Audio Engineering Society

Conference Paper 65

Presented at the Conference on
Immersive and Interactive Audio
2019 March 27 – 29, York, UK

Reproducing Real World Acoustics in Virtual Reality Using Spherical Cameras

Luca Remaggi, Hansung Kim, Philip J. B. Jackson, and Adrian Hilton

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

Virtual Reality (VR) systems have been intensely explored, with several research communities investigating the different modalities involved. Regarding the audio modality, one of the main issues is the generation of sound that is perceptually coherent with the visual reproduction. Here, we propose a pipeline for creating plausible interactive reverb using visual information: first, we characterize real environment acoustics given a pair of spherical cameras; then, we reproduce reverberant spatial sound, by using the estimated acoustics, within a VR scene. The evaluation is made by extracting the room impulse responses (RIRs) of four virtually rendered rooms. Results show agreement, in terms of objective metrics, between the synthesized acoustics and the ones calculated from RIRs recorded within the respective real rooms.

1 Introduction

Nowadays, virtual reality (VR) is one of the main technological research foci [1]. Since VR is able to bring the user into new fictional words, it is usually employed for recreational purposes such as gaming [2] and movie productions [3]. Nevertheless, VR can also aid health sciences, by generating new ways of performing invasive analysis and tests [4], or improving the life quality of patients [5]. Moreover, VR can be applied to other research areas which, at first glance, may seem to be completely unrelated such as psychology [6], manufacturing [7], tourism [8], and education [9].

During the last decades, researchers mainly focused on improving the visual side of VR experiences [10, 11]. The amount of research produced in this field has led to great achievements, in particular regarding 3D rendering and human-machine interaction [12]. However, the ability of producing sounds that are coherent with the virtual visual experience is a capability that must be addressed in collaboration with the computer vision community. In fact, a VR experience would never be perceived as being "real" if sounds are not reproduced in harmony with the human visual sensory system perception [13, 14, 15].

It is important to distinguish between "plausible" and "authentic" acoustic reproductions. In general, a reproduced sound that evokes auditory events that a listener perceives as having occurred in a real environment is typically defined as being plausible [16]. On the other hand, an existing, real environment reproduction is authentic where the same percepts as the real environment are evoked [17]. Strength and precision of spatial information vary between the audio and visual modalities, when they are analyzed individually [18]. Nonethe-

less, spatial information is acquired by humans as a multimodal audio-visual combination. For example, humans, as average, perform sound source localization with an error of about 2° when the reproduction is audio-only, however, for audio-visual reproductions, this error increases to about 10° [18]. This observation is typically exploited by ventriloquists. Additional studies, analyzing the effect of visual cues on plausibility of audio rendering, showed a dominance of vision over acoustic cues [19]. This means that, when in the presence of visual stimuli, the perceptual differences between a real and a synthetic acoustic environment are not as strictly defined as they are for unimodal scenarios. In VR, a visual component exists, hence, here, audio-visual plausibility is the perceptual target.

Creating models that describe room acoustics has been extensively investigated in the spatial audio community [20]. Particular attention has been given to propose RIR parameterization methods for spatial audio production [21, 22, 23, 24]. These methods have the ability of generating a set of parameters, able to characterize the room acoustics. This process is typically done given acoustic signals (i.e. room impulse responses (RIRs)) recorded in the environment under investigation. With focus on VR, approaches focusing on the reproduction side of the chain have been also proposed. For instance, it has been proposed how to reproduce binaural sounds given B-format signals [25, 26].

Recent studies demonstrated that, from vision, it is possible to determine, approximately, certain acoustic components, such as the reverberation time (RT60) [27]. Moreover, it has been demonstrated that from a couple of 360° cameras it is possible to estimate the acoustics [28]. This is done by determining the room geometry and recognizing the materials. Computer vision techniques using visual sensors have played a major role in geometry reconstruction. Recovering geometric information from a single perspective photograph or 360 image relies on geometrical cues such as lines and texture [29, 30, 31]. 3D reconstruction from stereo or multiple images is widely used for general scene reconstruction [32, 33]. Kinect-like Red Green Blue Depth (RGBD) sensors also provide good range information for an indoor scene estimation [34, 35]. Regarding the material recognition task, several vision-based approaches are available in the literature. In [36], Bayesian generative models were proposed to exploit features such as color and micro-texture. Kernel descriptors were then utilized in [37], together with a Nearest Neighbor (NN)-based approach. Later, accuracy was improved by employing convolutional neural networks [38]. These were also employed in [39], where multi-class classification was proposed to identify both the object type and attributes.

In this paper, we propose a novel pipeline to reproduce, in VR environments, the acoustics of real rooms given visual sensors. We first perform a geometry estimation of the room, which includes the pose of furniture and related material labels. The main reconstruction algorithm is extended from previous work proposed in [40], for estimating reverberant spatial audio objects [41]. The estimation pipeline has been modified by applying the new semantic segmentation algorithm and reconstructing the final model using a cuboid fitting method to generate more accurate room models. The VR implementation applies the estimated acoustics of real rooms into virtual environments using Google VR Resonance [42]. An additional novelty is the approach used for evaluating the VR acoustics. Objective metrics are utilized to compare RIRs that were recorded in the real rooms and the correspondent synthetic ones, produced by the proposed pipeline.

