Perceptual Evaluation of Modal Synthesis for Impact-Based Sounds

Adrián Barahona
Department of Computer Science,
University of York
ajbr501@york.ac.uk

Sandra Pauletto
Department of Media Technology
and Interaction Design,
KTH Royal Institute of Technology
pauletto@kth.se

ABSTRACT

The use of real-time sound synthesis for sound effects can improve the sound design of interactive experiences such as video games. However, synthesized sound effects can be often perceived as synthetic, which hampers their adoption. This paper aims to determine whether or not sounds synthesized using filter-based modal synthesis are perceptually comparable to sounds directly recorded. Sounds from 4 different materials that showed clear modes were recorded and synthesized using filter-based modal synthesis. Modes are the individual sinusoidal frequencies at which objects vibrate when excited. A listening test was conducted where participants were asked to identify, in isolation, whether a sample was recorded or synthesized. Results show that recorded and synthesized samples are indistinguishable from each other. The study outcome proves that, for the analysed materials, filter-based modal synthesis is a suitable technique to synthesize hit sound in real-time without perceptual compromises.

1. INTRODUCTION

Nowadays, most video games, films and pieces of media are sound designed using pre-recorded samples. Pre-recorded samples are obtained from direct audio recordings or layered sound effects and stored in audio files. However, pre-recorded samples have several limitations in interactive environments such as video games. As actions in games can be performed several times, if there is only one sample per action, the sound will be repeated, which can lead to listener fatigue and loss of authenticity [1]. To solve this, several samples can be assigned and shuffled played when a player performs an action, which the consequential increment in studio and implementation time, asset management and memory footprint problems.

Hit or impact-based sounds are the acoustic consequence of physical collisions. Changes in an object material or size will produce changes to the resulting impact sound. For games or interactive applications with hundreds or thousands of interactable assets, such as open world games or VR experiences, this will lead to an exponential growth in the need for different sound samples to sonify any particular scenario where two or more of those assets collide.

In the context of video games and interactive applications, an alternative solution is to use real-time sound synthesis during gameplay. This approach, known as procedural audio, is defined by Farnell [2] as a "non-linear, often synthetic sound, created in real time according to a set of programmatic rules and live input" .

Farnell [2] enumerates several benefits procedural audio has over pre-recorded samples. Among them, the programmatic nature of procedural audio allows sound designers to decide or change aesthetic considerations later in the development cycle. Moreover, as procedural audio is object based, it can automate part the sound design process, especially in the implementation stage, as a single procedural audio model can contain all the possible sonic interactions a player can perform. Procedural audio also offers more variety, versatility and adaptability than pre-recorded samples. While a pre-recorded sample will always play the same way, procedural audio can change dynamically.

However, Farnell [2] also identifies perceived realism as one of the problems in procedural audio. Synthesized sound effects can often be perceived as too synthetic compared to pre-recorded samples. One of the challenges is, then, to create procedural audio models that can be indistinguishable from pre-recorded samples.

This study aims to measure whether or not it is possible to identify synthesized hit sound effects using filter-based modal synthesis. Instead of comparing pre-recorded samples and their synthesized versions side by side, this study will present them independently. The motivation of taking this approach relies in the idea that if pre-recorded hit samples and efficient real-time synthesized hit sound effects are indistinguishable from each other, the synthesized version can be used without perceptual loss of authenticity and obtaining all the benefits procedural audio presents. The null hypothesis of this study is that the synthetic sound effects are easily recognizable from the recorded ones.

1.1 Previous work

There is a body of research on evaluating the perception of synthesized sounds. A recent study analysed different sound synthesis techniques for different sound classes (applause, babble, bees, fire, rain, stream, waves and wind), concluding that there is not substantial difference between the reference sample and the synthesized version when an appropriate synthesis method was used, except for additive

synthesis [3]. However, they did not focus on impact-based sounds.

Other studies evaluated the use of modal synthesis for the synthesis of weapons sounds [4]. The results showed a convincing result for weapons with resonant modes (such as axe, hammer or rapier). However, they took a rating approach for the experiment, where participants rated several versions of the sound effects (synthesized, synthesized and processed and pre-recorded) instead of evaluating them in isolation. The motivation of evaluating the sounds in isolation instead of rating several versions of them comes from avoiding any possible bias from comparing a specific sound to another. Real-time synthesized rolling sounds of glass and wood were also evaluated in other experiments [5]. Instead of identifying synthetic samples, participants rated the synthesized sounds in a 0-100 scale to evaluate the realism obtained in the rolling effect.

