Audio Engineering Society

Convention Paper 10711

Presented at the AES 156th Convention
2024 June 15–17, Madrid, Spain

This paper was peer-reviewed as a complete manuscript for presentation at this convention. This paper is available in the AES E-Library (http://www.aes.org/e-lib), all rights reserved. Reproduction of this paper, or any portion thereof, is not permitted without direct permission from the Journal of the Audio Engineering Society.

Assessing Perceptual Discrimination in Acoustic Rendering Methods with Reduced Computational Complexity

Haowen Zhao[1] and Damian Murphy[1]

[1]*AudioLab, School of Physics Engineering and Technology, University of York, York, UK*

Correspondence should be addressed to Haowen Zhao (haowen.zhao@york.ac.uk)

ABSTRACT

Virtual acoustic environments can aurally simulate an environment that is non-existent in reality. Achieving realistic and immersive audio in applications such as gaming and virtual reality (VR) requires high-quality virtual acoustic modelling. In real-time applications, the quality needs to be guaranteed while minimising computational cost. In this study, we generated two audio rendering systems with two stages of Freeverb: one with the full number of filters involved, and another with the half number of filters. An objective comparison analysis in acoustical difference and computational efficiency is conducted between results. A web-based ABX listening test is conducted as perceptual evaluation. The results indicate that, based on the proposed system, certain acoustic details are still overlooked in the identification of a single sound source in non-VR environments.

1 Introduction

Captivating audio effects are a vital part in achieving an immersive experience in virtual environments. Traditionally, this is done through a process of recording and post production. However, the emergence of more interactive applications e.g. virtual reality (VR) has led to the need for real-time processing. Simulation based solutions are also required because they can create scenes that do not exist in the real world, or can create acoustic environments in real-time with game world parameters as an input.

VR brings new challenges in audio quality evaluation as audio is only one part of the multimodal experience. Focusing on audio alone may bias participants' attention towards features unrelated to the quality of this multimodal experience [1]. Emphasis should be placed on the matching of visual and audio content perceptually [2]. In a 6 Degree-of-Freedom (DoF) VR system, the users' impression comes from multiple forms of synergistic sensory information integration. The integration of audio content, visual content [3], and users' actions [4] are supposed to meet the expectations of our real-life experience [5]. A reasonable criterion in virtual acoustic environments, 'plausibility', represents an expectation based on each listener's self-experience, instead of the actual identity of an actual object [6]. In game audio, a plausible simulation should only include appropriate reproduction of required features [7], without apparent contradiction in audio-visual coherence [8]. Two points need to be emphasised: Firstly, in virtual acoustic environment, high physical accuracy is not the most important requirement. Instead, it is sufficient to simulate the significant characteristics of the sound source to achieve a plausible simulation of the given environment. Secondly, the importance of multisensory consistency is highlighted; listeners' expectations of the acoustic environment will be influenced by other

perceptual modalities and the tasks they are supposed to perform. Thus, it is possible to reduce the detail of acoustic modelling, while plausibility remains.

In a complex VR world, cognitive overload impacts a player's ability to distinguish several sounds simultaneously as in real life [9]. Sound designers have categorised sounds in VR by priority, to highlight the main tasks while maintaining non-diegetic sounds with a sufficiently good impression [9][10]. The high-priority sound is reproduced at a higher level with low-priority sounds filtered by cutting off frequency components, or through attenuation greater than would be expected in a real world equivalent. This is a promising path for saving computational cost in simulation, especially in VR. When users are supposed to focus on high-priority sounds, less attention will be paid to low-priority sounds, and so simulations can be applied with lower acoustic detail. Such an object-based acoustic simulation system has the potential to be able to conserve computational resources while maintaining a higher overall impression of audio quality, and have plausibility.

Another challenge is to what extent the simplification of acoustic details as applied to sound objects can be aurally distinguished. Current research in room acoustic modelling involves simplification of acoustic models [11][12][13], specifically for real-time systems in VR [14][15], where a considerable sacrifice in acoustic accuracy is required to ensure computational efficiency. Although shoebox shaped geometries can be used for complex virtual environments [16][17], it remains unclear how far the level of acoustic detail can be reduced while remaining perceptually plausible [18]. Therefore, a framework for verifying that an audio system can meet the above criteria needs to be proposed.