The rest of the paper is structured as follows: Section 2 describes the proposed pipeline for reproducing real world acoustics into virtual reality given spherical camera images; Section 3 shows the experiments made to evaluate the plausibility of the reproduced sounds; finally, Section 4 draws the overall conclusions.

2 Proposed Pipeline

Fig. 1 shows the proposed pipeline to reconstruct an acoustic VR room from a pair of 360° camera images. A full surrounding scene is captured by a pair of vertically aligned 360° cameras at two different heights. These top and bottom images are used to estimate 3D depth information of the scene by disparity estimation. In addition, the top image is used for semantic scene segmentation and object recognition using SegNet [43]. Object-labelled cuboid structure is reconstructed from the depth estimation and semantic segmentation results. Acoustic properties for the detected objects are assigned by mapping to the acoustic material list in the Google Resonance Audio package. Finally, the acoustic VR environment is modelled by setting sound source and player models in Unity.

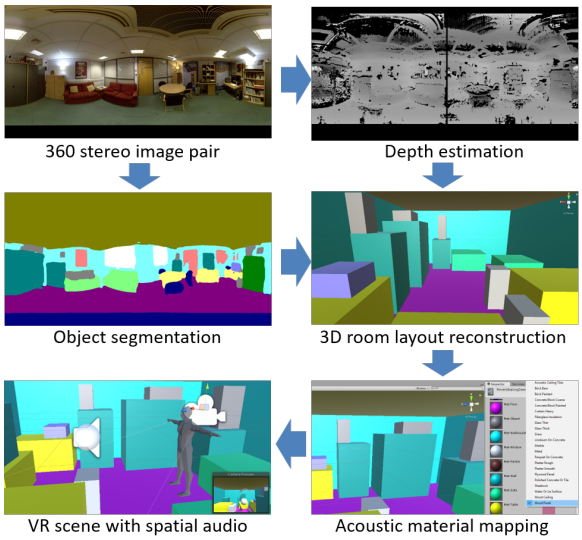
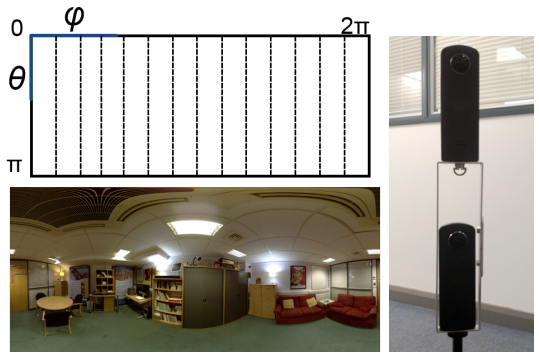


Fig. 1: Block diagram of the proposed pipeline.



(a) Equi-rectangular projection image (b) Cam set up

Fig. 2: Visual capture system.

2.1 Visual Capture and Depth Estimation

3D geometry reconstruction from 2D photos or Colour+Depth images has been widely investigated [44, 45, 46]. However, the limited field-of-view (FOV) of normal perspective cameras requires a large number of input images and long processing time. 3D reconstruction from 360° photos provides a solution to overcome this limitation. Inexpensive commercial 360° cameras[1] [2] are getting popular and providing well-calibrated equi-rectangular photos aligned to a spherical coordinate. In [40], a frequency-dependent acoustic prediction technique based on geometrical room modelling using 360° cameras was proposed. Based on this approach, we introduce a cuboid-based semantic room geometry modelling for room acoustics estimation using a pair of 360° cameras.

Two photos acquired from pre-calibrated fish-eye lenses in a 360° camera are stitched to each other to generate an equi-rectangular projection image as illustrated in Fig. 2 (a). Equi-rectangular projection images from two different heights are used to recover depth information in the scene by a stereo matching method [47]. Fig. 2 (b) shows the set up of two Ricoh Theta S cameras used in our experiments. We used these cameras because they provide well-aligned seamless stitching with less distortion in mapping textures from the fisheye lens to the Spherical coordinates.

3D depth information of the scene is estimated using dense stereo matching with spherical stereo geometry using a block matching method incorporating a region-diving technique [48] which produces quick and reliable disparity with occlusion region detection. In the spherical stereo geometry, the depth r_t from the top camera to the scene point P is calculated by triangulation as Eq. (1):

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

where θ_t and θ_b are angles of the projection of the point P onto the 360° image pairs, d is the disparity between two corresponding pixels, and B is the baseline distance between the cameras' center of projection.

