Audio Engineering Society

Convention Paper

Presented at the 142nd Convention
2017 May 20–23, Berlin, Germany

Acoustic Room Modelling using a Spherical Camera for Reverberant Spatial Audio Objects

Hansung Kim[1], Richard J. Hughes[2], Luca Remaggi[1], Philip JB Jackson[1], Adrian Hilton[1], Trevor J. Cox[2], and Ben Shirley[2]

[1]*Centre for Vision Speech and Signal Processing, University of Surrey, UK*
[2]*Acoustics Research Centre, University of Salford, UK*

Correspondence should be addressed to Hansung Kim (h.kim@surrey.ac.uk)

ABSTRACT

The ability to predict the acoustics of a room without acoustical measurements is a useful capability. The motivation here stems from spatial audio reproduction, where knowledge of the acoustics of a space could allow for more accurate reproduction of a captured environment, or for reproduction room compensation techniques to be applied. A cuboid-based room geometry estimation method using a spherical camera is proposed, assuming a room and objects inside can be represented as cuboids aligned to the main axes of the coordinate system. The estimated geometry is used to produce frequency-dependent acoustic predictions based on geometrical room modelling techniques. Results are compared to measurements through calculated reverberant spatial audio object parameters used for reverberation reproduction customized to the given loudspeaker set up.

1 Introduction

Accurate knowledge of the acoustics of an environment allows several advantages. In the acoustic design of spaces, either existing or at the planning stage, Room Impulse Responses (RIRs) can be used to predict aspects such as strong echoes, clarity or Reverberation Time (RT60) [1, 2, 3]. This can highlight potential issues and help inform solutions to improve the overall acoustic. Through application of spatial audio techniques these environments can also be reproduced/auralized [4], allowing the listener to experience a space without being there. Although measurements can provide this information, they are inherently restricted to pre-existing spaces, and the number of required measurements for some applications can rapidly become impractical. Consequently acoustic predictions offer an attractive alternative.

One area of interest is the application of room modelling to compensation techniques for spatial audio reproduction in rooms. If the RIR at the listening position for each loudspeaker is known, it is possible to adjust the loudspeaker signals to compensate for alterations in the frequency response, strong early reflections, or to some extent the level of reverberation [5, 6]. This is particularly the case in the context of recent interest in object-based audio, where more control is passed to a renderer at the listener end [7]. For instance, recorded

RIRs can be parameterized to generate Reverberant Spatial Audio Objects (RSAOs) [8]. However, by estimating the room geometry, predictions can be made when acoustic measurements are not available. This scenario also fits new research areas, such as mixed reality [9]. In this paper, RSAOs estimated from visual room geometry estimation are evaluated.

For simulation of an acoustic environment it is necessary to know the room geometry. Spaces such as domestic living rooms on the reproduction side, or recording environments on capture, ordinarily have unknown geometries and interior designs. Consequently, a robust method for obtaining the geometry is needed. There have been many studies into indoor scene geometry reconstruction from various visual sensors such as a normal camera, video camcorder and RGB+Depth (RGBD) camera [10, 11, 12, 13]. However, the limited field-of-view presents a challenging problem to ensure complete scene coverage; acquisition and processing of long video sequences is required for complete reconstruction. Another problem is that they produce a high-level of redundancy which is not necessary for acoustic room modelling.

Cuboid-based simplified room geometry modelling using a spherical camera provides a potential solution for the above problems. The room interiors are assumed to be composed of piecewise planar surfaces aligned to the main axes (Manhattan world) as introduced in [14]. Although not always the case, rooms - and the larger objects within - very often fit well with this assumption. In our previous research, we have used the Spheron VR[1], a mechanically calibrated line-scan camera for simplified scene modelling [15]. Recently various inexpensive off-the-shelf 360° cameras have become popular[2,3,4]. The room geometry modelling method used in this work extends the alignment process to 3 degree-of-freedom (DOF) for a commodity spherical camera (Ricoh Theta S[4]). In this paper an automatic cuboid-based room geometry estimation method is proposed. This produces a more complete scene model with a compact representation for acoustic predictions.

The main contributions of this paper are:

[1]Spheron, https://www.spheron.com/products.html [Accessed 27 Feb. 2017]

[2]LG 360, http://www.lg.com/uk/lg-friends/lg-LGR105/ [Accessed 27 Feb. 2017]

[3]Samsung Gear 360, http://www.samsung.com/uk/wearables/gear-360-c200/ [Accessed 27 Feb. 2017]

[4]Ricoh Theta, https://theta360.com/en/ [Accessed 27 Feb. 2017]

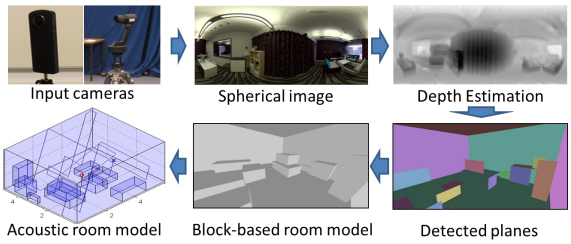


Fig. 1: Block diagram of the proposed system

- Estimation of complete room geometry, including internal objects, using off-the-shelf spherical cameras.
- Application of visual geometry estimates as inputs to a room acoustic model.
- Comparison of predicted and measured RIRs through derived RSAOs.