This paper details a framework for a two stage objective and perceptual evaluation of a simulation method, based on computational efficiency. The aim is to verify whether audio system qualities can be noticed by the listeners in a non-VR environment while reducing the computational cost. The objective evaluation is based on analysing the impulse responses of two systems with different room parameters. A web-based ABX listening test is built for this perceptual evaluation.

2 Audio engines

2.1 Acoustic modelling method

Traditional audio effect designed for a virtual acoustic environment is modified manually, relying on the sound designers' understanding of the audio effect, then approximated through appropriate parameter control. Typical acoustic modelling methods include (real-time) geometric acoustic (GA) methods [19], partially precalculated wave based methods [20], and artificial reverberation [21].

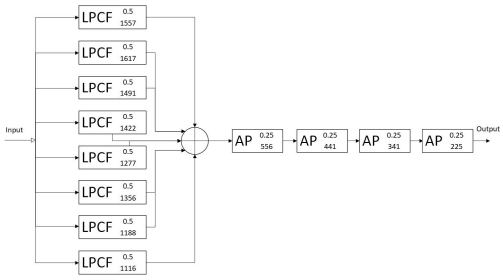
Wave based methods are more often used in architectural acoustics, as they require accuracy instead of computational efficiency, where offline non-realtime simulation is acceptable [20]. In game audio, sound sources, listeners and geometry are mostly dynamic and unpredictable in their movements. In this case, geometric acoustics and artificial reverberation are preferred for real-time applications, sacrificing more in physical accuracy to ensure computational efficiency.

The audio system used in this study is based on Freeverb [22]. It is a well known extension of the Schroeder reverberator [22], with eight Schroeder-Moorer filtered feedback comb filters and four Schroeder allpass filters. Although there are more options developed recently, e.g. feedback delay network (FDN), recent research has proved there is no significant perceptual difference between Freeverb and FDN implementation [23]. The advantage of Freeverb is the simple structure of Freeverb makes it easier to control the computational cost with respect to acoustic detail.

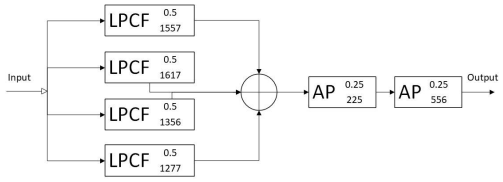
2.2 System design

The system is designed in Max/Msp [24]. All stimuli are generated offline and processed in Spat5 [25] for objective evaluation. All stimuli are also uploaded to Qualtrics [26], a cloud-based platform that enables users to conduct online surveys for an online listening test.

Two audio systems are involved in this study. The original Freeverb (Full-Freeverb) and the modified Freeverb (Semi-Freeverb), the block diagrams of which are shown in Figure 1. To save computational cost, only half the number of allpass and feedback comb filters are used in Semi-Freeverb. A pilot test was introduced with different numbers of filters involved (e.g., eight



(a) Block diagram of Full-Freeverb



(b) Block diagram of Semi-Freeverb

Fig. 1: Block diagram of the audio engines in this study. Parameters in the lowpass comb filters (LPCF) are damping (0.5) and centre frequency (in Hz). Parameters in the allpass filters (AP) are gain (0.25) and centre frequency (in Hz).

feedback comb filters with two allpass filters, six feedback comb filters with two allpass filters, etc.). However, it was shown that only a relevant combination of four feedback comb filters and two allpass filters can achieve a balance of perceived naturalness and efficient computation.

The lowpass feedback comb filter in Freeverb is built by employing a delay line, which undergoes lowpass filtering in the feedback loop. The transfer function is shown in (1), where d stands for damping, N is the delay time, and f is feedback respectively.

$$LBCF(z) = \frac{z^{-N}}{1 - f\frac{1-d}{1-dz^{-1}}z^{-N}} \tag{1}$$

The allpass filter implemented in Freeverb combines a feedback comb filter in series with a feedforward comb filter. In this case, it is only an approximation of allpass [22]. The transfer fuction is shown in (2), where g is feedback coefficient. All fixed parameters are presented in Figure 1. All variables are explained in next section.

$$AP(z) = \frac{-1 + (1+g)z^{-N}}{1 - gz^{-N}} \tag{2}$$

2.3 Parameter control

The damping parameter d controls the cutoff of the low pass filter in each of feedback comb filters. The value is between 0 to 1, causing each feedback loop of the comb filter to be more filtered than the previous one, resulting in increased high-frequency damping over time. This is a constant set as 0.5 in this system.

