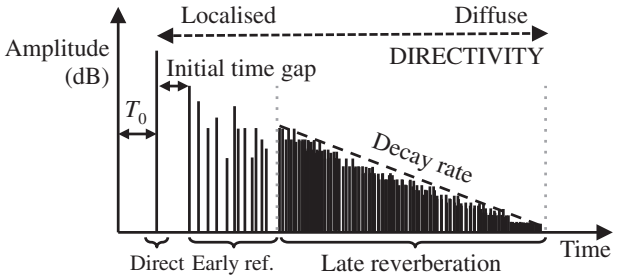


Fig. 1. Generic RIR model (based on [10], Fig. 1), showing the squared magnitude. Direct sound arrives after time T_0, early reflections begin after the initial time gap, and late reverberation decays exponentially. Sound becomes increasingly diffuse with time.

2 ROOM MODEL AND PERCEPTION

To represent the reverberation in an object-based manner, it is useful to consider the perceptual characteristics of the room alongside a physical room model. At low frequencies (i.e., below the Schroeder frequency [19]), modal behavior dominates, leading to monaural timbral effects. As frequency increases, reflections are often thought of (and modeled) as sound rays following the principles of geometrical acoustics [20]. Fig. 1 shows a generic room reverberation model representing the development of an RIR over time, from which the main perceptual properties can be explained. The model comprises the direct sound arriving after time T_0, a number of early reflections, and late reverberation characterized by an exponential decay curve. As marked on the figure, the diffuseness of the sound field generally increases with time. Following the direct sound, early room reflections are initially sparse in time, appearing as distinct contributions arriving from specific directions determined by the room geometry. Reflections arriving within the first 5–10 ms affect localization and are usually associated with a perceived image shift and broadening of the primary sound source [21]. These early reflections can also lead to coloration through comb-filtering [22]. The initial time gap separating the direct sound and the first reflection is thought to affect perception of the presence or intimacy of the room and its apparent size [23]. As time progresses, the sound field becomes a mix of diffuse and specular reflections of decaying level and increasing temporal density and spatial diffuseness, displaying behavior more statistically random in nature [10]. The diffuse reflections and the reverberant decay affect predominantly spatial attributes such as perceived envelopment and spaciousness [23] and provide cues to source distance (predominantly via the DRR [24]) and room size. In order to convey the key properties of the target room acoustics, object-based reverberation should therefore preserve the initial time gap, the spatial and timbral properties of the early reflections, and the diffuse and decay characteristics of the late reverberation.

3 PRODUCING REVERBERATION

Current approaches for recording and representing reverberation in object-based content fall into two main classes:

Table 1. Summary of parametric 3D reverberation in the context of an object-based production pipeline. The proposed RSAO (highlighted, Sec. 4) is included for comparison, referencing preliminary work in [17].

	Capture	Parameters	Editing	Rendering
MPEG-4 Physical [13, 14]	Estimate room dimensions and surface filtering properties	Source directivity and FIR surface filter or structured audio spec.	Edit room description	Computational acoustics room rendering or convolution
MPEG-4 Perceptual (Spat) [14,34]	Estimate physical room properties and perceptual correlates from mono RIR	Perceptual parameters (linked to delay network coefficients)	Modify source/room in perceptual parameters domain	Early reflection module and FDN coefficients
RSAO [17]	Circular array RIR	TOA/DOA for direct sound and discrete early reflections; octave-band decay for late reverb	Modify image sources (early) and octave-band reverb time (late)	Split signal: pan non-diffuse and decorrelate diffuse
SIRR [15]	B-Format RIR	Time-frequency-wise azimuth, elevation, and diffuseness	Edit TF-cell parameters	Split signal: pan non-diffuse and decorrelate diffuse
R-WFS [35]	Circular array RIR	High resolution plane wave RIR	Edit response in plane wave domain	Render source types by WFS, convolve with input audio
SDM [16]	RIRs from 4+ microphones in a 3D layout	Omni RIR plus DOA for each time segment	Low-level editing of omni RIR or DOA for an image source	Convolve and render to target DOA

signal-based approaches and parametric approaches. Here, we focus mainly on the parametric approaches that can be applied to user-end rendering.

3.1 Signal-Based Approaches

Signal-based approaches are those that record signals with a particular microphone array and include spatial microphone techniques (which directly record the reverberant signals) and convolution reverbs (which apply reverberant RIRs to a dry signal in post-production). These approaches can utilize channel-based or scene-based representations. Channel-based spatial microphone techniques are intended to be reproduced over a specific reproduction layout. Main microphone techniques have been developed for stereo (e.g., [25]), surround (e.g., [26]), and with-height systems (e.g., [27, 28]). A room microphone array (e.g., [29]) can be used to capture diffuse room sound. Captured signals are moderately editable by adjusting the relative mix of the microphone signals [30]. However, the spatial aspects of the recording are fixed in the channels and intended for a specific loudspeaker arrangement. Similarly, user interactivity is limited because the contributions of individual sources are not available.

Scene-based approaches (e.g., [8,31]) allow scene rotation, scaling, and spatial filtering to enhance or attenuate certain directions. User interaction capabilities are limited in the same way as for channel-based approaches. Convolution reverbs assume that the RIR of a reverberant space can be applied as a finite impulse response (FIR) filter, thus any dry signal can be made reverberant in post-production by convolution with a pre-recorded set of RIRs. Application of a convolution reverb is analogous to making a recording in the space where the RIR was recorded (or computationally generated), having limitations as described above, although RIRs for convolution can be edited (e.g., [32, 33]). Signal-based reverberation is, in general, difficult to represent in a format-agnostic way, limiting the potential benefits from using an object-based representation. An extended discussion on the signal-based approaches may be found in [18].

3.2 Parametric Approaches

In this section we provide an overview of parametric synthetic reverberation for object-based audio. In Sec. 3.2.1 we discuss approaches based on *high-level* parameters, and in Sec. 3.2.2 we discuss the parameterization and synthesis of rooms based on low-level parameters directly available to the renderer. Table 1 summarizes the discussion of the parametric approaches, which are also shown in Fig. 2. Parameters describing the reverberation are defined in the parameterization process, optionally using recordings from a real room. Then in production, the producer edits object and reverb parameters, using a local version of the renderer to *r*ender the reverberant scene and monitor the production. The reverberant signals are finally represented as audio and metadata streams and rendered to the end user's loudspeakers, accounting for user personalization input.

