# Immersive Spatial Audio Reproduction for VR/AR Using Room Acoustic Modelling from 360° Images

Hansung Kim, Luca Remaggi, Philip J.B. Jackson and Adrian Hilton*
CVSSP, University of Surrey
Guildford, UK

**ABSTRACT**

Recent progresses in Virtual Reality (VR) and Augmented Reality (AR) allow us to experience various VR/AR applications in our daily life. In order to maximise the immersiveness of user in VR/AR environments, a plausible spatial audio reproduction synchronised with visual information is essential. In this paper, we propose a simple and efficient system to estimate room acoustic for plausible reproducton of spatial audio using 360° cameras for VR/AR applications. A pair of 360° images is used for room geometry and acoustic property estimation. A simplified 3D geometric model of the scene is estimated by depth estimation from captured images and semantic labelling using a convolutional neural network (CNN). The real environment acoustics are characterised by frequency-dependent acoustic predictions of the scene. Spatially synchronised audio is reproduced based on the estimated geometric and acoustic properties in the scene. The reconstructed scenes are rendered with synthesised spatial audio as VR/AR content. The results of estimated room geometry and simulated spatial audio are evaluated against the actual measurements and audio calculated from ground-truth Room Impulse Responses (RIRs) recorded in the rooms.

**Index Terms:** Human-centered computing—Human computer interaction (HCI)—Interaction paradigms—Virtual reality; Computing methodologies—Artificial intelligence—Computer vision—Scene understanding

## 1 INTRODUCTION

Virtual reality (VR) and augmented reality (AR) are currently major topics in the media researches to create new immersive experiences [4, 30]. Spatial audio is important for the sense of immersion in a virtual or augmented space but previous research has primarily focused more on visual experience. Recent researches have been extended to include spatial audio because human perception relies on both audio and visual information to understand and interact with the environment [43, 44]. Previous research has shown that spatio-temporal synchronisation of sound with visual information improves the sense of presence in virtual and augmented environments [25].

A plausible and coherent audio-visual reproduction can be achieved by understanding the scene geometry and related materials. The best way to reproduce the acoustic design of spaces is to measure Room Impulse Responses (RIRs) for the space [19, 37, 39]. However, it is sometimes difficult to obtain actual acoustic measurements for a certain environment considering practical applications of VR and AR. For example, setting up microphones and speakers for acoustic measurements may be too invasive to be deployed at private spaces like living rooms or bedrooms. Furthermore, RIR is only valid at the single point of measurement and will change with location in the scene. What is required for immersive experiences where the user moves through the space is the acoustic modelling of the environment to allow rendering of spatial audio according to the listener location. It is impractical to measure or update RIRs according to the changes of geometry or user positions for interactive dynamic scene rendering. A few methods have been proposed to reproduce scene-aware spatial audio reproduction from a single recording [26] and self-supervised deep learning [32] but they were only for 360° video rendering.

Instead of direct RIR measurements using audio recording in the space, computer vision techniques can be utilised to predict room acoustics. Recently, several toolkits have been also developed to render spatial audio from the geometry and acoustic material information on VR/AR platforms [13, 31]. 3D Models describing both geometry and materials allow to approximate real room acoustics for VR/AR environments [17, 23]. It has been demonstrated that high-quality sound reproductions improves the perceived similarity to reference environments [6, 36].

For simulating an acoustic environment in those platforms, a robust recognition method for room geometry and object materials is required. The closest work for this goal is the work by Schissler et al. [38] using a Microsoft Kinect sensor. They built a dense 3D geometry using a Shape-from-Motion technique from several hundreds of RGB+Depth images [10] and proposed a two-step procedure using a convolutional neural network (CNN) to estimate acoustic material properties for sound rendering. However, this approach using a normal RGB-D camera has several drawbacks as follows: 1) It requires time and resource consuming multiple captures of the scene to cover a complete scene layout estimation due to the limited field-of-views of the camera. 2) Dense geometry makes the real-time acoustic simulation impractical because it drastically increases computational complexity and run-time for spatial audio rendering.

It is well-known that human audio perception is not sensitive enough to recognise differences of sound from the change of geometrical details as long as the change is within the just-noticeable difference (JND) level [21]. Therefore, we suggest to use approximated geometry which allows the use of simple acoustic models to generate synthetic versions of the environment acoustics in a more efficient way. In this paper, we propose a simple pipeline for acoustic room modelling with cuboid-based room and object representation from a single pair of spherical 360° images. For cuboid model reconstruction, room interiors are assumed to be composed of planar surfaces aligned to the main axes (Manhattan world), as introduced in [15]. Generally, room layouts and large objects often fit this assumption. Objects in the scene are segmented and classified by a CNN-based semantic segmentation (SegNet) [2], and estimated acoustic properties are assigned to each object to build VR environments or AR references using Google Resonance Audio [12].