The feedback coefficient g (gain) in the allpass filter controls the reverberation time at very high frequencies, it is a constant set as 0.25 in this system.

The feedback parameter f influences the feedback coefficient in the feedback comb filters. It can change the low-frequency boundary absorption coefficient, thereby altering the reverberation time, to simulate the sound effects of different room sizes. This parameter is a variable, and changes between 0.7, 0.84, and 0.92, producing simulated rooms with reverberation times of 0.5s, 1.0s, and 2.0s respectively.

Perceived source distance might also be an influencing factor on overall quality. An approximation of closer distance perception between source and receiver is produced by applying a dry/wet ratio of 50/50 for the comparison with the original 0/100 dry/wet output. To be more specific, the input signal and processed signal from the reverberation algorithm are halved in amplitude and added together to get the final output.

3 ABX listening test

A web-based ABX listening test is conducted for perceptual evaluation. 22 student participants from the Undergraduate, Master, and PhD from the University of York were recruited. Eight of them are marked as trained listeners, as they either have music experience or involved in audio engineering. The test aims to investigate if participants can distinguish the acoustical difference between stimuli produced from two audio systems in different scenarios. Ethical approval was obtained from the Physical Science Ethics committee, University of York.

3.1 Stimuli normalization

Four anechoic sound sources from [27][28] are used in this experiment: solo acoustic guitar, female speech, a cymbal, and an orchestra. All selected samples are saved as a 24-bit WAV file at the sampling rate of 48 kHz. To reduce influencing factors, all audio samples are downmixed to mono, and normalised to a target amplitude level of -23.0 LUFS.

3.2 Procedure

Participants are required to take the test in a quiet place with headphones. Firstly, a practice session is designed to guide them through an example task followed by 52 questions for the main test. In each question, participants are presented with three audio samples, A, B, and X. A and B represent the reference and the altered versions, generated from the full-Freeverb and semi-Freeverb engine at random, while X is a randomly selected sample from either A or B. The task is to identify whether X is the same as A or B. A twenty-minute break is set after finishing the first half of the trials.

The four sound sources are processed by the two Freeverb systems, with three room sizes, and two source distances, which leads to 24 combinations. All combinations are presented in random order twice in the test. Furthermore, to ensure that the results are reliable from a web-based survey, there are four extra questions, where a distinct difference exists between the two references (e.g., an anechoic sample and a sample with reverb). These questions are selected at random and inserted every thirteen questions. They are designed to filter out the unqualified results. If participants get any of these four questions wrong, their results will be considered invalid.

3.3 Evaluation

The response to the ABX listening test can be treated as a series of binary outcomes: participants can either identify the subject right or wrong, which means each trial can be considered as a binomial distribution. Assuming a participant will randomly choose an answer if they cannot distinguish the difference between both stimuli, this means that subject detection at guess rate is 50%, which is the null hypothesis. From the p-value set as 0.5, we can calculate a 95% confidence threshold based on the number of trials. If the correct amount exceeds this threshold, we can reject the null hypothesis and conclude that participants can distinguish a difference with 95% confidence. Furthermore, in order to analyse the different scenarios separately, data will be presented across various experimental variables. The possible reasons for any resulting differences will be analysed and discussed.

4 Results

Results include objective analysis based on room impulse response, acoustic parameters and computational efficiency first. This is followed by subjective results from the ABX listening test.

4.1 Objective evaluation

The goal for objective evaluation is to show the extent to which acoustical differences exist between impulse responses generated from both systems. Six room impulse responses have been processed with from the small-sized room, the middle-sized rooms, and the large-sized rooms. The IRs are obtained from a deconvolution of a 20-20000 Hz sine sweep audio signal as input to the Max/MSP system, as rendered by both implementations, to ensure the analysed results obtained are equivalent and comparable. The determination of parameters is based on selecting parameter sets that aurally and acoustically correspond to different room sized within the recommended parameter range of the original Freeverb algorithm.

4.2 Room impulse response

The echo density of various rooms/algorithms are depicted in Figure 2. It is clear that the IRs from Full-Freeverb have denser content than semi-Freeverb. This is because the feedback loop in the lowpass comb filter allows the delayed signal to feedback into the delay line, creating multiple reflections over time. However, fewer echoes are present in the very early stage of Full-Freeverb, they also appear later in time compared to those in the Semi-Freeverb. The reason is because employing a greater number of filters tends to smooth or obscure early-stage signals due to the introduction of more reflections and attenuation. Additionally, the delay line introduces a time delay, replicating the time it takes for sound to reflect off surfaces in a room. Furthermore, the build up of reflections over time has changed, resulting in variations in relative amplitude. This explains why there are more distinct peaks in the waveform from Semi-Freeverb.

