OPEN ACCESS Freely available online

C. Yeoward, R. Shukla, R. Stewart, M. Sandler AND J.D. Reiss, "Real-Time Binaural Room Modeling for Augmented Reality Applications"
J. Audio Eng. Soc., vol. 69, no. 11, pp. 818–833, (2021 November).
DOI: https://doi.org/10.17743/jaes.2021.0046

Real-Time Binaural Room Modeling for Augmented Reality Applications

CHRISTOPHER YEOWARD,[1] RISHI SHUKLA,[1] REBECCA STEWART,[2]
(chris.yeoward@gmail.com) (r.shukla@qmul.ac.uk) (r.stewart@imperial.ac.uk)
MARK SANDLER,[1] *AES Fellow* **AND JOSHUA D. REISS,[1]** *AES Fellow*
(mark.sandler@qmul.ac.uk) (joshua.reiss@qmul.ac.uk)

[1]*Centre for Digital Music, School of Electronic Engineering and Computer Science, Queen Mary University of London, London E1 4NS, UK*
[2]*Dyson School of Design Engineering, Faculty of Engineering, Imperial College London, London SW7 2AZ, UK*

This paper proposes and evaluates an integrated method for real-time, head-tracked, 3D binaural audio with synthetic reverberation. Virtual vector base amplitude panning is used to position the sound source and spatialize outputs from a scattering delay network reverb algorithm running in parallel. A unique feature of this approach is its realization of interactive auralization using vector base amplitude panning and a scattering delay network, within acceptable levels of latency, at low computational cost. The rendering model also allows direct parameterization of room geometry and absorption characteristics. Varying levels of reverb complexity can be implemented, and these were evaluated against two distinct aspects of perceived sonic immersion. Outcomes from the evaluation provide benchmarks for how the approach could be deployed adaptively, to balance three real-time spatial audio objectives of envelopment, naturalness, and efficiency, within contrasting physical spaces.

0 INTRODUCTION

Effective simulation of real-world acoustic spaces continues to challenge the field of spatial audio for immersive media. Algorithms strive toward physical and perceptual accuracy using implementations that also optimize efficiency and flexibility. In the case of augmented reality (AR), the ability to navigate one's environment is crucial to achieving an immersive experience. Audio-augmented reality (AAR) is a relatively new area of activity in both sonic interaction design research and commercial product development. It describes devices and applications that enhance real-world experiences with sound sources that seemingly emanate from the user's current environment [1–3]. AAR is therefore a very particular context in which managing the trade-off between fidelity and processing load becomes especially acute. Evidently in these instances it is desirable to implement solutions that run on compact, low-power devices but also produce a plausible acoustic effect customized to the listener's location. Not only must these algorithms be lightweight, they should also minimize configuration to different spaces, layouts, and orientations.

Scattering delay network (SDN) reverberation has been presented as a simplified means of simulating physical spaces, controlled by high-level parameters that determine properties of the room and source and listener position. It has been shown to produce outputs closely aligned with Sabine and Eyring predictions of reverb time, frequency response, and room surface absorption [4]. Additional analysis has shown closely matched behavior to an image source model (ISM) implementation across the same three metrics but at considerably lower computational expense [5]. Further investigation has also indicated that the algorithm synthesizes acoustics with perceptually favorable results. In evaluation based on noninteractive simulation of two listening rooms, SDN was judged more natural than binaural room impulse response, ray-tracing, and feedback delay network alternatives [6].

However, two areas require further exploration to establish the suitability of SDN for AR applications. First, a model for binaural SDN has not yet been presented within a real-time architecture suitable for mobile or wearable computing devices. Second, binaural SDN has not been perceptually assessed using fully immersive, head-tracked sound scene simulation.

This paper advances and evaluates an interactive, integrated method for rendering a sound source within SDN-modeled rooms—i.e., the process commonly known as *auralization* [7]. 3D spatialization of the source, combined with varying complexities of SDN reverb, is achieved using

pure vector base amplitude panning (VBAP) [8] and binauralization via virtual loudspeakers [9]. Low-latency tracking of head rotation is incorporated to enable full immersion within three degrees of freedom (3DoF). Two specific research questions are then addressed through a subjective listening test:

- How do the variations in SDN spatialization complexity affect perception of the resulting reverberation?
- What considerations for using SDN reverberation in AR contexts can be drawn from this investigation?

Sec. 1 reviews related work contextualizing the rationale for using VBAP, SDN, and virtual loudspeakers to implement an interactive, real-time, adaptive binaural rendering engine for AR applications. Sec. 2 outlines the software and hardware implementation used here for evaluating the real-time binaural SDN method. The design incorporated five reverberation algorithm variants for the subsequent investigation. The method used to evaluate their perceived spatial impression ("envelopment") and realism ("naturalness") is described in SEC. 3. Results from the assessment are presented in SEC. 4, followed by analysis and summary conclusions in SECS. 5 and 6.

1 RELATED WORK

Special focus is given to concepts related directly to the implementation described and evaluated in the latter half of this paper.

1.1 Ambisonics

For fully immersive media, such as 3D gaming or 360° cinematic experiences, Ambisonics has emerged as the dominant format for representing or designing spatial audio scenes [10–12]. Firstly, its inherent capacity to support either recording or synthesis of complex sound fields makes it an ideal format for these applications, which seek to reflect the detail of true or imagined acoustic environments in full. Secondly, Ambisonic B-format encoding permits flexible decoding and reproduction by varied loudspeaker configurations [13, 14]. This flexibility also allows Ambisonics to be realized binaurally by simulating virtual loudspeaker arrays over headphones [9].

It is now well established that the simplest implementation—first order Ambisonics (FOA)—does not adequately reproduce differences in inter-aural timing (ITD) or intensity and filtering (ILD), nor the monoaural spectral cues required to render spatial scenes with sufficient localization precision or tonal transparency [15–17]. Improving the performance of B-format audio either requires use of higher order Ambisonics (HOA) or more involved decoding processes. Use of HOA necessarily increases computational demands. Optimal application of more sophisticated decoding methods is an active area of research that includes approaches designed to improve binaural reproduction of Ambisonic signals at both first and higher orders [18–20].

1.2 Vector Base Amplitude Panning

Vector base amplitude panning (VBAP) is an alternative sound spatialization technique of particular relevance to virtual auditory displays (VAD) [8, 21]. The approach can be applied to maximize sharpness in sound localization and clarity of tone within predetermined spatial areas. Unlike Ambisonics, only the minimum number of speakers required to render a given location are ever deployed (either one, two, or three). VBAP also permits flexible positioning of loudspeakers in which higher concentrations improve fidelity. Ambisonics, on the other hand, is generally most effective when speakers are positioned uniformly and symmetrically around the listener, using the minimum number required to reproduce the given order [18]. Vertically sparse VBAP arrays are known to present errors in elevation cue representation [22], whereas laterally situated triplets distort ITD and ILD cues toward the median plane [23]. Nevertheless, specific configurations of VBAP speaker layouts have been shown to offer substantially improved horizontal localization cue accuracy over comparable first and also second order Ambisonics setups [24–26].

VBAP does not constitute or offer a surround sound audio interchange format like Ambisonics. Instead individual source positions are defined and processed according to spatial coordinates relative to the listener. This limits VBAP to rendering of individual point sources rather than enveloping sound fields. The directional audio coding (DirAC) algorithm proposed a means of bridging the generality and flexibility of Ambisonics B-format with the improved rendering precision of VBAP [27, 28]. However, if sounds are being generated and spatialized in a fully integrated VR or AR system—and where the output speaker array configuration is already known—there is no benefit to an encoding/decoding process. Doing so would introduce computational overhead and degraded spatial and tonal clarity that manifest with Ambisonics. If a pure VBAP rendering approach is used, one remaining obstacle is how to produce plausibly enveloping sound fields via point source rendering.

1.3 Binaural Synthesis of Spatial Audio Formats

Virtual loudspeaker spatialization was first applied as a binaural decoding method for B-format Ambisonics [9]. It uses a sparse selection of head-related impulse response (HRIR) measurements to simulate an array of loudspeakers. The benefits of applying a virtual loudspeaker approach to binaural VBAP rendering—specifically the ability to concentrate improved rendering resolution in the frontal field—are given detailed presentation in [24]. Because it is impractical to capture personalized HRIR measurements, binaural systems typically utilize head-related transfer function (HRTF) databases of either human or dummy head measurements, for example [29].

1.4 Head-Tracked Interactive Binaural Synthesis

3DoF head-tracking (pitch, yaw, and roll) is applied in binaural synthesis to counter-rotate the spatial sound scene against changes in the listener's orientation. Doing so fixes audio sources relative to the listener's surroundings, not the headphones [30, 31]. This interactive processing improves both sense of externalization [32] and localization of sources for the listener—particularly when differentiating between positions in front or behind [33]. Head-tracking has also been shown to have a greater effect on improving localization than either using individualized (rather than generic) HRTFs or applying reverberation [34]. Rendering a static virtual auditory scene, improving externalization, and enhancing perceived location are all crucial attributes to creating a coherent VR or AR experience.

With virtual loudspeaker implementations, head-tracking is usually realized by rotating the sound scene itself, with the head remaining fixed relative to the speakers. In the case of virtual FOA, the relative positions of sources within a sound field are encoded into the B-format signal. Rotation transformations can be applied to the Ambisonics channels before decoding [35]. However the mathematics required to rotate virtual HOA sound fields is nontrivial [12]. For VBAP, on the other hand, although each individual point source requires re-positioning to account for revised head positions, this only ever requires a simple vector rotation to update each source's position. VBAP arrays can therefore be concentrated and retained (irrespective of head movement) in the frontal hemisphere, where perceptual acuity is greater and where enhanced fidelity is typically most desirable in an interactive audio environment [24].