The rest of this paper is organised as follows: Section 2 outlines overview of the proposed system and describes details of the proposed methods. Section 3 presents system set up and datasets for experiments. Experimental results and discussion are given in Section 4, and Section 5 makes conclusions of this work.

2 Proposed method

2.1 System Overview

A pipeline is proposed for estimating acoustic RIRs from visual capture information. The examples presented are for loudspeaker based spatial audio setups in listening room environments, but the principle is extendible to other applications. Figure 1 shows a block diagram for the whole process. In 3D geometry estimation, a simplified structured room model is reconstructed using cuboids from spherical stereo image pairs. A full surrounding scene is captured by a spherical camera at two different heights and mapped to equirectangular images. They are aligned to the room coordinate axes by cubic projection and line alignment. Depth information of the scene is retrieved by disparity estimation and planar regions are detected. Cuboid elements are fitted to the detected planes to generate a complete cuboid-based room model. In addition, object classes for each region are predicted with a multi-scale Convolutional Neural Network (CNN). This geometry and object information is used as an input to acoustic

room modelling pipeline which is based on geometrical acoustics assumptions. Frequency dependent acoustic simulations are broken down into three sections: early reflections derived from an image source model (ISM) (providing a more deterministic early temporal response); later reflections and onset of the reverberant decay follow a ray tracing approach; and the late reverberant tail using Gaussian shaped and filtered white noise, with an envelope based on the decay of the preceding solution.

2.2 Visual Capture System and Image Alignment

To recover 3D scene information, the scene is captured as a vertical stereo image pair with the spherical camera. The spherical coordinate system of each camera can be misaligned relative to one another or to the room coordinate system. For image alignment to the room coordinate system, cubic projection and Hough-line based optimisation as proposed in [15] are utilised. However, since this was designed to find the optimal z-axis rotation for the mechanically tuned industrial camera, the method was extended to 3 DOF (x-axis (α), y-axis (β) and z-axis (γ)) optimisation for a normal spherical camera which can be less accurate in alignment. The optimal α , β and γ values are found by Eq. (1), where k represents the k-th face image in the cubic projection, H the lines detected by the Hough line detection, and C the cubic projection of the image I. The Hough lines are categorised into general Hough lines H, horizontal Hough lines H^h, and vertical Hough lines H^v, where horizontal and vertical Hough lines represent those detected parallel and perpendicular to the horizon within 1° of angle tolerance. Figure 2 shows an example of a stereo alignment result.

$$(\alpha_{opt},\beta_{opt},\gamma_{opt}) = \underset{\alpha,\beta,\gamma}{\operatorname{argmax}} \sum_{k=1}^{6} \frac{|H_k^h(\alpha,\beta,\gamma)\cup H_k^v(\alpha,\beta,\gamma)|}{|H_k(\alpha,\beta,\gamma)|} \tag{1}$$

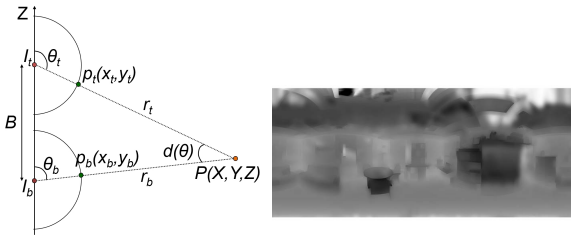
$$H_k(\alpha,\beta,\gamma) = H(C_k(R(\alpha,\beta,\gamma)I(x,y,z)))$$

2.3 Object Classification and 3D Geometry Reconstruction

Our CNN architecture for semantic object classification was built on the design of [16] and modified for colour, depth and surface normal inputs from stereo matching. Cubic projection images from the image alignment are used as the input of the CNN because the spherical

(a) Original spherical image pair (b) Aligned spherical image pair

Fig. 2: Example image alignment result (Meeting room (MR) data)



(a) Spherical stereo geometry (b) Disparity (depth) map

Fig. 3: Depth reconstruction

image is not appropriate for this architecture due to its distortion from the spherical coordinate. Top and bottom images of the cubic projection have very little information for object recognition so they are forced to be labelled as "ceiling" and "floor", respectively. Zheng et al. introduced a surface attributes detection method using CNN [17] but it is restricted to a few material categories and the result is not reliable. Therefore, acoustic coefficients are manually assigned from the object classification result.

3D geometry of the scene is reconstructed using correspondence matching with spherical stereo geometry illustrated in Fig. 3 (a). When disparity $d(\theta)$ is the angle difference between θ_b and θ_t, the distance of a certain 3D point P from the top camera is calculated:

$$r_t = B/\left(\frac{\sin\theta_t}{\tan(\theta_t+d)} - \cos\theta_t\right) \tag{2}$$

Any correspondence matching algorithm can be used, but here the variational approach [18] has been used as region-based matching methods suffer errors from spherical image distortion. The regions 5° from the epipoles are cropped because depth from disparity diverge near the epipole areas according to the spherical stereo geometry.