4.3 Room acoustic parameters

Standard room acoustic parameters Reverberation Time (RT60), Early Decay Time (EDT), and Clarity (C50) are calculated using Spat5 [25]. Analysis has been done for all six IRs (as Table 1). The results are compared between RIRs obtained with the same parameters in both systems.

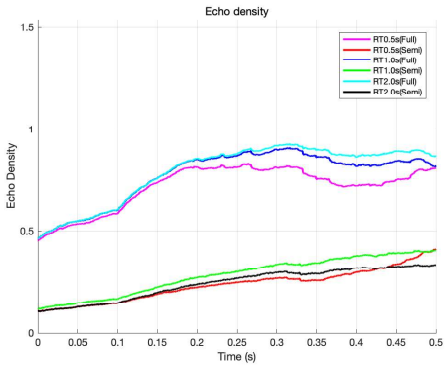


Fig. 2: Echo density profiles of produced impulse responses at different reverberation time (RT): RT at 0.5s (magenta, red), RT at 1.0s (blue, green), and RT at 2.0s (cyan, black).

Table 1: Acoustic parameters for different rooms

Room size	Alg.	RT60(s)	EDT(s)	C50(dB)
Small	Semi	0.511	0.388	9.0
	Full	0.507	0.359	8.6
Middle	Semi	1.04	0.755	4.8
	Full	1.02	0.730	4.4
Large	Semi	2.17	1.69	1.4
	Full	2.06	1.69	1.2

4.4 Computational efficiency

The computational efficiency of the system can be estimated theoretically or statistically. For the current pre-computed system with low computational complexity, statistical estimation is based on peak detection of internal and external CPU/RAM monitors in Max/MSP. The experiment is conducted on a PC equipped with an AMD Ryzen 9 5900X. The PC is configured with 32 GB of DRAM clocked at 3700 MHz. The operating system used for the experiment is Windows 10 Professional (64-bit).

In Freeverb, the main influencing factor in computational efficiency is the number of filters. With the CPU monitor in Max/MSP, a CPU usage difference between 69.62% (Full-Freeverb) and 52.07% (Semi-freeverb) exists within the audio system (this monitor only calculates the proportion of CPU usage at Max/MSP). Furthermore, We extracted the peak CPU and RAM usage of Max/MSP during computation with Windows Task Manager, with the input audio of a three-hour white noise in Table 2.

Table 2: Peak CPU/RAM usage from Windows monitor

Algorithms	CPU(%)	RAM(MB)
Semi-Freeverb	4.1	1421
Full-Freeverb	5.6	1526

Table 3: Comparison between operations required per sample for each module for the two systems

Module	Additions		Multiplications	
Algorithm	Semi	Full	Semi	Full
Comb	20	40	6	12
Allpass	16	32	2	4
Gain	0	0	1	1
Sum	3	7	0	0
Total	39	79	9	17

The comparison between number of floating-point operations (FLOPs) per sample required by the two audio systems are presented in Table 3. The FLOPs are quantified based on the number of additions and multiplications required by each module of the systems. In both filters, delay is treated as an addition. Gain is an overall gain modification applied when input the signal. Sum refers to the summation between comb filters in parallel. It can be seen that operations are mainly in filters.

Comparison is also made in memory cost in Table 4. We have calculated the memory use based delay time. Initially, essential parameters such as the sampling rate (48kHz), bit depth (32 bits float) were multiplied for memory demand per second per sample. As a mono sample, we don't need further calculation for channels. Furthermore, the memory necessary for each filter's delay line was assessed based on its respective delay time. Finally, the cumulative memory requirements of all delay lines were aggregated to establish the total memory needed for the system. In this study, the

Table 4: Comparison in delay line memory cost between both systems.

	Delay-Line Memory (MB)	
Algorithm	Semi-Freeverb	Full-Freeverb
Memory cost	1.06	2.02

memory required for the entire system is determined by the delay line with the largest requirement among the comb filters, not the allpass filters. This is because the memory requirements for the delay lines for the allpass filters are much lower compared to the comb filters. Table 4 shows Semi-Freeverb is more than twice computationally efficient than Full-Freeverb.