It is important to distinguish two terms "plausibity" and "authenticity" in acoustic reproductions. Plausibility describes the agreement of the heard scene with an inner reference (expectation) [27], while authenticity judges whether it is perceptually identical to an external reference [5]. Therefore it is considered that plausibility

*e-mail: h.kim@surrey.ac.uk. 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

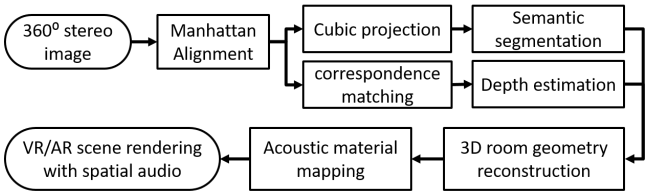



Figure 1: Block diagram of the proposed system

is more important for VR applications where real reference is not available for users, and authenticity is more important for AR applications where coherence of virtual sound with environmental sound can be easily noticeable. Some studies also showed that vision cues dominate over acoustic cues in human perception when they are simultaneously provided [3]. This means that the perceptual differences between real and synthetic acoustic environments in the presence of visual stimuli are not as strictly defined as they are for unimodal (sound only) scenarios.

In order to evaluate the "plausibity" and "authenticity" for VR and AR applications, RIRs generated in the VR environments are compared with the measured RIRs in the real environments. The acoustics reproduced in VR were evaluated by analysing the early decay time (EDT) and reverberation time (RT60) with the RIRs. We also provide a video of reconstructed VR scene rendering with spatial audio as a supplemental material.

## 2 System Overview

In this research, we propose a simple and efficient method to reproduce a 3D VR environment with acoustic properties from a vertical pair of 360° photos of a real scene. Figure 1 shows the block diagram for acoustic room modelling and VR reproduction with spatial audio in a normal room environment.

A full surrounding scene is captured by vertically aligned 360° cameras. Each camera has two fish-eye lens and two captured fish-eye images are mapped and stitched into an equirectangular image. They are aligned to the room coordinate axes by the Manhattan world alignment utilising cubic projection and the façade alignment techniques [22] which identify the principal directions. Then the process is split into two stages: semantic object classification and 3D scene reconstruction. Depth of the scene is estimated by dense correspondence matching between two images. For semantic scene segmentation and object classification, the equirectangular image is projected onto a unit cube centred on the camera to produce general perspective images and each projected image goes through the SegNet pipeline. The output labels from SegNet are back-projected to the original equirectangular format. Based on the object labels and depth information, object-labelled cuboids are reconstructed to represent the scene structure. Acoustic properties for the classified objects are assigned from the acoustic material list. Finally, the acoustic VR scene or AR sound is rendered by setting sound source and player models on the VR platform.

## 3 Proposed Methods

### 3.1 Visual capture and pre-processing

The scene is captured by two 360° cameras in order to recover 3D information from the pair. Previously this required accurately aligned high resolution spherical images from expensive industrial equipment such as Ladybug [33] and Spheron VR [41], but inexpensive off-the-shelf 360° cameras are now getting popular and provides good quality of scenes [14, 18, 35]. In this work we use two Ricoh Theta cameras [35] which provide accurately calibrated equirectangular photos aligned to the spherical coordinate system. We use a vertical stereo camera set up rather than typical horizontal stereo in order to decrease stereo matching errors due to texture distortions on the equirectangular images and minimise occlusions between cameras. Figure 2 shows the camera set up and captured top and bottom images of the Meeting Room (MR) scene used in the experiments.

(a) Camera set up (b) Captured images (Top and Bottom)

Figure 2: Visual capture system using a pair of Rico Theta cameras

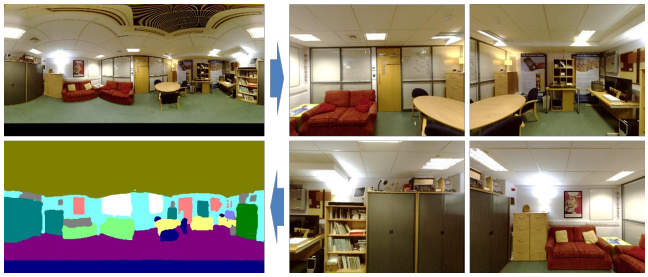
The two cameras may be slightly misaligned to each other, and resulting in stereo matching errors in high resolution capture. Images from the cameras also need to be aligned to the world (room) coordinate system. Cubic projection and Hough-line based façade alignment proposed in [22] are utilised to align both images to the room coordinate (Manhattan-world) system. This allows accurate matching between 360 image pairs even for consumer cameras.

### 3.2 Semantic segmentation

Semantic segmentation aims to segment the scene into semantically meaningful regions and label those regions with pre-defined classes. A good survey of semantic segmentation for RGB images is available in [47]. The traditional pipeline of semantic object classification has been recently replaced by CNN [8]. CNN-based semantic segmentation architectures are still actively being developed.

