Psychometric Mapping of Audio Features to Perceived Physical Characteristics of Virtual Objects

Mattia Colombo[†1], Alan Dolhasz[†2], Jason Hockman[†3] and Carlo Harvey[†4]

[†]DMT Lab, Birmingham City University
orcid.org/{[1]0000-0002-4169-2045, [2]0000-0002-6520-8094, [3]0000-0002-2911-6993, [4]0000-0002-4809-1592}

***Abstract*—Physically-based sound synthesis can simulate virtual sound sources whose audio features reflect the physical characteristics of corresponding objects displayed in a virtual environment, allowing for real-time generation of content without relying on pre-existing audio samples. This, however, requires efficient control strategies for sound synthesis models that, depending on the nature of the sounding objects, require to be mapped to varying physical characteristics displayed through visual information. In this experiment, participants were asked to adjust a set of sound synthesis parameters based on varying physical characteristics of a virtual bouncing ball: distance, elasticity and radius. Statistical analysis of recorded subject responses shows that object radius influences evaluation of pitch and amplitude for the object's representation. Similarly, distance influences user evaluation of both reverb and amplitude whilst elasticity doesn't influence user evaluation of the feature distributions. This result is consistent across user groups evaluated: audio experts and naïve listeners. Models are produced that encode these observations using linear regression, enabling automatic parameterisation of this feature space for audio synthesis engines.**

***Index Terms*—perceptual audio, sound synthesis, procedural audio**

I. Introduction

Perception in virtual environments (VEs) can be studied from an increasing number of perspectives thanks to advances in computer games technology. Modern game engines have made it increasingly cost-effective to generate multi-modal VEs and thus enable the perception of various characteristics of the environment through multiple senses, including hearing. Sound sources, even when only one is present, play a primary role in evoking presence in VEs [1]. Despite the common approach to the generation of such sounds is by processing samples of recorded audio, physical models, based on mathematical representations of natural phenomena, can be used to generate auditory events. The computational costs of these, however, are often prohibitively large for interactive applications.

Modal synthesis offers an alternative approach to the generation of auditory events required to be played in synchrony with visual events, e.g. collisions or footsteps. Modal synthesis allows for the generation of realistic auditory events for interactive applications without the need for prerecorded audio [2]. Farnell's [3] library of real-time implementations of physically-informed sound synthesis models enables the generation of a wide range of everyday sounds such as instruments, mechanical devices, ambiences, explosions and many more. Based on analysis of real auditory events, the models are built using source-filter designs. Although they are parameterised to extend to a range of varying physical characteristics and materials of virtual objects, they introduce the time-consuming and non-intuitive process of manual parameter tuning and mapping [4]. With listeners being the final target, models should only consider parameters directly mapping to relevant aspects of their perception.

To better understand how humans perceive such audio-visual parameterisation, this study examines the relationships between visual object properties and parameters of modal-synthesis-based audio engines, by showing participants a virtual object, with varying physical characteristics, whose sound is synthesized in real-time. Participants are asked to set parameters controlling a synthesizer, according to the physical characteristics of the virtual object. Using an 'off-the-shelf' model for impact sounds, we study how changes in the physical characteristics of virtual objects are reflected in subjective preferences of synthesis parameters, defined according to perceptual studies on synthesis for impact sounds [5]. The main contributions of this work are a model, built from subjective data, that describes how physical attributes of objects are reflected in parameters of an audio synthesis engine; a methodology for data collection in this task relating both human perception and physical characteristic mapping; insight into the difference between two groups, *experts* and *naïve* listeners and how these groups correlate in their opinions on this evaluation task.

II. Related Work

In this section, we summarise recent work in multi-modal displays and human perception in VEs.

Harvey *et al.*, studied the effects of spatialised directional sound on the perception of rendered images showing how rendering performance benefits from modelling relationships between features of stimuli and human perception. [6]. Doukakis *et al.*, conducted a series of perceptual experiments evaluating perceived quality of tri-modal (auditory, visual and olfactory) stimuli in a VE. By varying the availability of computational resources, they indicate how each stimulus affects human perception. The influence of auditory information was shown to have a positive relationship with budgets of computational resources [7], [8].

Boneel *et al.*, tested varying levels of detail in audio-visual displays, composed of a visual render of a virtual object and its relative synthesized sound [9]. They report that the quality of synthesized audio has a significant impact on the perception of details of visual displays. Hulusic *et al.*, review state-of-the-art audio-visual rendering methods discussing their efficiency and applications for VEs. [10]. The synopsis proposed allows determining which methods and algorithms can be used as foundations for VEs with respect to the human sensory system.