1.5 Scattering Delay Network Reverberation

SDN reverberation can be characterized as a combination of earlier digital waveguide network (DWN) and ray-tracing ISM approaches [4, 5]. A schematic is shown in Fig. 1. It consists of a DWN with scattering junction nodes at points of the first order reflections, positions of which are determined by basic geometric ray tracing calculations. SDN therefore accurately models early reflections and more coarsely approximates higher orders. In doing so it provides physical accuracy closer to room simulation models but maintains the lower processing requirements of other delay network approaches. The result is a computationally simplified, coherent reverberation model with early and late components governed by the same set of parameters [36].

The contribution from each node contains information about early reflections and diffusion generated by sound scattered between the junctions. The delay times and length of the reverberation are controlled by the room geometry and surface absorption characteristics. Lower absorption and larger rooms both result in slower energy reduction over time. Larger rooms will also increase the delay between the direct sound and first reflections.

The source signal is processed to arrive at the virtual microphone and each node location via a primary set of delay lines. This input is propagated within the network

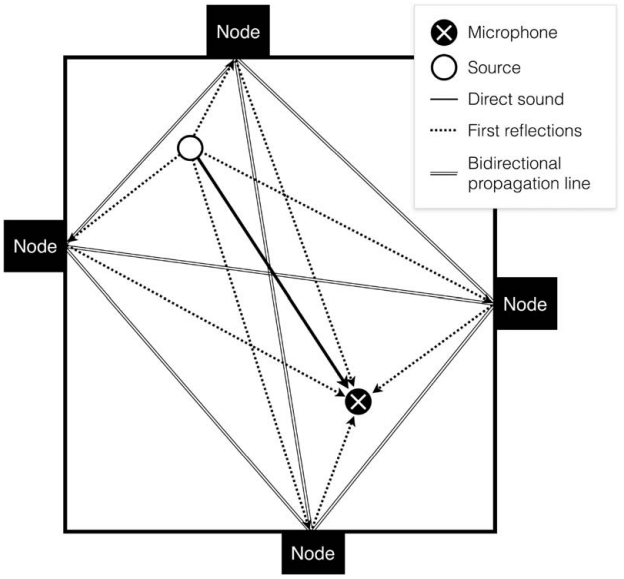

Fig. 1. Conceptual model of a scattering delay network. (Recreated here from the figures originally published in [4] and [5].)

via a second set of bidirectional delay lines. Each node therefore receives a summed input wave vector comprising the following:

1. a delayed source signal, and
2. delayed contributions from each of the other nodes.

The input wave vector for each node is multiplied by a scattering matrix, with the result sent to each other node in the network. Various scattering matrices can be used and these can differ between nodes. The choice of matrix affects the computational complexity of the scattering operation and the echo density of the resultant diffusion. After the scattering operation, the input wave vectors for each node are sent to a third set of delay lines, which the return node outputs to the microphone.

The length of each delay line is determined by the computed physical distance between the two respective points. For example, the delay from the source $\mathbf{x}_S$ to the microphone $\mathbf{x}_M$, $D_{S,M}$, is given by the following:

$$D_{S,M} = \lfloor \frac{F_s ||\mathbf{x_S} - \mathbf{x_M}||}{c} \rfloor$$

where c is the speed of sound in air and F_s is the sampling frequency.

A number of further propagation features are simulated in the SDN model, including the following:

- directivity shaping for the source orientation,
- distance attenuation due to air absorption,
- directivity shaping for the microphone orientation,
- surface absorption characteristics, and
- delay modulation to counteract *flutter* effect.

Each of these is described in the full definition of the SDN algorithm [5].

1.6 Summary

AR applications—and AAR or VAD use cases in particular—have a clear need for accurate spatial rendering of audio sources for sonic localization and interaction purposes, delivered using devices with constrained processing power. VBAP offers a level of spatial and timbral clarity that is not afforded by FOA and that might not be technically feasible to replicate via HOA on low-capability devices. The potential advantages of using a virtual loudspeaker approach to render binaural VBAP have been demonstrated in previous research [24]. However, an effective means of rendering plausible spatial reverberation with this approach, tailored to different physical environments, is yet to be established.

The SDN model marries well with VBAP's point-source-oriented spatialization method. Both direct sound and early reflections can be treated as individual sources, each panned within a virtual loudspeaker array that is convolved with corresponding HRIRs. Because the diffuse component is transmitted along the same paths as the first order reflections, the late element of the reverberation is spatialized in the same way. By incorporating low-latency head-tracking, the research that follows is the first known evaluation of SDN using a real-time, interactive model for rendering.

Furthermore separate evidence has shown that some reverberation algorithms do not necessarily yield significantly more convincing results when configured with more complex and detailed spatialization. For instance, mono reverberation was found to be almost indistinguishable from equivalent HOA spatialized renderings by groups of non-expert listeners using both static (non-head-tracked) [37] and interactive (head-tracked) [38] presentation. Because anticipated computational constraints are a motivation for the enquiry, this paper also investigates perceptual differences between five configurations of SDN reflection node spatialization.

2 TECHNICAL IMPLEMENTATION

The auralization system comprises four main components, which are reflected in Fig. 2:

1. Reverberation synthesis via SDN,
2. Spatialization of direct sound and reflections via VBAP,
3. Scene rotation based on head-tracking, and
4. Virtual loudspeaker binaural synthesis by HRIR convolution.

2.1 Reverberation Synthesis

The SDN reverb synthesis is written as an object-oriented C++ module and closely follows the implementation described in [5]. The *isotropic scattering matrix* is used as it offers improved computational efficiency alongside a high echo density, such that:

$$A = \frac{2}{K}\mathbf{1}\mathbf{1}^T - \mathbf{I}$$

where K is the network order, $\mathbf{1} = [1_1, ..., 1_k]^T$, and $\mathbf{I}$ is the identity matrix. This matrix uniformly processes and distributes a sample incoming from one node to each of the others, reflecting a small portion back to the originating node. A cuboidal room of six surfaces is modeled, where $K = 5$. With this configuration, the scattering operation can be executed with five additions, one multiplication, and five subtractions ($2K + 1$ operations):

```
sum = 0;
for all other nodes in network do
    sum = sum + input from current node;
end
scale sum by 2/K;
for all other nodes in network do
    output for current node = sum - input from
      current node;
end
```

To spatialize node outputs, their positions relative to the listener are required. All node azimuths and elevations are pre-calculated at system startup, ready for integration with VBAP spatial rendering.

2.1.1 Directivity Simulation

To limit computational load and for simplicity of implementation, the source is always assumed to emit sound uniformly. Microphone directivity is also treated as omnidirectional within the SDN algorithm. Both the direct path and reflection node outputs are subsequently spatialized using the VBAP configuration and HRIR processing. Thus all streams output from the SDN reverberation component are ultimately filtered to simulate the directive hearing of a human listener, rather than a microphone.

2.1.2 Surface Absorption

The inclusion of second or third order infinite impulse response (IIR) filters for each destination node, at each scattering junction, increases considerably the computational load, for limited perceptual gain. Surfaces are therefore assigned frequency independent scattering characteristics. Average values for random incidence absorption coefficients α are taken from [7] for various surfaces. The floor is modeled as carpet on concrete with $\alpha = 0.18$, walls as hard surfaces with $\alpha = 0.0343$, and ceiling as perforated tiles with $\alpha = 0.7$.

2.1.3 Air Absorption

For realistic attenuation of higher frequencies, air absorption must be taken into consideration. The specific SDN implementation is not detailed in [5] but instead Moorer's definition is referenced. Moorer describes using a first order low-pass filter as a rough approximation of air absorption [39], with the cutoff adjusted by the single parameter g. To

