PROCEEDINGS of the
23rd International Congress on Acoustics
9 to 13 September 2019 in Aachen, Germany

Perceived quality and spatial impression of room reverberation in VR reproduction from measured images and acoustics

Luca REMAGGI[(1)], Hansung KIM[(1)], Annika NEIDHARDT[(2)], Adrian HILTON[(1)], Philip, J. B. JACKSON[(1)]

[(1)]Centre for Vision, Speech and Signal Processing, University of Surrey, UK
[(2)]Technische Universität Ilmenau, Germany

Abstract

Virtual reality (VR) systems have emerged as platforms for personal interactive audio-visual media experiences. In order to have a real-world reference against which to evaluate the room acoustics reproduced within VR, methods are needed to compare the virtual experience with that of a real room. In this work, two pipelines were developed for evaluating room reverberation over headphones within VR, acoustically and visually. The acoustical pipeline involves parameterisation of B-format room impulse responses via the reverberant spatial audio object (RSAO) and subsequent binaural rendering. The visual pipeline uses stereo 360° images to reconstruct the room geometry and materials that are then translated into binaural via two popular VR software development kits. Audio-visual subjective tests were conducted to evaluate the perceived quality and spatial impression of real rooms, using hidden anchors. The results show RSAO to be the better reverb estimator. However, the vision-based pipeline proved to provide a better listening experience when compared to the use of sounds carrying wrong acoustics. Future work will examine interactivity with realistic spatial room reverberation.

Keywords: Acoustics, Reverb, VR, Audio-Visual

1 INTRODUCTION

Virtual reality (VR) is one of the most popular technologies. It allows users to be transported into imaginary worlds, thus it is widely utilized in industries such as gaming [29] and film production [10]. Moreover, real-world scenes can be reproduced in VR, by employing 360° cameras and spatial audio recordings [15]. This capability allows VR systems to be employed as an aid for health sciences, creating new non-invasive tests [22], or supporting long-term patients [18]. VR has also potential in education [8].

In recent years, researchers have mainly focused on improving the visual side of VR reproductions [34, 30]. However, a VR experience would never be perceived as "real" if sounds are not in harmony with the visual perception [11]. It is fundamental, for instance, that sounds carry reverb information matching what humans would expect by looking at the environment [2]. This not only makes a virtual environment to seem realistic, but it also allows a correct perception of the sound source distance [24]. Mathematical models that describe reverberation have been extensively investigated, with a particular attention to reproducing the physical characteristics of it [36]. Researchers developed methods to describe reverb through sets of parameters [35, 27, 25], which allow to recreate (and edit) real-world reverb in virtual scenes. The parameterization process is typically done from recorded acoustical room impulse responses (RIRs). In [35], RIRs were recorded by using an array composed of four microphones, in a 3D layout. Parameters described the direction of arrival (DOA) of the main sound component, by looking at several time windows. In [25], parameters were still employed to define the DOA of the temporal-dependent soundfield components, to improve high-order Ambisonics. Nevertheless, it also included the description of a diffuse part, following [26]. In [27], instead, we proposed a representation based on specular early reflections and diffuse late reverberation, aiming at object-based audio [4].

Recently, it has been demonstrated that also from vision it is possible to approximate real rooms' reverberation [17]. 3D room geometry reconstruction is a widely explored field in computer vision [31, 3, 39]. In [31], 3D reconstruction was performed from multiple images. In [39], instead, the more challenging scenario where only one camera position is available was tackled, by exploiting geometrical cues, such as lines and texture.

Kinect was then used in [3], to generate the depth-map of the room. The material recognition problem, on the other hand, is more recent. Researchers tried different approaches, for instance, by employing features such as colors and micro-textures [20]. In [14], we used two 360° cameras to generate synthetic RIRs having acoustical characteristics similar to the captured environment, by first estimating the room geometry and materials.

In [28, 15], we performed an objective evaluation of the quality of the rendered reverb in VR. Nevertheless, it is difficult to define an objective metric that could reliably describe the quality of a sound that is reproduced in VR. 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 [33]. In VR, a visual component exists, hence, here, we perform audio-visual subjective tests. We are interested in analyzing the sound quality and spatial impression, since they are two key features characterizing spatial audio in a virtual acoustic scene [38].