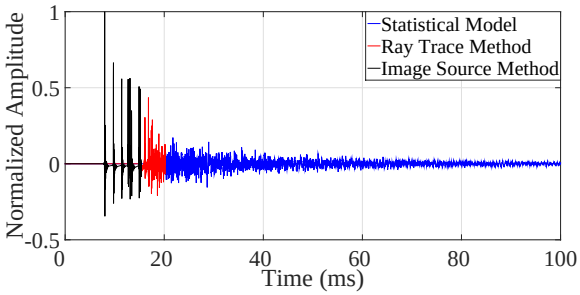


Fig. 4: RIR simulated by joint techniques.

In order to build a complete cuboid-based proxy structure, a block world reconstruction method is used [15]. Planes aligned to the main axes are firstly detected. Unreliable planes which are too distant from the camera or whose angle to the camera is too big are eliminated. Planes close each other are merged into one plane to simplify the scene. Finally cuboid structures fitted to the plane elements are reconstructed. In order to get a closed space for acoustic predictions, the largest and farthest planes in each direction are considered as walls for the room layout and their surface normals are set to the inside of the room layout. All other planes are used for cuboid structure generation by outward extrusion process from the camera capture position and the face normals are set outward (i.e. internal room objects).

2.4 Acoustic Modelling

The acoustic RIR modelling was achieved using a geometrical acoustic approach [3]. Whilst this method is generally more accurate for medium to large scale spaces, the technique is suited to medium to high frequencies and provides a useful estimate of time and direction of arrival of predicted reflections [3].

For each source and receiver pair the model was broken down into 3 sections for efficiency. The early reflections were modelled using an ISM technique [19], which provides a more deterministic estimation of the early temporal response than stochastic methods. The later reflections and onset of the reverberant decay were modelled stochastically using a ray tracing approach [20], with the scattering coefficient used to determine the probability of specular and non-specular reflections. The late reverberant tail was modelled as Gaussian shaped and filtered white noise, with an envelope based on the decay of the ray-traced solution. The response was calculated in octave bands from 63 Hz to 8.0 kHz, with a summation providing the end result. The temporal threshold separating early and later reflections was calculated as the median of the second order reflection times of arrival (TOAs). This was done to define as early most of the first order reflections. The reverberation onset time (i.e. the mixing time) was calculated from the visually estimated room geometry, exploiting a model-based perceptual mixing time [7, 21]:

$$T_{\text{mix}} = 20 \cdot \frac{V}{S} + 12, \tag{3}$$

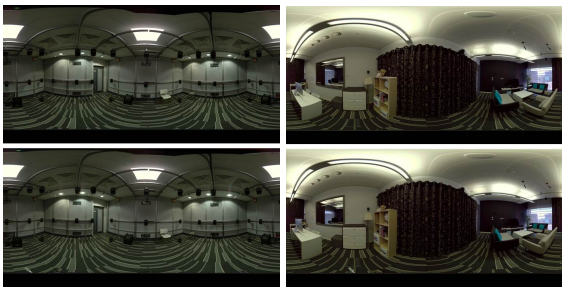
in milliseconds, where V is the room volume and S is the total reflective surface area.

Manual input of source and receiver coordinates (based on known distance from the back-right corner of the room) and surface materials are used at this stage, with the focus being on assessing the effect of the room geometry used for predictions. The materials selected were based on simple surface types for which acoustic absorption coefficient and approximate scattering coefficient data were available (e.g. plasterboard walls, carpet)[22]. The intention was to use material types which could more realistically be identified by visual techniques in future work. An example of a simulated RIR is reported in Figure 4.

3 System set up and datasets

3.1 Visual Capture

For the visual capture of the room, two different spherical cameras introduced in Section 1 were tested: Spheron and Theta S. The proposed pipeline was evaluated for three different spaces: BBC spatial audio listening room (LR, Fig. 5 (a), Spheron), BBC usability lab (UL, Fig. 5 (b), Spheron) and University of Surrey meeting room (MR, Fig. 2, Theta S). The LR is a more controlled listening environment approximately 5.6 m $\times$ 5.0 m $\times$ 2.9 m with loudspeakers surrounding a central listening position. The UL and MR are by design more representative of typical domestic living room environments, and are approximately 5.6 m $\times$ 5.2 m $\times$ 2.9 m and 5.6 m $\times$ 4.3 m $\times$ 2.3 m, respectively.