According to Rebillat, moving sound sources can be reproduced in auditory displays with various spatial synthesis algorithms. Tests with audio-visual stimuli on perceived distances of objects determine that there is no correlation between the presence and visual distance underestimation [11]. Further research on distance perception in VEs conducted by Gonzel *et al.*, demonstrated how, using Ambisonics audio renders, subjects are more inclined to accept larger incongruences between auditory and visual information of objects at far distances (farther than $8m$) as opposed to close distances (closer than $8m$) [12]. In addition, Tsingos *et al.*, proposed an effective pipeline for audio rendering of virtual sound sources based on their frequency content and perceptual importance to the global soundscape of the VEs [13].

III. Method

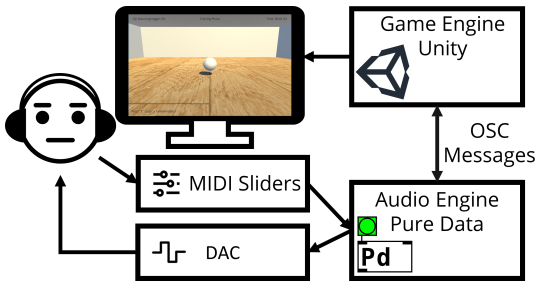

Fig. 1. Experimental set-up of the subjective experiment: the Unity game engine displays the scene to the participant, sets parameters for the virtual object every scene, and interfaces with the audio engine via Open Sound Control (OSC) messages. The audio engine displays synthesised sounds to the participant, through a Digital to Analogue Converter (DAC), based on parameters set from the MIDI sliders and trigger from the game engine.

We model relationships between visual and auditory object properties with respect to human perception, by allowing subjects to control parameters of a modal audio synthesis algorithm to reflect visual properties of a dynamic virtual object - *a bouncing ball*. Subjects have tactile control, via MIDI sliders, over three auditory parameters of the ball: *amplitude*, *reverb* and *pitch*. Subjects are not explicitly informed of what each slider controls — they are simply requested to set the auditory parameters in such a way that they best fit the visual stimulus, based on trial and error and their subjective preference.

A. Stimuli

We design all our experiments in Unity and use PureData for audio synthesis [14]. During the experiment, participants

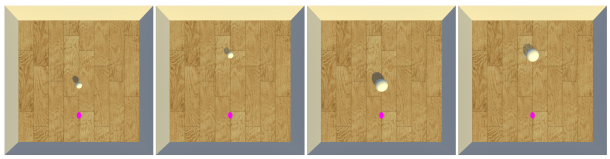

Fig. 2. Permutations of *low* and *high* levels of *distance* and *radius* in ascending order from left to right. Each of the 16 test scenes shows a random permutation of the physical parameters of the bouncing ball.

are shown a scene with a textureless matte ball bouncing on the wooden floor of a virtual room. In each trial, throughout the experiment we vary the physical properties of the ball: *radius*, *distance* and *elasticity*. Each property has two possible levels of magnitude: $radius \in \{0.75m, 1.5m\}$, $distance \in \{7m, 14m\}$ and $elasticity \in \{0.4, 0.85\}$, where 0 is maximum loss of energy after impact and 1 is no loss of energy.

B. Audio engine

For audio synthesis we use a bouncing ball Pure Data model, sourced from Farnell's library [3], [14]. User-controlled parameters affect the following properties of the modal synthesis:

- *pitch*: the fundamental tone of the source-filter algorithm affecting the overall frequency spectrum and thus, the perceived pitch. The fundamental frequency scales linearly in $(87Hz, 167Hz)$;
- *reverb*: output of a Schroeder reverb implementation in Pure Data with 2 as damping factor and 0.05 as room size (reproducing common medium-size room reverberations) — this parameter scales linearly between the original signal and the output of the reverb;
- *amplitude*: subjects have full control of the signal amplitude, scaled logarithmically between $-\infty$ to 0.0 dBFS.

C. Apparatus

We design an audio-visual subjective testing system, shown in Figure 1, comprising of a Unity virtual environment, where the event takes place, displaying visual information to participants via an Apple HD Cinema display with $1920 \times 1080px$ resolution. Audio synthesis is handled by the Pure Data patch, which synthesises audio depending on real-time object collision events sent from Unity via the Open Sound Control (OSC) protocol. The audio generated in Pure Data is played to participants using a pair of Beyerdynamic DT990 headphones. Participants control synthesis parameters using a MIDI device with 3 sliders having a resolution of 2^7 steps to modify *amplitude*, *pitch* and *reverb*.