2.2 Object Recognition and Room Geometry Reconstruction

To build the acoustic room model, acoustic attributes for all object surfaces in the scene are required. It is well known that it is difficult to estimate acoustic properties of the materials purely from visual inspection [49]. There have been some approaches to detect material attributes from images [50, 28], but their accuracy was typically below 50 %, for cross-dataset scenarios. Hence, they cannot be considered as suitable to estimate the absorption and scattering coefficients of objects. In this paper, we propose to use an object

[1] GoPro Fusion, https://shop.gopro.com/EMEA/cameras/fusion/CHDHZ-103-master.html

[2] Ricoh Theta, https://theta360.com/en/

recognition method and map the object categories to approximated acoustic properties of materials in Section 2.3, similar to the approach used in [40].

Semantic object recognition in the scene is performed with SegNet [43], a fully convolutional neural network (CNN) architecture for semantic segmentation. The results are labelled into 14 object categories. The labelled image through the SegNet is refined by morphological opening process [51] to separate partially connected objects with the same label and smooth object boundaries. Small regions are eliminated to generate simplified scene structure.

Approximated 3D geometry of the scene is reconstructed based on the semantic object segmentation and depth estimation results. The image points with estimated depth information on the image are projected to the 3D space. This point cloud is clustered into objects with the labels assigned in SegNet. Finally the cuboid fitting algorithm for 3D point clouds [52] is applied for each cluster to build the cuboid-based 3D scene model.

2.3 VR Scene Rendering with Spatial Audio

The results of geometry reconstruction in Section 2.2 are saved as OBJ geometry and JSON metadata files. The OBJ file includes 3D vertices, mesh groups, surface normal and texture information. The JSON file includes corresponding object information.

This output is directly imported to Unity[3] to build a VR model. To do so, several SDK tools for producing 3D audio in VR environment are available. For instance, Oculus' [53], SteamVR [54] and Google Resonance [42] are widely known, and they all implement an Ambisonics-based renderer. We chose the Resonance Audio package by Google, since it is representative of the above-listed group of SDK tools, and it fits well our needs. In Google Resonance, it is possible to have high-quality production of reverb, by precomputing acoustic probes. However, we chose not to employ probes, instead, we prioritized rel-time interaction. Resonance provides 22 types of materials with frequency-dependent acoustic attributes (i.e. absorption and scattering coefficients for the 9 octave bands between 63 Hz and 16 kHz). As it is difficult to directly detect acoustic materials from visual input, we map the recognized object labels by SegNet in the previous section to the material types in Resonance Audio as

[3]Unity, `https://unity3d.com/`

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

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 spatial audio in the reconstructed scene.

3 Experiments

In this section, we present the experiments, with related results, that evaluate the quality of the acoustics reproduced in VR environments. For the sake of this evaluation, we compared RIRs recorded in the real environments with the related synthetic RIRs produced in VR. We employed early decay times (EDTs) and RT60s as evaluation metrics, to directly analyse both the early reflections and the late reverberation.

3.1 Rooms and Their Reconstruction

Four rooms were employed for the analysis. Their dimensions as well as loudspeaker and microphone positions are reported in Table 2. The first, named as "LR", is a listening room at BBC R&D, in Salford, UK [55]. The second, named as "UL" [55], is a usability laboratory at BBC R&D. "S1" is a large recording studio at the University of Surrey [56]. The fourth is a meeting room "MR" at the University of Surrey [55]. RIRs were recorded in each room by employing the swept-sine method [57]: the rooms were excited through a 10 s swept-sine signal between 20 Hz and 24 kHz, with a sampling frequency of 48 kHz. In every room, the Countryman B3 Omnidirectional Lavalier microphone was used, chosen for its combination of high-quality and portability. For the same reasons, we also employed Genelec 8020B, as loudspeakers, in MR, UL and LR. Being S1 a larger room, there, we used a Genelec 1032B, to exploit its higher power and acoustically excite the whole room. Two 360° Ricoh Theta cameras were utilized to estimate the room geometry and perform the material classification.

Table 2: Properties of the recorded room setups.

Room	Size (m^3)	Loudsp. pos. (m)	Mic. pos. (m)
MR	4.25×5.61×2.33	[3.00, 2.12, 1.00]	[0.33, 2.12, 1.00]
UL	5.20×5.57×2.91	[2.24, 1.31, 1.07]	[2.27, 3.58, 1.07]
LR	5.64×5.05×2.90	[2.80, 0.51, 1.20]	[2.79, 2.55, 1.08]
S1	14.55×17.08×6.50	[5.00, 4.94, 1.50]	[5.00, 6.94, 1.50]

Fig. 3 shows the rendered scenes of the original 360° photos and reconstructed cuboid-based models with colour-coded object labels in Unity[4]. The 360° photos are simply mapped to a large sphere to give a look around effect on the left. The color texture is not visualized in the rendered virtual view in the full 3D environment on the right. The results for MR are shown in Fig. 1. The cuboid primitives represent the approximate geometrical structure of the scene.