In this paper, we describe two pipelines to render real-world reverb in VR. The first one is based on the analysis of acoustic signals, i.e. RIRs. It estimates parameters defining reverberant spatial audio objects (RSAOs) [27]. The second pipeline, instead, is based on visual captures made with 360° cameras, as we first proposed in [28, 15]: we perform a geometry and material estimation of the room, hence implicitly containing the reverb description. The information estimated via these two pipelines are rendered into VR using Google Resonance [12] and Steam VR [37]. For the audio-based pipeline, once obtained the RSAO parameters, we first record the rendered sounds via using the versatile interactive scene renderer framework (VISR) [7], then, we reproduce them through an array of 52 virtual loudspeakers, using Resonance. For the visual-based pipeline, instead, we try both Resonance and Steam VR as renderers for VR. In fact, we test two variants of the same pipeline. As main contribution we present subjective tests, looking at the sound quality and spatial impression of sounds reproduced, in VR, via the audio-based pipeline and the two vision-based pipeline variants.

The rest of the paper is structured as: Section 2 describes the proposed pipeline for reproducing real world acoustics given acoustic recordings; Section 3 introduces the other pipeline, based on visual captures; 4 describes the subjective experiments and discusses the results; finally, Section 5 draws the overall conclusions.

2 ESTIMATING REVERB FROM ACOUSTIC RECORDINGS

This section describes the first method, which parameterizes room reverberation, given recorded B-format RIRs.

2.1 Reverberant Spatial Audio Objects (RSAOs)

The RSAO, which we previously presented in [27, 5], models enclosed environment reverberation by defining a set of parameters describing the three RIR components [36]: the direct sound, revealing the position of the sound source; the early reflections, conveying a sense of the environmental geometry; and the late diffuse reverberation, indicating the size of the environment. In this paper, we employ RIRs recorded in B-format. The early reflections are encoded into parameters describing their times of arrival (TOAs), levels, DOAs, and frequency responses [27]. The late reverberation parameters, instead, describe its temporal envelope for a set of frequency bands. For each subband, the rate of decay is encoded together with the level in the neighbourhood of the mixing time [27, 4]. Direct sound's parameters follow the same flow as for the early reflections' [27].

2.1.1 Early Reflection Parameters

To estimate TOAs from RIRs, we developed, in [27], a method based on the dynamic programming projected phase-slope algorithm (DYPSA) [23]. DYPSA detects the early reflections in the W-channel of a B-format RIR by observing its group delay function $G(\omega)$, with ω representing the angular frequency. Sudden variations (i.e. peaks) correspond to zero crossings in $G(\omega)$. The DYPSA output is thus a sequence of non-zero values placed at the time samples n_k, with k being the reflection index. TOAs are calculated as $t_k = n_k/f_s$, where f_s is the sampling frequency. The second parameter describing the early reflections is the level [27]. The W-channel of a B-format RIR is segmented in the neighbourhood of the detected peaks, by using a Hamming window of length D. The energy is calculated as $E_k = \frac{1}{D}\sum_{n=n_k-(D/2)}^{n_k+(D/2)} |r(n)|^2$, where $r(n)$ is the RIR. E_k is then converted

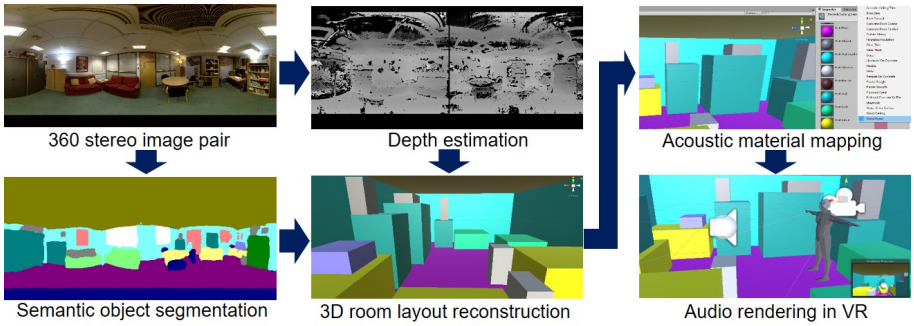

Figure 1. Overview of the audio rendering from visual captures.