Estimation of acoustic properties from visual information alone is a challenging problem due to the inherent ambiguity [20]. A number of approaches have been introduced to detect material attributes from images [22, 46], but their accuracy is typically below 50%, for cross-dataset scenarios, which is too low to be considered as a suitable methods for estimating absorption and scattering coefficients of objects. Even though materials can be predicted from the visual information, it is still hard to define the acoustic parameters of those materials such as roughness, density and thickness of the surface on which acoustic properties depend. For instance, we can detect a carpet on the floor, but there is no way to detect its pile type and thickness. Therefore, we propose to use an object recognition method and map the object categories to approximated acoustic properties of materials in Section 3.4.

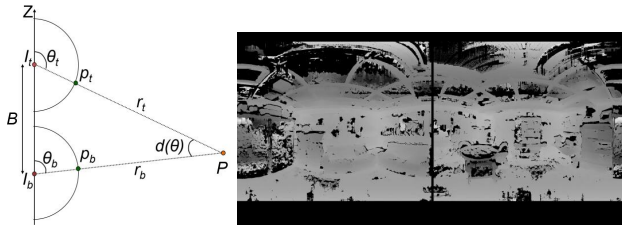
In the proposed pipeline, SegNet [2] is used for semantic segmentation and object labelling. SegNet provides a model trained on the SUN RGB-D indoor scenes dataset [40] to semantically segment structure and objects in indoor scene images. Four side images in the cubic projection in Section 3.1 are extended to 4:3 aspect ratio to be matched to the trained SUN RGB-D dataset format, and also to compensate recognition error at the image boundaries. the ceiling and floor face image are forced to be labelled as ceiling and floor. All output labels from the SegNet process are back-projected to provide a fully labelled equirectangular image. Finally, the labelled image is refined by a morphological opening process [11] to smooth object boundaries and remove small regions. Each labelled region



(a) Semantic segmentation process


(b) Colour code for object classes

Figure 3: Cubic projection and semantic segmentation




(a) Spherical stereo geometry



(b) Disparity (depth) map

Figure 4: Depth estimation using a pair of spherical stereo images

is considered as an independent object in 3D reconstruction. The cubic projection images and final label image are shown in Fig. 3.

### 3.3 Depth estimation

Depth information of the scene is estimated using correspondence matching with spherical stereo geometry illustrated in Fig. 4 (a). In the proposed vertical 360° stereo setup, real-scale depth can be directly estimated from simple stereo matching along 1D vertical lines in contrast with normal depth reconstruction from perspective stereo images which requires complex internal and external camera calibrations. Depth estimation in the proposed system requires only baseline distance $B$ and disparity information. When the angle disparity $d(\theta) = \theta_t - \theta_b$ for a certain 3D point $P$ is calculated from the two projected point $p_t$ and $p_b$ on the spherical coordinate, the real distance of the 3D point $P$ from the top camera is calculated as Eq. (1).

$$r_t = B / \left( \frac{\sin\theta_t}{\tan(\theta_t + d(\theta))} - \cos\theta_t \right) \quad (1)$$

For correspondence matching, any feature matching algorithm can be used for the image pair. We use the feature-based dense block matching method [24] which produces reliable disparity fields by detecting occlusion regions and ambiguous regions based on bi-directional consistency and the ordering constraint.

### 3.4 3D modelling and spatial audio rendering

All 2D points on the captured image are projected to the 3D space using the depth information estimated in the previous section. This

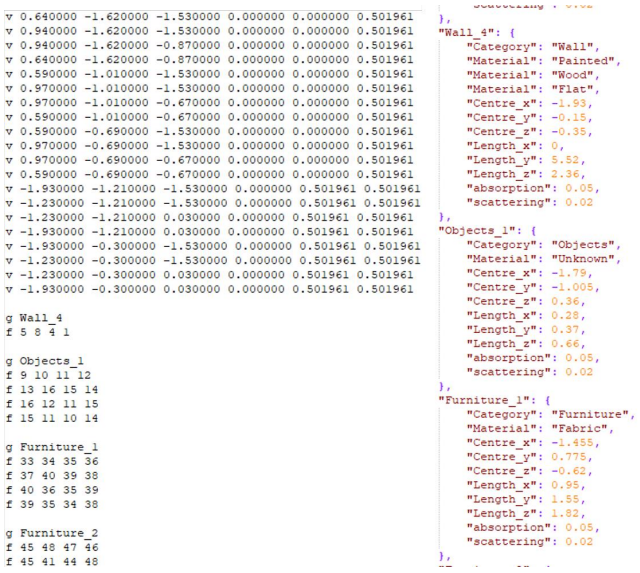



Figure 5: Examples of geometry OBJ (left) and metadata JSON (right) files

Table 1: Material matching to object.

| Object | Material | Object | Material |
|---|---|---|---|
| Ceiling | Wood panel | Furniture | Heavy curtain |
| Book | Sheetlock | Chair | Wood panel |
| Floor | Parquet | Object | Metal |
| Window | Thick Glass | Wall | Smooth Plaster |
| Sofa | Heavy curtain | Table | Wood panel |
| TV | Metal | Unknown | Transparent |

3D point cloud is segmented into clusters based on the object labels assigned in Section 3.2. From this object point clouds, block structures are reconstructed based on their point occupancy to build an approximated geometry of the scene. 10% of the farthest points from the centre of each cluster are eliminated as outliers to reduce errors from depth estimation and segmentation, then cuboid primitives aligned to a Manhattan world are fitted to the inlier point clouds. Finally the volume of the reconstructed cuboids are refined based on their physical stability [15]. Any floating primitives are extended to the ground and boxes near the wall are also extended to the wall because the narrow gap between objects increase the complexity in the sound field rendering.