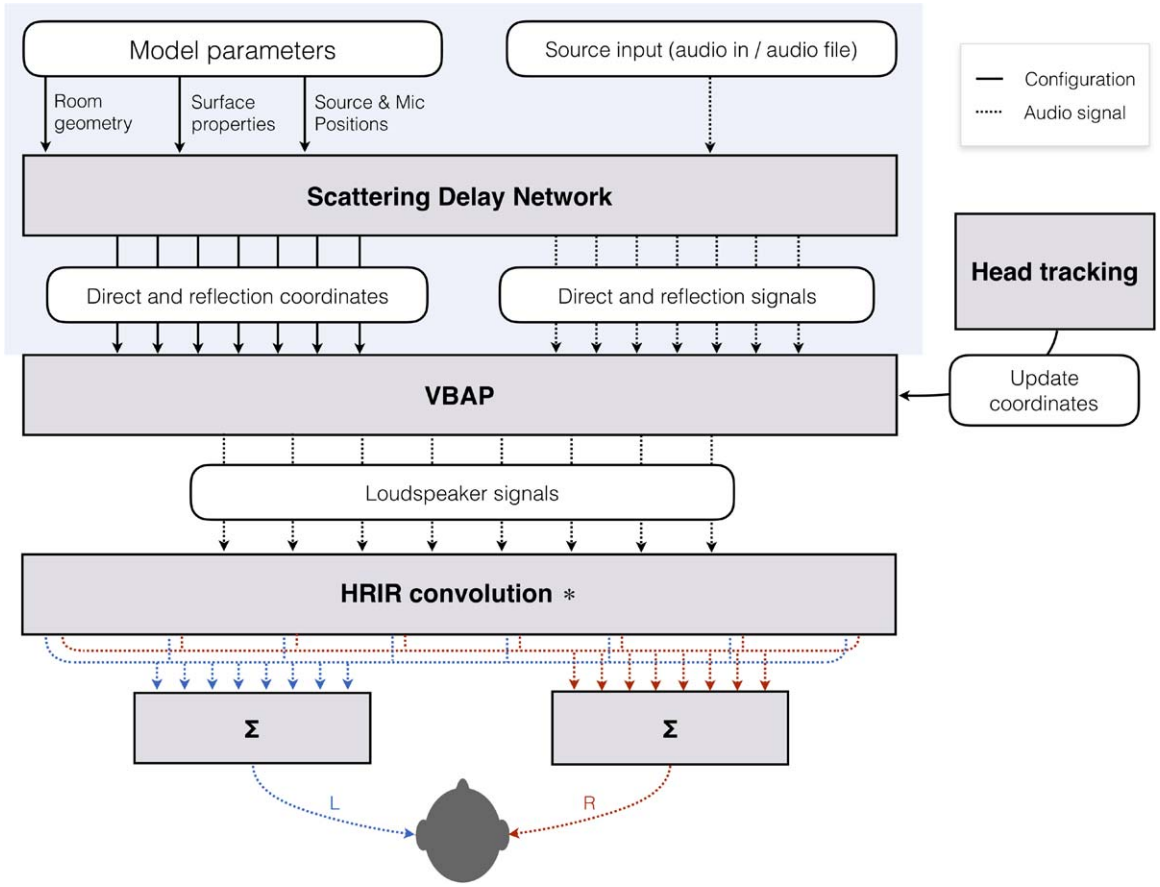

Fig. 2. Flow diagram of the system implementation.

ensure unconditional stability of the system, the output is normalized:

$$T(z) = \frac{1-g}{1-gz^{-1}}$$

where g is between 0 and 1. With humidity taken to be 50% and the sampling rate fixed at 44.1 kHz, the following function dynamically determines the value of g from the distance d:

$$g = \frac{1}{5}\log\left(\frac{d}{3}+1\right)$$

2.1.4 Modulated Delay Lines

To mitigate *flutter*, modulated delay lines are implemented between the nodes, as discussed and advocated in [40]. A continual variation is applied to the length of the propagation lines connecting each node (i.e., not any of those connected to the source or mic position). This is implemented using linearly interpolated fractional delay lines to avoid distortion and digital "zip noise." Modulation is applied with an amplitude value of 0.003 and frequency between 0 and 2 Hz, assigned randomly to each node.

2.1.5 Model Verification

Appendix A.2 outlines the three steps taken to ensure technical validation of the SDN implementation in line with the published specification.

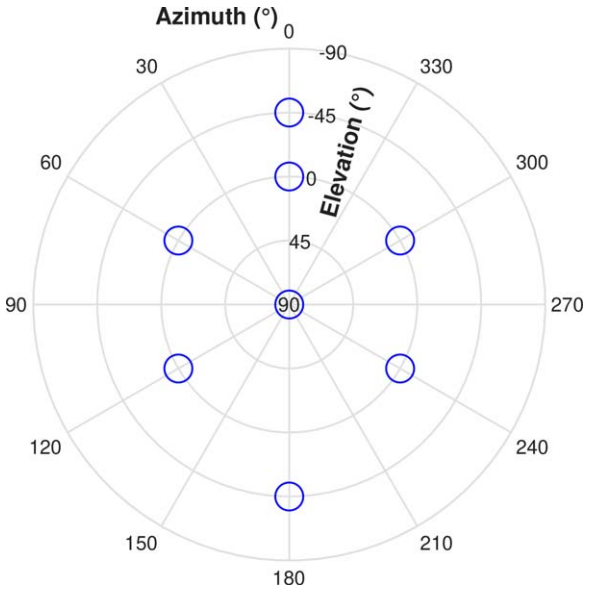

Fig. 3. Virtual VBAP loudspeaker positions (identified by O).

2.2 Virtual VBAP

The virtual VBAP model used in the system uses eight virtual loudspeakers, as shown in Fig. 3. The layout achieves a pseudospherical rendering capability, symmetrically arranged across all three axes, other than one speaker positioned at 0° azimuth and 0° elevation for improved

frontal resolution. This configuration was conceived with close reference to [25] and to enable the direct comparison against a first order Ambisonic cube layout outlined in [24].

Both the virtual VBAP and FOA systems were initially implemented without integrated reverberation on an embedded Linux platform with single 1 GHz ARM Cortex-A8 CPU core. A partial implementation of the full auralization system described here was successfully developed on the embedded device but with only five reflection surfaces and without air absorption or fractional delay lines. The essence of this program was ported to a bespoke VST plugin running on MacOS to immediately access a limited amount of additional computing resource that permitted six reflection surfaces, plus all the extended SDN features as described.

2.3 HRIR Convolution

The KEMAR dummy head HRTF measurements from the SADIE II database are used for virtual speaker simulation [29]. No binaural personalization process was included in the perceptual evaluation described in SEC. 3, so a generic HRTF set was deemed most suitable. The SADIE II database offers HRIRs that are 256 samples long (as opposed to the more standard length of 512). This mitigates the substantial computational costs incurred by convolution processes and so is more consistent with the overall investigation's focus on efficiency of implementation, albeit with an accompanying reduction in binaural fidelity.

2.4 Head-Tracking

Head-tracking for the VST implementation utilizes the open source Mr Headtracker solution [41]. A BNO055 inertial measurement unit sensor[1] is connected to an Arduino, which emits MIDI events handled by the VST plugin. The module produces yaw, pitch, and roll values used by the VBAP system to rotate initial source positions. The majority of the computational load incurred in the VBAP implementation arises from vector rotations required to manipulate sound source positions in line with head-tracking.

2.5 Variations in Complexity

Four variations from the fully spatialized reverberation model were developed for perceptual evaluation:

1. Full spatialization (FullSpat): Maximum spatialization, in which the output of each node is positioned as an independent source in the VBAP rendering component.
2. Lateral spatialization (LatSpat): Partial spatialization, in which the output of each wall node is spatialized as an independent source, but the contributions from floor and ceiling nodes are simply distributed equally to all loudspeakers.
3. Mono with independent streams (Mono IS): All node outputs are summed together, but this combined reverberation stream is then directed separately to each node position, as multiple independent sources in the VBAP rendering component.
4. Mono distributed (Mono D): All node outputs are summed together and distributed equally to all loudspeakers.
5. Mono panned (Mono P): All node outputs are summed together and panned with the spatialized anechoic source signal.

Table 1. Relative complexity for each of the algorithm variations.

Algorithm	VBAP Inputs
1 - FullSpat	7
2 - LatSpat	5
3 - Mono IS	7
4 - Mono D	1
5 - Mono P	1

The relative complexities for each of the algorithms are shown in Table 1. Each renders the reverberation in the same way, with the savings in computational load arising from reduction of inputs to the VBAP spatialization component. For the multiple input algorithms, complexity scales linearly with respect to the number of sources.

LatSpat offers a minimal simplification of the full model by collapsing the direction-specific simulation of reverberation to the horizontal plane only. *Mono IS* is included to consider whether spatializing any other mono artificial reverb model along the early reflection points might have a comparable effect to the primary method. *Mono D* is substantially simplified but should still exhibit a reasonable degree of envelopment, as the effect is generated from all loudspeakers surrounding the listener. It should also sound relatively natural. *Mono P* was chosen to be a noticeably less convincing version of the processed signal so was expected to score the lowest of all the reverberant signals against both criteria.

3 RESEARCH DESIGN

To evaluate the system and its variations in reverb complexity systematically two rooms, three sound sources, and two auditory characteristics were defined and selected.

3.1 Room Simulations

Two virtual rooms were specified to provide different reverberation lengths:

- Small office: 5 m × 5 m × 3 m
- Large library hall: 12 m × 10 m × 6 m

Sources were placed at a distance of 1.5 m from the listener in the small room and 3 m away in the large room. Speech and single instrument sources were placed directly in front of the listener, who was positioned in the center of the room. When two instruments played, these were offset by ± 30°. Fig. 4 illustrates the configuration of the small

[1] learn.adafruit.com/adafruit-bno055-absolute-orientation-sensor/overview.

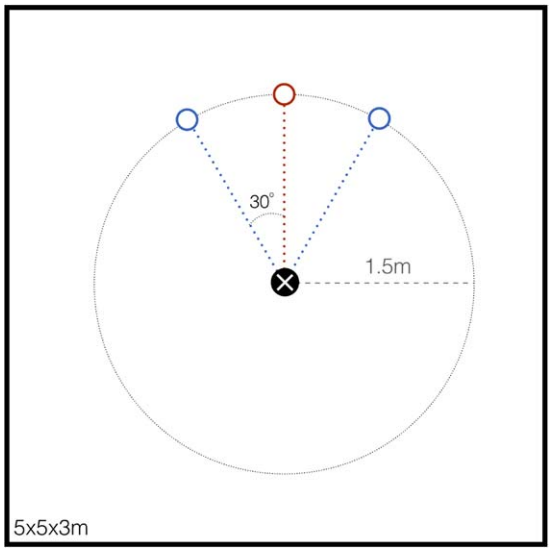

Fig. 4. Relative positions of sources in the small room.

room and locations used for both single and dual-source stimuli.

