MODAL SYNTHESIS OF WEAPON SOUNDS

LUCAS MENGUAL[1], DAVID MOFFAT[1], AND JOSHUA D. REISS[1]

[1] *Center for Digital Music, Queen Mary University of London, UK*
l.mengual92@gmail.com ; d.j.moffat@qmul.ac.uk ; joshua.reiss@qmul.ac.uk

Sound synthesis can be used as an effective tool in sound design. This paper presents an interactive model that synthesizes high quality, impact-based combat weapons and gunfire sound effects. A procedural audio approach was taken to compute the model. The model was devised by extracting the frequency peaks of the sound source. Sound variations were then created in real-time using additive synthesis and amplitude envelope generation. A subtractive method was implemented to recreate the signal envelope and residual background noise. Existing work is improved through the use of procedural audio methodologies and application of audio effects. Finally, a perceptual evaluation was undertaken by comparing the synthesis engine to some of the analyzed recorded samples. In 4 out of 7 cases, the synthesis engine generated sounds that were indistinguishable, in terms of perceived realism, from recorded samples.

1 INTRODUCTION

The current workflow for sound effect design within audio media production such as film, TV, and video games, typically involves the use of sound effects following tradional recording methods. Alternatively, the procedural audio approach [1] applies a real-time computer generated synthesized sound effect. Generated sounds vary through control parameters, and offer more flexibility for the designer to model a sound that is adaptive and unique for every instance.

Sound synthesis is an appealing approach to game audio since it allows the modelling and creation of dynamic sound effects at run time and does not rely on simple playback of a sample. The process of procedural audio can improve interaction and user experience and create a highly immersive experience. Procedurally generated sound synthesis is fundamental to improving human perception of human computer interactions from an audible perspective. Böttcher and Serafin demonstrated subjectively that, in an interactive gameplay environment, 71% of users found synthesis methods more entertaining than audio sampling. Users rated synthesised sound as higher quality, more realistic and preferable [2].

This paper presents various analysis techniques to obtain acoustical features from recorded samples. These acoustic features are then applied to a synthesis model which is able to produce a range of different weapon sounds. We choose to focus on weapon sounds since they are prevalent in sound design, especially in many computer games. Existing work is extended by applying some procedural methodologies, application of post-processing effects and perceptual evaluation of the different synthesised sounds.

The rest of this paper is presented as follows. Section 2, discusses the current work in modal synthesis. Section 3 presents the implementation undertaken within this project. Section 4 presents the results and perceptual evaluation of our synthesis model. Section 5 will then conclude this work.

2 PREVIOUS WORK

There has been significant research on the generation of impact sounds through a spectral modelling synthesis approach [3-5], and also applying residual noise generation with subtractive synthesis. Despite this and despite their pervasive use as sound effects in media production, there is little work on weapons sound synthesis, whether they are combat weapons or gunfire sounds.

Research in procedural audio has demonstrated its potential for use in game audio [6, 7], and recent research suggests that it may be deployed as a replacement to sound effect libraries [8].

2.1 Sinusoidal Additive Synthesis

Sinusoidal additive synthesis is often used in musical instrument synthesizers, such as piano and guitar emulators. Normally, these emulators work well on harmonic-based sinusoidal synthesis. This method alone has a low preference in the real world, due to the output audio sounding too artificial and harmonically clean [9]. Therefore, to obtain a more complete and realistic sound, it is common to apply some noise to the signal [3, 4, 11].

Results have shown that a listener receives a more pleasant perception of the sound when any sort of pink or white noise is added to the signal [10]. In this case, an adaptive method of subtractive synthesis is implemented, so that a series of notch filters opposing the sinusoidal modes leave space in the frequency spectrum, so that the stochastic and harmonic signals merge naturally and sound appealing.