D. Participants

We recruit $N = 27$ participants, 7 of whom are female, with an average age of 23.9 ($SD = 4.6$), comprising two groups: *naïve* ($N_n = 12$), having no prior experience with listening tests or digital audio technology and *expert* ($N_e = 15$) including researchers, audio engineers, and students from digital-audio-related undergraduate and postgraduate courses.

TABLE I
U AND W STATISTICAL VALUES OF RESPECTIVELY MANN WHITNEY-U AND WILCOXON T TESTS RESULTS SHOWN WITH RELATIVE Z SCORES AND p SIGNIFICANCE VALUES. DISTRIBUTIONS OF *pitch*, *reverb* AND *amplitude* ARE MEASURED FOR BOTH LEVELS OF *radius*, *distance* AND *elasticity*.

		expert				*naïve*				*aggregated*			
		U	W	Z	p	U	W	Z	p	U	W	Z	p
Radius	Pitch	2547.5	9807.5	-8.65	0.00	3629.5	8285.5	-2.54	0.01	12595	36031	-8.27	0.00
	Reverb	7065.5	14325.5	-0.25	0.80	3962	8618	-1.68	0.09	22038	45474	-0.99	0.32
	Amplitude	5391.5	12651.5	-3.36	0.00	3704	8360	-2.35	0.02	18009	41445	-4.1	0.00
Distance	Pitch	6654	13914	-1.02	0.31	4549	9205	-0.15	0.88	22451	45887	-0.68	0.50
	Reverb	2518.5	9778.5	-8.71	0.00	2774	7430	-4.77	0.00	11245	34681	-9.31	0.00
	Amplitude	3908	11168	-6.12	0.00	3238.5	7894.5	-3.56	0.00	14390	37826	-6.89	0.00
Elasticity	Pitch	7030.5	14290.5	-0.32	0.75	4156.5	8812.5	-1.17	0.24	22093.5	45529.5	-0.95	0.34
	Reverb	7119.5	14379.5	-0.15	0.88	4026	8682	-1.51	0.13	21982.5	45418.5	-1.04	0.30
	Amplitude	6252	13512	-1.76	0.08	4429.5	9085.5	-0.46	0.64	21179.5	44615.5	-1.66	0.10

Subjects participated in the study voluntarily and were not rewarded in any way.

E. Procedure

Participants are screened for normal hearing and requested to provide information about their expertise in audio-related fields, then assigned to one of the two experience-based groups. The experiment begins with a training scene, allowing participants to familiarise themselves with the controls and interfaces. To reduce potential learning effects, subjects are encouraged to experiment with the audio parameters using the MIDI sliders, familiarising themselves with the synthesis model. During this phase, they may also change the physical settings of the ball. Participants cannot control the camera position, Figure 1 shows the fixed viewpoint.

After the training phase, subjects are to complete multiple trials with scenes displaying different permutations of the stimuli. As each of the *three* physical characteristics has *two* possible levels of magnitude, a minimum of $\mathcal{P} = 2^3 = 8$ unique test scenes are required to display all their possible permutations. We repeat each permutation twice, resulting in $2 \times \mathcal{P} = 16$ trials per participant. The scene order is chosen at random for every test session. There is no time limit during the test scenes; participants are free to adjust the settings until they feel satisfied.

Once participants confirm parameter values at the end of every trial, a logger integrated into the game engine retrieves the latest information from the MIDI sliders, suspending their input and storing parameters with relative timestamps and physical characteristics of the ball for the current scene. Before proceeding to the successive trial, the engine reassigns a new random order of MIDI sliders to parameters of the audio engine, compelling participants to always understand controls, as well as a new permutation of visual physical characteristics of the virtual object. Finally, MIDI inputs are resumed. The procedure is repeated for the 16 trials of the experiment.