3.2 Sound Sources

Three contrasting sound sources were used in the evaluation process:

- *Female speech*—a 3-m 36-s semi-anechoic recording of selections from the Harvard list of phonetically balanced sentences, which is an established set of content used in speech quality assessment for audio systems [42].[2] Speech was judged an important element to include in the evaluation process because of listeners' strong familiarity with hearing it in contrasting spaces.
- *Solo cornet*—a 0-m 22-s excerpt from an anechoic recording of Haydn's *Trumpet Concerto in Eb* [43]. The passage edited from this performance contained several sustained notes, so any subtle artefacts generated by the reverberation algorithm would be more likely to present with these periodic waveforms.
- *Piano and drums*—dry and isolated recordings of either instrument performing The Dave Brubeck Quartet's "Take Five."[3] This combination was chosen to provide a form of broadband, harmonically rich frequency content, but which would still form a real world reference point against which listeners could evaluate the effect of different simulations.

3.3 Quality Assessment Terms

From a detailed review of existing literature on spatial audio evaluation approaches [44, 28] [45–47], *"naturalness"* and *"envelopment"* were selected as the attributes for assessment, with the following definitions:

- **Naturalness**: how realistic and natural the sound is; how well the sound conforms to what you would expect the sound in the room to be like.
- **Envelopment**: the degree to which the sound appears to come from all around you and not from a single point.

3.4 Study Design

Twenty participants aged from 20 to 45 were recruited for the listening tests—six female and the remainder male. The tests took place in a sound-proofed and acoustically deadened recording studio, using Sennheiser HD650 open-back headphones (without any equalization for binaural reproduction applied). Progress through both pre-task orientation and the study itself was conducted via a bespoke graphical user interface (GUI), connected to the VST plugin described in SEC. 2 running on a 2.8 GHz Core i5 MacBook Pro. The head-tracking unit was mounted securely and unobtrusively to the headphones' headband. Participants were free to turn and rotate their head as far as possible in all directions. They were instructed that only their head orientation would be tracked, not translation, so were asked to remain seated.

3.5 Evaluation Methodology

The five spatialized reverb variants defined in SEC. 2.5 were evaluated using a within-subjects design adapted from the "MUlti Stimulus test with Hidden Reference and Anchor (MUSHRA)" specification [48]. Each participant provided 12 sets of ratings for all five reverberation models. Ratings were conducted six times for both the "naturalness" and "envelopment" criteria. The rating iterations comprised the three different source samples detailed in SEC. 3.1, placed in the two rooms specified in SEC. 3.2, totaling six individual audio scenes. *FullSpat* represented the complete implementation of SDN reverberation, so effectively acted as a "hidden reference." An anechoic source provided a sixth "hidden anchor" stimulus in every iteration of the evaluation process.

The GUI enabled participants to freely switch between stimuli and rate each on a specified scale. Switching between methods resulted in no interruption to the source audio, so listeners could alternate between the reverberation methods as quickly as they desired. An example of the interface for the envelopment criterion is shown in Fig. 5. The interface for naturalness was identical but featured the accompanying text:

> *Grade your impression of naturalness. This attribute describes how natural and realitic the sound is. How well the sound conforms to what you would expect the sound in that room to be like.*

with the word "naturalness" substituted in place of "envelopment" on the rating scale.

The scales for each attribute were split into 5 sections, similar to other studies [38, 6], although descriptors were only provided for the least, middle, and uppermost values.

[2] odeon.dk/downloads/anechoic-recordings/.

[3] www.karaoke-version.com/custombackingtrack/the-dave-brubeck-quartet.

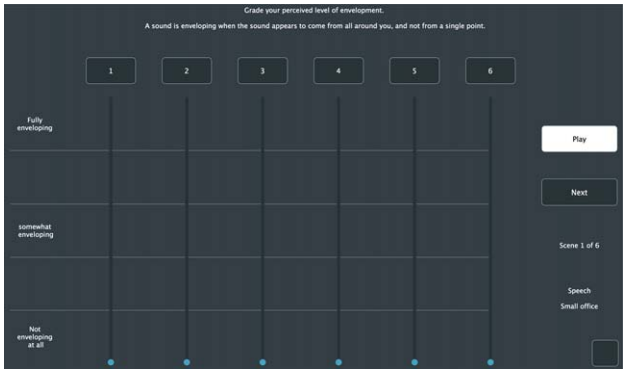

Fig. 5. GUI of the multiple stimulus test for envelopment.

No meaningful descriptions were determined for the interim values, as those considered were deemed likely to confuse more than aid the participants. On selection, the sliders provided dynamic presentation of their current value, within a range of 0 to 100.

3.6 Training

Before the evaluation phase, participants were first provided with written and oral explanations of the study procedure. The written description is provided in Appendix A.2. Following confirmation that they understood the task they were about to undertake, participants were presented with contrived audio examples, which were designed to convey the differences between the terms "naturalness" and "envelopment." To avoid bias these examples were generated via convolution using four different impulse responses from contrasting sources of reverberation.[4] These four examples were presented alongside the unprocessed input signal, which was of male speech and taken from the same source as the female voice identified in SEC. 3.2. The four artificial reverberation models were as follows:

- More natural, more enveloping: using an impulse response from a real room, presented in stereo.
- More natural, less enveloping: using an impulse response from a real room, presented in mono.
- Less natural, more enveloping: using an impulse response from a synthetic reverb unit, presented in stereo.
- Less natural, less enveloping: using an impulse response from a synthetic reverb unit, presented in mono.

Participants were allowed to review the audio examples and accompanying text until they were satisfied about the definitions of naturalness and envelopment.

[4]www.ableton.com/en/packs/convolution-reverb/. Examples were generated via the Max for Live Convolution reverb, using the 'Wood Room Small' (more natural) and 'UMSS282 Space Repeat 1' (less natural) patch impulse responses.

3.7 Rating Procedure

Participants were first presented with a practice test round, in which they were able to listen to each of the six stimuli (reverberation implementations), for all six scenes (room/source combinations). This ensured familiarity with the interface, head-tracking, auditory environment and each of the sound sources, before the evaluation started.

The evaluation took place over two rounds, with the participants first focusing on one attribute, followed by a short break, then the second attribute. Half of the participants evaluated naturalness first, the other half envelopment. Before each round participants were provided with a brief oral description of the attribute in question. A software rating procedure engine randomized the sequence in which the six scenes were presented, as well as the order of the stimuli within each evaluation iteration.

For each scene, written descriptions were provided defining the source and room (indicated in the bottom right of Fig. 5). Use of text was chosen over images of the rooms to prevent any prior experiences and associations clouding participants' judgments. All participants were encouraged to move their head and explore the virtual space, making full use of head-tracking. In total, the evaluation process took approximately an hour to complete: 15 minutes for training and familiarization and 20 minutes for each of the two rounds, plus a 5-minute break.

4 RESULTS

Although a within-subjects evaluation methodology was pursued, some sifting and separation of outlier data was deemed necessary for results analysis.

4.1 Analysis Rationale

Early analysis of the study data identified significant diversion in responses amongst participants. A significant minority scored the anechoic, anchor sound higher than 40 (i.e., the lower boundary of *somewhat enveloping/somewhat natural*) for two or more of the scenes. These participants were therefore judged to have misinterpreted that definition and so needed to be removed. For envelopment, three participants rated the anechoic anchor above 40 in two or more scenes, so 17 participants remained in that analysis.

For naturalness the outcome was more complex. Nine participants rated two or more scenes in the small room greater than 40 but no participants did so for the large room scenes. This significant anomaly suggests that a confounding factor affected participants' interpretation of the small room condition. The naturalness distributions were therefore analyzed in two separate sets. Set 1 comprised those who did not rate the anechoic source highly (11 participants) and Set 2 included all those who did (9 participants). Having served as an anchor and aiding the preliminary screening, the anechoic stimulus (method 6) was then excluded from analysis. The null hypothesis for statistical analysis was therefore that average scores for reverb spatialization methods 1 to 5 were the same.

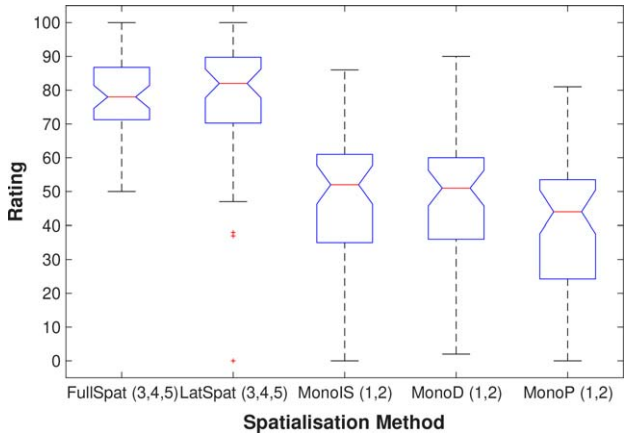

Fig. 6. Envelopment ratings for the large hall condition.

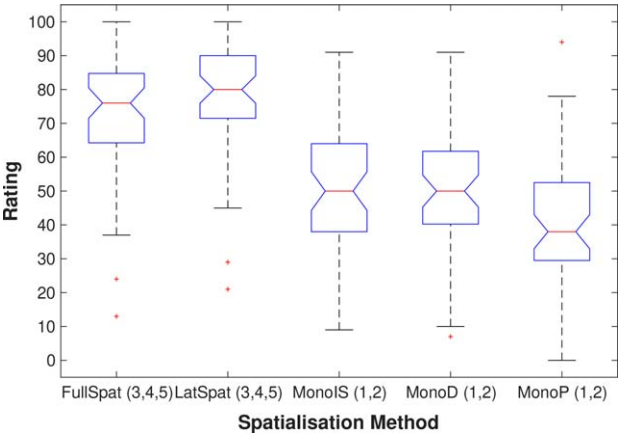

Fig. 7. Envelopment ratings for the small room condition.