3.2.1 High-Level Parameters and Synthesis

Reverberation can be synthesized based on high-level parameters that describe a room in physical or perceptual terms. The main examples of high-level parametric reverberation are found in the physical and perceptual parameters for room modeling in MPEG-4 v2 [13,36]. Physical parameters are specified in terms of the transmission paths in the environment and frequency-dependent directivity models for each sound source, and are rendered by computational acoustic modeling and convolution. For the perceptual parameters, some are specific to the source (e.g., presence, warmth, brilliance), while others describe overall reverberation (e.g., reverberance, envelopment, liveness) [37]. These (high-level) perceptual parameters map to low-level feedback delay network (FDN) coefficients (convert

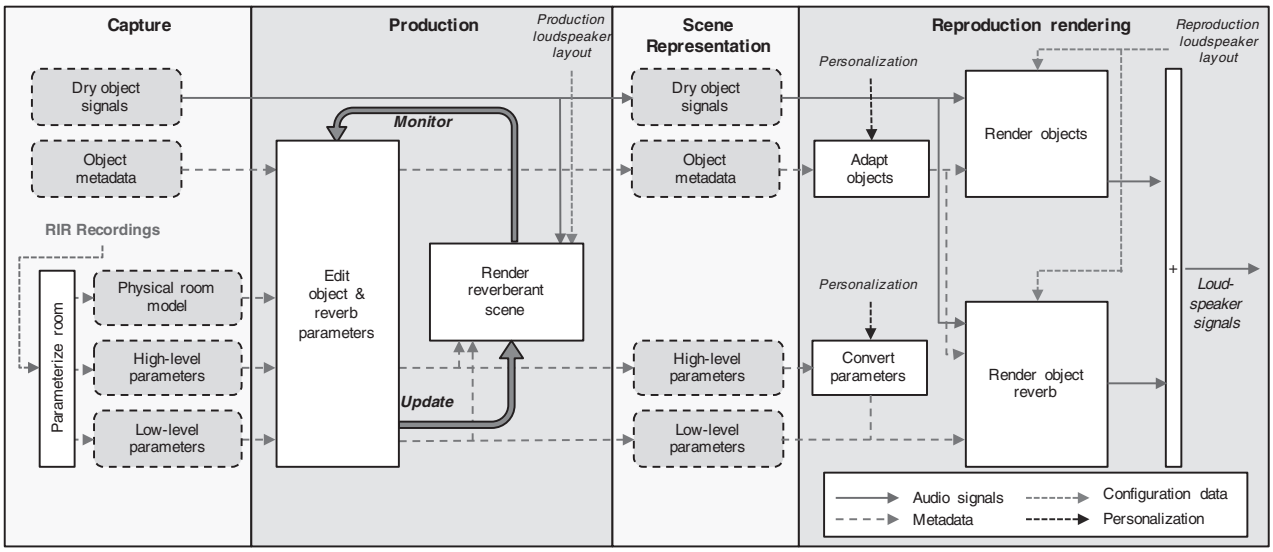


Fig. 2. Signal and metadata flows for capturing, editing, representing, and rendering parametric reverberation, accounting for the local loudspeaker arrangement and listener personalization input.

parameters block, Fig. 2) to control various portions of the RIR. The MPEG-4 perceptual approach is based on the Spat reverberator [34]. To synthesize the low level parameters, the direct portion is panned directly according to the source direction, the first reflections are created by panning delayed versions of the direct sound (the early reflection module), and the late portions are created using FDNs whose coefficients are linked to the mixing time and decay time. Although parameters are typically estimated from a mono RIR, recent work [38] has considered the problem of mapping a known room geometry to FDN coefficients. The FDN coefficients are, however, linked to a fixed number of virtual loudspeakers for reproduction, reducing the opportunities for format-agnostic rendering.

3.2.2 Low-Level Parameters and Synthesis

Alternatively, low-level parameters (directly interpreted at the renderer) may be used to define the desired room acoustic. One method of efficiently parameterizing an RIR is the SDM [16]. The SDM is based on the assumption that the RIR is composed of image sources in the far field. In each time segment, a microphone array is used to determine the direction of arrival (DOA) of the most prominent image source. This information is combined with the actual RIR recorded at an omnidirectional microphone at the center of the array, to spatialize the omnidirectional signal (in 3D). In the context of object-based audio, the omnidirectional RIR would be transmitted to the renderer, where a convolution engine would be implemented to spatialize the dry object audio.

SIRR [15] (which underpins the analysis and synthesis in directional audio coding (DirAC) [39]) is an alternative framework for analyzing, encoding, and synthezing a spatial RIR. The analysis part is based on a B-Format or higher-order recording [40], and represents the time-frequency-wise spatial response using three parameters: the DOA (azimuth and elevation) and a diffuseness coefficient. Editing this kind of metadata in production was considered in [41] to achieve effects including rotation, zoom, compression, and spatial filtering. This shows that, although the route from production tools to low-level parameters is not intuitive, tools could indeed be developed for producers to use. For synthesis, the direct sound component is panned via vector-base amplitude panning (VBAP) [42], while the diffuse portion is decorrelated and sent to all loudspeakers [43]. The parametric 3D nature of SIRR means that it can be synthesized flexibly, similar to point source or diffuse object types. Similarly to SDM, the SIRR represents a good approach to parameterize a specific room, but the parameters may not be straightforward to adjust in editing.

Another system for capturing, editing, and rendering room acoustics based on a plane wave description of the sound field was proposed in [35] with wave field synthesis (WFS) as the target rendering approach. We refer to this as reverberant WFS (R-WFS). First, the wave field is analyzed based on measurements from a circular microphone array [44, 45]. Then, RIRs in the plane wave domain are divided into an early part and late part. In addition, strong early reflections are extracted by spatio-temporal windowing. This leads to a representation of the room comprising discrete early reflections, the remaining early part (reflections and building diffuseness) and late part (reverberation tail) of the room response. The discrete early reflections may be modified based on the position and directivity of the direct sound, whereas the early part and late part of the reverberation are fixed for each room (see also [46]). Representation of the sound field as a sum of physical point sources and plane waves is inherently object-based. The point sources are used for early reflections, and at least ten plane wave sources [47] distributed evenly around the listener are used to render the late reverberation. The same approach has been used for surround sound and binaural reproduction [48].

3.3 Discussion

In contrast to signal-based approaches, parametric reverberation rendering gives opportunities for comprehensive producer control and user personalization. For automatic dialogue replacement (ADR), for example, the user would be able to select their own language and any reverberant sound would also change. Similarly, the user could adapt parameters relating to reverberation level or time, in order to improve speech intelligibility. For producers, elements of the audio scene could be emphasized, attenuated, or made to appear more distant by modifying the DRR, and objects could be moved in the scene while retaining intuitive discrete early reflection patterns. The viability of this kind of personalization depends on the abstraction level of the transmitted parameters. Low-level parameters, such as those describing recorded RIRs, might be harder to map to intuitive tools.