IV. RESULTS AND ANALYSIS

We analyse our results across the *naïve*, *expert*, and *aggregate* groups, which are composed of auditory parameters set by participants during the experiment. Overall, we collect $27 \times 16 \times 3 = 1296$ responses; 432 for each of the three continuous variables: *amplitude*, *pitch* and *reverb*. Responses are analysed with respect to three independent variables: *radius*, *distance* and *elasticity*. To understand whether visual parameters affect recorded subjective responses on audio synthesis parameters, we perform multiple linear regression analysis, modeling synthesis parameter values based on the physical characteristics of objects. Table II reports fitness and significance of models we fit. D'Agostino's K^2 tests conducted on recorded subjects' responses failed to prove that the responses are normally distributed. Hence, we measure statistical differences between distributions and groups of populations by performing repeated 2-tailed Mann-Whitney U tests. Each independent variable was tested against distributions of subjective responses, setting the following hypotheses:

1) $H_0^{'}$, all means under *distance* are equal;
2) $H_0^{''}$, all means under *elasticity* are equal;
3) $H_0^{'''}$, all means under *radius* are equal.

Table I details results of the tests.

V. DISCUSSION

This study reveals how visual visually perceived physical characteristics of objects are associated with subjective responses of parameters of the audio engine. Regression analysis indicates that *Pitch* is determined by *radius*. From a physical point of view, this aligns with modal synthesis models for impact sounds and with perceptual experiments associating pitch with size of visually displayed objects [5], [15]. However, changes in fitness of *pitch* models for *experts* ($R^2 = 33\%$) and *naïve* ($R^2 = 4\%$) indicate inconsistency across levels of expertise. Participants associate *amplitude* and *reverb* with distance of objects, proving validity of experiments conducted on depth perceptions in VE [16]. For the aggregated group, U and W scores determining significant differences between distributions of *reverb* and *amplitude* under the influence of *distance* confirm the validity of assumptions derived from the regression analysis. This also occurs for distributions of *pitch* and *amplitude* under the influence of *radius*. Considering a 5% level of confidence, hypotheses $H_0^{'''}$ and $H_0^{'}$ are rejected. Whereas $H_0^{''}$ is not rejected, and, according to the regression

TABLE II

ANALYSIS SCORES FOR NAÏVE, EXPERTS AND AGGREGATED SUBJECT GROUPS. LINEAR REGRESSION MODELS, FIT ACROSS BOTH LEVELS OF VISUAL PARAMETERS ARE DESCRIBED WITH m GRADIENTS AND THEIR RELATIVE b INTERCEPTS. THE FITNESS OF EACH MODEL AND ITS RELATIVE STATISTICAL SIGNIFICANCE ARE DESCRIBED BY R^2 SCORES AND p VALUES, RESPECTIVELY.

naïve	Pitch				Reverb				Amplitude			
	b	m	R^2	p	b	m	R^2	p	b	m	R^2	p
Distance	0.435	-0.006	0.035	0.88	0.262	0.195	0.152	0.00	0.365	-0.109	0.086	0.00
Elasticity	0.435	0.040	0.035	0.34	0.262	-0.071	0.152	0.06	0.365	-0.011	0.086	0.74
Radius	0.435	-0.101	0.035	0.02	0.262	0.065	0.152	0.08	0.365	0.086	0.086	0.01
experts	Pitch				Reverb				Amplitude			
	b	m	R^2	p	b	m	R^2	p	b	m	R^2	p
Distance	0.520	0.044	0.330	0.16	0.174	0.251	0.275	0.00	0.384	-0.181	0.199	0.00
Elasticity	0.520	0.023	0.330	0.47	0.174	-0.007	0.275	0.79	0.384	-0.059	0.199	0.12
Radius	0.520	-0.332	0.330	0.00	0.174	-0.015	0.275	0.57	0.384	0.112	0.199	0.06
aggregated	Pitch				Reverb				Amplitude			
	b	m	R^2	p	b	m	R^2	p	b	m	R^2	p
Distance	0.483	0.022	0.159	0.40	0.213	0.226	0.196	0.00	0.375	-0.149	0.142	0.00
Elasticity	0.483	0.030	0.159	0.24	0.213	-0.035	0.196	0.12	0.375	-0.037	0.142	0.09
Radius	0.483	-0.230	0.159	0.00	0.213	0.020	0.196	0.37	0.375	0.100	0.142	0.00

analysis scores, *elasticity* has no significant effect on any of the continuous variables as shown in table II and I.