Five data sets then existed to test for differences in scores between stimuli, for the following scenarios:

- envelopment in a large hall (17 participants)
- envelopment in a small room (17 participants)
- naturalness in a large hall (20 participants)
- naturalness in a small room Set 1 (11 participants)
- naturalness in a small room Set 2 (9 participants)

Individual analysis of these data sets showed that they did not meet the normality of distribution required for standard ANOVA analysis. A nonparametric Friedman test incorporating repeated measures analysis (to account for aggregation of source type ratings) was therefore used, with post-hoc Dunn-Sidak applied for pairwise comparison between stimuli (other than where stated).

4.2 Envelopment Ratings

Fig. 6 reflects the significantly enhanced sense of envelopment perceived for *FullSpat* and *LatSpat* in the large hall condition, which is confirmed by a Friedman test ($\chi^2 = 138.68$; $p < 0.001$). Post-hoc Dunn-Sidak analysis identifies that *FullSpat* ($p \leqslant 0.001$) and *LatSpat* ($p < 0.001$) were both significantly more enveloping than *MonoIS*, *MonoP,* and *MonoD*. There were no significant differences evident between the three contrasting mono models.

These trends are mirrored in Fig. 7 for the small room condition. A significantly greater sense of envelopment was provided by the two models that distributed reverberation reflections to dynamically rendered node positions ($\chi^2 = 115.76$; $p < 0.001$). There was again no statistical difference between the perceptual effect of *FullSpat* and *LatSpat* but each of these was rated significantly higher ($p \leqslant 0.040$ and $p \leqslant 0.002$, respectively) than the latter three mono models.

Analysis of envelopment ratings by each sound source showed agreement with overall trends, with two exceptions. In the small hall condition for speech, *FullSpat* was judged significantly more enveloping than only *MonoIS* ($p = 0.011$) and *MonoP* ($p < 0.001$), whereas with piano/drums the same algorithm showed no significant difference to any other method.

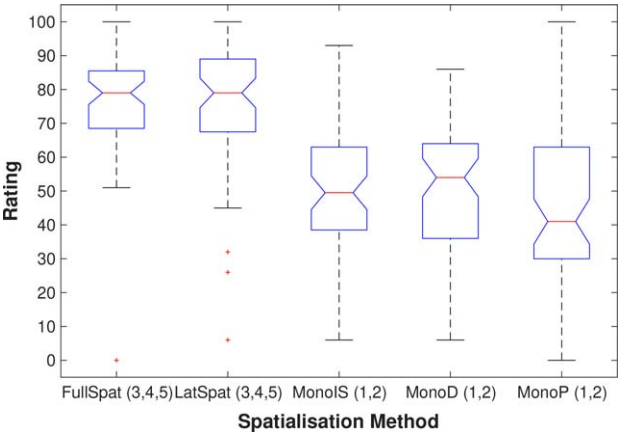

Fig. 8. Naturalness ratings for the large hall condition.

4.3 Naturalness Ratings

The same trend and relationship between methods seen for envelopment is replicated in naturalness ratings for the large hall condition, shown in Fig. 8. The two spatialized reverberation variants were perceived to be more natural sounding, to a significant degree ($\chi^2 = 104.49$; $p < 0.001$). Once again, no statistical difference between the performance of the *FullSpat* and *LatSpat* algorithms was reflected in participants' responses, but either one was again found to be significantly more natural (in turn $p \leqslant 0.005$ and $p \leqslant 0.003$) than each of the mono methods.

A more complex picture is presented for naturalness ratings in the small room condition. Figs. 9 and 10 show that the pattern of ratings between models differs clearly from the three data sets examined so far. However there is clear similarity between either cohort, despite Set 2's misattribution of preferential ratings to the anechoic anchor.

The Friedman analysis identifies significant differences between naturalness ratings of models within both Set 1 ($\chi^2 = 22.125$; $p < 0.001$) and Set 2 ($\chi^2 = 23.270$; $p < 0.001$). Post-hoc Dunn–Sidak analysis does not identify any significant pairwise differences between methods in either group. However Fisher's Least Significant Difference test highlights statistical power between Set 1's ratings of *MonoP* and both *MonoIS* ($p = 0.013$) and *MonoD* ($p = 0.040$) implementations. The same test identifies, in Set 2,

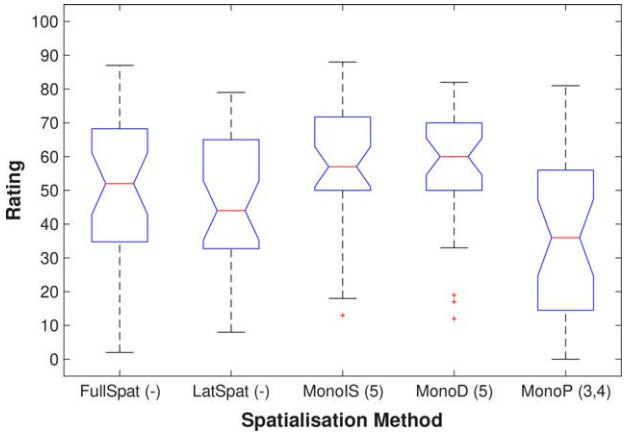

Fig. 9. Naturalness ratings for the small room condition (Set 1).

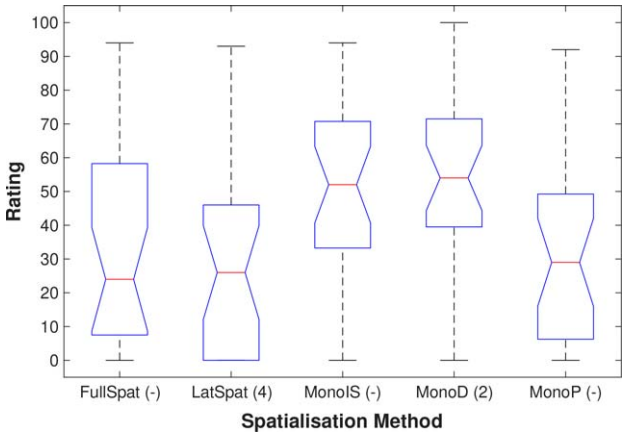

Fig. 10. Naturalness ratings for the small room condition (Set 2).

statistically significant ratings for *MonoD* over *LatSpat*. No other pairwise differences in either Set 1's (Fig. 9) or Set 2's (Fig. 10) naturalness ratings can be judged as significant.

Analysis of naturalness by each sound source showed complete agreement with overall trends for the large hall, but individual variations were present for the small room. In the latter condition, speech was judged most natural using *MonoIS* and significantly more so than *MonoP* by both Set 1 ($p = 0.018$) and Set 2 ($p = 0.009$). Set 1 also judged *MonoIS* significantly more natural than *LatSpat* ($p = 0.018$) for speech. For musical sound sources in the small room, distinction between algorithms proved more difficult. Each set found *MonoP* the least natural for cornet, but this was only significant in the case of Set 2, when compared with *MonoD* ($p = 0.032$). Likewise only Set 2 attributed significant difference between algorithms for piano/drums, for which *LatSpat* was judged as significantly less natural than *MonoD* ($p = 0.005$). *LatSpat* was also reported by Set 2 as having lower naturalness than both *MonoIS* and *MonoP*, at the threshold of significance ($p = 0.05$).

5 DISCUSSION

The implications of these results are discussed against the two research objectives stated in SEC. 0.

5.1 Perception of Variations in SDN Complexity

Full spatialization of SDN reverberation through virtual VBAP has been shown to produce an enhanced sense of envelopment, even when contrasting room sizes and varied source material are taken into account. This outcome is assumed to be a consequence of increased spatial accuracy for first order reflection rendering. The data further suggest no clear perceptual benefit to spatializing the floor and ceiling reflection streams to dedicated node positions. Merely summing the output of these streams and distributing the result evenly to each loudspeaker in the virtual VBAP array proved as effective as the fully spatialized implementation.

There was no statistical significance between envelopment ratings for any of the mono variants in either room condition. This is somewhat surprising, because there is clear audible contrast in the spatial character of *MonoP* compared with *MonoD* and *MonoIS*. However it is evident that, when presented against the two spatialized alternatives, each of the mono implementations proved similarly less satisfactory in their sense of envelopment.

For naturalness, in contrast, superior ratings for *FullSpat* and *LatSpat* are evident only for the large hall condition. The audible difference between mono implementations is clearly reflected in naturalness ratings for a small room. Perhaps of greater interest is that, in the small room simulation, the spatialized approaches (*FullSpat* and *LatSpat*) did not perform any better in terms of their naturalness than either *MonoIS* or *MonoD*. A prima facie reading of this data suggests that, for smaller virtual environments, a more natural sounding reverb is achieved either by rendering SDN without spatialization (*MonoD*), or potentially by spatializing the output of an alternative mono artificial reverberation algorithm to the reflection nodes computed via SDN (partially explored by *MonoIS*).

It is also worth noting that the contrasting ratings between the two evaluation criteria, for the small room simulation, provides potential validation of the assessment method. This suggests that participants were able to distinguish between the intended meaning of the terms envelopment and naturalness and evaluate either one on the required basis. In contrast, previous similar studies have either found considerable overlap in participant responses for multiple criteria assessment [37, 6] or used just a single metric to evaluate perceptual quality [38].

5.2 Considerations for Using SDN Reverb in AR

A mono approach to SDN spatialization appeared preferable for the small room condition. It is unclear why this was the case, but three factors could have played a part.