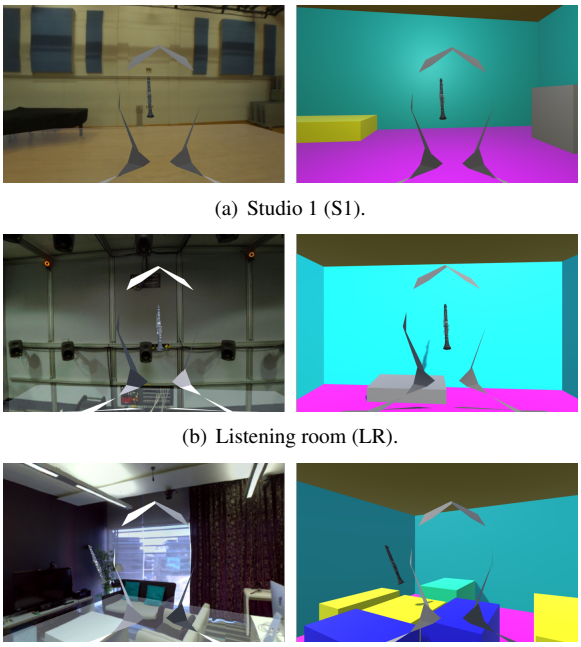
3.2 Extracting RIRs from VR Scene

To perform the pipeline evaluation, we need to compare RIRs recorded within the real environment with the related RIRs generated within VR. Although procedures to record RIRs in real environments are well established [58], extracting RIR information from VR has not been extensively explored in the literature yet. Therefore, we decided to follow an approach that treats the virtual environment as a real one. Since Google Resonance faces some difficulties when used for reproducing sinusoids [42], it was not possible to employ methods such as the swept-sine [57]. Instead, the sound of an anechoic gun-shot (normalized in the time domain) was placed at the source position [59]. By recording the response at the listening position, we obtained the virtual environment's RIR.

3.3 Evaluation Metrics

To evaluate the quality of the acoustics reproduced in VR we analyze the EDTs and RT60s. EDT is, in fact, a good metric to evaluate the acoustics from a point of view that is subjectively important, by taking into account the energy carried by the early reflections [60, 61]; on the other hand, RT60 relates to the average absorption and size of the room, making it important for describing the reverberance from a physical point of view, and it mainly takes into account the late diffuse reflections [60, 61]. EDT is calculated as the time

[4]The positions of sound sources and listeners in Fig. 3 are chosen for visualization purposes. They are different from the ones used for the RIR recordings and reported in Table 2.



(a) Studio 1 (S1).

(b) Listening room (LR).

(c) Usability lab (UL).

Fig. 3: Original 360 photos and reconstructed VR scenes with virtual sound source and listener (Left: 360 photo, Right: Reconstructed scene with object labels).

required for the energy to decay 10 dB after the direct sound, whereas RT60 defines the time employed by the energy to decay 60 dB. Both EDT and RT60 results are reported as the average over the 6 octave bands between 250 Hz and 8 kHz.

To better understand the perceptual similarity between the generated RIRs and the related recorded ones, we define the just noticeable differences (JNDs) for the evaluation metrics. The JND for the RT60 is chosen to be the 20 % [62], whereas the one for the EDT is the 5 % [63]. These JNDs are adopted from the literature, however, it is important to note that the same literature explains how these percentages vary with different types of sound. For instance, in [62] JNDs for RT60 were found up to about 30 % for musical signals [64]. Moreover, it is important to remark what was already discussed in Section 1: for VR purposes, authenticity is not the benchmark to target. It is, instead, widely recognized that plausibility should be achieved [16, 17]. However, to our knowledge, it has not been identified yet, in the literature, any threshold which defines plausi-

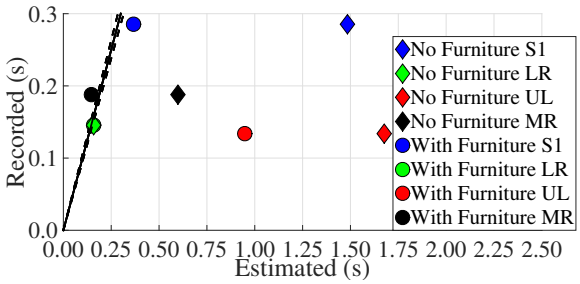


Fig. 4: Estimated EDTs for the four rooms, related to the estimated RIRs in VR environment.

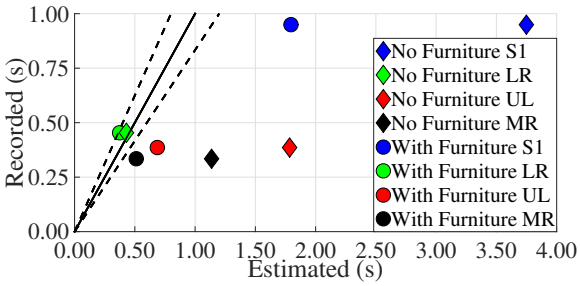


Fig. 5: Estimated RT60s for the four rooms, related to the estimated RIRs in VR environment.

bility limens, for the objective metrics employed. Previous studies typically focused on determining plausibility by observing the overall sound perception, without distinguishing between the perception of early reflections and late reverberation [65]. Furthermore, in the presence of visual stimuli, the perceptual differences between real and synthetic acoustic environment are not as strictly defined as they are for unimodal scenarios [19]. In this paper, we employ JNDs to be coherent with the available literature. However, JNDs typically refer to authenticity in audio-only scenarios.