VI. CONCLUSIONS AND LIMITATIONS

The linear regression model effectively describes how visual physical characteristics displayed can be mapped to parameters of the sound synthesis control strategies. Given the context of a bouncing ball, both regression analysis and non-parametric tests determined that:

- *elasticity* has no effect, within the range of parameters tested, on any of the perceived audio parameters;
- *reverb* correlates positively to *distance*, as well as *Pitch* to *radius*;
- *distance* and *radius* influence *amplitude*;
- *size* influences *pitch* and *reverb*;

Differences in statistical significance found between population groups can be meaningful for targeting interactive audio synthesis applications for differing audiences. However, our observation was that both audio experts and non-experts agreed consistently with one another:

The main limitation of this study is the focus on a single scenario, limiting the generalisability of the fit models. Future work will investigate the extension of the feature space exploration to a continuum rather than two discrete points on the feature scale, facilitating interrogation of the non-linearity of the feature space, better informing the models that represent these observations for audio synthesis engines. The analysis reveals relationships between elementary physical characteristics of objects and audio features that are already established. However, the reported perceptual study shows the statistical significance of each of the parameters considered determining their impact on sound synthesis parameters. This proves that perceptual experiments performed adopting similar methods could reveal the relevance of physical characteristics for designing control strategies of sound synthesis models. Thus, contributing to their optimisation and efficiency enabling their use in interactive applications in which complex scenes can be very demanding in terms of audio rendering.

REFERENCES

[1] Pontus Larsson, Daniel Västfjäll, and Mendel Kleiner. Perception of self-motion and presence in auditory virtual environments. In *Proceedings of seventh annual workshop presence*, pages 252–258, 2004.

[2] Perry R Cook. *Real sound synthesis for interactive applications*. AK Peters/CRC Press, 2002.

[3] Andy Farnell. *Designing sound*. Mit Press, 2010.

[4] Taesoo Kwon and Jessica K Hodgins. Momentum-mapped inverted pendulum models for controlling dynamic human motions. *ACM Transactions on Graphics (TOG)*, 36(1):10, 2017.

[5] Davide Rocchesso and Federico Fontana. *The sounding object*. Mondo estremo, 2003.

[6] Carlo Harvey, Kurt Debattista, Thomas Bashford-Rogers, and Alan Chalmers. Multi-modal perception for selective rendering. *Computer Graphics Forum*, 36(1):172–183, 2017.

[7] Efstratios Doukakis, Kurt Debattista, Thomas Bashford-Rogers, Amar Dhokia, Ali Asadipour, Alan Chalmers, and Carlo Harvey. Audio-visual-olfactory resource allocation for tri-modal virtual environments. *IEEE Transactions on Visualization and Computer Graphics*, 25:1865–1875, 2019.

[8] E. Doukakis, K. Debattista, C. Harvey, T. Bashford-Rogers, and A. Chalmers. Audiovisual resource allocation for bimodal virtual environments. *Computer Graphics Forum*, 37(1):172–183, 2018.

[9] Nicolas Bonneel, Clara Suied, Isabelle Viaud-Delmon, and George Drettakis. Bimodal perception of audio-visual material properties for virtual environments. *ACM Transactions on Applied Perception (TAP)*, 7(1):1, 2010.

[10] Vedad Hulusic, Carlo Harvey, Kurt Debattista, Nicolas Tsingos, Steve Walker, David Howard, and Alan Chalmers. Acoustic rendering and auditory–visual cross-modal perception and interaction. *Computer Graphics Forum*, 31(1):102–131, 2012.

[11] Marc Rébillat, Xavier Boutillon, Étienne Corteel, and Brian F. G. Katz. Audio, visual, and audio-visual egocentric distance perception by moving subjects in virtual environments. *ACM Trans. Appl. Percept.*, 9(4):19:1–19:17, October 2012.

[12] Gavin Kearney, Marcin Gorzel, David Corrigan, John Squires, and Frank Boland. Distance perception in virtual audio-visual environments. In *AES*. AES, 3 2012.

[13] Nicolas Tsingos, Emmanuel Gallo, and George Drettakis. Perceptual audio rendering of complex virtual environments. *ACM Transactions on Graphics (TOG)*, 23(3):249–258, 2004.

[14] Miller S Puckette et al. Pure data. In *ICMC*, 1997.

[15] Zohar Eitan, Asi Schupak, Alex Gotler, and Lawrence E Marks. Lower pitch is larger, yet falling pitches shrink: Interaction of pitch change and size change in speeded discrimination. *Proceedings of Fechner Day*, 27:81–88, 2011.

[16] Barbara G Shinn-Cunningham. Distance cues for virtual auditory space. In *Proceedings of the IEEE-PCM*, volume 2000, pages 227–230. Citeseer, 2000.