5.2.1 Possible Impact of Experimental Conditions

During the study sessions, some participants noted verbally that for all scenes one method always sounded like a small room. For large hall scenes this was not an expected response, because the anechoic sources should have sounded disassociated from any size of room. One participant mentioned that they had been comparing the natural-

ness of the small room simulation with the test environment, indicating a possible cause of this uncertainty. The venue was a recording studio with similar proportions to the simulated space, which included sound absorption treatment to dampen reverberation somewhat (but not completely). Other participants might have been influenced by the visual and acoustic impression of the physical surroundings and therefore felt that the anechoic sound was "natural."

It seems unlikely that this potential confounding factor alone accounts for the markedly different average ratings that resulted for small room naturalness, compared with all other data sets, for three reasons:

- The evaluation rubric outlined in SEC. 3.5, Fig. 5, and appendix A.2 expressly instructed participants to assess the effect of the simulations in the imagined context of a "small office" environment.
- The relative ratings between methods provided by Sets 1 and 2 (Figs. 9 and 10) are very similar. Any potential misinterpretation or misattribution of the anechoic sources by Set 2 did not appear to impact their discrimination between the comparative naturalness of the five SDN models.
- Average ratings of the models for small room envelopment (Fig. 7) are comparable with those for both qualities of the large hall models (Figs. 6 and 8). Incogruence between a binaurally synthesized acoustic space and physical surroundings are known to also affect sense of externalization [49]. Had the test environment introduced widespread confusion into judgment of the small room condition, some impact on the envelopment rating should also be expected.

5.2.2 Demands of Small Room Auralization

Generating convincing auralization of small room acoustics is known to offer particular challenges [50]. Precise and thorough head-tracked rendering of early reflections is more vital to achieving faithful simulation of smaller spaces than it is for larger ones (where diffuse reverb time is more influential in dictating apparent room size) [38]. Four approximations built-in to the low-complexity design of this renderer are therefore likely to be noticeably detrimental to small room simulation.

First, the sparsity of the chosen speaker array has been shown to introduce more prominent error in spatial cues at some vertical and extreme lateral locations [24]. Increasing the density and adapting the layout of the current array would result in more accurate localization [23], yielding perceptual benefits to early reflection simulation.

Second, it is well established that the use of a generic dummy head HRTF for virtual speaker binauralization increases confusion between front/back and up/down sound source localization [31]. Replacing the default KEMAR HRTF set with an option that is personalized to the user would enhance hemispherical discrimination (such as that explored and evidenced in [51]) and improve the representation of early reflection directionality.

Third, the system operates at a maximum latency of 59 ms, which is within the 75-ms threshold for maintaining sound source stability advocated by [52]. First order reflection times range between 10–20 ms for the small room configuration used in this experiment. Any potential effects of the latency on perceived acoustic integrity of small room simulation would require detailed examination.

Fourth, timbral coloration of both music and spoken sound sources is known to occur where the delay between direct and reflected sound is within 20 ms [53]. All six first order reflections in the small room test condition fall within this threshold. However only three of the six first order reflections in the large hall test condition occur less than 20 ms from the arrival of the direct sound. The computational basis of SDN reverberation will inevitably produce a more prominent degree of systematic comb-filtering in the small room condition that could be regarded as artificial or undesirable.

In the *MonoIS* and *MonoD* algorithms, early reflections from all surfaces are summed and rendered equally either to the node positions or to all virtual loudspeakers (respectively). With these two methods, a "smearing" might manifest across errors and artefacts that result from the four factors identified here, each of which would be more pronounced in small room simulation due to the greater perceptual import of early reflections. The result is less accurate spatially (lower envelopment) but perhaps more forgiving in timbre (higher naturalness).

5.2.3 Plausibility of Sources in a Small Room

The analysis of individual stimuli revealed that type of source was influential in the small room context. Small room naturalness ratings for both the solo cornet and piano/drums duo resulted in a lower level of discrimination between algorithms than every other experimental scenario. This suggests that the plausibility of acoustic musical performances in the simulated space (small office) influenced sense of naturalness to some degree, above and beyond the specific character of each reverb variant. It is not surprising that algorithm ratings for speech within the small room were more clearly distinguished by both Set 1 and 2 in favor of *MonoIS*, since this would be a more familiar combination of sound source and acoustic space for participants to evaluate.

6 CONCLUSION

A computationally efficient auralization system has been outlined for real-time, head-tracked binaural rendering via virtual VBAP and physically parameterized SDN reverberation. The system was evaluated by comparing algorithms with decreasing degrees of SDN reverb spatialization, against separate criteria of envelopment and naturalness. The two models using directed spatialization of the reverb were rated as significantly more enveloping and natural in the case of large hall synthesis. The same two models were also judged to be significantly more enveloping for small room auralization.

In all three listening contexts, the approach using directed spatialization of lateral-only reflections presents a more efficient algorithm, with no detrimental perceptual outcome. It makes a marginal computational saving for the context of AR applications and devices, avoiding real-time vector rotations that would otherwise be applied to nodes for the upper and lower room surfaces. A further (and more substantial) computational saving could potentially be achieved by removing the floor and/or ceiling reflection components altogether, thereby eliminating the associated matrix multiplications, filtering and modulation operations associated with these two delay lines. That method was not within the scope of this study, so it remains an open question whether or not such an approach would yield noticeably degraded results.

Findings for the naturalness of small room simulation suggest that an even distribution of the combined SDN reverberation streams is perceptually optimal in this iteration. As the system design stands, simplified spatialization such as *MonoD* seems to benefit the naturalness of small room auralization. This method has the added benefit of further computational gains but at noticeable expense to sense of envelopment. Alternatively, combining, for example, *LatSpat* with *MonoD* for virtualization of compact spaces might balance sense of envelopment and naturalness and at negligible additional expense compared with a pure *LatSpat* approach.

It is possible that ratings for small room naturalness were impacted by the physical environment of the study itself, or by the atypical combination of musical instrument sources with a compact virtual space. Nevertheless further investigation is necessary to establish the influence of specific design and perceptual factors that might particularly impact quality of small room auralization using SDN reverb with binaural VBAP—i.e., density and location of virtual loudspeakers, incorporation of HRTF personalization, and lowered head-tracking latency. Experimentation with these parameters might improve the naturalness of the more enveloping *FullSpat* or *LatSpat* algorithms. It is also worth noting that improving vertical rendering precision could also enhance the former method's degree of envelopment compared with the latter.

Small room environments are evidently a crucial scenario for developers of AAR or VAD applications, in which home living spaces and private work environments form the context for common use cases. Improving understanding of the relationship between envelopment, naturalness, room size, and algorithm complexity is therefore crucial in enabling AR engineers and designers to tailor auralization approaches to their development objectives.

Subsequent evaluation within virtual or physical spaces that correspond to the binaural simulations is also a logical next step. Doing so would avoid the potential for misapprehension between the auralized and described spaces when making judgments while also eliminating the possible confounding factor of disconnected physical surroundings. An evaluation methodology that incorporates comparison against other models of reverb generation would also be advocated. Although comparison of SDN against alternative approaches has been undertaken for static binaural evaluation [6], this has not been conducted with head-tracking incorporated nor using evaluation against a physical room reference. The correspondence of SDN simulations to the reverberation characteristics of real world rooms therefore also needs further validation to assist with its effective integration into AR applications.

7 ACKNOWLEDGMENT

This work was supported by EPSRC and AHRC under the grant EP/L01632X/1 (Centre for Doctoral Training in Media and Arts Technology).

Thanks are extended to Lorenzo Picianli and Isaac Engel, of Imperial College London, whose advice and expertise was particularly helpful and valued during the development of the research study design.

8 REFERENCES

[1] Y. Vazquez-Alvarez, I. Oakley, and S. A. Brewster, "Auditory Display Design for Exploration in Mobile Audio-Augmented Reality," *Pers. Ubiquit. Comput.*, vol. 16, pp. 987–999 (2012 Sep.). https://doi.org/10.1007/s00779-011-0459-0.

[2] Y. Vazquez-Alvarez, M. P. Aylett, S. A. Brewster, R. Von Jungenfeld, and A. Virolainen, "Designing Interactions With Multilevel Auditory Displays in Mobile Audio-Augmented Reality," *ACM Trans. Comput.-Hum. Interact.*, vol. 23, no. 1, pp. 1–30 (2016 Feb.). https://doi.org/10.1145/2829944.

[3] M. McGill, S. Brewster, D. McGookin, and G. Wilson, "Acoustic Transparency and the Changing Soundscape of Auditory Mixed Reality," in *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, pp. 1–16 (Honolulu, HI) (2020 Apr.). https://doi.org/10.1145/3313831.3376702.

[4] E. De Sena, H. Hacıhabiboğlu, and Z. Cvetkovic, "Scattering Delay Network: An Interactive Reverberator for Computer Games," in *Proceedings of the AES 41st International Conference: Audio for Games* (2011 Feb.), paper 3-1.

[5] E. De Sena, H. Hacıhabiboğlu, Z. Cvetković, and J. O. Smith, "Efficient Synthesis of Room Acoustics via Scattering Delay Networks," *IEEE Trans. Audio Speech Lang. Process.*, vol. 23, no. 9, pp. 1478–1492 (2015 Jun.). https://doi.org/10.1109/TASLP.2015.2438547.

[6] S. Djordjević, H. Hacıhabiboğlu, Z. Cvetković, and E. De Sena, "Evaluation of the Perceived Naturalness of Artificial Reverberation Algorithms," presented at the *148th Convention of the Audio Engineering Society* (2020 May), paper 10353.

[7] M. Vorländer, *Auralization* (Springer, Aachen, 2008).

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

[9] A. McKeag and D. S. McGrath, "Sound Field Format to Binaural Decoder With Head Tracking," in *Proceedings*

of the Audio Engineering Society 6th Australian Regional Convention (1996 Aug.), paper 4032.