Tests to discriminate whether a sound is recorded or synthetic have been developed in the field of synthesis of musical instruments. Wun, Horner and Ayers [6] proposed a discrimination factor, *d*, which measure the quality of a synthetic tone based on how often it can be distinguished from its recorded counterpart. The discrimination factor is similar to the *Accuracy* rate used in other works dealing with binary classifiers [7]. This metric has been used to evaluate the use of synthetic piano sustain-pedal effects [8] or a synthetic clavinet model [7].

2. METHOD

2.1 Sound Synthesis

Filter-based modal synthesis is a particular use of subtractive synthesis and it is especially indicated to synthesize impact-based sounds [9]. Filter-based modal synthesis has two components: a deterministic part, the model modes, and a stochastic part, the model noise envelope.

The process consisted in analysing a pre-recorded sample to extract its modes. Modes are the individual sinusoidal frequencies to which an object vibrates, in this case, when it is impacted. In the spectrogram of a hit sound, these modes are represented by straight horizontal lines (Figure 1). The modes are used as frequency bands in a series of resonant filters which band-pass an enveloped white noise signal. These modes represent the deterministic part of the sound.

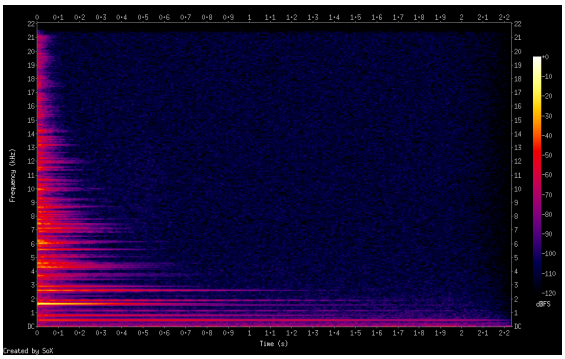

Figure 1. Recorded metal flask hit spectrogram.

The modes are then subtracted from the original sound, obtaining the noise envelope, also known as the residue. This represent the stochastic component of the signal. The residue is stored in an audio file and it is triggered with the deterministic component.

The ChucK programming language [10] was used for the modes extraction and residue generation. The extraction and residue generation code was written by Perry Cook [11].

For this study, nine sounds were recorded from materials that exhibit clear modes when excited. The choice of the materials was driven by the suitability of the synthesis method used. These materials were ceramic (2 samples: a plate and a mug), glass (3 samples: an empty bottle, a water glass and a pint glass), metal (3 samples: two lids of different sizes and a flask) and wood (1 sample: a short rod) (Figure 2). The hits were performed with a metal spoon. The recordings were made at 44.1kHz/24bit with a Zoom H6 recorder, using the built-in XY capsules. 100 modes were extracted from each material and subtracted from the original recording to generate the residue.

Figure 2. Materials used.

The modal synthesizer was also programmed in ChucK, using a modified version of a modal synthesizer created by Cook [11]. Each sound was synthesized by using an enveloped white noise signal thought 100 resonant band-pass filters plus the residue. Every time a hit is triggered, the individual mode frequencies, individual mode gains, individual filter Q, residue pitch (playback speed) and balance between the deterministic and stochastic components were randomized. The level of randomization differs slightly for each material. For context, in the case of the mug, the range of randomization of the individual frequencies was of frequency $\pm$ frequency/300, of the individual gains was of gain $\pm$ gain/5, the individual filter Q were randomized between 800 and 1200, the residue playback rate was randomized between 0.99 and 1.01, the gain of the deterministic component was randomized between 5 and 30 and finally the gain of the stochastic component was randomized between 0.7 and 1. The aim of this randomization is to create a natural variation between hits.

A live performance of the modal synthesizer was recorded. One audio file between 3 and 6 seconds with a series of hits was recorded for each object. The original non-synthetic recordings were sliced to generate one audio file for each object comparable to the synthesized version. The synthesized versions did not have any reverberation as they

were not recorded in any physical environment. To avoid any bias in the perceptual evaluation, an impulse response of the room where the original recordings were made was recorded and mixed with the synthesized samples. All audio files were normalized to 0LU and exported in Ogg Vorbis. The choice of Ogg Vorbis instead of uncompressed WAV was determined by the technical limitations of the questionnaire platform used, Qualtrics [1]. However, this should not cause a drastic perceptual variation [12]. No more processing was applied to the audio files.