The results of geometry reconstruction are saved as geometry and metadata files in "OBJ" and "JSON" formats, respectively, as shown in Fig. 5. This output is directly imported to Unity [42] to build a VR environment. The Resonance Audio package [12] by Google is used to simulate spatial audio in the Unity engine. Resonance Audio provides 22 types of materials with their acoustic attributes. As mentioned in Section 1, it is difficult to directly detect acoustic properties of materials from visual input. Therefore, we map the object labels to the material types in Resonance Audio as Table 1. We assign to the acoustically closest material when it was difficult to match the material for certain objects. Finally, a virtual listener and an audio source are placed in the scene to simulate the reconstructed virtual scene with spatial audio. Figure 6 illustrates the reconstructed simplified geometry of the MR scene in Fig. 2 and reproduced VR environment with virtual sound source and player. The reconstructed VR scene can be played with real-time interaction on any VR kit supported by the Unity engine. We used VIVE Pro [16], a VR headset supporting spatial audio in our experiments.

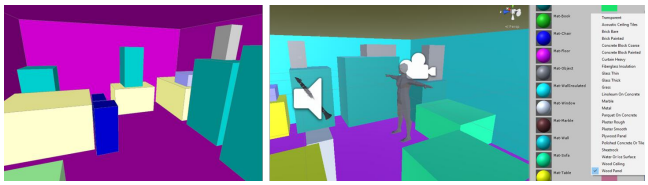
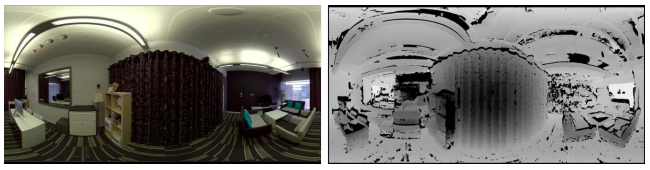
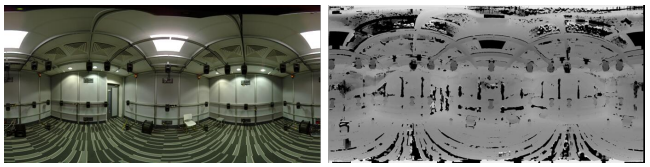


Figure 6: Reconstructed geometry (left) and reproduced VR environment with acoustic properties (right)



(a) Usability Lab (UL)



(b) Listening Room (LR)

(c) Studio Hall (ST)

Figure 7: Dataset used in the experiments

## 4 Experiments

The proposed system was evaluated for four different rooms. The Meeting Room (MR) and Usability Lab (UL) data are similar to typical domestic living room environments. Listening Room (LR) is an acoustically controlled room and Studio Hall (ST) is a large hall. The MR scene has been used as examples in Section 3, and all other datasets with their estimated depth maps are given in Fig. 7.

### 4.1 Geometry reconstruction

Figure 8 shows the results of semantic segmentation and 3D model reconstructions for the UL, LR and ST datasets. It is clear that the proposed SegNet-based method produced meaningful segmentation results for 360° image with cubic projection. Most objects were correctly classified including windows and mirrors. However some small objects were missing due to the postprocessing. These will not significantly affect the perceived acoustics. Snapshots of the reconstructed 3D models are visualised on the right column of Fig. 8 with colour-coded object labels and the full geometry are visible in the supplementary video. For efficient geometry representation, Pictures and Windows are merged to Wall, and Book labels to Furniture in the final scene reconstruction. In the snapshots, Ceilings and Floors were coloured with the Wall colour because they were represented as one cuboid, but they are decomposed into Ceiling and Floor in the acoustic material mapping.

The 3D reconstruction and recognition process has been run on a normal PC with a Intel Core i7 3.40 GHz CPU and 32G RAM. It took less than 5 mins for the whole geometry reconstruction process

