Audio Engineering Society

Convention Paper

Presented at the 138th Convention
2015 May 7–10 Warsaw, Poland

Estimation of room reflection parameters for a reverberant spatial audio object

Luca Remaggi, Philip J. B. Jackson, and Philip Coleman

Centre for Vision, Speech, and Signal Processing, University of Surrey, Guildford, UK

Correspondence should be addressed to Luca Remaggi (l.remaggi@surrey.ac.uk)

ABSTRACT

Estimating and parameterizing the early and late reflections of an enclosed space is an interesting topic in acoustics. With a suitable set of parameters, the current concept of a spatial audio object (SAO), which is typically limited to either direct (dry) sound or diffuse field components, could be extended to afford an editable spatial description of the room acoustics. In this paper, we present an analysis/synthesis method for parameterizing a set of measured room impulse responses (RIRs). RIRs were recorded in a medium-sized auditorium, using a uniform circular array of microphones representing the perspective of a listener in the front row. During the analysis process, these RIRs were decomposed, in time, into three parts: the direct sound, the early reflections and the late reflections. From the direct sound and early reflections, parameters were extracted for the length, amplitude and direction of arrival (DOA) of the propagation paths by exploiting the dynamic programming projected phase-slope algorithm (DYPSA) and classical delay-and-sum beamformer (DSB). Their spectral envelope was calculated using linear predictive coding (LPC). Late reflections were modelled by frequency-dependent decays excited by band-limited Gaussian noise. The combination of these parameters for a given source position and the direct source signal represents the reverberant or "wet" spatial audio object. RIRs synthesized for a specified rendering and reproduction arrangement were convolved with dry sources to form reverberant components of the sound scene. The resulting signals demonstrated potential for these techniques, e.g. in SAO reproduction over a 22.2 surround sound system.

1. INTRODUCTION

One major aim of spatial audio is to reproduce the characteristics of an indoor environment, with the intention of providing the listener with a sensation of being in the recorded environment. This research area can be defined as virtual acoustic environment

modelling. It can be subdivided into main tasks: source modelling (e.g. natural audio, synthetic audio, source directivity), room modelling (e.g. modelling of acoustic spaces, artificial reverberation) and listener modelling (e.g. head-related transfer functions (HRTFs), microphone directivity) [1]. This paper will be focused on room modelling, with the aim of SAO production [2].

1.1. Auralization

The process of characterizing closed environments, and convolving a "dry" source with synthesized RIRs, is called auralization [3]. Researchers attempted to approximate a RIR using multirate systems and discrete-time wavelet transform (DTWT) [4]. Another work tried to recreate the sound field of a given room using the so called "plenacoustic function" [5]. This was generated using RIRs produced through the image source method, but conceptually requires infinite RIRs in the continuous time domain. For this reason, the authors tried to achieve the best approximation by sampling the RIRs in time and using a large finite number of them. Other researchers exploited the image source theory to synthesize RIRs [6]. This model was based on analysing recorded RIRs to localize the sources and extract the pressure signals, giving good results. However, the performance of the sources localization techniques used is inversely proportional to the noise level.

Another approach generates virtual sources from their DOAs and amplitudes [7]. This method uses the so-called vector base amplitude panning (VBAP). An extension, the spatial impulse response rendering (SIRR) model [8, 9], analyses the time-dependent direction of arrival and diffuseness of recorded RIRs dividing them into frequency subbands. In SIRR, an impulsive source was used in each time-frequency element. Thereafter, directional audio coding (DirAC) [10] was developed, where multiple sources are allowed.

Another approach, based on the Kirchoff-Helmholtz integral, considers any point of the sound wave front as a secondary source [11]. To implement this theoretic concept, a uniform circular array (UCA) of microphones recorded multi-channel RIRs [3]. These RIRs were parameterized, and then synthesized. This process is called auralization through wave field synthesis (WFS) [12]. This technique allows the spatial reproduction of sound images for large areas [13].

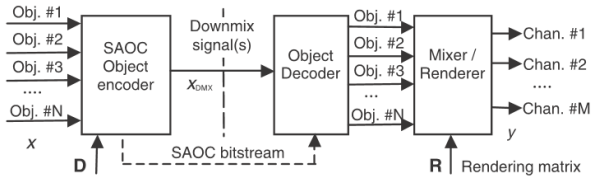

Fig. 1: General SAOC structure overview. Figure reproduced from [17].

In [14], a model was presented to render spatial sound using multiple RIRs. They approximated the direct sound and early late reflections through the image-source method and the reverberation using filters derived from the recorded signals. Here, we apply these ideas in the context of SAOs.