The recorded sounds, the ChucK code of the synthesizer used and the resulting synthesized sounds can be found in the online repository [2].

2.2 Experimental design

The test used was inspired by the RS -or real and synthetic- listening test proposed by Gabrielli, Squartini and Vlimki [7]. Although the test was originally used with musical instruments, it can be easily applied to sound effects. The RS listening test proposes a series guidelines and this study does not follow all of them.

In this study, as opposed to the RS guidelines, the test was not carried out in a controlled listening environment. The listening test was conducted online with no information of the playback device used by the participants, although the use of headphones was suggested. This helps to replicate more closely the playing environment, where the participants are likely to use their own equipment to play the game or interactive application. The participants were asked to identify, one by one, whether the sound played was recorded or synthesized (Figure 3). Participants classified the samples without being asked to specify to which material a sample belongs to. The order of the audio samples was randomized and the participants were asked to listen to each sound just once. As suggested in the RS test guidelines, an acid test (a clearly synthesized sound) was included to acts as a control.

The participants were also asked to introduce their level of expertise in sound design, ranked from 1 (no expertise) to 5 (professional).

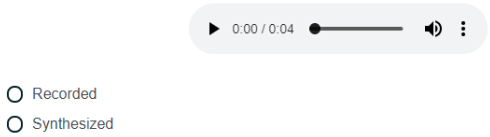

Figure 3. Online test interface.

The metrics used were the discrimination factor and the F-measure.

The discrimination factor, *d*, is described as [6]:

$$d = \frac{P_{CS} - P_{FP} + 1}{2} \tag{1}$$

With P_{CS} being the percentage of correctly detected synthesized sounds and P_{FP} the false positives (recorded samples identified as synthetic). Following the RS test criteria, *d* values below 0.75 mean the sounds compared are considered indistinguishable from each other. Values of *d* around 0.5 are not different from random guessing.

As suggested in the RS guidelines, the F-measure was also evaluated. F-measure is described as [7]:

$$F - measure = \frac{(\beta^2 + 1) \times Precision \times Recall}{\beta^2 \times Precision + Recall} \tag{2}$$

With *Recall* defined as:

$$Recall = \frac{P_{CS}}{P_{CS} + P_{FN}} \tag{3}$$

And *Precision* defined as:

$$Precision = \frac{P_{CS}}{P_{CS} + P_{FP}} \tag{4}$$

Being P_{FN} the percentage of false negatives (synthetic sounds labeled as recorded) and $\beta = 1$. The interpretation of the F-measure values is similar to the *d* values.

A total of 19 participants, 12 males and 6 females (1 participant did not disclose this information) with ages between 18 and 58 took the test. A breakdown of the participants level of expertise in sound design is showed in Figure 4.

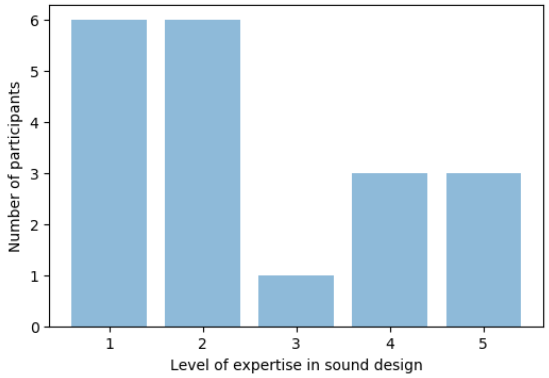

Figure 4. Participants level of expertise in sound design.

3. RESULTS

All participants correctly labelled the acid test as synthetic, so it was removed from the analysis of results. The purpose of this sample was to filter out participants that were random guessing or not paying attention to the test.

The mean results clearly show that recorded and synthetic hit sound effects are indistinguishable from each other. All scores are below the 0.75 threshold and closer to the random guessing mark of 0.5. There is also a low fluctuation

[1] https://www.qualtrics.com

[2] Repository containing the recorded sounds, the ChucK code of the synthesizer used and the resulting synthesized sounds: https://github.com/adrianbarahona/SMC-Conference-2019_Perceptual-Evaluation-of-Modal-Synthesis-for-Impact-Based-Sounds

Avg *d*	σ	σ^2	Max d	**Avg F-measure**	σ	σ^2	Max F-measure
0.5	0.12	0.02	0.72	0.43	0.16	0.02	0.73

Table 1. d and F-measure values across all participants.