2.2 Modal Synthesis

Modal synthesis is a specific application of sinusoidal modelling. For a given impact, a number of modes are detected and the impact is then synthesised through sinusoidal modelling. The output sound may be obtained through a precise recombination of such frequencies, depending on excitation and output parameters. Modal synthesis can be calculated as:

$$x(t) = \sum_{m}^{X^{M}} g_m A_m(t) \sin(2\pi f_m(t) + \varphi_m) + r(t)$$

Where the synthesized output signal $x(t)$ is a summation of the M modes, and g_m, $f_m(t)$, ϕ_m, $A_m(t)$ are the gain, frequency (Hz), phase, and amplitude envelope of each mode, respectively. $r(t)$ represents the residual noise that is added in order to complete the computational representation of the output signal. An example spectrogram of the residual noise from an impact sound is given in Fig. 1, and the power spectrum of the modal output and the original signal is given in Fig. 2.

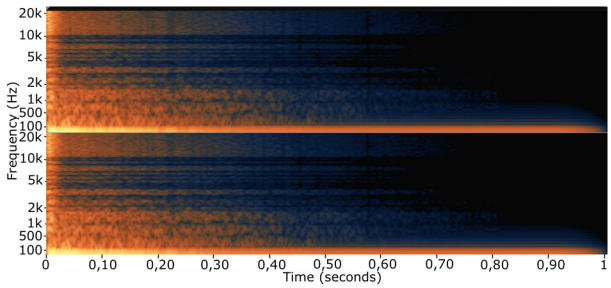

Figure 1: Noise Residual (stochastic part) representation of a Hammer signal. Left (top) and Right (bottom).

3 IMPLEMENTATION

A synthesis method was developed to simulate gun (Beretta M9 pistol, Winchester 1300 shotgun, and AK47 machine gun) and combat weapon sounds (Axe, Hammer and Rapier) all taken from the *Boom Library*. Samples of sound effects were analysed to interpret features from audio signals. This analysis, shown in Fig. 2, consisted of a Short-Time Fourier Transform (SFTF) to collect frequency information from the sound signal. Spectral Modelling Synthesis (SMS) was used to recreate the detected frequency peaks, using an additive synthesis method. The envelope of the signal is then analysed. Finally, a noise signal is generated and shaped with the input envelope and a subtractive synthesis method is used to shape the noise background and tail of the sound effect.

3.1 Analysis

The choice of analysis window establishes the trade-off between the frequency and time resolutions. The selection of the window used determines two key features; the width of the main lobe, representing the number of bins (frequency), and the height of the side lobes. A Blackman window was used since it has a wide main lobe, but the height difference between the main lobe and sides lobes is approximately 60dB. Also, a resolution of 4096 samples was used in order to get a better frequency resolution.

It is challenging to distinguish between frequency or noise-based (gunshots) peaks for signals that have high noise content, due to a dynamic range restriction of the sound pressure level (dB) captured by a microphone and spatial location. As a result, the peaks were chosen based on an overall energy grouping in the STFT windows and peaks were determined to be within the frequency range of the peak group.