The target application might also influence which reverberation representation is most appropriate. High-level parametric approaches might be useful for applications such as sonic art (e.g., popular music or radio drama production), where reverberation is predominantly used as a creative effect. In this case, there is not necessarily a real-world reference room, different components of the acoustic scene might have different reverberation, and the rooms might not be architecturally practical in practice. On the other hand, a producer may wish to directly capture and edit reverberation parameters. For an application such as recording a classical music concert, the acoustic space is an inherent part of the performance, affecting the conductor and musicians [49, pp. 3–15] as well as the audience. Thus, the recording process should attempt to faithfully reproduce an impression of the room acoustics, while allowing the engineer to modify the room impression if desired. Signal-based room techniques might be appropriate, but editing the reverberant content of the room depends on the skill of the recording engineer (i.e., by judicious microphone placement in-situ). Furthermore, the resulting mix would be inflexible over different reproduction systems. An alternative approach would therefore be to combine close-microphone recordings of the orchestra sections with a parametric description of the acoustic space. The choice of parametric approach depends on the degree to which the producer wishes to edit the room. The Royal Albert Hall in London, for example, has a strong echo that may be undesirable [49, p. 237]. The R-WFS and RSAO approaches would both allow this reflector to be removed or attenuated in the parameter domain. On the other hand, if minimal editing of the space were required then the SIRR or SDM approaches might be most appropriate. These methods are straightforward to capture and reproduce, but are not as intuitive to edit (although production tools could be envisaged to this end). Overall, using a parametric representation rather than a signal-based recording would allow the impression of the concert hall to be conveyed flexibly over a range of reproduction systems.

A further consideration for object-based reverberation is the required computation, power, and bandwidth for rendering. While it could be assumed that professional studio environments, and probably home systems, are able to supply sufficient resources to render parametric reverberation, the same cannot be assumed for mobile devices. In particular, convolutions for many objects might not be feasible. However, there exist computationally efficient techniques for convolution, and the parameters require less bandwidth than additional channels dedicated to ambience or reverberation. Another general risk with parametric approaches compared to signal-based approaches is that the producer must rely on the renderer to faithfully preserve their intentions across many different kinds of reproduction systems. However this might be outweighed by the opportunity to flexibly render reverberant content over many different loudspeaker layouts or headphones. Finally, depending on the application, the renderer might produce an intermediate output format, for instance, a new set of "standard" audio objects, or a scene-based representation.

4 THE REVERBERANT SPATIAL AUDIO OBJECT

As an alternative to the parameteric approaches described in Sec. 3.2, the RSAO framework describes a compact set of reverberation parameters that may be estimated from measured RIRs. Moreover, the parameters can be rendered using the same signal processing steps required to reproduce common object metadata such as point sources and diffuse objects. The parameters are based on the common approach of separating the RIR into the direct sound, L specular early reflections, and late reverberation, similar to the model in Fig. 1. Fig. 3 shows a block diagram of the system, showing the capture, parameterization, object parameters, and rendering stages for the direct sound and early reflections (see Sec. 4.1) and the late reverberation (see Sec. 4.2).

4.1 Direct Sound and Early Reflections

The direct sound and early reflections characterize the source direction, source width, and room geometry. The parameters encoding these portions of the RIR are based around a listener-centered image source approach, similar to SDM and the MPEG-4 physical method. The direct sound is characterized by the DOA of the source with respect to the listening position, together with an optional spectral filter (giving opportunities to encode source directivity, obstructing objects, and monaural timbral effects due to modal behavior at low frequencies). Early reflections are modeled as peaks in the time domain RIR, generated by directional specular reflections. In addition to the direct sound parameters, the early reflections also have a parameter for the delay of the image source with respect to the direct sound. The parameters for the direct sound and early reflections can be edited based on a physical model of the listener in the room or by abstracting the parameters to higher-level descriptions such as envelopment or intimacy. In the renderer, both the direct sound and the early reflections are rendered as point-like sources, assuming that the DOA of each reflection is consistent across the full frequency range.

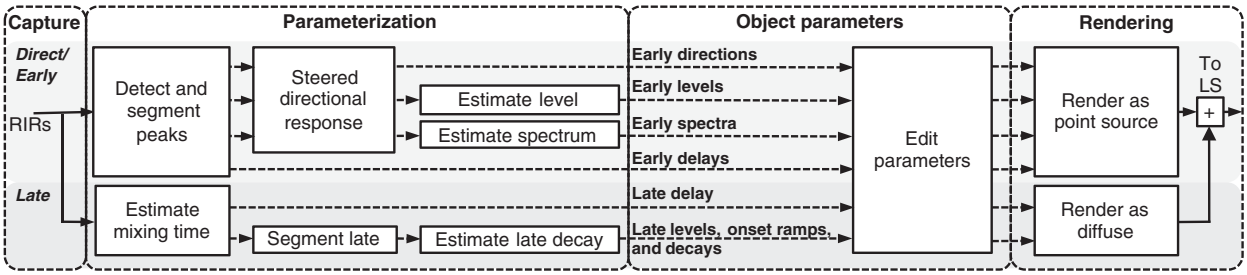


Fig. 3. Overview of the proposed RSAO framework, showing the capture, parameterization, object parameters, and rendering stages and illustrating the direct/early reflection and late reverberation processing paths.

4.2 Late Reverberation

Late reverberation is modeled in the RSAO as exponentially decaying white noise, which can characterize the superposition of high-order specular and diffuse reflections [50], and convey room size, spaciousness, and distance. Parameters for the late reverberation are a late delay, corresponding to the mixing time, together with a level, decay constant, and onset ramp length for a number of subbands. The onset ramp aims to build diffuseness through the early part of the RIR, coinciding with the later early reflections. These parameters can be edited based on a physical model, mapping to perceptual parameters, or DRR control. The late reverberation is rendered by producing a single channel FIR filter, convolving it with the incoming object audio and rendering to all loudspeakers via diffusion filters.

5 RSAO IMPLEMENTATION EXAMPLE

In this section we expand on the RSAO concept with an implementation example where we estimate parameters from measured RIR data and render them with an object-based renderer. We also present results of a listening test, showing that the RSAO framework can be used for format-agnostic, editable reverberation rendering.

5.1 Datasets

Multichannel RIRs measured in four rooms in Guildford, UK, were used for the parameterization experiments [51]. The rooms are: Vislab, an acoustically treated lab (room volume V=240 m^3; RT60=0.80 s (0.5–4 kHz); source distance d=1.7 m); Studio 1, a classical recording studio (V=1615 m^3; RT60=1.11 s; d=3 m); Church I, a Victorian church (1904) with thick concrete walls and thinly carpeted floor (V=1027 m^3; RT60=1.31 s; d=3 m); and Church II, a modern church (1991) with brick walls, large wooden roof, and carpeted floor (V=2857 m^3; RT60=1.41 s; d=5 m)[1]. Each recording was made with a bi-circular array of 48 Countryman B3 omni lavalier microphones [52], and with Genelec 8020B (Vislab), 8030A (Church I and II), and 1030A (Studio 1) loudspeakers, respectively. All sources used for this experiment were placed directly in front of the microphone array. An example of a single RIR recorded in each room is shown in Fig. 4.