Samples	% Correctly labeled as recorded	Samples	% Correctly labeled as synthesized
Ceramic recording	66%	Ceramic synthesized	29%
Glass recording	54%	Glass synthesized	37%
Metal recording	60%	Metal synthesized	40%
Wood recording	68%	Wood synthesized	68%

Table 2. Different materials raw results.

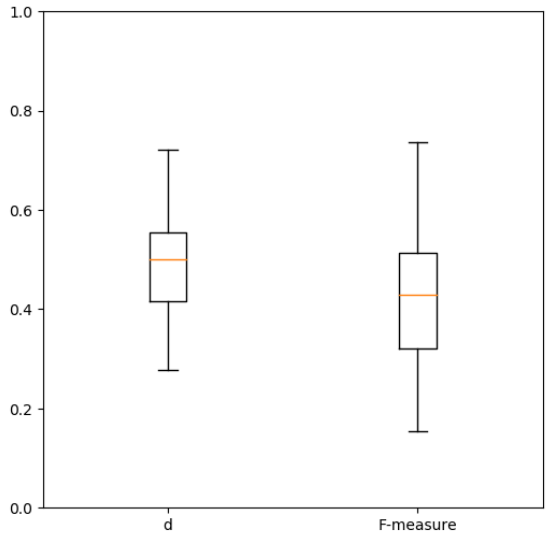

Figure 5. d and F-measure values across all participants.

between the participants (Figure 5). The results are shown in the Table 1.

Some samples performed better than others. The synthetic version of the wood rod was successfully identified by 13 of the 19 participants. In the other hand, the synthetic version of the mug was only identified by 4 of 19. An overall breakdown of the different materials performance is shown in Table 2. The table shows the different objects grouped in the correspondent material.

The level of expertise in sound design was not a decisive factor to spotting synthetic sounds. Participants with an expertise in sound design ranked between 1 and 2 (out of 5) scored a d and F-measure of d = 0.48 ($\sigma = 0.10$) and F-measure = 0.42 ($\sigma = 0.14$). Participants with the highest level of expertise in sound design, 4 and 5, scored d = 0.55 ($\sigma = 0.11$) and F-measure = 0.46 ($\sigma = 0.17$). In fact, the participant with the highest d and F-measure (0.72 and 0.73 respectively), had no expertise in sound design.

4. DISCUSSION

This paper presented the perceptual evaluation of filter-based modal synthesis hit sounds using a method inspired in the RS listening test. The aim was to measure whether or not listeners can identify synthesized hit sound effects using filter-based modal synthesis. Results showed that, for the analysed materials, recorded and synthetic samples are indistinguishable from each other.

The different performance among the materials suggest a further study focused on what specific materials are more suitable for this particular synthesis method. Moreover, a bank of modes and their relative residue files could be created, removing the need of analysing a new audio file for each new model. The bank of modes can be created by analysing several recordings from the same material to establish a range of common modes for each material analysed.

Filter-based modal synthesis is comparable to the method used by Mengual, Moffat and Reiss to synthesize weapon sounds [4]. In their case, they took an additive approach, using spectral modelling synthesis [13]. Instead of using filtered white noise for the deterministic component, they used sinusoidal waves and the noise component is also modelled instead of triggered from the residue file. Filter-based modal synthesis offers more control over the deterministic component as the parameters of the filters, such as the Q, can be controlled in real time. However, filter-based modal synthesis offers less control over the stochastic component in this case, as it is extracted directly from the original recording and stored as an audio file.

There are some improvements to the modal synthesiser than can be implemented and evaluated. First, to ease the CPU usage, a test measuring the perceptual impact of using less modes can be performed. The synthesizer programmed for this paper uses 100 modes for each material, but it would be beneficial to draw threshold in the number of modes where models start losing authenticity. This could be also applied to implement dynamic levels of audio detail in video games. Another test could measure where that threshold is situated when the procedural models are not played in isolation but as part of a soundscape, given that frequency masking could hide their synthetic nature.