(a) Listening room (LR) (b) Usability lab (UL)

Fig. 5: Visual capture dataset images

3.2 Acoustic Measurements

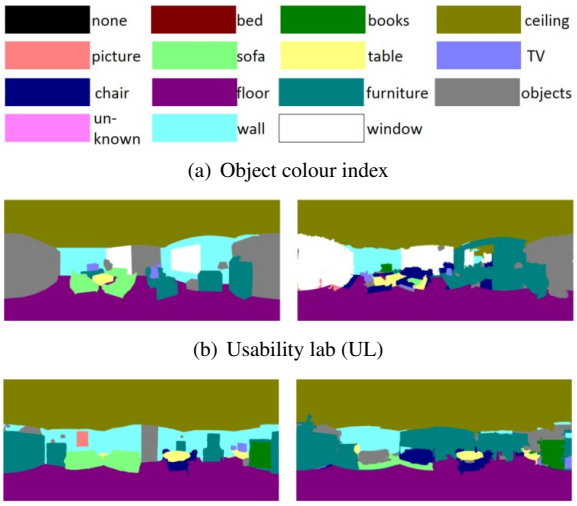
For each test environment a series of RIR measurements were taken using a swept sine method [23]. Both living room style environments had loudspeaker setups based on an ITU 5.0 surround sound setup [24], whilst the LR included a high channel count setup, formed by 32 loudspeakers. 48 microphone positions were recorded, evenly spaced around two concentric circles of radii 8.5 cm and 10.6 cm, respectively, to form a custom array [25]. Furthermore, at the center of the circular array, it was placed a soundfield microphone, recording additional RIRs.

4 Experimental Evaluation

RIRs generated from the room geometry modelling were parameterized following the RSAO concept [7, 8]. These parameters were then compared to the ones extracted from the recorded RIRs, to evaluate the room geometry estimation accuracy. Object materials were labelled to the most probable frequency-dependent absorption coefficient [22].

4.1 Visual Geometry Modelling

In object classification, we used the model of Eigen and Fergus [16] trained for the version of NYUDepth v2 dataset which was labelled with the 14 classes indexed in Fig. 6 (a). The training set consists of a set of 795 RGBD images, which was augmented using random transformations. Fig. 6 (b)-(c) show manually annotated ground-truth and estimated object class labels for the UL and MR. Most of the objects have been correctly classified but some false labels are observed in Sofa/Chair, Object/Furniture, Object/Wall, Picture/Wall and Wall/Furniture.



(a) Object colour index

(b) Usability lab (UL)

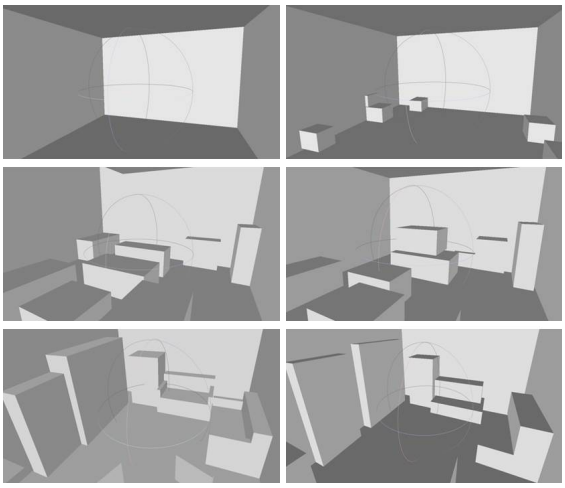
(c) Meeting room (MR)

Fig. 6: Object classification results (Left: Ground-truth, Right: Estimated labels)

Figure 7 (a) and (b) show the manually generated ground-truth models of the rooms from the actual measurements and the reconstructed cuboid-based models, respectively. In the ground-truth model of the LR, most small objects are eliminated to simplify the acoustic model, since their effect will be small and their modelling inaccurate. The size of the reconstructed LR is 5.85 m $\times$ 5.1 m $\times$ 2.9 m and includes a few cuboids representing loudspeakers and chairs. For the UL, the thin monitor on the table which was neglected in the ground-truth model was reconstructed as a thick cuboid because the thickness could not be estimated from the images, and the table in the corner was missing because it was occluded by the monitor. The MR was captured by the Theta camera which is less accurately rectified and aligned but the cuboid primitives represent the approximate structure of the scene well. The estimated room sizes are 6.1 m $\times$ 5.0 m $\times$ 2.9 m for the UL and 6.15 m $\times$ 4.7 m $\times$ 2.45 m for the MR, which are close (<10% error) to the original sizes.

4.2 RSAO Parameters

The RSAO parameters can be divided into two groups, depending on the RIR part they belong to [8]: the direct sound and the early reflection parameters are described by their TOAs and Directions Of Arrival (DOAs); the reverberation parameters are the late energy decays.



(a) Ground-truth model (b) Reconstructed model

Fig. 7: Room geometry estimation results (Top: LR, Middle: UL, Bottom: MR)

The mixing time and frequency-dependent RT60s, calculated from the energy decay, were analyzed for a comprehensive comparison [7].