3.4 Results

In Fig. 4 and 5, we compare the EDTs and RT60s, respectively, of recorded and generated RIRs. These results are reported as average among the 6 octave bands between 250 Hz and 8 kHz. Two ways of using the proposed visual methods for modeling the acoustics are tested. The first one synthesizes the room acoustics in VR without considering any furniture: only the room boundaries are modeled. The second way, instead, calculates the acoustics of the rooms, in VR, by also including the estimated furniture. We decided to compare these two variants of the proposed pipeline to analyze the importance of furniture information, while constructing a room acoustic model for VR.

Let us start from Fig. 4. The EDT results related to RIRs generated without modeling the furniture are, in general, large for every dataset, a part for LR where the estimated EDT falls close to the related JND band, being the error equal to 8 %. S1 and UL present the largest errors, with the EDT of the estimated RIRs being 420 % and 1183 % larger than the recorded RIRs', respectively. In MR, instead, the EDT error is 218 %. Considering then the acoustic model including the furniture, there is a great improvement with respect to the boundary-only-based acoustic model. In S1, MR and UL, the errors drastically drop down to 29 %, 22 % and 609 %, respectively. The only room that seems not to benefit from this improvement is LR, which error remains 8 %. However, LR was actually an empty room, thus there is no substantial difference between the two tested room models. In general, the EDTs of the synthetic RIRs fall now close to the JND bands, with the average error among S1, LR and MR being 20 %. The only outlier is UL, where the EDT is overestimated. This is due to issues in recognizing the materials of sofas lying next to the listening position, hence generating wrong early reflections.

Regarding the RT60 results in Fig. 5, there are similar trends to the ones observed for the EDTs. Observing first the errors produced by the model that considers only the room boundaries, S1, UL and MR present poor performance. Their errors are 295 %, 363 % and 239 %, respectively. However, as for the EDTs, the RT60 errors are dramatically reduced when also furniture is modeled. Only LR does not present any improvement, since, as described above for the EDTs, it is an empty room. Nevertheless, for both models the synthetic RIR RT60s fall inside the JND band, making the estimations indistinguishable from the recordings. The error percentage drops to 78 % in UL, and to 52 % in MR. The largest error, i.e. 89 %, is seen in S1. The reason for overestimating the RT60, there, is the acoustic panels placed along the walls for acoustical treatment. The material recognition algorithm does not recognize them, and labels the whole planar reflectors as concrete (see Fig. 3 (a)). To quantitatively analyze this issue, we manually edited the wall labels from concrete to the material considered as being the closer to groundtruth (i.e. heavy curtains in Google Resonance). By doing so, the absorption coefficient, averaged over all the room

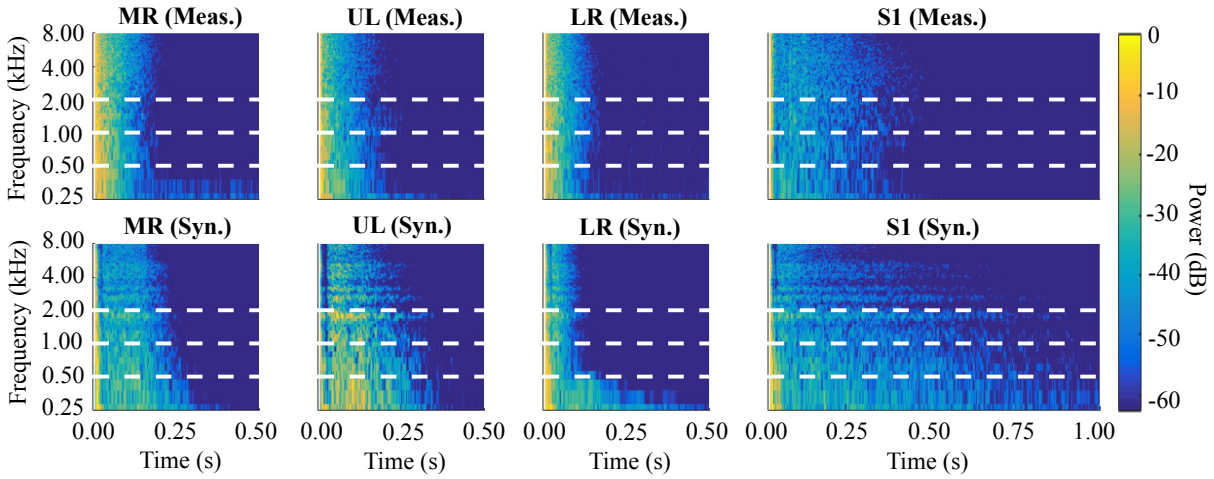


Fig. 6: Short time Fourier transform of the measured (top row) and synthesized (bottom row) RIRs, for the four datasets. The white dashed lines indicate the middle frequencies, i.e. 500 Hz, 1 kHz and 2 kHz.

surfaces, increased 4.31 times. Due to the Sabine's equation [66], this led the RT60 error to drop from 89 % to 53 %. Concluding, these results prove that, when furniture is modeled, the generated RIR RT60s are similar to the respective recorded ones, with errors falling close to the JND bands. Therefore, we can confirm that the interaction of the sound with the objects lying inside the environment is a fundamental concept to be considered during the definition of an acoustic room model for VR. It is also interesting to report that informal listening tests were undertaken, highlighting the plausibility of the synthesized VR acoustics.