Another improvement can be the use of individual filter frequency decay time as every individual mode decay time is different for each material. This effect can be clearly appreciated in Figure 1. A test could measure the perceptual impact of this feature, comparing models with and without it. This is especially interesting for materials that performed worse using the current synthesizer, such as wood. In addition, the residual component could be also modeled

with filtered white-noise by taking the most relevant modes from the residue file.

The study results combined with the fact that the synthesizer runs in real time encourages the use of the models within a game engine. This can be done in the Unity game engine [14] by using the Chunity plugin [15]. Chunity is a package for Unity that integrates ChucK within the game engine. Tests to measure the synthesizer impact on the CPU at runtime could be done to determine whether the models are ready for production.

Controlling the synthesizer in real-time can also be a direction for a second stage of this study. Physical parameters within the game engine such as stiffness, size or geometry have been already used to control the sound of modal synthesizers in real-time [16]. This opens the possibility of expanding the use of the models to perform other actions apart from hits, such as scratching or brushing, or using haptic controllers to interact with the synthesizer.

Acknowledgments

This work was supported by the EPSRC Centre for Doctoral Training in Intelligent Games & Game Intelligence (IGGI) [EP/L015846/1].

5. REFERENCES

[1] R. Selfridge, D. Moffat, E. Avital, and J. Reiss, "Creating Real-Time Aeroacoustic Sound Effects Using Physically Informed Models," *Journal of the Audio Engineering Society*, vol. 66, no. 7/8, pp. 594–607, 2018.

[2] A. Farnell, "An Introduction to Procedural Audio and Its Application in Computer Games," in *Audio Mostly Conference*, 2007, pp. 1–31.

[3] D. Moffat and J. D. Reiss, "Perceptual Evaluation of Synthesized Sound Effects," *ACM Transactions on Applied Perception*, vol. 15, no. 2, pp. 1–19, 2018.

[4] L. Mengual, D. Moffat, and J. Reiss, "Modal Synthesis of Weapon Sounds," in *Audio Engineering Society Conference: 61st International Conference: Audio for Games*, London, UK., 2016.

[5] E. Murphy, M. Lagrange, G. Scavone, P. Depalle, and C. Guastavino, "Perceptual Evaluation of a Real-time Synthesis Technique for Rolling Sounds," in *Conference on Enactive Interfaces*, Grenoble, France, 2007.

[6] C.-W. Wun, A. Horner, and L. Ayers, "Perceptual Wavetable Matching for Synthesis of Musical Instrument Tones," *Journal of the Audio Engineering Society*, vol. Vol.49, no. 4, pp. 250–262, 2001.

[7] L. Gabrielli, S. Squartini, and V. Välimäki, "A Subjective Validation Method for Musical Instrument Emulation," in *AES 131st Convention*, New York, USA, 2011.

[8] H.-M. Lehtonen, H. Penttinen, J. Rauhala, and V. Välimäki, "Analysis and Modeling of Piano Sustain-Pedal Effects," *The Journal of the Acoustical Society of America*, vol. 122, no. 3, pp. 1787–1797, 2007.

[9] P. R. Cook, *Real Sound Synthesis for Interactive Applications*. Natick, Massachusetts: A K Peters, 2002.

[10] G. Wang, "ChucK," 2018.

[11] P. Cook and J. O. Smith, "Physics-Based Sound Synthesis for Games and Interactive Systems," 2018. [Online]. Available: https://www.kadenze.com/courses/physics-based-sound-synthesis-for-games-and-interactive-systems-iv

[12] H. P. S. Selasky, "Evaluation of Perceptual Sound Compression with Regard to Perceived Quality and Compression Methods," Ph.D. dissertation, Høgskolen i Agder ; Agder University College, 2006.

[13] X. Serra and J. O. Smith, "Spectral Modeling Synthesis: A Sound Analysis/Synthesis System Based on a Deterministic Plus Stochastic Decomposition," *Computer Music Journal*, vol. 14, no. 4, p. 12, 1990.

[14] Unity Technologies, "Unity," 2018.

[15] J. Atherton and G. Wang, "Chunity: Integrated Audiovisual Programming in Unity," in *NIME*, Virginia, US, 2018.

[16] L. Pruvost, B. Scherrer, M. Aramaki, S. Ystad, and R. Kronland-Martinet, "Perception-Based Interactive Sound Synthesis of Morphing Solids' Interactions," in *SIGGRAPH Asia 2015 Technical Briefs*, 2015, p. 17.