To extract the TOAs from measurements, the clustered-dynamic programming projected phase-slope algorithm (C-DYPSA) was employed [25]. It is based on the DYPSA algorithm [26], and was used to calculate the time domain peaks on each of the 48 microphone RIRs. A clustering technique was then utilized to eliminate outliers, considering every k-th reflection, over the 48 DYPSA outputs. The mean of the inlier TOAs corresponds to the TOA parameter $\overline{\tau}_{k,l}$, where l is the loudspeaker index. Both azimuth and elevation DOAs were extracted from the RIRs using a delay-and-sum beamformer (DSB) technique [27]. To avoid the up-down ambiguity given by the planar microphone array, the soundfield microphone at the center of it was employed. To apply the DSB, the RIRs were first segmented, by applying a Hamming window (heuristically obtained length of 2.5 ms for UL and LR, 0.8 ms for MR), for each RIR, centered at $\overline{\tau}_{k,l}$ [8]. The simulated RIRs were generated by virtually placing a single microphone at the center of the microphone array used for the recordings. Thus, TOAs and DOAs were calculated by directly observing the image source positions.

For the mixing time, Eq (3) was used [21]. The RT60 was calculated for each octave band between 125 Hz and 8 kHz, by analyzing the first 20 dB of decaying late energy after T_{mix} [7].

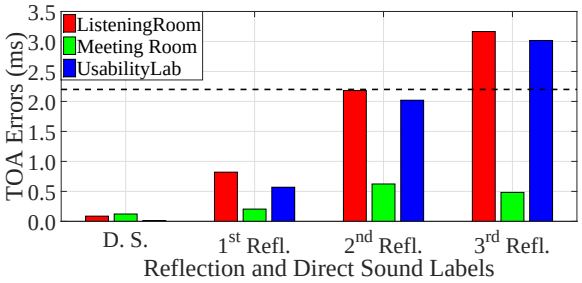


Fig. 8: The median TOA errors E_k^{TOA}. The dashed line indicates the JND.

4.3 Evaluation Metrics

For comparison, the RSAO parameters were calculated for both simulated and recorded RIRs. The TOA parameter errors $\varepsilon_{k,l}^{\text{TOA}}$ were calculated as the absolute value of the difference between the TOA obtained from the simulated and the recorded RIRs, considering the direct sound ($k = 0$) and each k-th early reflection, separately. The evaluated error E_k^{TOA} was then calculated as the median over the L loudspeakers available. Similarly, the DOA errors $\varepsilon_{k,l}^{\text{DOA}}$ were obtained as the absolute value of the difference between the DOA calculated from the simulated and recorded RIRs. This was done for both azimuth and elevation, separately. As for E_k^{TOA}, also the evaluated error E_k^{DOA} was obtained as the median over the L loudspeakers. The mixing time error E^{MT} was calculated as the difference between the mixing times obtained through the estimated room geometry and the geometry groundtruth. Here, the absolute value was not calculated. Finally, the RT60s for both simulated and recorded RIRs were estimated, for each octave bands. The median of the simulated RT60s was then calculated over the L loudspeakers.

4.4 Results and Discussion

The three different rooms LR, UL and MR are evaluated. The early reflections, that were analyzed, were the ones simulated through the image source method. Whenever it occurred that multiple reflections were detected within an interval of 1 ms, only the reflection with the least TOA error $\varepsilon_{k,l}^{\text{TOA}}$ was included into the analysis. Following this removal process, only the direct sound and the first three reflections were analyzed. Acoustic simulations were run for the manually labelled surface material types. An additional experiment on MR was performed: a vision-based algorithm

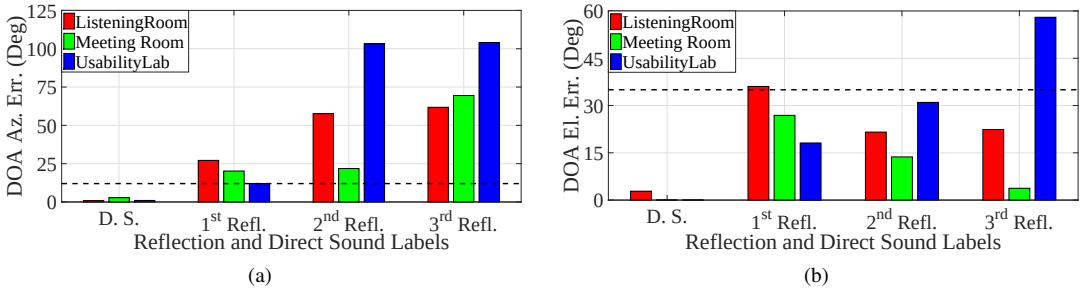


Fig. 9: The DOA azimuth (a) and elevation (b) errors. The dashed lines indicate the JNDs.