1.2. Spatial audio object coding

During the last decades, the intention of providing the sensation of being inside the recorded environment has been extended to domestic rooms. One problem was the transmission of high-quality multichannel data through band-limited channels. Parametric coding techniques based on spatial audio coding (SAC) were then studied. The MPEG group defined a standard for spatial audio called MPEG Surround [15]. Recently, their activities turned into the so-called spatial audio object coding (SAOC) [16, 17], a coding technique, that exploits rendering of multiple SAOs [2]. An overview of the SAOC conceptual structure is shown in Figure 1. The idea of creating SAOs for sound scene reproduction is included in the MPEG-4 standard. Specifically, a scene description language called Binary Format for Scenes (BIFS) is defined [18]. In [19] the authors introduced a subdivision of the early reflections in two parts to modify the MPEG-4 BIFS "perceptual" approach, whereas a 3D audio object generation has been presented in [20] following the "physical" approach. The MPEG group recently started the MPEG-H Audio Coding development. The current status of the standardization project has been reported in [2]. This new standard will allow different input formats. Regarding the audio objects, the decoding part will be the one extended from the previous standards, expanding the Unified Speech and Audio Coding stage (USAC) for 3D audio and defining VBAP as the algorithm used to render SAOs.

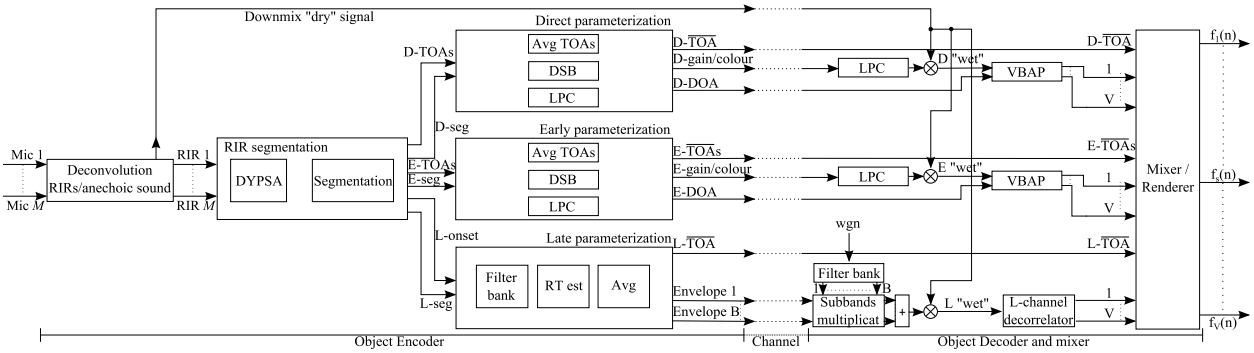

Fig. 2: System diagram overview of the RSAO encoder and the combination of RSAO decoder and mixer. The encoder is composed by deconvolution, segmentation, parameterization and dry object encapsulation; B subbands are used to parameterize the late reverberation. The decoder converts dry objects into wet objects before generating V signals (V is the number of loudspeakers) per each part (direct sound, early reflections and late reverberation). Then, the mixer renders the signals $f_s(n)$, where s indicates the s-th loudspeaker. "wgn" denotes white Gaussian noise, "D", "E" and "L" stand for direct, early and late respectively.

1.3. Reverberant spatial audio objects

In this paper, we present a parametric RIR model to be transmitted as part of an audio object. We refer to this as a reverberant spatial audio object (RSAO) and it is based on the physical representation of the sound scene. Our method synthesizes RIRs from measured ones, recorded through a UCA of microphones. The novelties are given by the new combination of methods for extracting parameters, i.e. DYPSA [21], DSB [22] and LPC [23]. The RIR component parts are selected: direct sound, early reflections and late reverberation [24]. Each of these is subjected to analysis to extract parameters. Four parameters are extracted: source range and the amplitude exploiting DYPSA, the DOAs using the DSB, and the colouration through LPC. The direct sound and the early reflections are directly extracted from the RIRs, and the late reverberation is modelled, using exponentially decaying Gaussian noise. Once the RIRs are parameterized and sent to the decoder, they can be combined with anechoic signals to create the RSAO and spatialized, e.g. by WFS, VBAP or planarity panning [25]. The overview of the components is reported in Figure 2. Preliminary results show that this approach allows the production of signals with the acoustic characteristics of the reference room.

The article is structured as follows: in Section 2 the encoding part of the reverberant object is presented; Section 3 reports the decoding and rendering part; in Section 4 the experiments performed are explained; finally Section 5 draws the overall conclusions.

2. ROOM OBJECT ENCODING

Recording signals throughout a specific environment can generate a huge amount of data which can be difficult to transmit. Furthermore, it does not allow the final user to interact with the virtual scene. For this reason, the SAOC has been defined in MPEG-4, treating different audio signals as different audio objects [2]. In our paper, the SAOs before the decoding part are thought of as implicitly hidden within each signal recorded by the UCA. In fact, the SAOs considered are the direct sound source, the early reflection image sources and the late reverberation diffuse sources. The main contribution in this article is the creation of a metadata package (bitstream) defining perceptually-relevant parameters representing the room acoustics. This package, sent through the data transmission channel, together with the anechoic signal defines the RSAO.

As a starting point, a deconvolution can be applied to the M recorded signals, where M is the number of microphones in the UCA. In this way, RIRs are obtained and separated from the anechoic signal. In our experiments RIRs are directly recorded from the field. Specific parameters are then extracted from the RIRs, packed and sent as bitstreams.