into level as $A_k = \sqrt{E_k}$. DOAs are then estimated by steering a virtual cardioid microphone, considering each W-channel RIR segment containing the reflections [5]. The output of the steered cardioid can be written as $S(\mathbf{v},n) = \frac{1}{2}[r^W(n) + v_x r^X(n) + v_y r^Y(n) + v_z r^Z(n)]$, where $\mathbf{v}$ is a spatial vector containing $v_x = \cos(\theta)\cos(\phi)$, $v_y = \sin(\theta)\cos(\phi)$, $v_z = \sin(\phi)$, for $0 \leq \theta < 2\pi$ and $\pi/2 \leq \phi \leq \pi/2$, with θ and ϕ being azimuth and elevation angles, respectively. $r^W(n)$, $r^X(n)$, $r^Y(n)$, and $r^Z(n)$ are the four channels of a B-format RIR. The DOA is estimated as the steered peak containing the most energy: $\text{argmax}_{\mathbf{v}} \sum_n S(\mathbf{v},n)^2$. The frequency content of the early reflections is also parameterized by applying the linear predictive coding (LPC) [21] to the W-channel RIR segments. The output is a 9-th order infinite impulse response filter approximating the reflection frequency response [27].

2.1.2 Late Reverberation Parameters

The exponential decay at different octave bands is the key parameter chosen in RSAO to describe the late reverberation [27]. To estimate the starting point of this decay, we first need to determine the mixing time [4]. To do so, we calculate the normalized echo density, as was first proposed in [19]. We defined the mixing time as when the echo density became greater than 1. The late reverberation (i.e. the portion of RIR after the mixing time) recorded at the W-channel of the B-format RIR is then passed through an octave filter bank, with the final aim of determining the frequency-dependent decay [27]. The Schroeder energy decay curve is estimated for each band and an exponential curve fitted to it, based on the decay over the first 20 dB (after the mixing time). The exponential coefficients are the RSAO late reverb parameters. For each subband, the length of an onset ramp rising linearly from zero at the first early reflection to the noise gain in the neighbourhood of the mixing time is also encoded. This allows an increase of diffuse energy even before the mixing time [5].

3 ESTIMATING REVERB FROM VISUAL CAPTURES

This section describes the system implemented for audio rendering from visual captures with 360° cameras, based on [15]. Figure 1 shows the pipeline for spatial audio reproduction in VR of a real environment.

A scene is captured by vertically aligned 360° cameras. Two captured fish-eye images are mapped and stitched into an equirectangular image. They are aligned to the room coordinate axes to identify the principal room directions. Then the process is split into two stages: semantic object classification and 3D scene reconstruction. Depth of the scene is estimated by using correspondence matching with spherical stereo geometry [15]. For correspondence matching, any feature matching algorithm can be used for the image pair. We use the feature-based dense block matching method [16] which detects occlusion regions and ambiguous regions. 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 semantic object labelling pipeline. SegNet [1] is used in the proposed pipeline. SegNet provides a model trained on the SUN RGB-D indoor scenes dataset [32] to semantically segment structure and objects in indoor scene images. The

Table 1. Size, acoustic properties, and recording setup information of the four rooms.

Room	Dimensions (m^3)	Mic/Camera Position (m)	Loudspeaker Position (m)	RT60 (ms)
MR	$5.61 \times 4.28 \times 2.33$	[0.33,2.12,1.00]	[3.00,2.12,1.00]	270
ST	$14.55 \times 17.08 \times 6.50$	[5.00,6.94,1.50]	[5.00,4.94,1.50]	913
CY	$10.10 \times 19.00 \times -$	[3.35,12.63,1.70]	[5.42,14.99,1.69]	688
KT	$6.64 \times 3.46 \times 2.67$	[1.59,4.10,1.71]	[1.98,1.95,1.68]	350

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. All 2D points on the captured image are projected to the 3D space using the depth information estimated in the previous section. This 3D point cloud is segmented into clusters based on the object labels. From this object point clouds, block structures are reconstructed based on their point occupancy to build an approximated geometry of the scene. This output is directly imported to Unity to build a VR environment. The Google Resonance [12] or Steam [37] Audio package is used to simulate spatial audio in the Unity engine. Resonance Audio provides 22 types of materials with their acoustic attributes, and Steam provides 11 preset materials and 1 custom material. We map the object labels to the acoustically closest material types in the provided audio package. Both Resonance and Steam calculate the early reflections via using head-related transfer functions (HRTFs) belonging to the closest DOA estimated via ray tracing. However, Resonance renders the reverberation through a set of virtual loudspeakers, whereas Steam calculates a single binaural RIR to convolve with the sound: Steam's goal is to generate an accurate acoustic simulation; Resonance aims to bring the spatial audio experience to mobile devices, thus reducing the computational complexity.