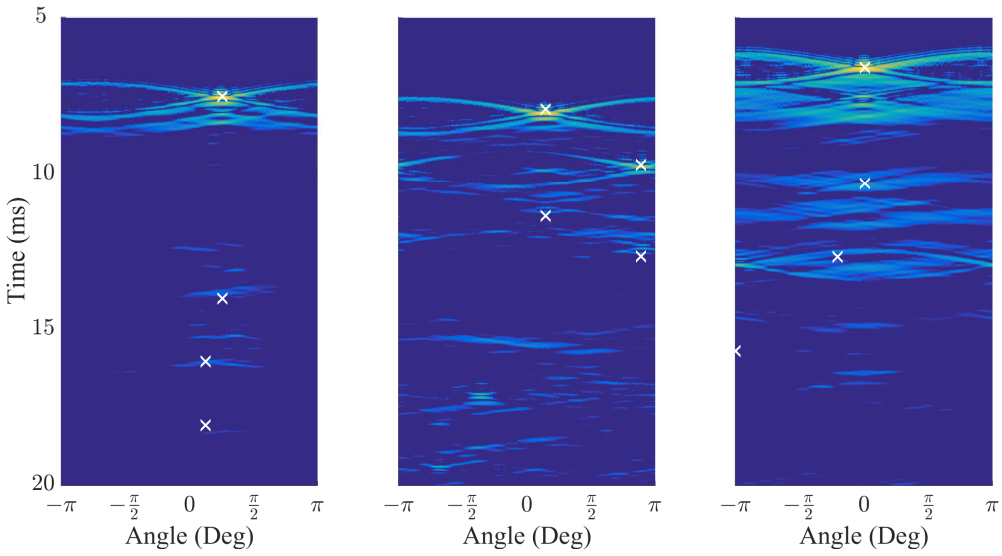


Fig. 10: The temporal evolution of the measured RIR DOAs, for LR (left), MR (center), and UL (right), selected as example. The red crosses indicate the TOA-DOA positions of the simulated RIR early reflections.

was used to recognize the materials [28], thus estimating the absorption coefficients automatically.

4.4.1 TOA and DOA Results

TOA errors E_k^{TOA} are shown in Fig. 8. It is evident that, as was expected, the error increases with the reflection order. This is due to the propagation of the errors introduced by the room geometry estimation. Among the three datasets, MR is the one where the performance in terms of TOAs is best, whereas LR is the worst. However, all the datasets have, up to the second reflection, errors lower than 2.2 ms, corresponding to 75 cm, and it is the minimum error perceived by humans [29].

Figure 9 (a) and (b) shows DOA azimuth and elevation errors, respectively. The DOA errors E_k^{DOA} increase with the reflection order. DOA results are also affected when multiple reflections arrive at the microphones within the same time window. In fact, the RIR segment that was extracted to calculate the DOA, usually contains more than one reflection, not allowing the DSB to determine a specific DOA. This problem reduced the performance, in particular for reflections later than the first. This observation follows the findings in [30], where it was described how the first reflection plays a major role in the human auditors perception of rooms. For the first reflection, both azimuth and elevation have errors in UL lower than their just-noticeable differences (JNDs), that are defined as 12° and 35°, respectively [29]. Furthermore, although in MR and LR the errors are slightly above the limit set by the azimuth JND,

when a more detailed evolution of the wavefronts is investigated, as shown in Fig. 10, the DOAs appear consistent with the measurements, perhaps suggesting errors in the measured DOAs. In this figure example, the three dataset TOAs and DOAs are compared, by plotting the temporal evolution of the energy arriving from each direction at the microphone array position [31]. These beamformed signals were obtained by steering the DSB towards each azimuth direction with a resolution of one degree. The TOA-DOA positions of the simulated RIR direct sound and early reflections were then overlaid as white crosses. From this figure, it can be seen that direct sound and first reflection are generally well estimated for every dataset. However, although TOAs are still well estimated, for higher order reflections, the simulated DOAs accuracy drop.

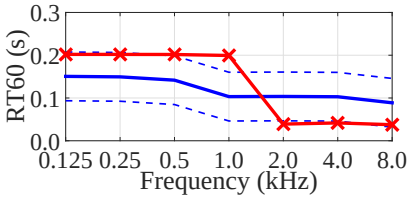
4.4.2 Mixing Time and RT60 Results

The mixing time errors E^{MT} are shown in Table 1. Since this parameter is calculated from the volume and reflective surface values, UL and MR produced larger error than LR. In fact, they are living room-like environments, containing several objects that the visual geometry estimation method could not accurately reconstruct. Furthermore, their negative values indicate the general trend of overestimating the room object sizes. On the other hand, LR is an empty room, thus, it was easier for the reconstruction method to estimate the geometry.

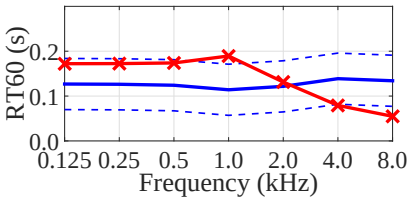
The RT60 results are shown in Figure 11. Produced by manually labelling the materials, the RT60s of all the datasets are well estimated and values at most of the frequencies are within the band defined by the JND. For the RT60, we assumed the JND as 57 ms [32].