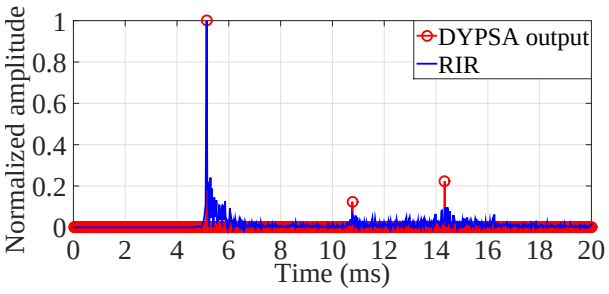

Fig. 3: The peaks detected by DYPSA are reported (red circled line) and compared to the RIR under investigation (blue line).

2.1. RIR definition

A signal $x(n)$ sent by a source is received by the i-th sensor as $y_i(n) = x(n) * r_i(n) + w(n)$, where $r_i(n)$ is the i-th RIR, $w(n)$ is the assumed Gaussian measurement noise and the symbol "$*$" denotes convolution. In general, a RIR is formed of infinite replicas of the source signal, with some additive noise. Each k-th replica has a path dependent attenuation $A_{k,i}$ and time of arrival (TOA) delay $n_{k,i}$ [26]:

$$r_i(n) = \sum_k A_{k,i} \cdot \delta(n - n_{k,i}) = \sum_k h_{k,i}(n - n_{k,i}), \quad (1)$$

where $\delta(n)$ is the discrete-time n dependent Dirac function and $n_{k,i}$ is the TOA relative to the k-th peak of the i-th microphone. The RIR $r_i(n)$ can be decomposed into direct sound $h^D(n)$, early reflections $h^E(n)$ and late reverberation $h^L(n)$ [24]:

$$\begin{aligned} r_i(n) =& h_i^D(n) + h_i^E(n) + h_i^L(n) = \\ =& h_{0,i}(n - n_{0,i}) + \sum_{k=1}^{L_1} h_{k,i}(n - n_{k,i}) + \\ &+ \sum_{k=L_1+1}^{L_2} h_{k,i}(n - n_{k,i}), \end{aligned} \quad (2)$$

where L_1 is the last peak before the reverberation time (RT60) [26] and L_2 the last peak of the recorded RIR. The direct sound is defined for $k = 0$, the early reflections for $1 \leq k \leq L_1$ and the late reverberation for $L_1 + 1 \leq k \leq L_2$.

In the analysis method proposed in the following subsections, for the direct sound and the first 2 early reflections ($L_1 = 2$), the analysis has been performed extracting TOAs, amplitude parameters, DOAs and frequency content, whereas for reflections with higher order ($k \geq 3$) a global frequency dependent analysis in the time domain has been performed. This different approach is due to early reflections appearing as delayed impulses in the RIR, whilst late reflections appear as a continuum. Furthermore, it is important to note that the energy of the reflections decays at an exponential rate [24], related to the RT60.

2.2. TOAs and segmentation

To extract TOAs from RIRs, a method for selecting the peaks of the signal has been developed based on DYPSA. This was designed to estimate glottal closure instances (GCIs) from speech signals, and has been modified to make it applicable RIRs [27].

The phased-slope function $S(\omega)$ is the average slope of the unwrapped phase spectrum of the short-time Fourier transform of the linear prediction residual [21]. In other words, it is the group delay function $G(\omega)$ of the signal, but with the opposite sign $S(\omega) = -G(\omega) = \mathrm{d}\Phi(\omega)/\mathrm{d}\omega$, where $\Phi(\omega)$ is the phase shift. Variations in the time domain (i.e. peaks) correspond to positive-going zero crossings in $S(\omega)$. To reliably select the instants where $S(\omega)$ has these zero crossings, it is smoothed using a Hann window. Finally, two main processes are applied, the first is to compare an ideal slope function creating a level of confidence and the second is to calculate the weighted gain of each peak considering its importance on the original signal. To adapt the algorithm to the purposes of this article, a threshold is defined on $S(\omega)$ in order to take only the most significant peaks of $r_i(n)$. The slope threshold is set to 0.2. Another threshold is applied on the time domain amplitude, to eliminate the peaks that are more than 25 dB below the main one. These thresholds are heuristically derived.

The DYPSA output is a sequence of non-zero values placed on the time samples corresponding to the peaks of $r_i(n)$. TOAs are calculated as $n_{k,i} = s_{k,i}/F_s$, where $s_{k,i}$ is the k-th non-zero sample and F_s the sampling frequency. To maintain the reflection energy as the RIR, the signal is windowed in the neighbourhood of the detected peaks. The energy $E_{k,i} = \frac{1}{D} \sum_{n=1}^{D} ||r_i(n)||^2$ is calculated for these time intervals, where D is the number of samples se-

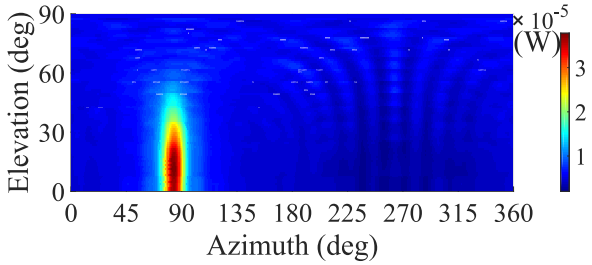

Fig. 4: Power in output of the DSB for a source positioned at 83° azimuth and 11° elevation with respect to the center of the microphone array.