The normalized power of the short time Fourier transform calculated from both recorded and synthesized RIRs (for the model including furniture) are reported in Fig. 6. The upper row shows the spectrograms of the recorded RIRs, whereas the lower row that of the synthetic RIRs. Columns from the left refer to MR, UL, LR and S1, respectively. As it was expected from the previous paragraphs' discussion, the energy of the RIRs generated through the proposed pipeline has a similar decay to the respective recorded RIRs, especially for frequencies above 500 Hz. Fig. 6 also shows that, in general, the low frequencies are overestimated. Issues at these frequencies were, however, expected, since the technology employed in Google Resonance to produce 3D sound is based on image sources [67]. A finite difference time domain (FDTD) method could be used, instead, to enhance the performance at the low-frequencies. Nonetheless, as it was also discussed in [40], FDTD has a high computational cost, making it impractical for real-time applications, such as VR.

4 Conclusion

We proposed a novel pipeline able to generate synthetic RIRs from 360 images, for VR productions, as approximation of real environment acoustics. The first part of the pipeline was composed of vision methods to estimate the room geometry and determine the materials. This information was used as input to the acoustic simulation stage of the pipeline, which was represented by the Google Resonance SDK, running in Unity.

Experiments were performed for four different rooms. The EDTs and RT60s calculated from RIRs recorded in these rooms were compared to the EDTs and RT60s related to RIRs generated in VR by the proposed pipeline. The results proved the importance of including furniture within acoustic room models. In fact, errors drastically dropped down with respect to a situation where only room boundaries were modeled. Furthermore, it was also shown that, in general, the errors resulted to be close to the JND bands. This means that, although the reproduction cannot be defined as being authentic, it is plausible, thus achieving the perceptual target typically defined for audio-visual reproductions, such as VR. Furthermore, the analysis highlighted a gap in the literature, regarding thresholds for the evaluation of acoustics' plausibility in VR reproductions, that could be investigated in future work.

Future work may also look at improving the material recognition method. Perhaps, acoustic sensors may be employed with this regard, to create a stronger relationship between objects and related acoustic properties. Moreover, formal listening tests may be undertaken, to subjectively analyze the acoustic plausibility of the proposed audio-visual method's output.

5 Acknowledgments

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.

References

[1] van Krevelen, D. W. F. and Poelman, R., "A Survey of Augmented Reality Technologies, Applications and Limitations," *International Journal of Virtual Reality*, 9(2), pp. 1–20, 2010.

[2] Rendon, A. A., Lohman, E. B., Thorpe, D., Johnson, E. G., Medina, E., and Bradley, B., "The effect of virtual reality gaming on dynamic balance in older adults," *Age and Aging*, 41(4), pp. 549–552, 2012.

[3] Gilbert, A., Volino, M., Collomosse, J., and Hilton, A., "Volumetric performance capture from minimal camera viewpoints," in *Proc. of ECCV 2018: European Conference on Computer Vision*, Munich, Germany, 2018.

[4] Malloy, K. M. and Milling, L. S., "The effectiveness of virtual reality distraction for pain reduction: A systematic review," *Clinical Psychology Review*, 30(8), pp. 1011–1018, 2010.

[5] Laver, K. E., George, S., Thomas, S., Deutsch, J. E., and Crotty, M., "Virtual reality for stroke rehabilitation," *The Cochrane Collaboration*, 2015(2), pp. 1–27, 2015.

[6] Wilson, C. J. and Soranzo, A., "The Use of Virtual Reality in Psychology: A Case Study in Visual Perception," *Hindawi Computational and Mathematical Methods in Medicine*, 2015(1), pp. 1–7, 2015.

[7] Ong, S. K. and Nee, A. Y. C., *Virtual and Augmented Reality Applications in Manufacturing*, Springer, 2004.

[8] Guttentag, D. A., "Virtual reality: Applications and implications for tourism," *Tourism Management*, 31(5), pp. 637–651, 2010.

[9] Freina, L. and Ott, M., "A Literature Review on Immersive Virtual Reality in Education : State Of The Art and Perspectives," in *Proc. of the International Scientific Conference - ELearning and Software Education*, Bucharest, Romania, 2015.

[10] Sun, B. and Saenko, K., "From Virtual to Reality: Fast Adaptation of Virtual Object Detectors to Real Domains," in *Proc. of the British Machine Vision Conference (BMVC)*, Nottingham, UK, 2014.