4.5 Subjective evaluation

Results from the listening test are presented in Figure 3. A red horizontal line is draw in each result figure, representing a significance level at 95%. Three results are invalid due to misjudgements on the extra questions. Two results are discarded because of incomplete questionnaire collection. Finally, seven results from trained listeners and ten results from untrained listeners are considered valid for analysis.

Results show that participants can reliably distinguish the difference between stimuli in all cases. With the highest correct rate of 79.38% and average correct rate at 72.92%, correct rate in all cases are higher than the significance level at 95% in binomial distribution. However, certain acoustic details are still overlooked, with the minimum correct rate for a particular scenario being 65.20%.

The blue data points is based on audio experience of listeners. It indicates that with the existing sample, experience level is not an effective influencing factor. It can be seen that both trained and untrained listeners have achieved similar correct rates, without statistically significant differences. Untrained listeners can distinguish approximately the same number of stimuli as trained listeners.

The red data points shows results varying with RT60 (room size). Although participants can reliably identify differences in all rooms over a guess rate, a trend can be concluded from the plot that the correct rate decreases as the size of the room increases and longer the reverberation time. It is interesting that, although the difference in RT60 between the middle sized room and the large room is more significant, the difference in results between them is much smaller than that between the small and the middle sized room. This suggests that while room size (RT difference) is an influencing factor, its impact is not linear. Once the RT reaches a certain threshold, its influence on distinguishing acoustic details tends to stabilize gradually.

The green data points highlights the difference with sound source. Speech and cymbal shows the same results, which is above the average performance. The correct rate of the orchestra source is the lowest, just above the significance level at 95%. A possible reason from participants' feedback is that the audio content (music) attracted their attention and makes it harder to distinguish the difference in acoustic details.

The yellow data points highlights the difference with dry/wet (source distance) mix. Participants performed better when dealing with close sound sources.

5 Discussion

With the algorithm input tuning parameters constant in Freeverb and the number of feedback low-pass filters increased, we should obtain higher echo and modal density, while reverberation time and early decay time remain. However, in this experiment, we observed an different trend. Semi-Freeverb exhibits longer RT compared to Full-Freeverb. Table 5 shows RT60, EDT and C50 obtained, using 6 lowpass comb filters plus 2 allpass filters, 4 lowpass filters plus 4 allpass filters, Semi-Freeverb and Full-Freeverb. The just noticeable difference (JND) of reverberation time and EDT are calculated based on ISO 3382-1 at 5% [29] based on values from Full-Freeverb. The JND for C50 is set based on Bradley's work as 1.1dB [30].

It can be observed that in RT60, a trend is observable that the value of Full-Freeverb consistently remains the smallest. One possible reason is the changes in number of filters changes the reflection patterns in the simulation. It takes longer in Full-Freeverb in reflection and has influence in RT calculations with different regression line calculations. Furthermore, the change in reflection patterns also changes the timbre of the reverberated sounds. The energy decay relief (EDR) of the impulse responses for the middle-sized room are showed in Figure 4. It can be seen that in the Semi-Freeverb, especially in low frequencies, more