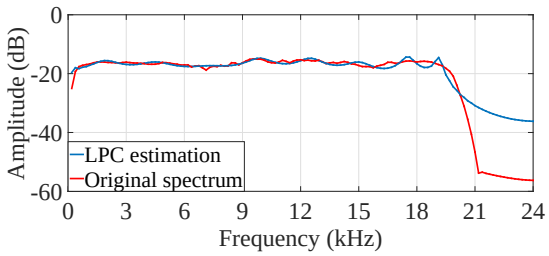

Fig. 5: Frequency spectrum amplitude in dB for the direct sound (red line) compared with the approximation made by LPC (blue line).

lected in the neighbourhood of the k-th peak. The k energies are used to obtain the amplitude $A_{k,i}$ of the output pulses using the equation $A_{k,i} = \sqrt{E_{k,i}}$. Figure 3 shows the output of the DYPSA algorithm for a measured RIR.

In measured RIRs, peaks are not distinguishable from the measurement noise after a certain time. For this reason, DYPSA is used to detect the direct sound and the first 2 early reflections. Consequently, in Equation 2 $L_1 = 2$. Knowing these TOAs, a segmentation of the RIRs is performed, placing Hamming windows with a size of 256 samples for the direct sound and a size of 128 samples for the early reflections. Finally, the mean of the TOAs of the M microphones in the array, $n_k^{'}$, is calculated and included in the RSAO.

2.3. Direction of arrival extraction

Another set of parameters to extract from the RIRs is the DOA of the direct sound and early reflections. Several algorithms can be used to reach this goal [28], however, we decided to choose the DSB [22], since it gives adequate performance for our purposes and is simple. Since the purpose of this sub-algorithm is to estimate the DOAs for the direct sound and the early reflections, the signals segmented exploiting DYPSA are evaluated.

The DSB calculates DOA of a signal exploiting the TOA between the specific source and each microphone of the array. Applying a phase shift to each signal received by each microphone, only the signals from a particular direction are aligned when they are finally summed. The square of this sum is then calculated to obtain the power for the specific angle. With prototype delays relative to every angle under investigation, the angle that yields the maximum power in the output signal gives the DOA. The vector containing the phase shifts can be defined as $\Phi_{dsb}(\omega) = [e^{j\omega n_1}, ..., e^{j\omega n_i}, ..., e^{j\omega n_M}]^T$, where M is the number of microphones and n_i the time delay applied, relative to the i-th microphone. A 3D version of DSB was used, exploiting two UCAs of the same radius, lying on two planes parallel to the floor and having the centre in points with the same $x - y$ coordinate but a different z. In this way, the angle and elevation were both estimated using the DSB. In Figure 4, an example of angles dependent power in output of a DSB is reported.

2.4. Colour extraction

The perception of a sound inside a room is not only provided by time delays, as the frequency content has an important role. For this reason, the frequency content of direct sound and early reflections was analysed. This analysis is based, as for the DSB, on the DYPSA-based segmentation. The pulses selected by this epoch detection algorithm are windowed from the RIRs using Hamming windows, as already explained in Section 2.3. In this way, the analysis in the frequency domain can be performed for these parts of the signals only.

The well-established LPC [23] is the method chosen to estimate the spectral envelope. Applying it to every acoustic event in $h_i^D(n)$ and $h_i^E(n)$, 16-th order finite impulse response (FIR) filters $z_{k,i}(n)$ are generated. The 17 filter coefficients are averaged over the M microphones. The results are those parameters that encapsulate the frequency content in the bitstream. An example of spectrum estimation through LPC for an $h_i^D(n)$ is shown in Figure 5.

2.5. Late reverberation parameterization

In the human auditory system, the sound cues processing is performed on a non-uniform frequency scale. Hence, it is important to transform the time-domain representation to a representation that resembles this non-uniform scale by using an appropriate filter bank [15]. We chose to divide $h_i^L(n)$ in $B = 9$ subbands, through the implementation of a filter bank composed by octave band FIR filters. The cut off frequency for the low-pass filter is 88 Hz, and the one for the high-pass filter is 11.3 kHz.

Analysis in the time domain is then performed. The parameter to extract is the energy decay $e_{b,i}(n)$, where i indicates the microphone under investigation and b is the subband index. To reach this aim, Schroeder's algorithm is used to estimate the reverberation time given an impulse response [29]. The envelopes are then averaged over all M microphones:

$$e_b^{'}(n) = \frac{1}{M}\sum_{i=1}^{M} e_{b,i}(n). \tag{3}$$

Each envelope is then encapsulated to be sent through the bitstream as 9 coefficients, representing an 8-th order polynomial.