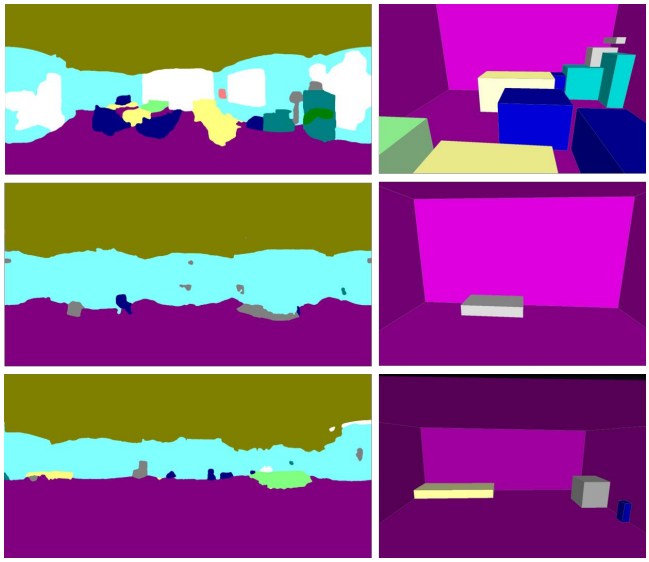


Figure 8: Semantic segmentation results (left) and snapshots of reconstructed 3D models (right) (Top: UL, Middle: LR, Bottom: ST)

Table 2: Evaluation of room layout reconstruction

| Data | Ground-truth ($m^3$) | Estimated ($m^3$) | Error (%) |
|---|---|---|---|
| MR | 5.61×4.28×2.33 | 5.52×4.35×2.36 | 1.3 |
| UL | 5.57×5.20×2.91 | 5.92×4.95×2.95 | 27.0 |
| LR | 5.64×5.05×2.90 | 5.77×5.17×2.98 | 7.6 |
| ST | 17.08×14.55×6.50 | 16.53×14.87×5.70 | 13.2 |

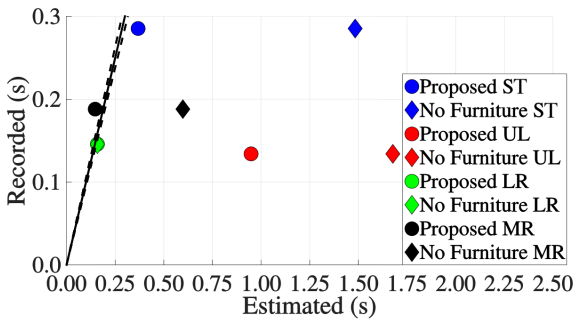
including pre-processing, depth estimation and cuboid reconstruction for any data set. The semantic segmentation took around 3 mins on an NVIDIA Tesla M2090 GPU with 5GB memory run in parallel. In a real environment, the whole process from camera setting to the final model output can be done within half an hour, which is much simpler and faster than audio-based approaches.

Table 2 shows the evaluation of the estimated dimensions against measured ground-truth for the rooms. The layout estimation errors vary according to the room characteristics. The UL data shows relatively large error due to the windows, mirror and dark wall which induce errors in correspondence matching. In the ST scene, the height of the room was incorrectly estimated due to the rails on the ceiling. It is difficult to quantitatively evaluate the reconstruction of objects in the scenes, but the proposed method recovered most of major objects in the scenes.

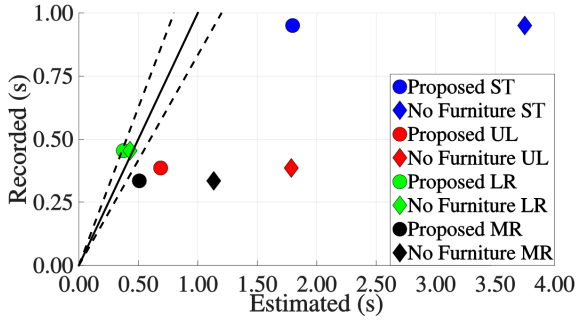
### 4.2 Room acoustics

For authenticity evaluation of the sound reproduced in the reconstructed scene, RIRs estimated in the VR environments are compared with the measured RIRs in the real environments. The real RIRs were measured in the test rooms with loudspeaker setups and microphone arrays which have 48 microphones evenly spaced around two concentric circles of radii 8.5 cm and 10.6 cm, respectively, to form a custom array [34] and one additional soundfield microphone at the center of the circular array. RIRs in the VR environment are measured using a virtual microphones and the sound of an anechoic gun-shot normalized in the time domain [9]. They are obtained by recording the responses at the same positions as in the real environment. We employed Google Resonance to render the sounds in VR. In Google Resonance, HRTFs are used to create virtual loudspeakers in a sphere around the listening position. Ambisonics as one way to reproduce the soundfield.