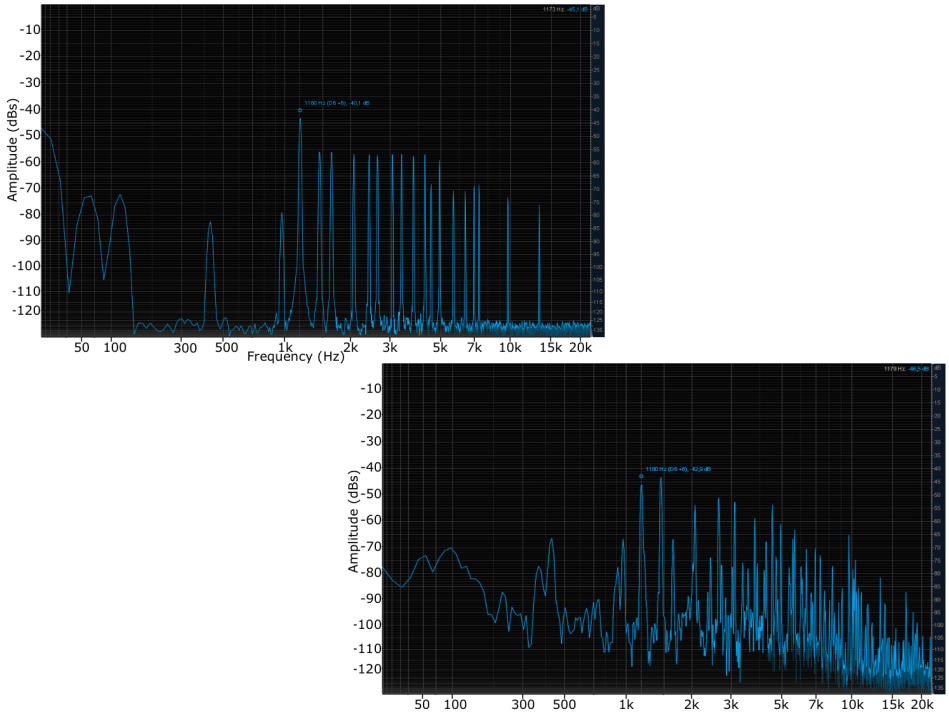

Figure 2: Spectrum of Hammer modal model (top) and original signal (bottom).

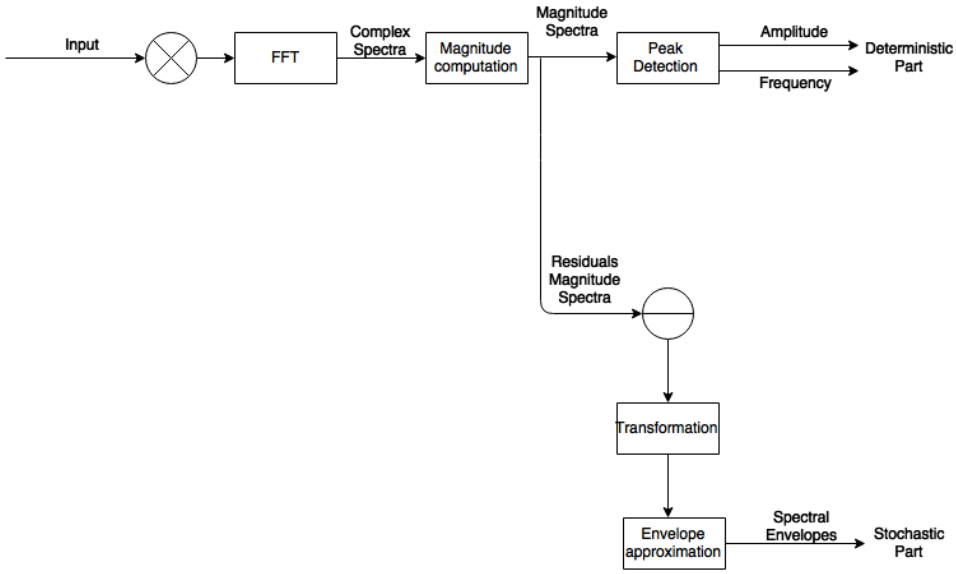

Figure 3: Block diagram of the analysis part of the sinusoidal modelling.

The STFT analysis is limited to sinusoids that vary slowly over amplitude and frequency range. However, the assumption of distinct frequency peaks can lead to a high quality representation of the impact-based sound being analysed. Therefore, harmonically structured sounds can be better represented using modal synthesis [9]. The frequency peaks (modes) were collected based on an analysis rule of the highest local maximum in the spectrum. Note that not all frequency peaks are equally perceived in volume (dB), due to natural physical and residual noise factors; the analysis was relevant to detect the frequency and amplitude of the peaks.

The analysed sound samples were from a sound library company called *Boom Library*[1], and were the 'Natural' recording from the Guns package and 'Medium' sounds from Medieval Weapons.

3.2 Deterministic Signal Modelling

The deterministic signal is modelled though the analysis of existing sound samples. Each peak is formed by an amplitude and frequency function and an envelope (ADSR) shaper that corresponds to the original signal. The signal is then constructed as a sum of cosine signals. The exact sinusoids produced are different every time, where the sinusoids are modulated slightly within a range determined by the analysis of existing samples.

The synthesised sounds have between 16 to 22 modes, each with random initial phase, in their deterministic construction. The waveform amplitude envelope is visually identified through an energy over time spectrogram. These captured energy levels are then spread out individually to each sinusoidal generator via a linear gain function.