The late reverberation starting time is known from the DYPSA segmentation. This time can be called late reverberation time of arrival (LR-TOA), and each i-th microphone has its own defined as $n_{k,i}$, with $k = L_1 + 1 = 3$ ($L_1 = 2$ as in Section 2). Also in this case, the LR-TOAs are averaged:

$$n_k^{'} = \frac{1}{M}\sum_{i=1}^{M} n_{k,i}, \qquad \text{for } k = L_1 + 1, \tag{4}$$

and $n_k^{'}$ is the parameter sent to the decoder.

3. ROOM OBJECT DECODING

Once all the parameters have been extracted from the measured RIRs, they are transmitted as a bitstream, together with the anechoic signal $x(n)$ defined in Section 2.1. The decoder, using specific algorithms for rendering sources and diffuseness, will convert them into RSAOs. Then, every signal is directed to the correct loudspeaker through an integration system, which is the responsible of handling a surround reproduction system composed of V loudspeakers.

3.1. Object decoder

Due to the different nature of the SAOs, two different approaches are used, depending on whether the decoded object represents the source position (direct sound and early reflections) or the room effect (late reverberation) [30]. The direct sound and early reflections can be rendered as independent sources (main and image sources), whereas the late reverberation is reproduced as a diffuse source.

3.1.1. Direct and early source objects

The SAOs received by the decoder are $L_1 + 2$. The first contains the direct sound impulse $h_0^{'}(n)$ (where the symbol $'$ means estimated), the other L_1 are composed by impulses approximating the early reflections $h_k^{'}(n)$ (with $1 \leq k \leq L_1$), and the last one the late reverberation part $h_k^{'}(n)$ (with $k = L_1 + 1$). Each SAO is coded as the combination of the anechoic signal and packages of metadata containing information relative to TOA, DOA and frequency content in the form of LPC coefficients.

The direct sound and early reflections are treated on the same way. Using the amplitude of the peaks A_k extracted from DYPSA, impulses are generated by the decoder and filtered using the filters $z_k(n)$ given by the LPC parameters coded in the bitstream. The resulting signals are the estimated impulse responses for direct sound and early reflections:

$$h_k^{'}(n) = A_k[\delta(n) * z_k(n)] \qquad \text{for } 0 \leq k \leq L_1 \tag{5}$$

where the symbol "$*$" stands for convolution.

At this point, the SAOs are converted into "wet" SAOs, convolving $x(n)$ defined in Section 2.1 by the synthesized impulse responses in Equation 5:

$$y_k^{'}(n) = x(n) * h_k^{'}(n) \qquad \text{for } 0 \leq k \leq L_1. \tag{6}$$

3.1.2. Direct and early object spatialization

Since V loudspeakers placed in the 3D space do not always coincide with the directions indicated by the DOA parameters, the VBAP algorithm [7] is exploited to create virtual sources for the main and image sources. The idea is based on panning between the three channels closest to the intended DOA of the source and leaving out the others. Different weights l_s are applied to the amplitude of the sound produced by these loudspeakers to create the impression of the virtual source. The k-th source

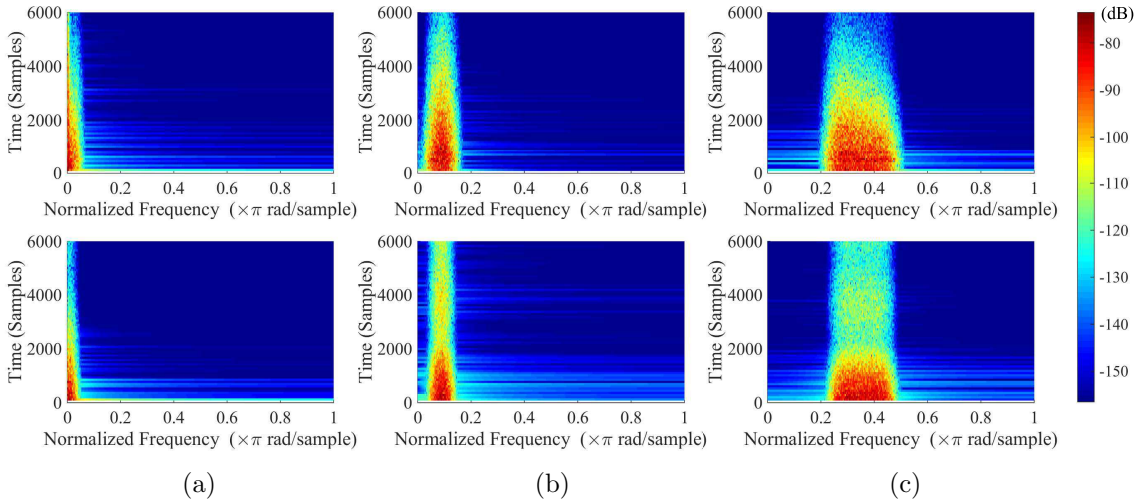

Fig. 6: Spectrogram of three late reverberation subbands, (a) the first (0-88 Hz), (b) the sixth (1.4-2.8 kHz) and (c) the eighth (5.7-11.3 kHz). The top figures represent the subbands relative to one of the recorded RIRs, the bottom ones the respective subbands for the decoded late reverberation object.

output can be so defined as $c_{s,k}(n) = l_{s,k} \cdot y_k^{'}(n)$, where $y_k^{'}(n)$ is defined in Equation 6 and $l_{s,k}$ indicates the weight applied to the s-th channel of the k-th source, for $0 \leq k \leq L_1$. It is important to note that for each k, just three values of s give $c_{s,k}(n) \neq 0$.