4 SUBJECTIVE EXPERIMENTS

The aim is to evaluate the perceived quality and spatial impression (compared to the visual reproduction) of the reverb estimated via the two presented pipelines. Static headphones were used for the sound reproduction, and a screen for observing the related room images.

4.1 Listening Test Setup

Four rooms were tested: a meeting room (MR), a large recording studio (ST), a courtyard (CT), and a kitchen (KT). The 360° images were captured using a couple of Ricoh Theta cameras. RIRs were recorded by using the swept-sine method with a sampling frequency of 48 kHz [6], employing a Genelec 8020B as sound source. Regarding the microphone, we used a Soundfield MK5 B-format microphone in ST and MR; whereas in CY and KT the B-format Ricoh TA-1 3D Audio Microphone. The rooms' dimensions, recording setups, and reverberation times (RT60s) are reported in Table 1. Two sounds were employed: an anechoic speech from the TIMIT dataset [9]; and a clarinet recorded in anechoic environment, downloaded from the OpenAirLib library.

Two tests were carried out, looking at subjective attributes that are known to determine a reproduced sound quality [13]: the "spatial impression" and the "overall quality". Listeners were presented with MUSHRA interfaces having multiple sliders to rank each stimulus against the attribute under test (see Figure 3), within a discrete scale of integer numbers between 0 and 5. We built three methods to produce the stimuli. The first one, here referred as "Resonance", estimated the room acoustics by employing the visual-based pipeline in Section 3, and reproduced using Google Resonance as renderer in VR [12]. Also the second one, named here as "Steam", followed the visual-based pipeline in Section 3, nonetheless, it employed the Steam Audio package, to reproduce sounds in VR [37]. The third method is based on the RSAO [27]. To reproduce in VR the sounds generated from the RSAO parameters estimated following Section 2, we first rendered them via VISR, the S3A project's object-based renderer [7], defining 52 virtual loudspeakers on a sphere around the listening position. These speakers were equispaced on 3 rings: 24 at elevations 0°, 12 at 30°, and 12 at −30°. The same virtual

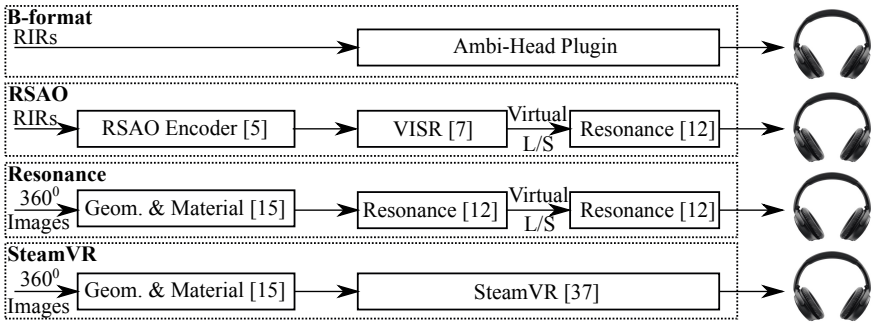

Figure 2. Overview of the four methods used to generate the stimuli.