Bonneel et al. noted that "*spreading impacts out over multiple frames can also help avoid peaks in computational load*" [5]. Thus, the model uses small delays between 1 to 4 milliseconds to eliminate any possible computational overload and digital clipping.

3.3 Stochastic Signal Modelling

The residual component of the signal is constructed via subtractive synthesis based on the deterministic signal. The residual component of the model relies only on frequency spectrum characteristics and amplitude level over time. "*The computation of the stochastic representation involves the subtraction of each magnitude spectrum of the deterministic component of the original sound, and the approximation of each residual spectrum with an envelope*" [11]. Pure white noise is generated to recreate the residual background signal of the original sound. But in some circumstances, pink noise is used for sounds that require less high frequency content (shotgun and combat weapons). This was identified manually based on perceptual analysis of the samples and synthesised sounds.

For the noise residual, we grouped a series of second order band-pass filters in a filter bank. The concept of this algorithm is easy to add into our model, and very attractive for the overall CPU usage.

At the end of the stochastic signal, there is a low-pass filter in which the centre frequency is automated to cut the high frequency tail as soon as possible. These automations vary depending on the residual analysis from the impact-based sounds. Other optimization methods applied in the residual signal are the multiplication function for the noise signal generator, so that memory is not wasted in summing noise signals, and a hard limiter to avoid any unwanted digital clipping.

Gunfire sounds show a larger dependency on the stochastic model than the combat weapon sounds. As shown in Fig. 4, gunfire sounds are noise-based sounds, so the sounds are more reliant on the stochastic components, rather than the deterministic.

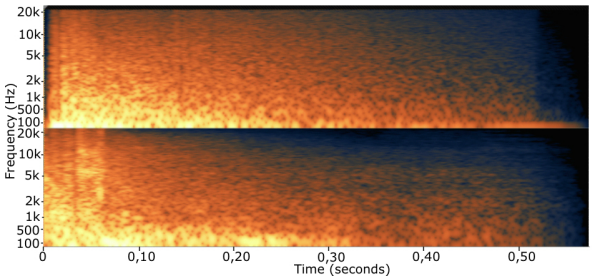

Figure 4: Computed stochastic noise signal of the Winchester 1300 shotgun (top) and original (bottom)

Sound samples were analysed and the characteristics of the spectral envelopes of the stochastic features, frequency and amplitude were used to synthesise the stochastic signal. The inverse Fourier Transform for each spectral envelope was computed in the deterministic model, resulting in a stochastic noise signal.

[1] http://www.boomlibrary.com/

3.4 Processed Effects

Our model also implemented some extra effects as the reverberation and saturator to achieve a different result from the initial, dry generated sound (dry signal). The goal of processing the real-time synthesized signal through these effects was to obtain a more realistic and enjoyable output.

We implemented a pre-built reverb function that utilizes a similar approach to Moorer´s early-reflections reverberator [12], plus advanced filtering options. Inside the reverb function, we were able to modify the liveliness, crossover frequency, high frequency damping, and a main dry/wet level. We adjusted the reverb into a general medium room environment, and kept the same values for all the different samples.

Saturation, or soft clipping distortion (see Fig. 5), was also applied. This results in additional harmonics, thus creating a richer signal. We processed the input signal by applying a distortion effect based on the arctangent function, $y[n]=5\arctan(x[n])$, and calculating a new output for each input sample.

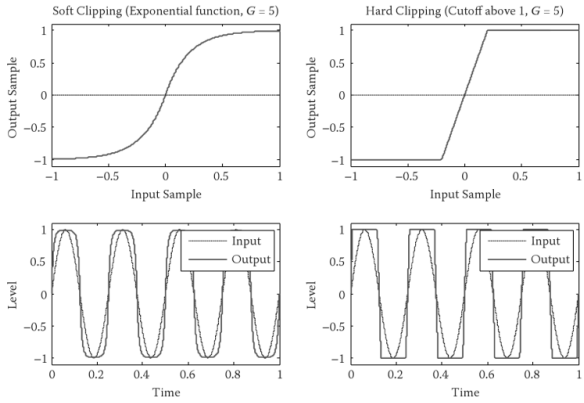

Figure 5: Illustrations depicting soft clipping (left) and hard clipping (right) [13]

4 RESULTS

We applied a web browser-based audio perceptual evaluation tool, shown in Fig. 6, in order for subjects to rate certain qualities of different audio fragments [14, 15]. The test rated the realism and desirability of the sounds.