3.1.3. Late diffuse object

The B envelopes $e_b^{'}(n)$ are received and multiplied by Gaussian noise, produced by a filtered pseudo-random sequence generator. Defining the b-th subband of the Gaussian noise as $w_b(n)$, the resulting subband signal is given by $h_{k,b}^{'}(n) = e_b^{'}(n) \cdot w_b(n)$, with $k = L_1 + 1$. The subbands are then summed:

$$h_k^{'}(n) = \sum_{b=1}^{B} h_{k,b}^{'}(n) \qquad \text{for } k = L_1 + 1. \tag{7}$$

In Figure 6 three late reverberation subbands relative to one of the measured RIRs (top) are compared to respective subbands for the decoded late reverberation object (bottom). At this point, the SAO is converted to a "wet" SAO:

$$y_k^{'}(n) = x(n) * h_k^{'}(n) \qquad \text{for } k = L_1 + 1. \tag{8}$$

The signal $y_k^{'}(n)$ is then sent to a V-channel decorrelator, which generates V different signals by convolving $y_k^{'}(n)$ by V all pass filters having poles randomly distributed within the unit circle [31]. In this way, different random noise is sent to each channel, although the energy decay is maintained. The s-th output signal is so defined as $t_{s,k}(n)$, with $k = L_1+1$. In contrast to the main and image sources, in this case for every $1 \leq s \leq V$, we have $t_{s,k}(n) \neq 0$.

3.2. Mixer

The last step in the MPEG-4 standard [15] is the mixer block. This block receives the signals placed in the channels by the object decoding algorithms and combines them to create the right connections with the available loudspeakers. Depending on the surround system implemented, V is the number of channels and loudspeakers available (e.g. for a 22.2 surround system $V = 24$). It also receives the TOAs n_k extracted from the SAO by the decoder and uses them to give the right time shift to each part of the RIR. The virtual sources produced by VBAP and the V decaying noises are combined into the final signals sent to the V loudspeakers:

$$f_s(n) = \sum_{k=1}^{L_1+1} c_{s,k}(n - n_k) + t_{s,k}(n - n_k), \tag{9}$$

where $1 \leq s \leq V$ indicates the s-th channel.

4. EXPERIMENTS

This section describes the sound recording setup,

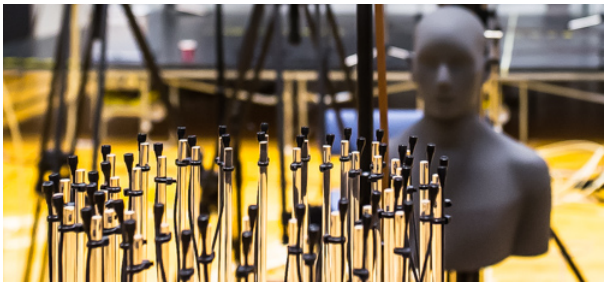

Fig. 7: The recording setup with the UCA close-up.

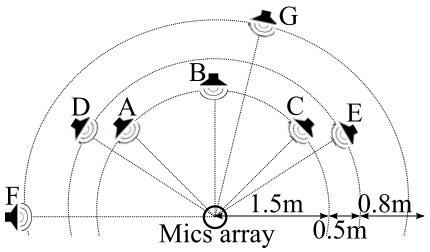

Fig. 8: The 7 loudspeaker used to record the RIRs (from "A" to "G") and the UCA of microphones. A, B, C and the UCA have an height of 1.50 m, D and F 1.12 m, F and G 0.30 m.

the listening test setup and the results of informal listening tests. Two rooms in the University of Surrey were used, one to record the signals and another to reproduce the output of the model presented. In the following subsections, the recording hardware and sound scene are described.

4.1. Microphone array

The microphone array consisted of two concentric UCAs, each with 24 omnidirectional capsules (Countryman B3) spaced evenly around the circle. The radii of the inner and outer circles were 85 mm and 107 mm, respectively. This configuration was adopted to allow for robust beamforming with equal resolution in all azimuths. To perform 3D beamforming avoiding elevation ambiguity, two different heights for the microphone array were used for the recordings, 1.50 m and 1.54 m. A photograph of the recording setup with the double UCA is shown in Figure 7. Level calibration was performed by recording a 1 kHz tone at 94 dB SPL, and scaling the recordings for each channel in software.