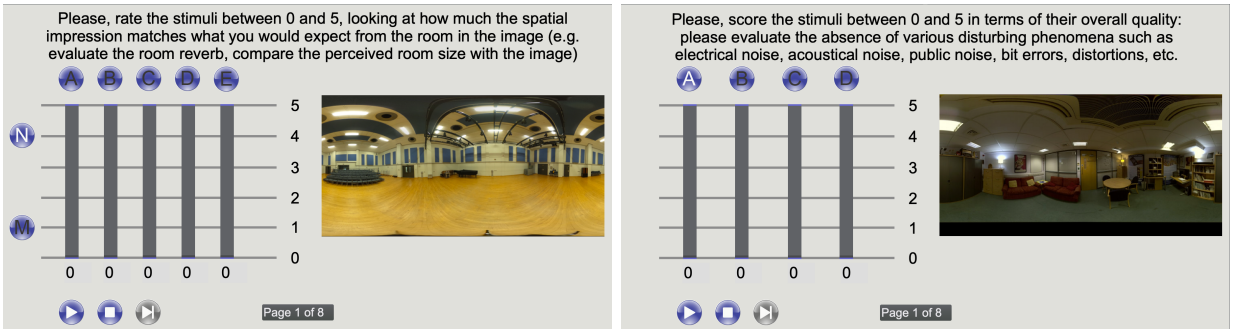

Figure 3. Screenshots of the MUSHRA interfaces for the two tests. The spatial impression test (left) presents 5 stimuli, named between A and E (randomly assigned to the RSAO, Resonance, Steam, B-format, and Wrong room sounds), and two anchors named as N and M. The overall quality test (right) presents 4 stimuli, named between A and D (randomly assigned to the RSAO, Resonance, Steam, and B-format sounds).

loudspeakers were then defined using Google Resonance for the playback. All the sounds were recorded, in VR, at the listening position, as .wav files. They were then embedded into the MUSHRA interfaces, developed in Max MSP. An overview of the methods used for producing sounds in VR is in Figure 2.

To have hidden anchors defining high quality reproduction, we also generated binaural sounds from B-format recordings (named during the analysis as "B-format"). This conversion was done using the NoiseMakers Ambi-Head plugin, in Reaper. As low anchor, we simply used a sound carrying a wrong reverberation (i.e. related to another room). Twenty listeners were tested. However, four of them were discarded from the analysis, having been identified as unreliable: by observing just the anchors' scores, listeners who did not correctly identify them more than two times were discarded. This was considered as an indicator of misunderstanding about the task.

4.2 Spatial Impression

This test was performed by reproducing sounds through headphones, while providing the panoramic photos of the related room under investigation. The MUSHRA interface used is depicted in Figure 3 left. Eight pages were tested, one for each combination of tested room and sound. Listeners were asked to: "Please, rate the stimuli between 0 and 5, looking at how much the spatial impression matches what you would expect from the room in the image". For this test, two hidden anchors were used: one, to be rated as 4, was the binaural sound obtained from B-format, as discussed in Section 4.1; whereas, to be rated as 1, it was a sound related to a room having completely wrong reverb. Although being hidden among the three pipelines to test, these two anchors were also explicitly provided to the listeners as reference for stimuli to rate as 1 and 4, respectively.