[10] Facebook, "Powered by Audio360: Spatial Workstation User Guide," https://facebook360.fb.com/spatial-workstation/ (Accessed Jan. 13, 2021).

[11] Unity, "Ambisonic Audio," https://docs.unity3d.com (Accessed Jan. 13, 2021).

[12] M. Gorzel, A. Allen, I. Kelly, et al., "Efficient Encoding and Decoding of binaural Sound With Resonance Audio," in *Proceedings of the AES International Conference on Immersive and Interactive Audio* (2019 Mar.), paper 68.

[13] M. Gerzon, "Ambisonics. Part Two: Studio Techniques," *Studio Sound*, vol. 17, no. 40, pp. 24–28 (1975 Aug.).

[14] P. Fellgett, "Ambisonics. Part One: General System Description," *Studio Sound*, vol. 17, no. 40, pp. 20–22 (1975 Aug.).

[15] S. Bertet, J. Daniel, E. Parizet, and O. Warusfel, "Investigation on Localisation Accuracy for First and Higher Order Ambisonics Reproduced Sound Sources." *Acta Acust. United Acust.*, vol. 99, no. 4, pp. 642–657 (2013 Jul./Aug.). https://doi.org/10.3813/AAA.918643.

[16] L. Thresh, C. Armstrong, and G. Kearney, "A Direct Comparison of Localisation Performance When Using First, Third and Fifth Order Ambisonics for Real Loudspeaker and Virtual Loudspeaker Rendering," presented at the *143rd Convention of the Audio Engineering Society* (2017 Oct.), paper 9864.

[17] B. Wiggins, "Analysis of Binaural Cue Matching Using Ambisonics to Binaural Decoding Techniques," in *Proceedings of the 4th International Conference on Spatial Audio* (Graz, Austria) (2017 Sep.).

[18] R. Nicol, "Sound Field," in A. Roginska and P. Geluso, *Immersive Sound: The Art and Science of Binaural and Multi-channel Audio*, pp. 276–310 (Routledge, New York, New York, 2018), 1st ed.

[19] G. Kearney and T. Doyle, "Height Perception in Ambisonic Based Binaural Decoding," presented at the *139th Convention of the Audio Engineering Society* (2015 Oct.), paper 9423.

[20] C. Schörkhuber, M. Zaunschirm, and R. Höldrich, "Binaural Rendering of Ambisonic Signals via Magnitude Least Squares," in *Proceedings of DAGA 2018* (Munich, Germany) (2018 Mar.).

[21] V. Pulkki and T. Lokki, "Creating Auditory Displays With Multiple Loudspeakers Using VBAP: A Case Study With DIVA Project," in *Proceedings of the 1998 International Conference on Auditory Display* (Glasglow, UK) (1998).

[22] V. Pulkki, "Localization of Amplitude-Panned Virtual Sources. II: Two- and Three-Dimensional Panning." *J. Audio Eng. Soc.*, vol. 49, no. 9, pp. 753–767 (2001 Sep.).

[23] R. Baumgartner and P. Majdak, "Modeling Localization of Amplitude-Panned Virtual Sources in Sagittal Planes," *J. Audio Eng. Soc.*, vol. 63, no. 7–8, pp. 562–569 (2015 Jul.). https://doi.org/10.17743/jaes.2015.0063.

[24] R. Shukla, I. T. Radu, M. Sandler, and R. Stewart, "Real-Time Binaural Rendering with Virtual Vector Base Amplitude Panning," in *Proceedings of the AES International Conference on Immersive and Interactive Audio* (2019 Mar.), paper 66.

[25] V. Pulkki, "Evaluating Spatial Sound with Binaural Auditory Model," in *Proceedings of the 2001 International Computer Music Conference*, pp. 73–76 (Havana, Cuba) (2001 Sep.).

[26] D. Satongar, C. Dunn, Y. Lam, and F. Li, "Localisation Performance of Higher-Order Ambisonics for Off-Centre Listening," Tech. Rep. WHP 254 (2013 Oct.). https://www.bbc.co.uk/rd/publications/whitepaper254.

[27] J. Vilkamo, T. Lokki, V. Pulkki and, "Directional Audio Coding: Virtual Microphone-Based Synthesis and Subjective Evaluation," *J. Audio Eng. Soc.*, vol. 57, no. 9, pp. 709–724 (2009 Sep.).

[28] M. V. Laitinen and V. Pulkki, "Binaural Reproduction for Directional Audio Coding," in *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, pp. 337–340 (New Paltz, NY) (2009 Oct.). https://doi.org/10.1109/ASPAA.2009.5346545.

[29] C. Armstrong, L. Thresh, D. Murphy, and G. Kearney, "A Perceptual Evaluation of Individual and Non-individual HRTFs: A Case Study of the SADIE II Database," *Appl. Sci.*, vol. 8, no. 11, paper 2029 (2018 Oct.). https://doi.org/10.3390/app8112029.

[30] D. R. Begault, *3D Sound for Virtual Reality and Multimedia* (Academic Press Limited, London, UK, 1994), 1st ed.

[31] A. Roginska, "Binaural Audio Through Headphones," in A. Roginska and P. Geluso (Eds.), *Immersive Sound: The Art and Science of Binaural and Multi-channel Audio*, pp. 88–123 (Routledge, New York, NY, 2018), 1st ed.

[32] W. O. Brimijoin, A. W. Boyd, and M. A. Akeroyd, "The Contribution of Head Movement to the Externalization and Internalization of Sounds," *PLoS ONE*, vol. 8, no. 12, pp. 1–12 (2013 Dec.). https://doi.org/10.1371/journal.pone.0083068.

[33] Y. Iwaya, Y. Suzuki, and D. Kimura, "Effects of Head Movement on Front-Back Error in Sound Localization," *Acoust. Sci. Technol.*, vol. 24, no. 5, pp. 322–324 (2003 Sep.). https://doi.org/10.1250/ast.24.322.

[34] D. R. Begault, E. M. Wenzel, and M. R. Anderson, "Direct Comparison of the Impact of Head Tracking, Reverberation, and Individualized Head-Related Transfer Functions on the Spatial Perception of a Virtual Speech Source," *J. Audio Eng. Soc.*, vol. 49, no. 10, pp. 904–916 (2001 Oct.).

[35] M. Kronlachner and F. Zotter, "Spatial Transformations for the Enhancement of Ambisonic Recordings," in *Proceedings of the International Conference on Spatial Audio 2014* (Erlangen, Germany) (2014).

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

[37] L. Picinali, A. Wallin, Y. Levtov, and D. Poirier-Quinot, "Comparative Perceptual Evaluation Between Dif-

ferent Methods for Implementing Reverberation in a Binaural Context," presented at the *142nd Convention of the Audio Engineering Society* (2017 May), paper 9742.

[38] I. Engel, C. Henry, S. V. A. Garí, et al., "Perceptual Comparison of Ambisonics-Based Reverberation Methods in Binaural Listening," in *Proceedings of the EAA Spatial Audio Signal Processing Symposium*, pp. 121–126 (Paris, France) (2019 Sep.). https://doi.org/10.25836/sasp.2019.11.

[39] J. A. Moorer, "About This Reverberation Business." *Comput. Music J.*, vol. 3, no. 2, pp. 13–28 (1979 Jun.). https://doi.org/10.2307/3680280.

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

[41] M. Romanov, P. Berghold, M. Frank, et al., "Implementation and Evaluation of a Low-Cost Headtracker for Binaural Synthesis," presented at the *142nd Convention of the Audio Engineering Society* (2017 May), paper 9689.

[42] H. R. Silbiger, W. D. Chapman, E. H. Rothauser, N. Guttman, and M. H. L. Hecker, "IEEE Recommended Practice for Speech Quality Measurements," *IEEE Trans Audio Electroacoust.* (1969 Sep.). https://doi.org/10.1109/TAU.1969.1162058.

[43] V. Hansen and G. Munch, "Making Recordings for Simulation Tests in the Archimedes Project," *J. Audio Eng. Soc.*, vol. 39, no. 10, pp. 768–774 (1991 Oct.).

[44] R. L. King, B. Leonard, W. Howie, and J. Kelly, "Real Rooms vs. Artificial Reverberation: An Evaluation of Actual Source Audio vs. Artificial Ambience," in *Proc. Meetings Acoust.*, vol. 29, no. 1, pp. 1–9 (2016 Nov.). https://doi.org/10.1121/2.0000515.

[45] G. Reardon, A. Roginska, P. Flanagan, et al., "Evaluation of Binaural Renderers: A Methodology," presented at the *143rd Convention of the Audio Engineering Society*, pp. 1–6 (2017 Oct.), paper 359.

[46] F. Rumsey and J. Berg, "Verification and Correlation of Attributes Used for Describing the Spatial Quality of Reproduced Sound," in *Proceedings of the AES 19th International Conference: Surround Sound Techniques, Technology and Perception* (2001 Jun.), paper 1932.

[47] S. Le Bagousse, C. Colomes, and M. Paquier, "State of the Art on Subjective Assessment of Spatial Audio Quality," in *Proceedings of the AES 38th International Conference: Sound Quality Evaluation* (2010 Jun.), paper 5-3.

[48] International Telecommunication Union, "ITU-R BS.1543-2: Method for the Subjective Assessment of Intermediate Quality Level of Audio Systems," Tech. Rep. BS. 1534-2 (2014 Jun.) https://www.itu.int/dms_pubrec/itu-r/rec/bs/R-REC-BS.1534-2-201406-S!!PDF-E.pdf.