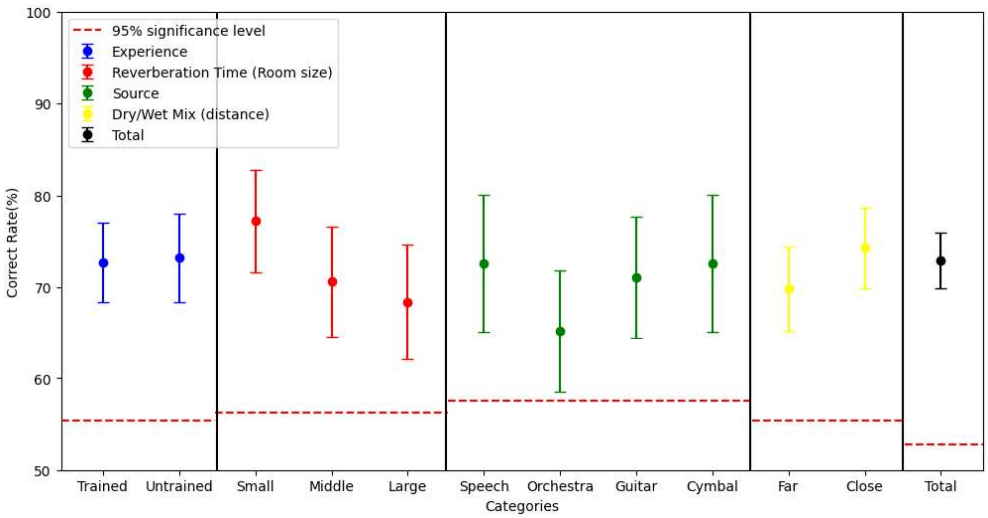
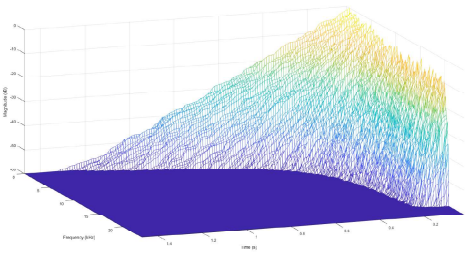
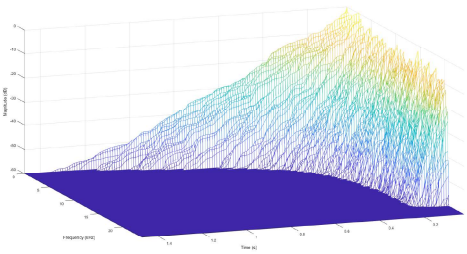


Fig. 3: Results and binomial confidence intervals for different scenario groups. The variables from left to right are: listener's auditory experience (blue), room reverberation time (red), sound source (green), dry-wet ratio (yellow), and total performance (black). The red dash line represents the 95% significance level. Error bars represent the 95% confidence interval of the binomial distribution.



(a)



(b)

Fig. 4: Early decay relief (EDR) for IRs from Semi-Freeverb (a) and Full-Freeverb (b). Both generated from the middle-sized room.

densely distributed energy decay lines are shown, indicating that its energy is more dispersed compared to the Full-Freeverb. Lastly, only the differences in the low-frequencies range exceed the threshold of the JND in part of combinations. This explains why some participants reported that they could hear difference in timbre in some low frequencies cases. However, although measurable differences exist in EDT and C50, these differences make themselves known in different ways which have influence on perceptual results. Manipulation of parameters and filters needs further research to ensure closer matching parameters in future.

With an average error rate of 27.08%, this experiment suggests that even in non-VR scenarios for single-source identification tasks, there is a significant amount of acoustic detail being overlooked. The acoustic differences in these stimuli are subtle, requiring considerable effort to distinguish. This is a rare situation in real life because we tend to be distracted by multisensory experiences. On the contrary, in this experiment, participants can freely switch between the stimuli with known differences and have a clear reference in the ABX test to help locate some acoustic differences that may be overlooked in scenes with higher complexity. In a multi-modal VR experience, the proposed acoustic

Table 5: Approximated just noticeable difference (JND) for IRs from combinations of 6 comb filters + 2 allpass filters, 4 comb filters + 4 allpass filters, Semi-Freeverb and Full-Freeverb. All generated from the middle-sized room. The JND is calculated at 5% based on Full-Freeverb. Values are in bold if they are greater than 1JND.

	RT60 (s)						
Octave bands (Hz)	125	250	500	1000	2000	4000	8000
6 CBs + 2 APs	1.2207	1.1900	1.1562	1.1021	0.9008	0.6108	0.3280
4 CBs + 4 APs	**1.2555**	**1.2521**	1.2032	1.1279	0.9336	**0.6273**	0.3281
Semi-Freeverb	**1.2455**	**1.2447**	1.2075	1.1404	0.9153	0.6184	0.3314
Full-Freeverb	1.1731	1.1696	1.1773	1.0992	0.8941	0.5961	0.3189
	EDT (s)						
6 CBs + 2 APs	**1.2010**	1.3651	0.9882	1.1337	0.8219	**0.5123**	**0.2845**
4 CBs + 4 APs	1.4331	1.2989	**1.1960**	**1.1884**	**0.8754**	0.5534	**0.3392**
Semi-Freeverb	1.3700	1.3287	**1.1173**	1.1436	**0.8448**	0.5177	**0.2888**
Full-Freeverb	1.4299	1.3130	0.9953	1.1213	0.7804	0.5437	0.3494
	C50 (dB)						
6 CBs + 2 APs	**-2.43**	**-3.40**	**-0.37**	**-0.62**	0.93	**4.54**	**9.74**
4 CBs + 4 APs	**-0.26**	-5.19	**-2.53**	-1.50	**-1.52**	2.65	6.59
Semi-Freeverb	**-0.70**	-4.95	**-0.14**	**-0.39**	0.03	**4.63**	**9.44**
Full-Freeverb	-9.13	-5.86	-6.16	-2.15	-0.42	2.44	6.34