Figure 6: HTML5 perceptual evaluation tool user interface.

In total, 15 test subjects completed the evaluation inside a professional mixing studio with high-end speakers. None of the test subjects had weapon-related expertise or hearing disorder. Each participant was asked to rate each group of sounds based on how “realistic” they perceived the sounds to be. All sounds were rated on a single horizontal scale, to encourage inter-sample comparison. Between 6 and 10 samples were evaluated in each test case. Each sound category included two original samples, two dry synthesized samples, two processed synthesized samples, separate deterministic and stochastic samples, and processed versions of these last. Every participant undertook all seven sound category evaluations, one per weapon sound. Information about the age, sex orientation, and previous audio-related experience was also retrieved during the test, but was not relevant to the results.

The average test duration was 15 minutes per participant. The sound categories and sample orders were both randomised by the test environment and the samples were normalised at an equal loudness level, and leaving the option for the participants to adjust the volume knob manually.

For the experimental results presented below, to name the audio samples we used the following naming convention. The original samples are tagged with “V1”and “V2”. If a sample was processed with both a saturator and a reverb effect, then it is tagged as “Wet”, otherwise, it will be tagged as “DryReverb” or “DrySaturator” effect, which means only the specified effect is applied. If the audio sample is tagged as “Dry” with no mention of reverb or saturator, then no additional processing is applied. Lastly, to categorise the samples we tagged “Full” for the samples including both stochastic and deterministic signals, “Noise” only stochastic, and “Modes” only the deterministic part.

There are two types of generated AK47 sounds, varying in the frequency peak (modes) selection and also the residual noise envelope. Since there are two synthesized versions of the AK47, we used the same sample, “V2” to compare them and different “V1” and “V1.2” samples. Also exclusively in the AK47 versions, we had “Single” samples, which means it plays a single shot, and “Burst” having multiple shots in the same sample.

Figures 7 to 13 show the confidence interval plots of the user preference based on the rating experiment undertaken.