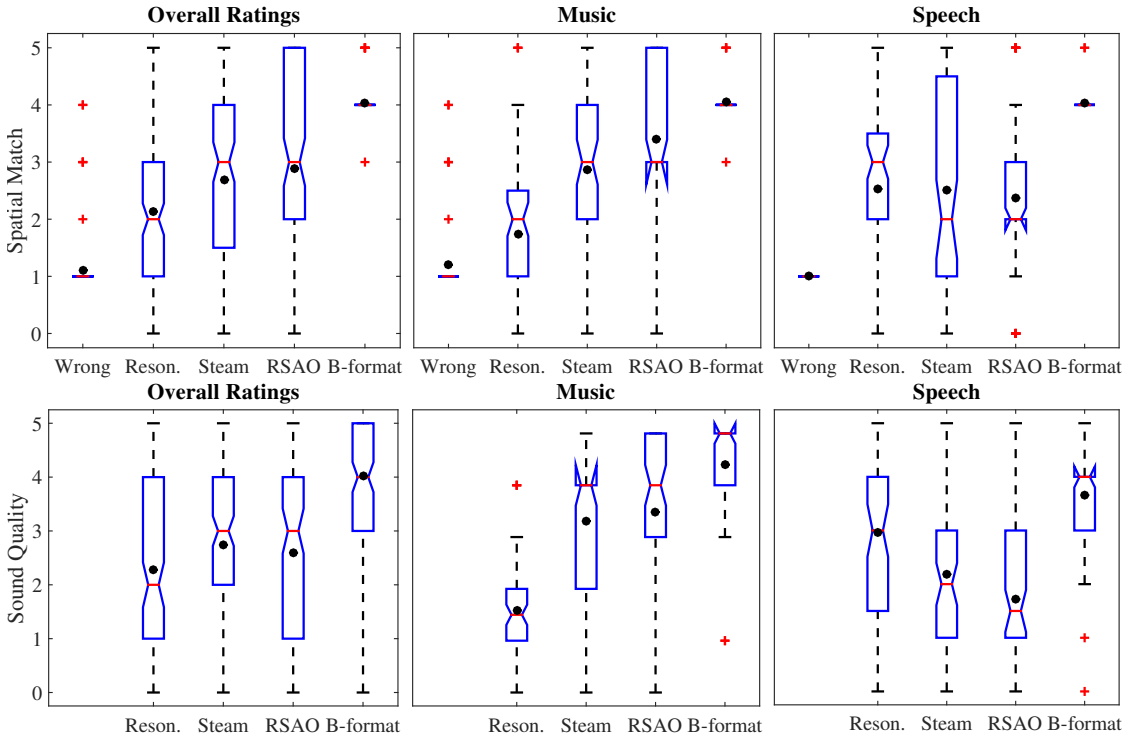

Figure 4. Subjective scores for spatial impression (top three) and overall quality (bottom three). On each box, the horizontal red line represents the median of the distribution, the bottom and top edges of the boxes are the 25$^{\text{th}}$ and 75$^{\text{th}}$ percentiles, respectively. The whiskers show the most extreme non-outlier data points, whereas the outliers are depicted as red crosses. The black circles are the means of the distributions.

Table 2. P-values of paired t-tests between the different methods' ratings across both music and speech. $h = 1$ means that the test rejects the null hypothesis of the two results belonging to normal distributions with equal means (95 % of confidence). When $h = 0$, it cannot reject the null hypothesis with a 95 % of confidence.

	Wrong vs Resonance	Resonance vs Steam	Steam vs RSAO	RSAO vs B-format
Spatial Impression	$8 \cdot 10^{-12}$ $(h = 1)$	0.007 $(h = 1)$	0.340 $(h = 0)$	$7 \cdot 10^{-13}$ $(h = 1)$
Overall quality	–	0.014 $(h = 1)$	0.457 $(h = 0)$	$2 \cdot 10^{-14}$ $(h = 1)$

The results are reported in Figure 4, top row. As expected, the overall rating shows both means and medians of the three methods to be lower than the anchor given by the binaural sound. Furthermore, all their spatial impressions are rated higher than the wrong room reverb. Comparing the three methods, RSAO presents the best mean and median, in fact it estimates the reverb directly from acoustic recordings, rather than from an image as the other two. Between Resonance and Steam, subjects seem to prefer Steam, in terms of spatial impression, when comparing the reproduced sound to the related room image. The other two figures split between the two pieces of content: music and speech. The trend observed in the overall score is mainly given by the music results. Instead, for speech, every method seems to provide similar spatial impression. This is related to the speech frequency spectrum being broader than the clarinet's. This masks better issues already found in the past, when Google Resonance was employed to reproduce single sinusoids [28]. In Table 2, we report the results of paired t-tests. With a 95 % confidence, the distribution of Resonance's rates has different mean from Steam's. RSAO rates, instead, cannot be claimed to be statistically different from Steam's, given the 16 samples.

4.3 Overall Quality

Here, the photos of the room under investigation were shown to the listeners just to provide an idea of the acoustics they should expect. Nevertheless, they only needed to compare the stimuli overall sound quality, scoring them between 0 and 5. The interface used in depicted in Figure 3 right. Eight pages were tested, one for each combination of tested room and sound. Only the B-format signal was used as hidden anchor (a wrong room reverb does not necessarily mean a lower sound quality), but not provided as explicit reference.

The results are reported in Figure 4, bottom row. As expected, the overall rating shows both means and medians of the three methods (i.e. RSAO, Resonance, and Steam) to be lower than the anchor. The interesting part of these results is that, comparing the three methods, RSAO and Steam presents better means and medians than Resonance. Nonetheless, since the percentiles clearly overlap, we reported in Table 2 the results of paired t-tests. From them, it is possible to claim that, with a 95 % confidence, the distribution of Resonance is different from Steam's. Whereas, as for the spatial impression, RSAO rates cannot be claimed to be statistically different Steam's. However, in general, it has to be taken into account that, to render RSAO, virtual loudspeakers were utilised with Google Resonance. Therefore, RSAO's results may be, perhaps, biased towards a lower rate. This is going to be better tested in the future, by running additional subjective tests, where RSAO will be reproduced in VR also using Steam. Looking at the other two figures, with music, RSAO performs better than both Steam and Resonance. On the other hand, for speech, Resonance seems to be the best. As mentioned above, the frequency spectra of the two sounds makes the difference: when Resonance plays the clarinet, a white background noise is audible. With speech, instead, the broader range of frequencies masks this problem [28]. This difference in quality may be also generated by the differences between Resonance and Steam during the rendering stage. As already discussed in Section 3, Resonance renders the reverberation through a set of virtual loudspeakers, whereas Steam calculates a binaural RIR. This makes Steam preferable when a high sound quality is required; nonetheless, Resonance is preferable for fast rendering applications.