[1] Accessible via http://cvssp.org/data/s3a/

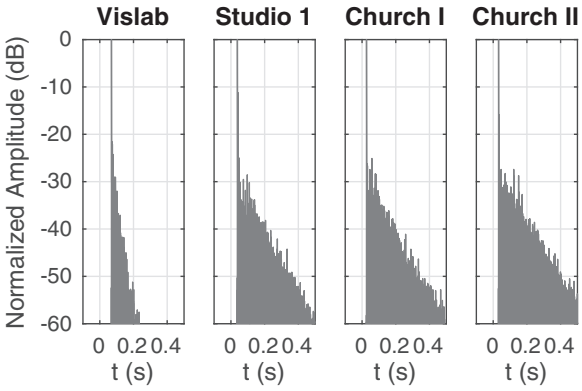
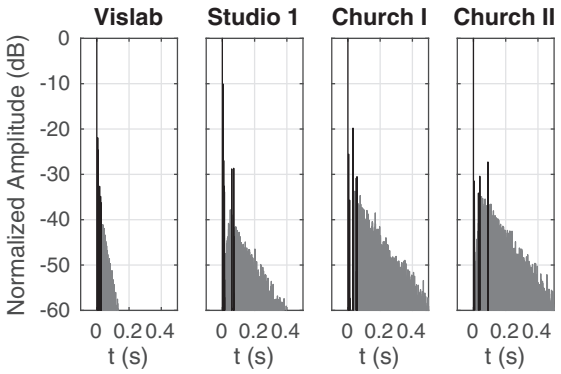


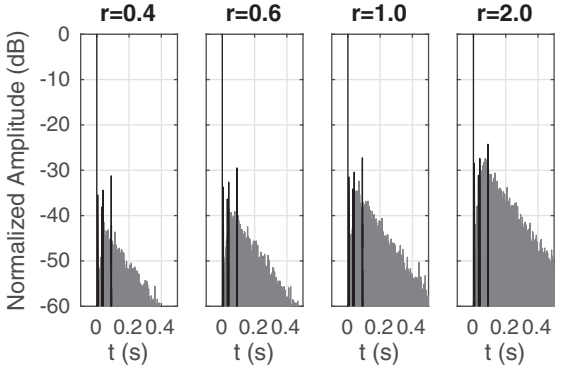
Fig. 4. An example RIR from each dataset showing the squared magnitude.

5.2 Parameter Estimation and Synthesis

The parameter estimation stages are shown in Fig. 3. Early reflection parameters were estimated with a refined implementation of [17]. In particular, the peak detection, frequency filter estimation, and mixing time estimation were improved. Peak detection used the Clustered Dynamic Programming Projected Phase-Slope Algorithm (C-DYPSA) [53] to extract the six strongest peaks (ranked by ampltiude) detected across all 48 RIR channels. Each detected peak was segmented with a window of $L = 64$ samples, and then a delay and sum beamformer (DSB) was steered in 3D to estimate the DOA. The level was then estimated from the DSB output using the noise gain, i.e., $\sqrt{\sum_n^L h_b(n)^2}$, where h_b is the segmented, steered reflection. Finally, the spectrum of h_b was estimated using 8 linear predictive coding (LPC) coefficients [54]. No frequency filtering was applied to the direct sound in this implementation. The late onset (mixing) time t_{mix} used was the model-based perceptual mixing time t_{mp50} via the regression formula given in [55, Eq.(12)], based on the room dimensions. Octave subbands were used for the late decays and levels; after subband filtering the exponential decay constant was estimated using the first 20 dB of decaying late energy after t_{mp50} and the subband level was estimated by calculating the noise gain (as above) in the 1 ms neighborhood of t_{mix}. In each subband, the onset ramp length (to t_{mix}) was specified as a linear ramp from zero to the late level, starting at the first reflection. All in



(a) Parametric RIRs for each room



(b) Parameteric RIRs for Church II with adjusted distance

Fig. 5. Representation of the parametric RIRs prior to rendering, showing the squared magnitude of the direct pulse, the delayed, filtered, and attenuated specular reflections (black) and the diffuse late reverberation filter (grey).

all, the following parameters were sent to the renderer as a JSON-formatted text string over UDP: direct sound level and position; six early reflection levels, delays, positions, and sets of LPC coefficients factorized into second-order filter sections; late delay; and nine late subband onset ramp lengths, levels, and exponential decay constants.

The parameters were received by an object-based renderer developed as part of the S3A project. Early reflections were delayed, scaled, and filtered based on their metadata and panned to the available loudspeakers via VBAP. For each late reverberation subband, an envelope was constructed as an initial delay modeled by zeros, a linear ramp up to the mixing time, and an exponential decay to the maximum number of samples. These envelopes were pointwise multiplied by subband-filtered white noise sequences, and the final broadband FIR filter was generated by summing the subband contributions. The incoming object audio was convolved with the FIR filter and sent to all channels via diffusion filters. The FIR filter used was 2 s long. The RIRs constructed from the parameters for each room under test are shown in Fig. 5(a), following temporal processing but before spatialization.

5.3 Listening Tests

In order to demonstrate the potential of the RSAO framework to achieve editable, format-agnostic reverberation, a set of pilot listening tests were carried out in the ITU-R BS.1116 standard listening room at the University of Surrey (described in [56]). Eleven listeners were tested: five experienced listeners and six inexperienced listeners. Eventually, two of the inexperienced listeners were removed from subsequent analysis: one reported having hearing difficulties on the day and another reported making random ratings because the task was too difficult. In each of three tests, described below, listeners were presented with a MUSHRA-style interface and used multiple sliders to rank each stimulus against the attribute under test. The scales were unmarked and listeners were asked to rate at least one item on each page at the bottom of the scale (0) and at least one item at the top of the scale (100). Three program items were used: an anechoic hand clap from the *Freesound project*[2] and anechoic male speech, and guitar recordings from the Bang & Olufsen *Music for Archimedes* CD [57]. Four possible types of rendering were used: 22chan, object rendering to ITU 22.0 loudspeakers; stereo, object rendering to stereo; mono, the stereo object render summed to mono (center channel); and, meas*, a timbral reference created by convolving one of the original omnidirectional microphone recordings with the source signal and replaying it in mono. In each case, the direct sound DOA was at 0 degrees azimuth and elevation with respect to the listening position. In reproduction, all channel layouts used a bass management system, using two frontal subwoofers, for bass content. In addition to the parameters estimated from the real room, edited versions of the parameters, described below, were used in some of the tests. The program items were manually loudness matched prior to processing; any loudness differences for different test stimuli were due to the reverberation rendering under test.