system may experience higher error rates if subjected to multi-source testing in VR, as it is a reference-free system with higher scene complexity. In other words, maintaining plausibility while retaining lower acoustic details could be possible.

6 Summary and future work

A framework for evaluating perceptual differences between levels of acoustic details has been introduced. Parametric and subjective listening experiment were carried out on two audio systems using the same algorithm but different levels of computational complexity. Different testing conditions were selected, and the results were analysed from various perspectives. The results show that, in the proposed audio system, individuals tend to overlook acoustic details in single-source distinguishing tasks outside of VR. Despite the unknown impact of the filters on audio timbre, both trained and untrained listeners still exhibit a 27% error rate.

In VR, multimodal experience make us to reconsider the level of acoustic detail needed for immersive audio. The shift from passive listeners to active testers enables us to create a plausible acoustic environment with varying levels of acoustic detail. While the proposed system has simple computational complexity and quality, challenges arise. What levels of acoustic details will be overlooked in multimodal experiences? Additionally, multiple sound sources can increase the JND threshold. To reduce possible influencing factors, the current experiment is based on two stages of a reverberation algorithm. More up-to-date acoustic modelling methods can be considered in future studies, ranging from GA methods, or hybrid modelling methods combining GA and artificial reverberation. In multi source experiments, the priority sequence of sounds can also be designed. Participants may overlook more acoustic details when focusing on certain primary tasks.

References

[1] Rummukainen, O., Robotham, T., Schlecht, S. J., Plinge, A., Herre, J., and Habels, E. A., "Audio quality evaluation in virtual reality: multiple stimulus ranking with behavior tracking," in *Audio Engineering Society Conference: 2018 AES International Conference on Audio for Virtual and Augmented Reality*, Audio Engineering Society, 2018.

[2] Larsson, P., Västfjäll, D., Olsson, P., Kleiner, M., et al., "When what you hear is what you see: Presence and auditory-visual integration in virtual environments," in *Proceedings of the 10th annual international workshop on presence*, pp. 11–18, 2007.

[3] Rummukainen, O., Wang, J., Li, Z., Robotham, T., Yan, Z., Li, Z., Xie, X., Nagel, F., and Habets, E. A., "Influence of visual content on the perceived audio quality in virtual reality," in *Audio Engineering Society Convention 145*, Audio Engineering Society, 2018.

[4] Wozny, D. R., Beierholm, U. R., and Shams, L., "Human trimodal perception follows optimal statistical inference," *Journal of vision*, 8(3), pp. 24–24, 2008.

[5] Slater, M., "Place illusion and plausibility can lead to realistic behaviour in immersive virtual environments," *Philosophical Transactions of the Royal Society B: Biological Sciences*, 364(1535), pp. 3549–3557, 2009.

[6] Burnett, B., Neidhardt, A., Cvetković, Z., Hacıhabiboğlu, H., and De Sena, E., "User Expectation of Room Acoustic Parameters in Virtual Reality Environments," in *2023 Immersive and 3D Audio: from Architecture to Automotive (I3DA)*, pp. 1–10, IEEE, 2023.

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

[8] Pellegrini, R. S., *A Virtual Reference Listening Room as an Application of Auditory Virtual Environments: Ein Virtueller Referenz-Abhörraum Als Anwendung Auditiver, Virtueller Umgebungen*, dissertation. de, 2002.

[9] Jain, D., Junuzovic, S., Ofek, E., Sinclair, M., Porter, J., Yoon, C., Machanavajhala, S., and Ringel Morris, M., "A taxonomy of sounds in virtual reality," in *Designing Interactive Systems Conference 2021*, pp. 160–170, 2021.

[10] Jain, D., Junuzovic, S., Ofek, E., Sinclair, M., R. Porter, J., Yoon, C., Machanavajhala, S., and Ringel Morris, M., "Towards sound accessibility in virtual reality," in *Proceedings of the 2021 International Conference on Multimodal Interaction*, pp. 80–91, 2021.