5 CONCLUSION

We evaluated two pipelines for the generation of spatial reverb in VR: one based on the RSAO parameterization; the other on the calculation of room geometry and materials, from images. For the visual-based, we employed either Google Resonance or Steam Audio to renderer the acoustics, thus we tested a total of three methods.

Two subjective tests were undertaken. The first one looked at the spatial impression, whereas the second one aimed to evaluate the overall quality of the sound reproduction. In general, the results proved RSAO to create a better representation of the reverb, since it estimates it by looking directly at acoustic signals. However, in scenarios where RIRs are not available, the visual pipeline is a good alternative, performing better than just assigning a wrong reverb. The method employing Steam showed to create a more accurate reverb with respect to Google Resonance, in general. However, this is due to the different goals of the two tools: Steam's aim is to produce high fidelity reverb; nonetheless, Resonance was mainly developed for fast rendering applications.

Future work will look at methods to reproduce RSAO in VR. Moreover, more participants will be included into the analysis. Additional tests will be also run, evaluating dynamic rendering and reproduction of reverb.

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.

REFERENCES

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

[2] W. Bailey and B. M. Fazenda. The effect of reverberation and audio spatialization on egocentric distance estimation of objects in stereoscopic

virtual reality. *J. Acoustical Society of America*, 141(5):3510, 2017.
[3] S. Choi, Q.-Y. Zhou, and V. Koltun. Robust reconstruction of indoor scenes. In *Proc. CVPR*, 2015.
[4] P. Coleman, A. Franck, P. J. B. Jackson, R. J. Hughes, L. Remaggi, and F. Melchior. Object-based reverberation for spatial audio. *J. of the Audio Engineering Society*, 65(1/2):66–77, 2017.
[5] P. Coleman, A. Franck, D. Menzies, and P. J. B. Jackson. Object-based reverberation encoding from first-order ambisonic rirs. In *Proc. of the 142nd AES Convention*, Berlin, Germany, 2017.
[6] A. Farina. Simultaneous measurement of impulse response and distortion with a swept-sine technique. In *Proc. of the AES Convention*, 2000.
[7] A. Franck and F. M. Fazi. VISR: A versatile open software framework for audio signal processing. In *Proc. of the AES International Conference on Spatial Reproduction - Aesthetics and Science*, Tokyo, Japan, 2018.
[8] L. Freina and M. Ott. 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.
[9] J. S. Garofolo, L. F. Lamel, W. M. Fisher, J. G. Fiscus, D. S. Pallet, and N. L. Dahlgren. DARPA TIMIT acoustic phonetic continuous speech corpus CDROM. Technical report, NIST Interagency, 1993.
[10] A. Gilbert, M. Volino, J. Collomosse, and A. Hilton. Volumetric performance capture from minimal camera viewpoints. In *Proc. of ECCV 2018: European Conference on Computer Vision*, Munich, Germany, 2018.
[11] M. Gonzalez-Franco and J. Lanier. Model of illusions and virtual reality. *Frontiers in Psychology*, 8(1):1125, 2017.
[12] Google. Google vr sdk. https://developers.google.com/resonance-audio/, 2017.
[13] W. Hoeg, L. Christensen, and R. Walker. Subjective assessment of audio quality - the means and methods within the EBU. Technical report, EBU Technical Review, 1997.
[14] 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 *Proc. of the 142nd AES Convention*, Berlin, Germany, 2017.
[15] H. Kim, L. Remaggi, P. J. B. Jackson, and Hilton. Immersive spatial audio reproduction for vr/ar using room acoustic modelling from 360 images. In *Proc. of the IEEE VR Conference*, Osaka, Japan, 2019.