4.4.3 Reflector Material Experiment Results

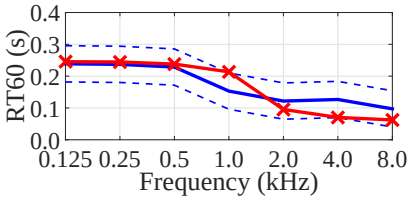
The last experiment's results are reported in Figure 12, where early decay time (EDT) and RT60 are reported, for the MR dataset only. Comparisons of the simulated quality were made by: manually assigning the reflectors' absorption coefficients; estimating the reflector materials using a visual signal processing method [28]. Results show again that, for manual labels, both the RT60 and EDT appear to be well estimated, being coherent with their JNDs (for EDT, the JND is 5 % [33]). Moreover, although simulations with estimated materials had lower performance, they are still inside their JND range for RT60, and close to the EDT perceptual band, thus shaping interesting paths for future work.



(a) LR dataset



(b) MR dataset



(c) UL dataset

Fig. 11: The RT60s for the simulated (red curves) and recorded (blue curves) RIRs. The object materials were manually labelled. The dashed lines are the JNDs.

Table 1: Mixing time errors, in ms and %.

LR	MR	UL
-0.07 ms (0.3 %)	-1.39 ms (6.5 %)	-3.39 ms (13.2 %)

5 Conclusions

A method has been outlined for prediction of room acoustic RIRs based on visually captured information. Room geometry was estimated through vertical spherical stereo systems using commercial off-the-shelf cameras. Aligned cuboid representations of the room were reconstructed using spherical stereo geometry. Experiments were conducted by comparing the RSAO parameters of the estimated RIRs with the ones extracted from recordings. Results show plausible agreement between predictions obtained using estimated geometry and measurements.

Future extension of this research will include robust material detection in the room geometry modelling to

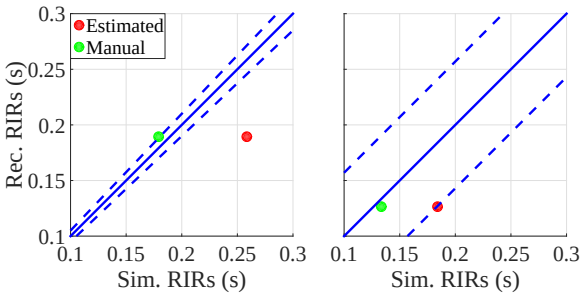


Fig. 12: Comparison between EDTs (left) and RT60s (right) calculated from the recorded and simulated RIRs, of MR. The different colors indicate different methods to label the object materials. The dashed lines are the JNDs.

replace the current manually defined surface materials. The room model is also being extended to include more accurate wave-based methods (e.g. Finite Difference Time Domain (FDTD)) which are inherently suited to the cuboid based geometry method presented. This work, still in progress, provides a step toward acoustic room model reconstruction using audio-visual data.

6 Acknowledgements

This work was supported by the EPSRC Programme Grant S3A: Future Spatial Audio for an Immersive Listener Experience at Home (EP/L000539/1) and the BBC as part of the BBC Audio Research Partnership. Details about the data underlying this work, along with the terms for data access, are available from: `http://dx.doi.org/10.15126/surreydata.00812228`. The authors would like to thank Philip Coleman and Sam Fowler for their role in capturing the measurement data.

References

[1] ISO 3382-2, "Acoustics - Measurement of room acoustic parameters - Part 2: Reverberation time in ordinary rooms," Standard, International Organization for Standardization, Geneva, Switzerland, 2008.

[2] Bork, I., "Report on the 3rd Round Robin on Room Acoustical Computer Simulation Part II: Calculations," *Acta Acustica united with Acustica*, 91(4), pp. 753–763, 2005, ISSN 1610-1928.

[3] Savioja, L. and Svensson, U. P., "Overview of geometrical room acoustic modeling techniques," *The Journal of the Acoustical Society of America*, 138(2), pp. 708–730, 2015, doi:http://dx.doi.org/10.1121/1.4926438.

[4] Vorländer, M., *Auralization – Fundamentals of Acoustics, Modelling, Simulation, Algorithms and Acoustic Virtual Reality*, Springer-Verlag, Berlin, Heidelberg, 2008, ISBN 978-3-540-48829-3.

[5] Kajita, S., Takeda, K., and Itakura, F., "Compensation of Room Acoustic Transfer Functions Affected by the Change of Room Temperature," in *Audio Engineering Society Convention 107*, 1999.

[6] Poletti, M. A., Betlehem, T., and Abhayapala, T. D., "Higher-Order Loudspeakers and Active Compensation for Improved 2D Sound Field Reproduction in Rooms," *J. Audio Eng. Soc.*, 63(1/2), pp. 31–45, 2015.

[7] Coleman, P., Franck, A., Jackson, P. J. B., Hughes, R., Remaggi, L., and Melchior, F., "Object-based reverberation for spatial audio," *J. Audio Eng. Soc.*, 65(1/2), pp. 66–77, 2017.

[8] Remaggi, L., Jackson, P. J. B., and Coleman, P., "Estimation of room reflection parameters for a reverberant spatial audio object," in *Audio Engineering Society Convention 138*, 2015.