5.3.1 Apparent Source Distance

The first test asked the listeners to *please rate the following stimuli according to how far they appear to be from you, rating at least one stimulus as "Farthest" and at least one stimulus as "Nearest."* For this test, the guitar and speech programs were used, and the stimuli comprised the original and three modified versions of the Church II parameters, rendered over mono, stereo, and 22chan, together with the original Meas* recording. Parameters were altered based on a set of simple rules and a relative distance coefficient r, as: $l_d = l_d/r$; $l_e = l_e/\sqrt{r}$, where l_d and l_e are the direct and early reflection levels, respectively. Fig. 5(b) shows the resultant RIRs. Although not edited here, the scheme equally allows for adjustment of reflection delay and direction to account for source and receiver position changes. The results are shown in Fig. 6, top. Overall, the listeners were most uncertain about rating the original distance (between $r = 0.6$ and $r = 2.0$), yet overall it is clear that by modifying the DRR in the parameter domain the listeners' perception of distance was altered. In general, the listeners rated the

[2] The clap was selected from a sequence recorded in the anechoic chamber at HKU, contributed by "Anton." URL https://www.freesound.org/people/Anton/sounds/345/

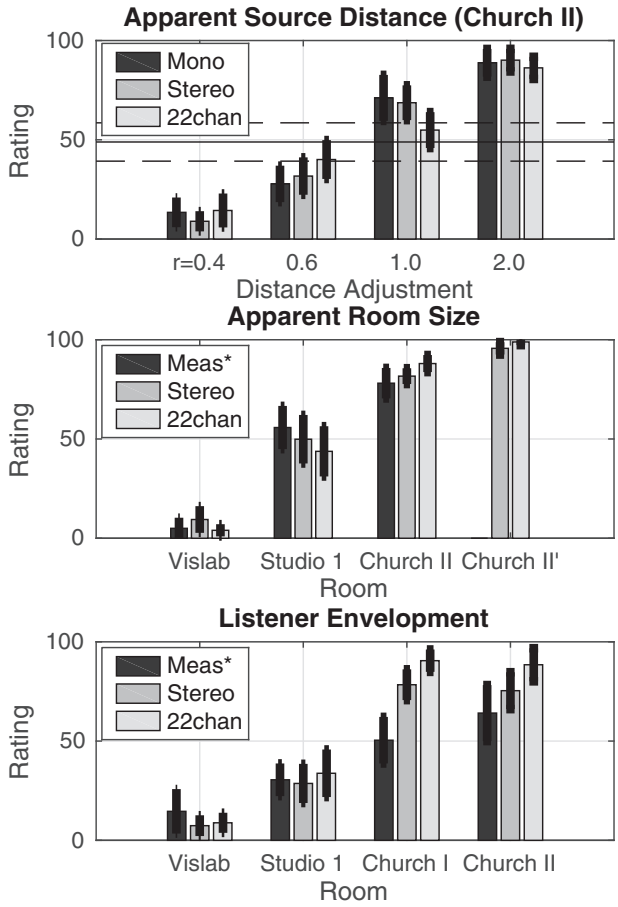


Fig. 6. Perceptual scores for apparent source distance (top), apparent room size (middle), and listener envelopment (bottom), averaged over program item, together with 95% CIs. Results are shown for Mono (left), Stereo (center), and 22 Channel Surround (right) reproduction. For mono reproduction, Mono (top) denotes a sum of the stereo object render, and Meas* denotes an original omni RIR convolved with the program. The room Church II′ was created by editing parameters, so there is no measured reference in this case. In the uppermost plot the Meas* score is shown as horizontal black lines (solid: mean; dashed: 95% CIs).

source to appear to be at the same distance across the three reproduction systems. There is also good agreement between the distance ratings for meas* (horizontal lines, Fig. 6, top) and the original parameters over 22chan, although meas* was rated closer than the mono and stereo object renders.

5.3.2 Apparent Room Size

The second test asked listeners to *please rate the following stimuli according to how large the room appears to be, rating at least one stimulus as "Largest" and at least one stimulus as "Smallest."* Here, only the speech program was used, and the original sets of parameters for Vislab, Studio 1, and Church II were rendered over stereo and 22chan together with meas* for each room. In addition, a modified version of Church II, where the late energy decayed 20% slower and t_{mix} was 50 ms later, was used (we refer to this as Church II′). The results (Fig. 6, middle) show that the listeners were able to rank the parameteric rooms in the same order as the real rooms, with no significant differences between the ratings. This implies that the parameters can properly convey the sense of the size of the target room. Similarly, the edited parameters were able to increase the perceived size of the target space.

5.3.3 Listener Envelopment (LEV)

Finally, listeners were asked to *please rate the following stimuli according to how surrounded you feel by them, rating at least one stimulus as "Most enveloping" and at least one stimulus as "Least enveloping."* For this test, the clap and guitar programs were used, and the original sets of parameters for all rooms were rendered over stereo and 22chan and compared with the measured reference meas*. In general, LEV increased with room size, although the ratings were similar between the two church buildings (Fig. 6, bottom). This similarity might be explained by Church I having stronger early reflection parameters, yet Church II being slightly more reverberant (see Fig. 5(a)). The LEV results across systems were similar across all reproduction methods for the smaller rooms, but for the larger spaces listeners rated the 22chan reproductions to be more enveloping than the mono or stereo. This result is not entirely surprising; nevertheless, it illustrates that, by using an object-based approach to reverberation, the renderer can improve LEV where more loudspeakers are available. When combined with the results above, the ratings suggest that this increase in envelopment may not in general come at the cost of altering the apparent size of the reverberant space.

6 DISCUSSION

The overall aim of the RSAO is to give a plausible room impression based on the desired acoustic environment. To this end, the RSAO represents reverberation with a set of intuitive low-level parameters. These parameters are linked to the key perceptual cues to give a plausible impression of the acoustic properties of a room, including the initial time gap, surface materials, and room size. Moreover, the object-based nature of the parameters gives opportunities for the renderer to generate loudspeaker or headphone signals based on local knowledge of the reproduction system. The listening test results presented here show that, overall, the parameterization maintains the sense of room size across reproduction systems and with respect to a measured reference case, being plausible at least in this regard. The RSAO parameters can also be edited for creative effects. Our tests demonstrated that a source could be made to appear more distant by using a basic level adjustment to modify the DRR, and that a space could be made to appear larger by modifying the decay constant governing the late reverberation filters. In future, using a low-level parametric reference such as the SDM would allow the RSAO approach to be perceptually evaluated in terms of the realism of the reconstructed room.

The parameters themselves were outlined in high-level terms in Sec. 4, allowing for other implementation options

within the RSAO framework. The detected early reflection peaks were visualized and shown to correspond to the ground-truth room geometry in [52]. However, more research is necessary to conclude, perceptually, how many early reflections should be encoded. Indeed, the question of whether these should be the *first N* reflections or the *strongest N* reflections is also open. The choice of six strongest early reflections was motivated here by the desire to find all first order reflections, however strong second order reflections might also be detected. The encoding of the early coloration filters by LPC is another implementation choice that may be refined. For the late part, the mixing time could be estimated from the data, for instance from the echo density [55,58], however we found that the estimation was not robust across all 48 channels of our data without further refinement. Similarly, the late level estimation might be made more robust in future. For the late FIR filter, efficient convolution techniques such as the velvet noise approach [59] might be adopted. Futhermore, various microphone arrays, including B-Format and higher-order microphones, could also be used to acquire RIRs for parameterization. Beyond this, one significant challenge will be to derive the parameters from speech or music signals, first for a single source and eventually for a mixture.