[11] Brinkmann, F., Aspöck, L., Ackermann, D., Lepa, S., Vorländer, M., and Weinzierl, S., "A round robin on room acoustical simulation and auralization," *The Journal of the Acoustical Society of America*, 145(4), pp. 2746–2760, 2019.

[12] Eley, N., Mullins, S., Stitt, P., and Katz, B. F., "Virtual Notre-Dame: Preliminary results of real-time auralization with choir members," *2021 Immersive and 3D Audio: from Architecture to Automotive (I3DA)*, pp. 1–6, 2021.

[13] Postma, B. N. and Katz, B. F., "Creation and calibration method of acoustical models for historic virtual reality auralizations," *Virtual Reality*, 19, pp. 161–180, 2015.

[14] Schröder, D. and Vorländer, M., "RAVEN: A real-time framework for the auralization of interactive virtual environments," in *Forum acusticum*, pp. 1541–1546, Aalborg Denmark, 2011.

[15] Kirsch, C., Poppitz, J., Wendt, T., van de Par, S., and Ewert, S. D., "Computationally efficient spatial rendering of late reverberation in virtual acoustic environments," in *2021 Immersive and 3D Audio: from Architecture to Automotive (I3DA)*, pp. 1–8, IEEE, 2021.

[16] Blau, M., Budnik, A., Fallahi, M., Steffens, H., Ewert, S. D., and Van de Par, S., "Toward realistic binaural auralizations–perceptual comparison between measurement and simulation-based auralizations and the real room for a classroom scenario," *Acta Acustica*, 5, p. 8, 2021.

[17] Kirsch, C., Wendt, T., Van De Par, S., Hu, H., and Ewert, S. D., "Computationally-Efficient Simulation of Late Reverberation for Inhomogeneous

Boundary Conditions and Coupled Rooms," *Journal of the Audio Engineering Society*, 71(4), pp. 186–201, 2023.

[18] Fichna, S., van de Par, S., and Ewert, S. D., "Evaluation of Virtual Acoustic Environments with Different Acoustic Level of Detail," in *2023 Immersive and 3D Audio: from Architecture to Automotive (I3DA)*, pp. 1–6, IEEE, 2023.

[19] Savioja, L. and Svensson, U. P., "Overview of geometrical room acoustic modeling techniques," *The Journal of the Acoustical Society of America*, 138(2), pp. 708–730, 2015.

[20] Rosen, M., Godin, K. W., and Raghuvanshi, N., "Interactive sound propagation for dynamic scenes using 2D wave simulation," in *Computer Graphics Forum*, volume 39, pp. 39–46, Wiley Online Library, 2020.

[21] Djordjevic, S., Hacihabiboglu, H., Cvetkovic, Z., and De Sena, E., "Evaluation of the perceived naturalness of artificial reverberation algorithms," in *Audio Engineering Society Convention 148*, Audio Engineering Society, 2020.

[22] Smith, J. O., *Physical Audio Signal Processing*, `http`:http://ccrma.stanford.edu/ jos/pasp//-`ccrma.stanford.edu/~jos/pasp/`, accessed <date>, online book, 2010 edition.

[23] Bona, R., Fantini, D., Presti, G., Tiraboschi, M., Engel Alonso-Martinez, J. I., and Avanzini, F., "Automatic Parameters Tuning of Late Reverberation Algorithms for Audio Augmented Reality," in *Proceedings of the 17th International Audio Mostly Conference*, pp. 36–43, 2022.

[24] "Max/MSP," Cycling '74, Max8, accessed: February 2024.

[25] IRCAM, "IRCAM SPAT," IRCAM, v5.3.1, accessed: February 2024.

[26] "Qualtrics," Qualtrics, 2024, accessed: February 2024.

[27] Olufsen, B., "Music for archimedes," *Compact disc CD B&O*, 101, 1992.

[28] Howard, D. and Angus, J., *Acoustics and psychoacoustics*, Routledge, 2013.

[29] International Organization for Standardization, "Acoustics – Measurement of room acoustic parameters – Part 1: Performance spaces," ISO standard ISO 3382-1:2009, International Organization for Standardization, 2009.

[30] Bradley, J. S., Reich, R., and Norcross, S., "A just noticeable difference in C50 for speech," *Applied Acoustics*, 58(2), pp. 99–108, 1999.