4.2. RIR capture

RIRs were recorded in a large recording studio with dimensions $17.08 \times 14.55 \times 6.50$ m and a reverberation time of 1.1–1.5 s. 15 different loudspeaker positions were used and 7 of them were selected for the purposes of this article, named from "A" to "G". The 3 loudspeakers between A and C were positioned at a height of 1.5 m, lying on a circle around the UCA with radius of 1.5 m; defining the loudspeaker B as the one at 0°, A was positioned at −45° and C at 45°. The loudspeakers D and E were positioned at a height of 1.18 m, at a distance of 2 m from the UCA with angles of −60° and 60° respect to the loudspeaker B respectively. Loudspeakers F and G were at 0.3 m height, 2.8 m away from the UCA and creating angles of −90° and 15° with the B loudspeaker respectively. The positions of microphones and loudspeakers are schematically reported in Figure 8. The sample frequency used was 48 kHz and the swept-sine technique was used to measure RIRs.

4.3. Reproduction system

A reproduction system was mounted on a spherical structure of radius 1.68 m, the "Surrey Sound Sphere" [32] (see Figure 9). It was placed in an acoustically treated room, with dimensions $7.90 \times 6.00 \times 3.98\,\mathrm{m}^3$, and RT60 of 215 ms. 10 loudspeakers (Genelec 8020b) were clamped to the equator to form a circular array, and another 12 were clamped to the sphere structure at different heights. A chair was positioned in the middle of the sphere to allow informal listening tests.

4.4. Listening tests

The model presented in Sections 2 and 3 has been subject to simulation analysis. To do this, the encoder/decoder was implemented in Matlab. Informal listening tests were performed by listening to the audio signals reproduced by the decoder. 22 signals were connected to the correct channels to implement a 22.0 surround system in the Surrey Sound Sphere. Those tests have provided to the listeners the sensation of being in an environment other the one where the sphere is placed.

5. CONCLUSION

A model to generate reverberant RSAOs, composed by an encoding and a decoding part, has been presented. Novelties for the parameter extraction have been introduced: the DYPSA algorithm estimated the TOAs and hence the energies of the

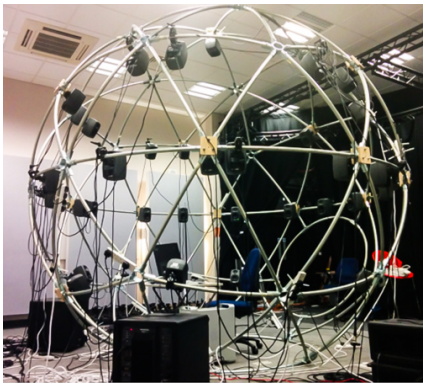

Fig. 9: The "Surrey Sound Sphere".

RIR components, the DSB estimated the DOAs, and LPC analysed the frequency content. RIRs were recorded from a small concert hall and used to test the model. Informal listening tests demonstrated this new model's capacity to give the listener a sense of being in the original recorded acoustic environment.

6. ACKNOWLEDGEMENTS

This work was supported by the Engineering and Physical Sciences Research Council (EPSRC) Grant EP/K014307/1, the MOD University Defence Research Collaboration in Signal Processing, EPSRC "S3A" Programme Grant EP/L000539/1, and the BBC Audio Research Partnership. Thanks to Jon Francombe for help with listening tests.

7. REFERENCES

[1] L. Saviola, J. Huopaniemi, T. Lokki, and R. Väänänen, "Creating interactive virtual acoustic environments," *J. Audio Engineering Society*, vol. 47, no. 9, pp. 675–705, 1999.

[2] J. Herre, J. Hilpert, A. Kuntz, and J. Plogsties, "MPEG-H audio - the new standard for universal spatial/3D audio coding," *J. Audio Engineering Society*, vol. 62, no. 12, pp. 821–830, 2014.

[3] E. Hulsebos, *Auralization using wave field synthesis*, Ph.D. thesis, Technische Universiteit Delft, 2004.

[4] M. Shoenle, N. Fliege, and U. Zoelzer, "Parametric approximation of room impulse responses by multirate systems," in *Proc. of the IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, New Paltz, USA, 1993.

[5] T. Ajdler and M. Vetterli, "The plenacoustic function, sampling and reconstruction," in *Proc. of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Hong Kong, 2003.

[6] S. Tervo, J. Pätynen, A. Kuusinen, and T. Lokki, "Spatial decomposition method for room impulse responses," *J. Audio Engineering Society*, vol. 61, no. 1/2, pp. 17–28, 2013.

[7] V. Pulkki, "Virtual sound source positioning using vector base amplitude panning," *J. Audio Engineering Society*, vol. 45, no. 6, pp. 456–466, 1997.

[8] J. Merimaa and V. Pulkki, "Spatial impulse response rendering i: analysis and synthesis," *J. Audio Engineering Society*, vol. 53, no. 12, pp. 1115–1127, 2005.

[9] V. Pulkki and J. Merimaa, "Spatial impulse response rendering ii: reproduction of diffuse sound and listening tests," *J. Audio Engineering Society*, vol. 54, no. 1/2, pp. 3–20, 2006.

[10] V. Pulkki, "Spatial sound reproduction with directional audio coding," *J. Audio Engineering Society*, vol. 55, no. 6, pp. 503–516, 2007.