In general, the RSAO has implications across the production chain. Unlike the high-level parametric methods standardized in MPEG-4, there is a clear route from RIR recordings in a real room to a set of parameters. Unlike low-level methods to parameterize an RIR (e.g., SIRR and SDM), the parameters map intuitively onto the room geometry, which may open new opportunities for room editing. Unlike the R-WFS approach, the late reverberation is represented by a generalized decay time instead of a sum of plane wave components, meaning that the renderer has more freedom to render diffuse sound based on the target reproduction system. We have not yet implemented these reference techniques and therefore cannot comment at this time on their relative perceptual quality or plausibility; however, on a conceptual level, the RSAO has advantages over these other methods.

7 SUMMARY

In this paper approaches to creating reverberation for object-based audio were discussed, focusing on parametric approaches. Parameters for these approaches can be efficiently captured, edited, transmitted, and modified at the renderer to suit the actual loudspeaker locations and listener personalization settings. The RSAO was proposed as an object type with a representation suitable for recording in a real room, extracting a set of parameters for plausible room reproduction, editing for creative effects, and rendering using signal processing techniques readily available in an object renderer. An implementation example was provided and evaluated with listening tests showing that: the RSAO correctly retained the sense of room size compared to a reference convolution reproduction; editing the parameters altered the perceived room size and source distance; and, that greater envelopment can be achieved where the reproduction system allows. Future work will refine the parameterization process based on perceptual testing, build production tools for parameter editing, and evaluate the overall quality of the synthesized reverberation.

8 ACKNOWLEDGMENTS

This work was supported by EPSRC Programme Grant S3A: Future Spatial Audio for an Immersive Listener Experience at Home (EP/L000539/1). Supporting data is available via https://doi.org/10.15126/surreydata.00812476.

9 REFERENCES

[1] S. Spors, H. Wierstorf, A. Raake, F. Melchior, M. Frank, and F. Zotter, "Spatial Sound with Loudspeakers and its Perception: A Review of the Current State," *Proc. IEEE*, vol. 101, no. 9, pp. 1920–1938 (2013), https://doi.org/10.1109/jproc.2013.2264784.

[2] S. Füg, A. Hölzer, C. Borß, C. Ertel, M. Kratschmer, and J. Plogsties, "Design, Coding and Processing of Metadata for Object-Based Interactive Audio," presented at the *137th Convention of the Audio Engineering Society* (2014 Oct.), convention paper 9097.

[3] B. Shirley, R. Oldfield, F. Melchior, and J.-M. Batke, "Platform Independent Audio," in *Media Production, Delivery and Interaction for Platform Independent Systems*, pp. 130–165 (John Wiley & Sons, Ltd., 2013), https://doi.org/10.1002/9781118706350.ch4.

[4] ITU-R, "Recommendation BS.2076-0, Audio Definition Model," International Telecommunication Union (ITU) (June, 2015).

[5] J. Herre, J. Hilpert, A. Kuntz, and J. Plogsties, "MPEG-H 3D Audio—The New Standard for Coding of Immersive Spatial Audio," *IEEE J. Sel. Topics Signal Process.*, vol. 9, no. 5, pp. 770–779 (May 2015), https://doi.org/10.1109/jstsp.2015.2411578.

[6] G. Potard and I. Burnett, "An XML-Based 3D Audio Scene Metadata Scheme," presented at the *AES 25th International Conference: Metadata for Audio* (2004 Jun.), conference paper 3-3.

[7] N. Peters, T. Lossius, and J. C. Schacher, "The Spatial Sound Description Interchange Format: Principles, Specification, and Examples," *Computer Music J.*, vol. 37, no. 1, pp. 11–22 (2013), https://doi.org/10.1162/comj_a_00167.

[8] H. Stenzel and U. Scuda, "Producing Interactive Immersive Sound for MPEG-H: A Field Test for Sports Broadcasting," presented at the *137th Convention of the Audio Engineering Society* (2014 Oct.), convention paper 9211.

[9] R. Oldfield, B. Shirley, and J. Spille, "An Object-Based Audio System for Interactive Broadcasting," presented at the *137th Convention of the Audio Engineering Society* (2014 Oct.), convention paper 9148.

[10] V. Välimäki, J. D. Parker, L. Savioja, J. O. Smith, and J. S. Abel, "Fifty Years of Artificial Reverberation," *IEEE Trans. Audio Speech Lang. Proc.*, vol. 20, no. 5, pp. 1421–1448 (2012), https://doi.org/10.1109/tasl.2012.2189567.

[11] B. A. Blesser, "An Interdisciplinary Synthesis of Reverberation Viewpoints," *J. Audio Eng. Soc.*, vol. 49, pp. 867–903 (2001 Oct.).

[12] V. Välimäki, J. Parker, L. Savioja, J. O. Smith, and J. Abel, "More Than 50 Years of Artificial Reverberation," presented at the *AES 60th International Conference: DREAMS (Dereverberation and Reverberation of Audio, Music, and Speech)* (2016 Feb.), conference paper K-1.

[13] E. D. Scheirer, R. Väänänen, and J. Huopaniemi, "AudioBIFS: Describing Audio Scenes with the MPEG-4 Multimedia Standard," *IEEE Trans. Multimedia*, vol. 1, no. 32, pp. 237–250 (1999), https://doi.org/10.1109/6046.784463.

[14] R. Väänänen and J. Huopaniemi, "Advanced AudioBIFS: Virtual Acoustics Modeling in MPEG-4 Scene Description," *IEEE Trans. Multimedia*, vol. 6, no. 32, pp. 661–675 (2004), https://doi.org/10.1109/tmm.2004.834864.

[15] J. Merimaa and V. Pulkki, "Spatial Impulse Response Rendering I: Analysis and Synthesis," *J. Audio Eng. Soc.*, vol. 53, pp. 1115–1127 (2005 Dec.).

[16] S. Tervo, J. Pätynen, A. Kuusinen, and T. Lokki, "Spatial Decomposition Method for Room Impulse Responses," *J. Audio Eng. Soc.*, vol. 61, pp. 17–28, (2013 Jan./Feb.).

[17] L. Remaggi, P. J. B. Jackson, and P. Coleman, "Estimation of Room Reflection Parameters for a Reverberant Spatial Audio Object," presented at the *138th Convention of the Audio Engineering Society* (2015 May), convention paper 9258.

[18] P. Coleman, A. Franck, P. Jackson, R. Hughes, L. Remaggi, and F. Melchior, "On Object-Based Audio with Reverberation," presented at the *AES 60th International Conference: DREAMS (Dereverberation and Reverberation of Audio, Music, and Speech)* (2016 Feb.), conference paper 2-3.

[19] M. R. Schroeder, "Statistical Parameters of the Frequency Response Curves of Large Rooms," *J. Audio Eng. Soc.*, vol. 35, pp. 299–306 (1987 May).

[20] M. Vorländer, "Simulation of the Transient and Steady-State Sound Propagation in Rooms Using a New Combined Ray-Tracing/Image-Source Algorithm," *J. Acoust. Soc. Am.*, vol. 86, no. 32, pp. 172–178 (1989), https://doi.org/10.1121/1.398336.