The acoustics reproduced in VR were evaluated by analysing

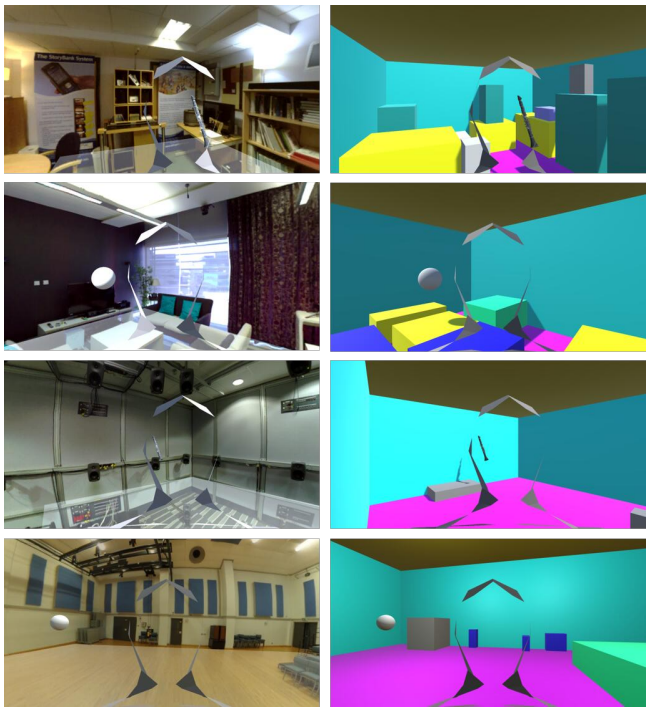



(a) EDT

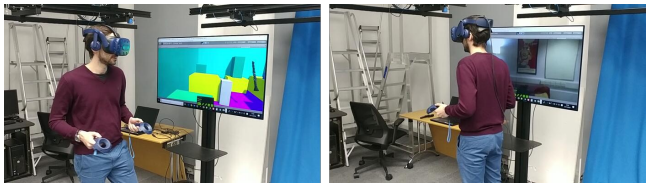



(b) RT60

Figure 9: Evaluation of simulated room acoustics in the VR environment against ground-truth. (a) shows the EDTs (Early Decay Times) and (b) the RT60s (Reverberation Times).



(a) Snapshots of rendered scenes (Left: 360° texture, Right: 3D model with object labels)



(b) Application on VR headset

Figure 10: Interactive VR rendering

EDT and RT60. EDT takes into account the energy carried by the early reflections and RT60 relates to the late diffuse reflections [7]. EDT is measured as the time from the arrival of direct sound to decay 10 dB, and RT60 defines the time employed by the energy to decay 60 dB. Both EDTs and RT60s are evaluated with the average over the octave bands between 250 Hz and 8 kHz.

We also define JNDs to understand how the estimated RIRs are perceptually similar to the recorded ones. The JNDs were chosen to be the 20 % for the RT60 [29] and 5 % for the EDT [45] based on the literature.

Fig. 9 illustrates the comparison of the EDTs and RT60s, i.e. Fig. 9(a) and Fig. 9(b) respectively, for the ground-truth and estimated RIRs. Both EDTs and RT60s of the estimated RIRs were close similar to the ground-truth ones, but the UL data showed a large error in the EDT and the ST data in the RT60. We guess the EDT was overestimated in the UL scene due to the errors in recognising the material of sofas near the microphone position (false early reflections), and the RT60 in ST was wrong because each wall was modelled as a whole concrete and the ceiling as a wooden panel while the real walls and ceiling in the ST scene have large soft panels to absorb sound as seen in Fig. 7 (c).

In Fig. 9(a), also the blue circle is outside the JND region. However, we use JNDs as ideal reference for authenticity, due to a gap in the literature about metrics for plausibility in VR environments. In fact, the EDT JND region was defined in the literature (cited in the paper as [40]) by looking at the output of subjective tests, undertaken by using only audio as reference. Nonetheless, it is well-known that for audio-visual reproductions, such as VR, audio perception is biased by the visual side (e.g. the McGurk effect [28]). Therefore, the plausibility region would be greater than the authenticity one (i.e. defined by JND), and contain the ST EDT in Fig. 9(a). Formal subjective tests, to define the plausibility region for a sound reproduced in VR, is currently in our future work plan. Moreover, from informal listening tests, the RT60 related to the blue circle in Fig. 9(b) (i.e. the ST dataset) sounded much more off than the one related to red circle (i.e. the UL dataset), when compared to the respective ground-truth. The comparison of the sounds rendered using the groud-truth RIRs and the sounds rendered in the reconstructed VR environments are given in the supplementary video: `https://youtu.be/bYJ7cSRGoWk`. We also compared the results with estimated room layouts without any object and included in Fig. 9 to show the importance of object recognition and classification in the scene. It is clear that the interaction of the sound with the objects lying inside the environment must be considered to accurately define the acoustic room model.

These comparisons were performed only at one listening position for each room, since only one microphone recording was available per dataset. More comparisons will be performed, in future work, by recording additional datasets with multiple microphone positions. There, we will compare the EDTs (that measure the early reflections' energy) of RIRs measured at different locations. However, it is well-known that RT60 (that measures the late reverberation energy) is typically constant at every position within an enclosed environment. Therefore, the analysis made during this paper already provides an extensive understanding, for the rooms that have been investigated, of the simulated room reverberation.