[16] 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.
[17] H. Kon and H. Koike. 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.
[18] K. E. Laver, S. George, S. Thomas, J. E. Deutsch, and M. Crotty. Virtual reality for stroke rehabilitation. *The Cochrane Collaboration*, 2015(2):1–27, 2015.
[19] A. Lindau, L. Kosanke, and S. Weinzierl. Perceptual evaluation of model- and signal-based predictors of the mixing time in binaural room impulse responses. *J of the Audio Engineering Society*, 60(11):887–898, 2012.
[20] C. Liu, L. Sharan, E. H. Adelson, and R. Rosenholtz. Exploring features in a bayesian framework for material recognition. In *Proc. CVPR*, pages 239–246, 2010.
[21] J. Makhoul. Linear prediction: a tutorial review. *Proc. of the IEEE*, 63(4):561–580, 1975.
[22] K. M. Malloy and L. S. Milling. The effectiveness of virtual reality distraction for pain reduction: A systematic review. *Clinical Psychology Review*, 30(8):1011–1018, 2010.
[23] P. A. Naylor, A. Kounoudes, J. Gudnason, and M. Brookes. Estimation of glottal closure instants in voiced speech using the DYPSA algorithm. *IEEE Transactions on Audio, Speech, and Language Processing*, 15(1):34–43, 2007.
[24] A. Neidhardt, A. I. Tommy, and A. D. Pereppadan. 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.
[25] A. Politis, S. Tervo, T. Lokki, and V. Pulkki. Parametric multidirectional decomposition of microphone recordings for broadband high-order ambisonic encoding. In *Proc. of the 144th AES Convention*, Milan, Italy, 2018.
[26] V. Pulkki. Spatial sound reproduction with directional audio coding. *J. of the Audio Engineering Society*, 55(6):503–516, 2007.
[27] L. Remaggi, P. J. B. Jackson, and P. Coleman. Estimation of room reflection parameters for a reverberant spatial audio object. In *Proc. of the 138th AES Convention*, Warsaw, Poland, 2015.
[28] L. Remaggi, H. Kim, P. J. B. Jackson, and A. Hilton. Reproducing real world acoustics in virtual reality using spherical cameras. In *Proc. of the AES International Conference on Immersive and Interactive Audio*, York, UK, 2019.
[29] A. A. Rendon, E. B. Lohman, D. Thorpe, E. G. Johnson, E. Medina, and B. Bradley. The effect of virtual reality gaming on dynamic balance in older adults. *Age and Aging*, 41(4):549–552, 2012.
[30] J. Rix, S. Haas, and J. Teixeira. *Virtual Prototyping: Virtual environments and the product design process*. Springer Int. Publishing, 2016.
[31] D. Scharstein and R. Szeliski. A taxonomy and evaluation of dense two-frame stereo correspondence algorithms. *Int. J. of Computer Vision*, 47(1):7–42, 2002.
[32] S. Song, S. Lichtenberg, and J. Xiao. Sun rgb-d: A rgb-d scene understanding benchmark suite. In *Proc. CVPR*, 2015.
[33] B. E. Stein and T. R. Stanford. Multisensory integration: current issues from the perspective of the single neuron. *Nature Reviews Neuroscience*, 9(5):255–266, 2008.
[34] B. Sun and K. Saenko. From virtual to reality: Fast adaptation of virtual object detectors to real domains. In *BMVC*, Nottingham, UK, 2014.
[35] S. Tervo, J. Patynen, A. Kuusinen, and T. Lokki. Spatial decomposition method for room impulse responses. *J. AES*, 61(1/2):17–28, 2013.
[36] V. Välimäki, J. D. Parker, L. Savioja, J. O. Smith, and J. S. Abel. Fifty years of artificial reverberation. *IEEE TASLP*, 20(5), 2012.
[37] Valve. Steamvr sdk. https://steamcommunity.com/steamvr, 2017.
[38] M. Vorländer. Virtual acoustics: Opportunities and limits of spatial sound reproduction. *Archives of Acoustics*, 33(4):413–422, 2008.
[39] X. Yan, J. Yang, E. Yumer, Y. Guo, and H. Lee. Perspective transformer nets: Learning single-view 3d object reconstruction without 3d supervision. In *Proc. NIPS*, 2016.