[21] S. E. Olive and F. E. Toole, "The Detection of Reflections in Typical Rooms," *J. Audio Eng. Soc.*, vol. 37, pp. 539–553 (1989 Jul./Aug.).

[22] S. Bech, "Timbral Aspects of Reproduced Sound in Small Rooms II," *J. Acoust. Soc. Am.*, vol. 99, no. 32, pp. 3539–3550 (1996), https://doi.org/10.1121/1.414952.

[23] N. Kaplanis, S. Bech, S. H. Jensen, and T. van Waterschoot, "Perception of Reverberation in Small Rooms: A Literature Study," presented at the *AES 55th International Conference: Spatial Audio* (2014 Aug.), conference paper P-3.

[24] P. Zahorik, D. S. Brungart, and A. W. Bronkhorst, "Auditory Distance Perception in Humans: A Summary of Past and Present Research," *Acta. Acust. united Ac.*, vol. 91, no. 32, pp. 409–420 (2005).

[25] F. Rumsey, *Spatial Audio* (Oxford, UK: Focal Press, 2001), https://doi.org/10.4324/9780080498195.

[26] G. Theile, "Multichannel Natural Recording Based on Psychoacoustic Principles," presented at the *108th Convention of the Audio Engineering Society* (2000 Feb.), convention paper 5156.

[27] G. Theile and H. Wittek, "Principles in Surround Recordings with Height," presented at the *130th Convention of the Audio Engineering Society* (2011 May), convention paper 8403.

[28] H. Lee and C. Gribben, "On the Optimum Microphone Array Configuration for Height Channels," presented at the *134th Convention of the Audio Engineering Society* (2013 May), eBrief 93.

[29] K. Hamasaki, T. Shinmura, S. Akita, and K. Hiyama, "Approach and Mixing Technique for Natural Sound Recording of Multichannel Audio," presented at the *AES 19th International Conference: Surround Sound—Techniques, Technology, and Perception* (2001 Jun.), conference paper 1878.

[30] J. Francombe, T. Brookes, R. Mason, R. Flindt, P. Coleman, Q. Liu, and P. Jackson, "Production and Reproduction of Program Material for a Variety of Spatial Audio Formats," presented at the *138th Convention of the Audio Engineering Society* (2015 May), eBrief 199.

[31] G. Thomas, A. Engström, J.-F. Macq, O. A. Niamut, B. Shirley, and R. Salmon, "State of the Art and Challenges in Media Production, Broadcast and Delivery," in *Media Production, Delivery and Interaction for Platform Independent Systems*, pp. 5–73 (John Wiley & Sons, Ltd., 2013), https://doi.org/10.1002/9781118706350.ch2.

[32] R. Albrecht and T. Lokki, "Adjusting the Perceived Distance of Virtual Speech Sources by Modifying Binaural Room Impulse Responses," *Proc. 19th Int. Conf. on Auditory Display (ICAD2013). Lodz, Poland. 6–9 July* (2013).

[33] T. Carpentier, T. Szpruch, M. Noisternig, and O. Warusfel, "Parametric Control of Convolution Based Room Simulators," *International Symposium on Room Acoustics (ISRA)* (2013).

[34] J.-M. Jot, "Efficient Models for Reverberation and Distance Rendering in Computer Music and Virtual Audio Reality," *Proc. Int. Computer Music Conference*, Thessaloniki, Greece (1997).

[35] F. Melchior, C. Sladeczek, A. Partzsch, and S. Brix, "Design and Implementation of an Interactive Room Simulation for Wave Field Synthesis," presented at the *AES 40th International Conference: Spatial Audio: Sense the Sound of Space* (2010 Oct.), conference paper 7-5.

[36] J. Schmidt and E. F. Schroeder, "New and Advanced Features for Audio Presentation in the MPEG-4 Standard," presented at the *116th Convention of the Audio Engineering Society* (2004 May), convention paper 6058.

[37] J.-M. Trivi and J.-M. Jot, "Rendering MPEG-4 AABIFS Content through a Low-Level Cross-Platform 3D Audio API," *Proc. ICME'02*, Lausanne, Switzerland, vol. 1, pp. 513–516 (2002), https://doi.org/10.1109/icme.2002.1035831.

[38] H. Bai, G. Richard, and L. Daudet, "Late Reverberation Synthesis: From Radiance Transfer to Feedback Delay Networks," *IEEE/ACM Trans. Audio. Speech Lang. Proc.*, vol. 23, no. 32, pp. 2260–2271 (2015), https://doi.org/10.1109/taslp.2015.2478116.

[39] V. Pulkki, "Spatial Sound Reproduction with Directional Audio Coding," *J. Audio Eng. Soc.*, vol. 55, pp. 503–516 (2007 Jun.).

[40] A. Politis, J. Vilkamo, and V. Pulkki, "Sector-Based Parametric Sound Field Reproduction in the Spherical Harmonic Domain," *IEEE J. Sel. Topics Signal Process.*, vol. 9, no. 5, pp. 852–866 (Aug. 2015), https://doi.org/10.1109/jstsp.2015.2415762.

[41] A. Politis, T. Pihlajamäki, and V. Pulkki, "Parametric Spatial Audio Effects," in *Proc. 15th Int. Conf. Digital Audio Effects (DAFx-12)*, York, UK (2012).

[42] V. Pulkki, "Virtual Sound Source Positioning Using Vector Base Amplitude Panning," *J. Audio Eng. Soc.*, vol. 45, pp. 456–466 (1997 Jun.).

[43] V. Pulkki and J. Merimaa, "Spatial Impulse Response Rendering II: Reproduction of Diffuse Sound and Listening Tests," *J. Audio Eng. Soc.*, vol. 54, pp. 3–20 (2006 Jan./Feb.).

[44] E. M. Hulsebos and D. de Vries, "Parameterization and Reproduction of Concert Hall Acoustics Measured with a Circular Microphone Array," presented at the *112th Convention of the Audio Engineering Society* (2002 Apr.), convention paper 5579.

[45] E. M. Hulsebos, *Auralization Using Wave Field Synthesis*, Ph.D. thesis, Delft University of Technology (2004).

[46] L. Savioja, J. Huopaniemi, T. Lokki, and R. Väänänen, "Creating Interactive Virtual Acoustic Environments," *J. Audio Eng. Soc.*, vol. 47, pp. 675–705 (1999 Sep.).

[47] J. Nowak, J. Liebetrau, and T. Sporer, "On the Perception of Apparent Source Width and Listener Envelopment in Wave Field Synthesis," *5th Workshop on Quality of Multimedia Experience (QoMEX)*, Klagenfurt, Austria, pp. 82–87 (2013), https://doi.org/10.1109/qomex.2013.6603215.

[48] F. Melchior, *Investigations on Spatial Sound Design Based on Measured Room Impulse Responses*, Ph.D. thesis, Delft University of Technology (2011).

[49] L. Beranek, *Concert Halls and Opera Houses: Music, Acoustics, and Architecture*, 2nd ed. (Springer-Verlag, New York, 2004), https://doi.org/10.1007/978-0-387-21636-2.