### 4.3 VR scene rendering

Plausibility is more important than authenticity for VR applications [5, 27]. However, to the best of our knowledge, it has not been clearly identified yet, in the literature, how to evaluate plausibility with objective metrics. Moreover, the perceptual differences between a real and a synthetic acoustic environment in the presence of visual stimuli are not as strict as they are for audio-only scenarios [3]. Therefore, subjective listening tests have been carried out to confirm the plausibility of the acoustics generated in VR environment. In this VR application, the user can freely navigate in the scene and switch the mode between: (1) original 360° photos mapped to a large sphere (2D) with the original sound source recorded in an anechoic chamber; (2) simplified 3D structure with coherent spatial audio rendering (Proposed method); and (3) Room layout only and spatial audio rendering in the empty room. Figure 10 shows some snapshots of the implemented real-time interactive VR scene rendering. In Fig. 10 (a), the rendered 360° textures are not exactly matched to the rendered 3D geometry because the textures have been simple mapped on a sphere as a 2D texture. The 360° texture just give a reference for the original scene. The full version of user interactive scene rendering with spatial audio can be found in the supplementary video. We received verbal feedback on four test scenes from two experienced and two untrained subjects. The implemented VR scenes produced plausible sound effects compared with the original source and the sound rendered in the empty room. However, the rendered sounds have some noise in the low frequencies because the Google Resonance produces 3D sound based on image sources [1]. Some error factors can be also identified in the rendered scene due to the material labeling errors.

## 5 Conclusion

In this work, the vision-based 3D structure and acoustic property estimation system has been proposed to provide plausible spatial audio in VR/AR environments. The approach requires only one pair of photos of the scene from commercial off-the-shelf 360° cameras. A simplified 3D geometry model of the scene is reconstructed by depth estimation and semantic segmentation of objects with labels is performed using a CNN. This visual information is used to predict acoustic properties within the scene, which allows perceptually plausible acoustic reproduction. This also allows the user correctly associate the sound with the respective room environment. The estimated room geometry and simulated spatial audio are evaluated against ground-truth data from actual measurements and recordings in the rooms. Experimental results showed a general agreement between the real and simulated acoustics.

The approach enables the simulation of plausible spatial audio renderings which match the acoustics of the room environment, however a number of limitations should be addressed in future work. Future extension of this research will include robust material detection using audio-visual sensors to compensate the current surface material mapping. Only rendered audio has been evaluated without visual cues in this experiment. Objective evaluation of plausibility in VR reproductions should be also accompanied as well as subjective evaluations with combined audio-visual cues. Another factor we did not deal with in this paper is "coherence" issue in AR application. Audio-visual coherence between virtual and real scenes will be investigated in our future study.

## References


[1] J. B. Allen and D. A. Berkley. Image method for efficiently simulating smallroom acoustics. *The Journal of the Acoustical Society of America*, 65(4):943–950, 1979.

[2] V. Badrinarayanan, A. Kendall, and R. Cipolla. Segnet: A deep convolutional encoder-decoder architecture for image segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2017.

[3] W. Bailey and B. M. Fazenda. The effect of visual cues and binaural rendering method on plausibility in virtual environments. In *Proc. of the 144th AES Convention*, Milan, Italy, 2018.

[4] L. P. Berg and J. M. Vance. Industry use of virtual reality in product design and manufacturing: a survey. *Virtual Reality*, 21(1):1–17, 2017.

[5] J. Blauert. *Communication Acoustics*. Springer-Verlag Berlin Heidelberg, 2005.

[6] N. Bonneel, C. Suied, I. Viaud-Delmon, and G. Drettakis. Bimodal perception of audio-visual material properties for virtual environments. *ACM Transacions on Applied Perception*, 7(1):1:1–1:16, 2010.

[7] J. S. Bradley. Review of objective room acoustics measures and future needs. *Applied Acoustics*, 72(10):713–720, 2011.

[8] K. Chatfield, K. Simonyan, A. Vedaldi, and A. Zisserman. Return of the devil in the details: Delving deep into convolutional nets. In *Proc. BMVC*, 2014.

[9] T. Cox. Gun shot in anechoic chamber. Freesound: `https://freesound.org/people/acs272/sounds/210766/`, 2013.

[10] M. Dou, L. Guan, J.-M. Frahm, and H. Fuchs. Exploring high-level plane primitives for indoor 3d reconstruction with a hand-held rgb-d camera. In *Computer Vision - ACCV 2012 Workshops*, pp. 94–108. Springer Berlin Heidelberg, Berlin, Heidelberg, 2013.

[11] R. C. Gonzalez and R. E. Woods. *Digital Image Processing*. Pearson, 2017.

[12] Google. Google resonance audio. `https://developers.google.com/resonance-audio/develop/overview`, 2018.

[13] Google. Google vr sdk. `https://developers.google.com/vr/`, 2018.

[14] GoPro. Gopro fusion. `https://shop.gopro.com/EMEA/cameras/fusion/CHDHZ-103-master.html`, 2018.

[15] A. Gupta, A. A. Efros, and M. Hebert. Blocks world revisited: Image understanding using qualitative geometry and mechanics. In *Proc. ECCV*, 2010.

[16] HTC. Vive pro. `https://www.vive.com/uk/product/vive-pro-full-kit/`, 2018.