[11] Rix, J., Haas, S., and Teixeira, J., *Virtual Prototyping: Virtual environments and the product design process*, Springer International Publishing, 2016.

[12] Zhang, Z., "Microsoft Kinect Sensor and Its Effect," *IEEE MultiMedia*, 19(2), pp. 4–10, 2012.

[13] Rummukainen, O., Robotham, T., Schlecht, S. J., Plinge, A., Herre, J., and Habets, E. A. P., "Audio Quality Evaluation in Virtual Reality: Multiple Stimulus Ranking with Behavior Tracking," in *Proc. of the AES Conference on Audio for Virtual and Augmented Reality*, Redmond, USA, 2018.

[14] Stecker, G. C., Moore, T. M., Folkerts, M., Zotkin, D., and Duraiswami, R., "Toward Objective Measure of Auditory Co-Immersion in Virtual and Augmented Reality," in *Proc. of the AES Conference on Audio for Virtual and Augmented Reality*, Redmond, USA, 2018.

[15] McArthur, A., Sandler, M., and Stewart, R., "Perception of Mismatched Auditory Distance - Cinematic VR," in *Proc. of the AES Conference on Audio for Virtual and Augmented Reality*, Redmond, USA, 2018.

[16] Lindau, A. and Weinzierl, S., "Assessing the Plausibility of Virtual Acoustic Environments," *Acta Acustica united with Acustica*, 98(5), pp. 804–810, 2012.

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

[18] Stenzel, H. and Jackson, P. J. B., "Perceptual Thresholds of Audio-Visual Spatial Coherence for a Variety of Audio-Visual Objects," in *Proc. of the AES Conference on Audio for Virtual and Augmented Reality*, Redmond, USA, 2018.

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

[20] Vorländer, M., *Auralization: Fundamentals of Acoustics, Modelling, Simulation, Algorithms and*

Acoustic Virtual Reality, Springer, 2008.

[21] Pulkki, V., "Spatial Sound Reproduction with Directional Audio Coding," *J. of the Audio Engineering Society*, 55(6), pp. 503–516, 2007.

[22] Tervo, S., Patynen, J., Kuusinen, A., and Lokki, T., "Spatial Decomposition Method for Room Impulse Responses," *J. of the Audio Engineering Society*, 61(1/2), pp. 17–28, 2013.

[23] Coleman, P., Franck, A., Jackson, P. J. B., Hughes, R. J., Remaggi, L., and Melchior, F., "Object-Based Reverberation for Spatial Audio," *J. of the Audio Engineering Society*, 65(1/2), pp. 66–77, 2017.

[24] Politis, A., Tervo, S., Lokki, T., and Pulkki, V., "Parametric multidirectional decomposition of microphone recordings for broadband high-order ambisonic encoding," in *Proc. of the 144th AES Convention*, Milan, Italy, 2018.

[25] T. McKenzie, D. M. and Kearney, G., "Directional Bias Equalisation of First-Order Binaural Ambisonic Rendering," in *Proc. of the AES Conference on Audio for Virtual and Augmented Reality*, Redmond, USA, 2018.

[26] Binelli, M., Pinardi, D., Nili, T., and Farina, A., "Individualized HRTF for playing VR videos with Ambisonics spatial audio on HMDs," in *Proc. of the AES Conference on Audio for Virtual and Augmented Reality*, Redmond, USA, 2018.

[27] Kon, H. and Koike, H., "Deep neural networks for cross-modal estimations of acoustic reverberation characteristics from two-dimensional images," in *Proc. of the 144th AES Convention*, Milan, Italy, 2018.

[28] Kim, H., Campos, T., and Hilton, A., "Room Layout Estimation with Object and Material Attributes Information using a Spherical Camera," in *Proc. 3DV*, 2016.

[29] Yan, X., Yang, J., Yumer, E., Guo, Y., and Lee, H., "Perspective Transformer Nets: Learning Single-View 3D Object Reconstruction without 3D Supervision," in *Proc. NIPS*, 2016.

[30] Su, H., Fan, H., and Guibas, L., "A Point Set Generation Network for 3D Object Reconstruction from a Single Image," in *Proc. CVPR*, 2017.

[31] Xu, J., Stenger, B., Kerola, T., and Tung, T., "Pano2CAD: Room Layout from a Single Panorama Image," in *2017 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pp. 354–362, 2017.

[32] Scharstein, D. and Szeliski, R., "A Taxonomy and Evaluation of Dense Two-frame Stereo Correspondence Algorithms," *International Journal of Computer Vision*, 47(1), pp. 7–42, 2002.

[33] Seitz, S. M., Curless, B., Diebel, J., Scharstein, D., and Szeliski, R., "A Comparison and Evaluation of Multi-View Stereo Reconstruction Algorithms," in *Proc. CVPR*, pp. 519–528, 2006.

[34] 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.