[50] J.-M. Jot, "An Analysis/Synthesis Approach to Real-Time Artificial Reverberation," *Proc. ICASSP'92*, San Francisco, CA, USA, pp. 221–224 (1992), https://doi.org/10.1109/icassp.1992.226080.

[51] P. Coleman, L. Remaggi, and P. J. B. Jackson, "S3A Room Impulse Responses [dataset]," https://doi.org/10.15126/surreydata.00808465 (2015).

[52] L. Remaggi, P. J. B. Jackson, P. Coleman, and J. Francombe, "Visualization of Compact Microphone Array Room Impulse Responses," presented at the *139th Convention of the Audio Engineering Society* (2015 Oct.), eBrief 218.

[53] L. Remaggi, P. J. B. Jackson, P. Coleman, and W. Wang "Acoustic Reflector Localization: Novel Image Source Reversion and Direct Localization Methods," *IEEE/ACM Trans. Audio. Speech Lang. Proc.*, vol. 25, no. 2, pp. 296–309 (2017), https://doi.org/10.1109/TASLP.2016.2633802.

[54] J. Makhoul, "Linear Prediction: A Tutorial Review," *Proc. of the IEEE*, vol. 63, no. 32, pp. 561–580 (1975), https://doi.org/10.1109/proc.1975.9792.

[55] A. Lindau, L. Kosanke, and S. Weinzierl, "Perceptual Evaluation of Model- and Signal-Based Predictors of the Mixing Time in Binaural Room Impulse Responses," *J. Audio Eng. Soc.*, vol. 60, pp. 887–898 (2012 Nov.).

[56] R. Mason, "Installation of a Flexible 3D Audio Reproduction System into a Standardized Listening Room," presented at the *140th Convention of the Audio Engineering Society* (2016 May), eBrief 256.

[57] Bang & Olufsen, "Music for Archimedes [Compact Disc]," *CD B&O*, vol. 101 (1992).

[58] J. S. Abel and P. Huang, "A Simple, Robust Measure of Reverberation Echo Density," presented at the *121st Convention of the Audio Engineering Society* (2006 Oct.), convention paper 6985.

[59] V. Valimaki, H.-M. Lehtonen, and M. Takanen, "A Perceptual Study on Velvet Noise and its Variants at Different Pulse Densities," *IEEE Trans. Audio Speech Lang. Proc.*, vol. 21, no. 32, pp. 1481–1488 (2013), https://doi.org/10.1109/tasl.2013.2255281.

THE AUTHORS


Philip Coleman


Andreas Franck


Philip J. B. Jackson


Richard J. Hughes


Luca Remaggi


Frank Melchior

Philip Coleman joined the Centre for Vision, Speech and Signal Processing, University of Surrey, UK, in 2010, earning his Ph.D. in 2014 on the topic of personal sound zones. He is currently working in the centre as a research fellow on the project S3A: Future Spatial Audio for an Immersive Listening Experience at Home, with a focus on recording and editing object-based content. His research interests include sound field control, loudspeaker and microphone array processing, and spatial audio. Previously, he received the B.Eng. degree in electronic engineering with music technology systems in 2008 from the University of York, UK, and M.Sc. with distinction in multimedia signal processing and communication from the University of Surrey, UK, in 2010.

•

Andreas Franck received the Diploma degree in computer science and the Ph.D. degree in electrical engineering, both from the Ilmenau University of Technology, Germany. Since 2004 he has been with the Fraunhofer Institute for Digital Media Technology IDMT, Ilmenau, Germany. In 2014 he joined the Institute of Sound and Vibration research, University of Southampton, UK, as a postdoctoral research fellow. He is currently working in the EPSRC-funded project S3A: Future Spatial Audio for an Immersive Listening Experience at Home. His research interests include spatial and object-based audio, efficient reproduction algorithms, audio signal processing, and architecture and implementation of audio software. Dr. Franck is a member of the Audio Engineering Society, IEEE, and IEEE Signal Processing Society.

•

Philip Jackson is Senior Lecturer in speech and audio processing at the Centre for Vision, Speech & Signal Processing (University of Surrey, UK) which he joined as in 2002, following a postdoctoral research fellowship (University of Birmingham, UK), with MA in Engineering (Cambridge University, UK) and Ph.D. in electronic engineering (University of Southampton, UK). With Prof Mark Plumbley and Dr. Wenwu Wang in CVSSP, he leads the machine audition research group (A-lab) of approximately 25 researchers. His acoustical and spoken-language processing research contributions span projects in active noise control, speech production, source separation, automatic speech recognition, articulation modeling, audio-visual speech enhancement, visual speech synthesis, and spatial audio quality evaluation (BALTHASAR, DANSA, Dynamic Faces, QESTRAL, UDRC, POSZ and S3A). He has over 100 academic publications in journals, conference proceedings and books, and several patents (Google h-index=15). He is associate editor for *Computer Speech & Language* (Elsevier) and reviews, e.g., for *Journal of the Acoustical Society of America* and *IEEE Transactions on Audio, Speech & Language Processing*.

•

Richard Hughes received a B.Sc. first-class honors degree in audio technology from the University of Salford, UK, in 2006, before completing a Ph.D. at the same institution in 2011 in the area of architectural acoustics. From 2012 to 2014 he worked as a Research Associate at the University of Manchester, UK. He is currently working in the Acoustics Research Centre at Salford as a research fellow as part of the EPSRC-funded project S3A: Future Spatial Audio for an Immersive Listener Experience at Home. His research interests include among others: room acoustics; acoustic modeling of rooms, diffuser design and application; array theory and multichannel systems; and digital signal processing.

•

Luca Remaggi received the B.Sc. and M.Sc. degrees in electronic engineering from Universitá Politecnica delle Marche, Italy, in 2009 and 2012 respectively, having prepared his final dissertation on a placement with Aalto University School of Science and Technology, Finland. He then worked for one year at the Loccioni Group, Italy. He is currently pursuing the Ph.D. degree at the Centre for Vision, Speech and Signal Processing, at the University of Surrey, UK. He is investigating spatial audio and source separation, exploiting information given by multipath propagation analysis.

•

Frank Melchior received the Dipl.-Ing. degree in media technology from the Ilmenau University of Technology, Germany in 2003 and the Dr ing. degree from Delft University of Technology, The Netherlands, in 2011. Since 2012 he is leading the audio research group and the BBC Audio Research Partnership at BBC Research and Development. From 2009 to 2012 he was the Chief Technical Officer and Director Research and Development at IOSONO GmbH, Germany. From 2003 to 2009 he worked as a researcher at the Fraunhofer Institute Digital Media Technology, Germany. He holds several patents and has authored and co-authored a number of papers in international journals and conference proceedings. His research is currently focused on next generation audio for broadcast and interdisciplinary innovations for new audience experiences in an IP based broadcast world of the future. Dr. Melchior is member of the Audio Engineering Society, the German Acoustical Society and represents the BBC in the International Telecommunication Union and the European Broadcasting Union.