[9] Ohta, Y. and Tamura, H., *Mixed reality: mergin real and virtual worlds*, Springer Publishing Company, Incorporated, 2014, ISBN 3642875149, 9783642875144.

[10] Furukawa, Y., Curless, B., Seitz, S. M., and Szeliski, R., "Reconstructing building interiors from images," in *Proceedings of ICCV*, 2009.

[11] Xiao, J. and Furukawa, Y., "Reconstructing the World's Museums," *International Journal of Computer Vision*, 110(3), pp. 243–258, 2014.

[12] Choi, S., Zhou, Q.-Y., and Koltun, V., "Robust Reconstruction of Indoor Scenes," in *Proceedings of CVPR*, 2015.

[13] Chen, K., Lai, Y.-K., and Hu, S.-M., "3D indoor scene modeling from RGB-D data: a survey," *Computational Visual Media*, 1(4), pp. 267–278, 2015.

[14] Gupta, A., Efros, A. A., and Hebert, M., "Blocks World Revisited: Image Understanding Using Qualitative Geometry and Mechanics," in *Proceedings of ECCV*, 2010.

[15] Kim, H. and Hilton, A., "Block world reconstruction from spherical stereo image pairs," *Computer Vision and Image Understanding*, 139, pp. 104–121, 2015.

[16] Eigen, D. and Fergus, R., "Predicting Depth, Surface Normals and Semantic Labels with a Common Multi-Scale Convolutional Architecture," in *Proc. ICCV*, 2015.

[17] Zheng, S., Cheng, M.-M., Warrell, J., Sturgess, P., Vineet, V., Rother, C., and Torr, P. H. S., "Dense Semantic Image Segmentation with Objects and Attributes," in *Proc. CVPR*, 2014.

[18] Kim, H. and Hilton, A., "3D Scene Reconstruction from Multiple Spherical Stereo Pairs," *International Journal of Computer Vision*, 104(1), pp. 94–116, 2013.

[19] Borish, J., "Extension of the image model to arbitrary polyhedra," *The Journal of the Acoustical Society of America*, 75(6), pp. 1827–1836, 1984, doi:http://dx.doi.org/10.1121/1.390983.

[20] Vorländer, M., "Simulation of the transient and steady-state sound propagation in rooms using a new combined ray-tracing/image-source algorithm," *The Journal of the Acoustical Society of America*, 86(1), pp. 172–178, 1989, doi:http://dx.doi.org/10.1121/1.398336.

[21] Lindau, A., Kosanke, L., and Weinzierl, S., "Perceptual evaluation of model- and signal-based predictors of the mixing time in binaural room impulse responses," *J. Audio Engineering Society*, 60(11), pp. 887–898, 2012.

[22] Cox, T. and D'Antonio, P., *Acoustic absorbers and diffusers, third edition: theory, design and application*, CRC Press, 2016, ISBN 9781498740999.

[23] Farina, A., "Simultaneous Measurement of Impulse Response and Distortion with a Swept-Sine Technique," in *Audio Engineering Society Convention 108*, 2000.

[24] Rec. ITU-R BS.775-3, "Multichannel stereophonic sound system with and without accompanying picture," Recommendation, International Telecommunication Union, Geneva, Switzerland, 2012.

[25] Remaggi, L., Jackson, P. J. B., Coleman, P., and Wang, W., "Acoustic reflector localization: novel image source reversion and direct localization methods," *IEEE/ACM Transactions on Audio, Speech and Language Processing*, 25(2), pp. 296–309, 2017.

[26] Naylor, P. A., Kounoudes, A., Gudnason, J., and Brookes, M., "Estimation of glottal closure instants in voiced speech using the DYPSA algorithm," *IEEE Transactions on Audio, Speech, and Language Processing*, 15(1), pp. 34–43, 2007.

[27] VanVeen, B. D. and Buckley, K. M., "Beamforming: a versatile approach to spatial filtering," *IEEE Acoustic, Speech and Signal Processing Magazine*, 5(2), pp. 4–24, 1988.

[28] Kim, H., d. Campos, T., and Hilton, A., "Room Layout Estimation with Object and Material Attributes Information Using a Spherical Camera," in *Fourth International Conference on 3D Vision (3DV)*, 2016.

[29] Middlebrookes, J. C. and Green, D. M., "Source localization by human listeners," *Annual Review of Psycology*, 42(1), pp. 135–159, 1991.

[30] Bech, S., "Spatial aspects of reproduced sound in small rooms," *The Journal of the Acoustical Society of America*, 103(1), pp. 434–445, 1998.

[31] Remaggi, L., Jackson, P. J. B., Coleman, P., and Francombe, J., "Visualization of compact microphone array room impulse responses," in *Proc. of the 139th Audio Engineering Society Convention (AES)*, New York, USA, 2015.

[32] Niaounakis, T. I. and Davis, W. J., "Perception of reverberation time in small listening rooms," *J. Audio Engineering Society*, 50(5), pp. 343–350, 2002.

[33] Vorländer, M., "International round robin on room acoustical computer simulations," in *Internaltional Congress on Acoustics 15*, 1998.