[49] S. Werner, F. Klein, T. Mayenfels, and K. Brandenburg, "A Summary on Acoustic Room Divergence and Its Effect on Externalization of Auditory Events," in *Proceedings of the 8th International Conference on Quality of Multimedia Experience*, pp. 1–6 (Lisbon, Portugal) (2016 Jun.). https://doi.org/10.1109/QoMEX.2016.7498973.

[50] C. Pike, F. Melchior, and T. Tew, "Assessing the Plausibility of Non-individualised Dynamic binaural Synthesis in a Small Room," in *Proceedings of the 55th AES International Conference: Spatial Audio* (2014 Aug.), paper 6-1.

[51] B. F. G. Katz and G. Parseihian, "Perceptually Based Head-Related Transfer Function Database Optimization," *J Acoust. Soc. Am.*, vol. 131, no. 2, pp. EL99–EL105 (2012 Jan.). https://doi.org/10.1121/1.3672641.

[52] Y. Iwaya, "Individualization of Head-Related Transfer Functions With Tournament-Style Listening Test: Listening With Other's Ears," *Acoust. Sci. Technol.*, vol. 27, no. 6, pp. 340–343 (2006 Nov.). https://doi.org/10.1250/ast.27.340.

[53] H. Hacıhabiboğlu, E. De Sena, Z. Cvetkovic, J. Johnston, and J.III O. Smith, "Perceptual Spatial Audio Recording, Simulation, and Rendering: An Overview of Spatial-Audio Techniques Based on Psychoacoustics," *IEEE Signal Process. Mag.*, vol. 34, no. 3, pp. 36–54 (2017 Apr.). https://doi.org/1 0.1109/MSP.2017.2666081.

[54] W. C. Sabine, *Collected Papers on Acoustics* (Harvard University Press, Cambridge, Massachusetts, 1922).

APPENDIX A

A.1 SDN Verification Process

Implementation of the SDN model was verified in three stages.

A.1.1 Stage 1: Manual Calculation and Comparison of System Outputs

Manual verification was conducted with two reference sets of room dimensions, specific source/microphone positions, with floor surface absorption set to 0 and the other five surfaces set to 1. Three metrics were checked:

1. node azimuth/elevation points were confirmed through manual geometrical calculation of correct positions
2. the amplitude of the first reflection from a Dirac impulse reflection was measured to be consistent with manually calculated inverse distance law attenuation, in which $a_s = 1/r_s$ and $a_r = 1/r_r$
3. the delay of the first reflection from a Dirac impulse reflection was measured to be consistent with the manually calculated interval, in which:

$$t_d = \frac{F_s(r_r - r_s)}{c}$$

where r_s is direct sound radiation distance, r_r is first reflection radiation distance, F_s is sample rate, and c the speed of sound.

A.1.2 Stage 2: Comparison of Energy Decay Rates to Source Implementation

The energy decay rate of the SDN implementation from a Dirac impulse, for the room size, source/microphone configuration, and two absorption coefficients defined in SEC.

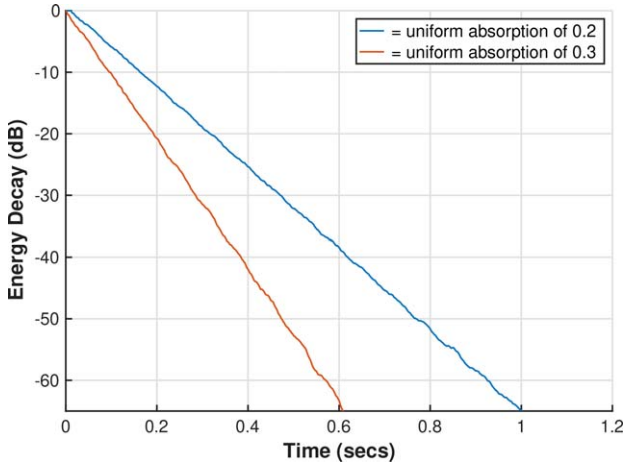

Fig. 11. Energy decay curve outputs from a Dirac impulse input for the two virtual rooms models specified in [4] and replicated here for verification of the SDN code.

Table 2. Mean T60 measurement values for the small and large rooms, compared with Sabine predictions.

Room	Predicted T60 (s)	Measured T60 (s)
Small room	0.5019	0.804
Large hall	1.0110	1.4693

4.1 of [4] was computed. The energy decay curves shown in Fig. 11 match closely those of Fig. 9 in [4], thus confirming acceptable convergence of the implementation used in this research with the original SDN specification.

A.1.3 Stage 3: Comparison of T60 Times to Standard Predictions

T60 times for the two rooms used in the listening tests were compared against the equivalent Sabine estimation [54] given by:

$$T60 = \frac{0.161V}{\sum_i A_i \alpha_i}$$

where V is the total volume of the room, and A_i and α_i the area and absorption coefficient of surface i. The source and microphone were placed in the volumetric center of each virtual room and three measurements were taken using a Dirac impulse in each. Means of the results are displayed in Table 2.

The reason measured decay times are 45%–60% greater than the predicted values is unclear. We note that previous comparisons of SDN suggest much closer adherence to Sabine predictions but also that each of these uses a simpler configuration of cubic geometry and uniform absorption coefficients across surfaces to demonstrate alignment [4, 5]. For the purpose of this study the reverberation times were not required to conform to any exact measure. It was deemed sufficient to have acoustic models of the rooms that produced a clear sense of distinction and contrast between a smaller and larger space.

A.2 Study Script: Perceptual Evaluation of Synthesized Reverberation in Spatial Audio

In this listening test you will be asked to evaluate the properties of sound produced in an interactive acoustic simulation. 6 methods for creating the sound are being tested for their perceptual qualities.

There are two attributes of the sound to test:

- ***Envelopment**: how much the sound appears to come from all around you, and not from a single point.*
- ***Naturalness**: how realistic and natural the sound is. How well the sound conforms to what you would expect the sound in the room to be like.*

As this test is about acoustics, focus more on the properties of the simulated room, rather than those of the sources.

You will measure each attribute separately, in two different rounds with a short break between. Each round should take approximately 20 minutes.

There are 6 scenes in each round, consisting of 3 source types in 2 different room models.

The source types are:

- *Person speaking*
- *Single monophonic instrument*
- *Multi-instrument*

The two rooms are:

- *Small office*
- *Large library hall*

For each scene, you may switch between the different methods freely, and are asked to rate each of them relative to the specified scale before proceeding to the next scene.

The sources will remain fixed relative to your position, so you are encouraged to move your head around to explore the virtual space.

THE AUTHORS


Christopher Yeoward


Rishi Shukla


Rebecca Stewart


Mark Sandler


Joshua D. Reiss

Christopher Yeoward is currently leading the engineering team at the adaptive, generative music company Wavepaths, building a system capable of delivering personally meaningful musical experiences at scale. He received an M.Sc. in Sound and Music Computing from Queen Mary University of London in 2019, where his research and projects explored various aspects of interactive digital media applications, and a B.Sc. in Physics from Bristol University in 2014.

•

Rishi Shukla is a Ph.D. researcher in the Centre for Digital Music at Queen Mary University of London. His research focuses on design and evaluation of interactive means for exploring, navigating, and arranging music content via binaural spatial audio. This work has included an investigation with BBC R&D on voice-led music discovery for audio-only devices. He has published papers on the topics of tangible interaction with digital music and optimization of binaural rendering for auditory display purposes. Rishi received an M.Sc. in Music Information Technology from City University, London in 2004. Prior to doctoral research, his career was dedicated toward supporting learning of music through technology, including five years overseeing iOS, Android, and web app development at The Associated Board of the Royal Schools of Music.

•

Dr. Rebecca Stewart is currently a Lecturer in the Dyson School of Design Engineering at Imperial College London, where she leads the e-Body Lab. Her research encompasses wearable technology with a focus on audio and textile interfaces. Extending beyond technical development, it also examines how to best communicate technical research findings to applications designers. She received her Ph.D. in spatial audio signal processing from the Centre for Digital Music at Queen Mary University of London, UK in 2010, M.Sc. in music technology from the University of York, UK in 2006, and B.M. in music engineering technology and computer science from the University of Miami, FL, USA in 2005.

•

Professor Mark Sandler, FREng holds a Ph.D. from University of Essex on Digital Audio Power Amplifiers (1984). He founded Queen Mary's Centre for Digital Music (C4DM) in 2003, where he is currently Director. He is a Fellow of the Royal Academy of Engineering and also a Fellow of IEEE, IET, and AES. He is author of nearly 500 papers, has supervised over 40 successful Ph.D. candidates, and has won more than £20M in research grants. He is co-Investigator on C4DM's new Centre for Doctoral Training in AI and Music. His research has covered many aspects of digital audio and digital music, including digital power amplifiers, data converters, coding, compression, streaming, drum synthesis, fractals and chaos, EQ, binaural and Ambisonics, music informatics, computational acoustics, transcription, audio for AR/VR, music linked data, segmentation, key and chord analysis, and more. More recent interests include applications of deep learning and graph theory.

•

Josh Reiss is Professor of Audio Engineering with the Centre for Digital Music at Queen Mary University of London. He has published more than 200 scientific papers (including over 50 in premier journals and 6 best paper awards) and co-authored two books. His research has been featured in dozens of original articles and interviews on TV, radio, and in the press. He is a Fellow and President-Elect of the Audio Engineering Society (AES) and chair of their Publications Policy Committee. He co-founded the highly successful spin-out company, LandR, and recently co-founded Tonz and Nemisindo, also based on his team's research. He maintains a popular blog, YouTube channel, and Twitter feed for scientific education and dissemination of research activities.