[11] A. J. Berkhout, D. de Vries, and P. Vogel, "Acoustic control by wave field synthesis," *J. Audio Engineering Society*, vol. 93, no. 5, pp. 2764–2778, 1993.

[12] E. Hulsebos and D. de Vries, "Parameterization and reproduction of concert hall acoustics measured with a circular microphone array," in *Proc. of the 112th Audio Engineering Society Convention (AES)*, Munich, Germany, 2002.

[13] U. Horbach, E. Corteel, and D. de Vries, "Spatial audio reproduction using distributed mode loudspeaker array," in *Proc. of the 21st Audio Engineering Society Conference (AES)*, St. Petersburg, Russia, 2002.

[14] Y. Li, P. F. Driessen, G. Tzanetakis, and S. Bellamy, "Spatial audio rendering using measured room impulse responses," in *Proc. of the IEEE International Symposium on Signal Processing and Information Technology (ISSPIT)*, Vancouver, Canada, 2006.

[15] J. Herre, K. Kjörling, J. Breebaart, C. Faller, S. Disch, H. Purnhagen, J. Koppens, J. Hilpert, J. Rödén, W. Oomen, K. Linzmeier, and K. S. Chong, "MPEG surround - The ISO/MPEG standard for efficient and compatible multichannel audio coding," *J. Audio Engineering Society*, vol. 56, no. 11, pp. 932–955, 2008.

[16] J. Engdegård, B. Resch, C. Falch, O. Hellmuth, J. Hilpert, J. Höelzer, L. Terentiev, J. Breebaart, J. Koppens, E. Schuijers, and W. Oomen, "Spatial audio object coding (SAOC) - The upcoming MPEG standard on parametric object based audio coding," in *Proc. of the 124th Audio Engineering Society Convention (AES)*, Amsterdam, The Netherlands, 2008.

[17] J. Herre, H. Purnhagen, J. Koppens, O. Hellmuth, J. Engdegård, J. Hilpert, L. Villemoes, L Terentiv, C. Falch, A. Höelzer, M. L. Valero, B. Resch, H. Mundt, and H. Oh, "MPEG spatial audio object coding - The ISO/MPEG standard for efficient coding of interactive audio scenes," *J. Audio Engineering Society*, vol. 60, no. 9, pp. 655–673, 2012.

[18] R. Väänänen and J. Huopaniemi, "Advanced audioBIFS: virtual acoustics modelling in MPEG-4 scene description," *IEEE Transaction on Multimedia*, vol. 6, no. 5, pp. 661–675, 2004.

[19] U. Reiter, A. Partzsch, and M. Weitzel, "Modifications of the MPEG-4 AABIFS perceptual approach: assessed for the use with interactive audio-visual application systems," in *Proc. of the 28th Audio Engineering Society Conference (AES)*, Piteå, Sweden, 2006.

[20] G. Potard, *3D-audio object oriented coding*, Ph.D. thesis, University of Wollongong, 2006.

[21] P. A. Naylor, A. Kounoudes, J. Gudnason, and M. Brookes, "Estimation of glottal closure instants in voiced speech using the DYPSA algorithm," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 1, pp. 34–43, 2007.

[22] B. D. VanVeen and K. M. Buckley, "Beamforming: a versatile approach to spatial filtering," *IEEE Acoustic, Speech and Signal Processing Magazine*, vol. 5, no. 2, pp. 4–24, 1988.

[23] J. Makhoul, "Linear prediction: a tutorial review," *Proc. of the IEEE*, vol. 63, no. 4, pp. 561–580, 1975.

[24] E. A. P. Habets, *Single- and Multi-Microphone Speech Dereverberation using Spectral Enhancement*, Ph.D. thesis, Eindhoven University, 2007.

[25] P. Coleman and P. J. B. Jackson, "Planarity panning for listener-centered spatial audio," in *Proc. of the 55th Audio Engineering Society Conference (AES)*, Helsinki, Finland, 2014.

[26] H. Kuttruff, *Room acoustics - Fifth edition*, Spon press, 2009.

[27] L. Remaggi, P. J. B. Jackson, W. Wang, and J. A. Chambers, "A 3D model for room boundary estimation," in *Proc. of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Brisbane, Australia, 2015.

[28] H. L. Van Trees, *Optimum Array Processing - Part IV of Detection, Estimation and Modulation Theory*, Wiley-Interscience, 2002.

[29] M. R. Schroeder, "New method of measuring reverberation time," *J. of the Acoustical society of America*, vol. 37, pp. 409–412, 1965.

[30] J. M. Jot, "Real-time spatial processing of sounds for music, multimedia and interactive human-computer interfaces," *Multimedia Systems, Springer-Verlag*, vol. 7, no. 1, pp. 55–69, 1999.

[31] U. Zölzer, *DAFX-Digital Audio Effects*, John Wiley & Sons, Ltd, 2002.

[32] P. Coleman, P. J. B. Jackson, M. Olik, and J. A. Pedersen, "Personal audio with a planar bright zone," *J. Acoustic Society of America*, vol. 136, no. 4, pp. 1725–1735, 2014.