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

[36] Liu, C., Sharan, L., Adelson, E. H., and Rosenholtz, R., "Exploring features in a Bayesian framework for material recognition," in *Proc. CVPR*, pp. 239–246, 2010.

[37] Hun, D., Bo, L., and Ren, X., "Toward Robust Material Recognition for Everyday Objects," in *Proc. of the British Machine Vision Conference (BMVC)*, pp. 48.1–48.11, 2011.

[38] Bell, S., Upchurch, P., Snavely, N., and Bala, K., "Material Recognition in the Wild With the Materials in Context Database," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015.

[39] Kim, H., de 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)*, pp. 519–527, 2016.

[40] Kim, H., Hough, R. J., Remaggi, L., Jackson, P. B. J., Hilton, A., Cox, T. J., and Shirley, B., "Acoustic Room Modelling using a Spherical Camera for Reverberant Spatial Audio Objects," in *Proc. 142nd AES*, Berlin, Germany, 2017.

[41] Remaggi, L., Jackson, P. J. B., and Coleman, P., "Estimation of Room Reflection Parameters for a Reverberant Spatial Audio Object," in *Proc. of the 138th AES Convention*, Warsaw, Poland, 2015.

[42] Google, "Google VR SDK," `https://developers.google.com/resonance-audio/`, 2017.

[43] Badrinarayanan, V., Kendall, A., and Cipolla, R., "SegNet: A Deep Convolutional Encoder-Decoder Architecture for Image Segmentation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2017.

[44] Sinha, S. N., Steedly, D., Szeliski, R., Agrawala,

M., and Pollefeys, M., "Interactive 3D Architectural Modeling from Unordered Photo Collections," in *Proceedings of SIGGRAPH ASIA*, 2008.

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

[46] Newcombe, R., Izadi, S., Hilliges, O., Molyneaux, D., Kim, D., Davison, A., Kohli, P., Shotton, J., Hodges, S., and Fitzgibbon, A., "KinectFusion: Real-Time Dense Surface Mapping and Tracking," in *Proceedings of ISMAR*, 2011.

[47] Kim, H., Remaggi, L., Jackson, P. J. B., Fazi, F. M., and Hilton, A., "3D Room Geometry Reconstruction Using Audio-Visual Sensors," in *2017 International Conference on 3D Vision (3DV)*, Qingdao, China, 2017.

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

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

[50] 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.

[51] Gonzalez, R. C. and Woods, R. E., *Digital Image Processing*, Pearson, 2017, ISBN 9781292223049.

[52] Kwon, S.-W., Bosche, F., Kim, C., Haas, C., and Liapi, K., "Fitting range data to primitives for rapid local 3D modeling using sparse range point clouds," *Automation in Construction*, 13(1), pp. 67–81, 2004.

[53] Oculus, "Oculus SDK," `https://developer.oculus.com/`, 2017.

[54] Valve, "SteamVR SDK," `https://steamcommunity.com/steamvr`, 2017.

[55] Kim, H., Remaggi, L., Jackson, P. J. B., and Hilton, A., "S3A audio-visual captures," `https://doi.org/10.15126/surreydata.00812228`, 2017.

[56] Coleman, P., Remaggi, L., and Jackson, P. J. B., "S3A room impulse responses," `http://dx.doi.org/10.15126/surreydata.00808465`, 2015.

[57] Farina, A., "Simultaneous measurement of impulse response and distortion with a swept-sine technique," in *Proc. of the 108th Audio Engineering Society Convention*, 2000.

[58] Stan, G. B., Embrechts, J. J., and Archambeau, D., "Comparison of different impulse response measurement techniques," *J. of the Audio Engineering Society*, 50(4), pp. 249–262, 2002.

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

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

[61] Rossing, T. D., *Springer Handbook of Acoustics - 2nd Ed.*, Springer-Verlag Berlin Heidelberg, 2014.

[62] Meng, Z., Zhao, F., and He, M., "The Just Noticeable Difference of Noise Length and Reverberation Perception," in *Proc. of the International Symposium on Communications and Information Technologies*, Bangkok, Thailand, 2006.

[63] Vorländer, M., "International round robin on room acoustical computer simulations," in *Proc. of the 15th ICA*, Trondheim, Norway, 1995.

[64] Wendt, T., de Par, S. V., and Ewert, S. D., "A Computationally-Efficient and Perceptually-Plausible Algorithm for Binaural Room Impulse Response Simulation," *J. of the Audio Engineering Society*, 62(11), pp. 748–766, 2014.

[65] Neidhardt, A., Tommy, A. I., and Pereppadan, A. D., "Plausibility of an interactive approaching motion towards a virtual sound source based on simplified BRIR sets," in *Proc. of the 144th AES Convention*, Milan, Italy, 2018.

[66] Kuttruff, H., *Room Acoustics - 5th Edition*, Spon Press, 2009.

[67] Allen, J. B. and Berkley, D. A., "Image method for efficiently simulating small-room acoustics," *The Journal of the Acoustical Society of America*, 65(4), pp. 943–950, 1979.