[17] V. Hulusic, C. Harvey, K. Debattista, N. Tsingos, S. Walker, D. Howard, and A. Chalmers. Acoustic rendering and auditory-visual cross-modal perception and interaction. *Journal of Computer Graphics Forum*, 31(1):102–131, 2012.

[18] Insta360. Insta360 one x. `https://www.insta360.com/product/insta360-onex`, 2018.

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

[20] C.-H. Jeong, G. Marbjerg, and J. Brunskog. Uncertainty of input data for room acoustic simulations. In *Proc. of bi-annual Baltic-Nordic Acoustic Meeting*, 2016.

[21] D. B. JUDD. Chromaticity sensibility to stimulus differences. *Journal of the Optical Society of America*, 22(2):72–72, Feb 1932. doi: 10.1364/JOSA.22.000072

[22] H. Kim, T. Campos, and A. Hilton. Room layout estimation with object and material attributes information using a spherical camera. In *Proc. 3DV*, 2016.

[23] H. Kim, R. J. Hughes, L. Remaggi, P. J. B. Jackson, A. Hilton, T. J. Cox, and B. Shirley. Acoustic room modelling using a spherical camera for reverberant spatial audio objects. In *Audio Engineering Society Convention 142*, Berlin, Germany, 2017.

[24] H. Kim and K. Sohn. 3d reconstruction from stereo images for interactions between real and virtual objects. *Signal Processing: Image Communication*, 20(1):61–75, 2005.

[25] P. Larsson, A. Väljamäe, D. Västfjäll, A. Tajadura-Jiménez, and M. Kleiner. *Auditory-Induced Presence in Mixed Reality Environments and Related Technology*. 2010. doi: 10.1007/978-1-84882-733-2_8

[26] D. Li, T. R. Langlois, and C. Zheng. Scene-aware audio for 360° videos. *ACM Transactions on Graphics*, 37(4), 2018.

[27] A. Lindau and S. Weinzierl. Assessing the plausibility of virtual acoustic environments. *Acta Acustica united with Acustica*, 98(5):804–810, 2012.

[28] H. McGurk and J. MacDonald. Hearing lips and seeing voices. *Nature*, 264(5588):746–748, 1976.

[29] Z. Meng, F. Zhao, and M. He. The just noticeable difference of noise

length and reverberation perception. In *Proc. of the International Symposium on Communications and Information Technologies*, Bangkok, Thailand, 2006.

[30] S. Mori, S. Ikeda, and H. Saito. A survey of diminished reality: Techniques for visually concealing, eliminating, and seeing through real objects. *IPSJ Transactions on Computer Vision and Applications*, 9(17), 2017.

[31] Oculus. Oculus sdk. `https://developer.oculus.com/`, 2018.

[32] T. L. Pedro Morgado, Nuno Vasconcelos and O. Wang. Self-supervised generation of spatial audio for 3600° video. In *Proc. NIPS*, 2018.

[33] Pointgrey. Ladybug. `https://www.ptgrey.com/360-degree-spherical-camera-systems`, 2018.

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

[35] Ricoh. Ricoh theta v. `https://theta360.com/en/about/theta/v.html`, 2018.

[36] F. Rumsey. Spatial quality evaluation for reproduced sound: Terminology, meaning, and a scene-based paradigm. *Journal of Audio Engineering Society*, 50(9):651–666, 2002.

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

[38] C. Schissler, C. Loftin, and D. Manocha. Acoustic classification and optimization for multi-modal rendering of real-world scenes. *IEEE Transactions on Visualization and Computer Graphics*, 24(3):1246–1259, March 2018.

[39] C. Schissler, P. Stirling, and R. Mehra. Efficient construction of the spatial room impulse response. In *Proc. IEEE VR*, Mar. 2017.

[40] S. Song, S. Lichtenberg, and J. Xiao. Sun rgb-d: A rgb-d scene understanding benchmark suite. In *Proc. CVPR*, 2015.

[41] Spheron. Spheron vr. `https://www.spheron.com/products.html`, 2018.

[42] U. Technologies. Unity. `https://unity3d.com/`, 2018.

[43] N. Tsingos, T. Funkhouser, A. Ngan, and I. Carlbom. Modeling acoustics in virtual environments using the uniform theory of diffraction. In *Proc. ACM SIGGRAPH 2001*, pp. 545–552, Aug. 2001.

[44] M. Turk. Multimodal interaction: A review. *Pattern Recognition Letters*, 36:189–195, 2014.

[45] M. Vorländer. International round robin on room acoustical computer simulations. In *Proc. ICA*, Trondheim, Norway, 1995.

[46] S. Zheng, M.-M. Cheng, J. Warrell, P. Sturgess, V. Vineet, C. Rother, and P. H. S. Torr. Dense semantic image segmentation with objects and attributes. In *Proc. CVPR*, 2014.

[47] H. Zhu, F. Meng, J. Cai, and S. Lu. Beyond pixels: A comprehensive survey from bottom-up to semantic image segmentation and cosegmentation. *Journal of Visual Communication and Image Representation*, 34:12–27, 2016.