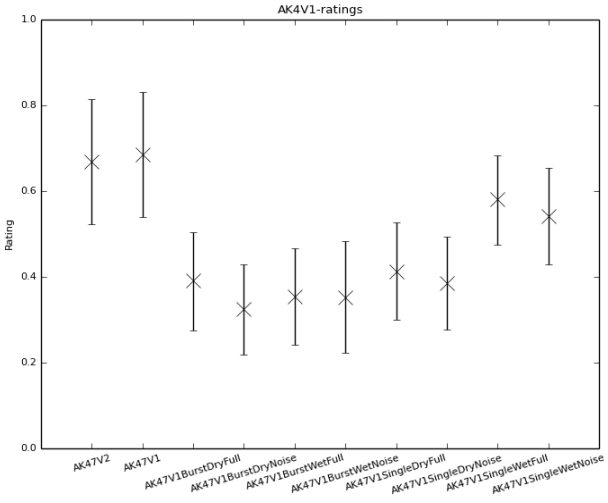

Figure 7: Ratings of AK47 version 1

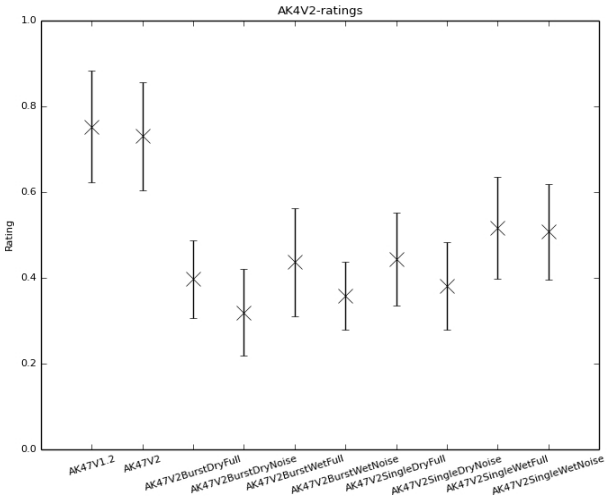

Figure 8: Ratings of AK47 version 2

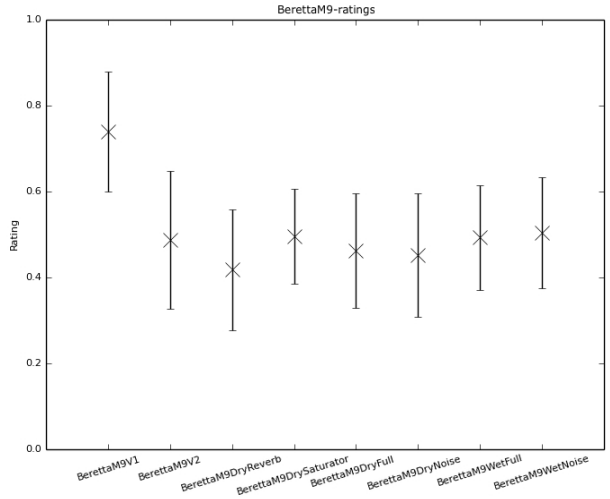

Figure 9: Ratings of Beretta M9 pistol

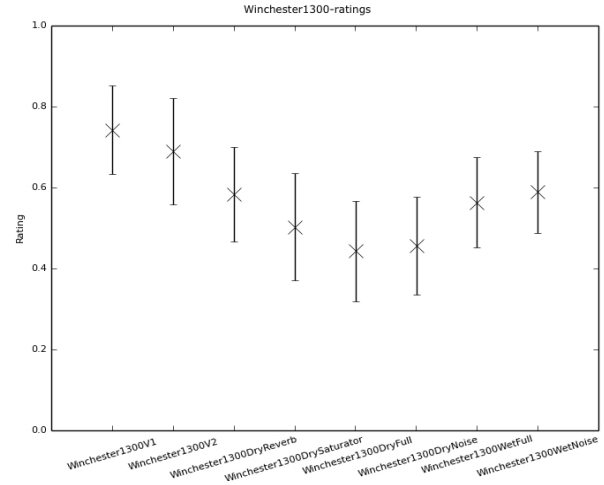

Figure 10: Ratings of Winchester 1300 shotgun

Figures 7 to 10 clearly show a strong tendency to like resonance-based (combat weapon) sounds over noise-based (gunfire) sounds. Gunfire groups, represented in both AK47 plots (Fig. 7 and Fig. 8), showed that the recorded samples offer a better outcome. It also shows that perceptually, one cannot easily distinguish between good synthesis and recorded samples, so the synthesis is ‘as good as’ samples for the Winchester 1300 (Fig. 10), and combat weapons (Fig. 11 to 13).

Nevertheless, results were more levelled or even managed to surpass the original resonance-based sound samples based on the participant´s preference for the most appealing and realistic sound.

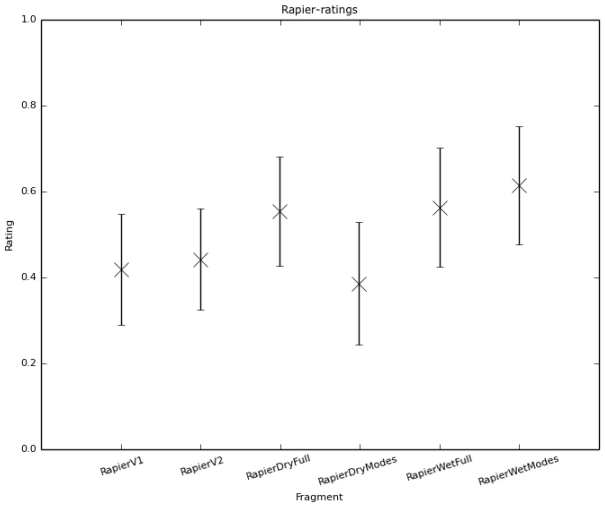

Figure 11: Ratings of Rapier

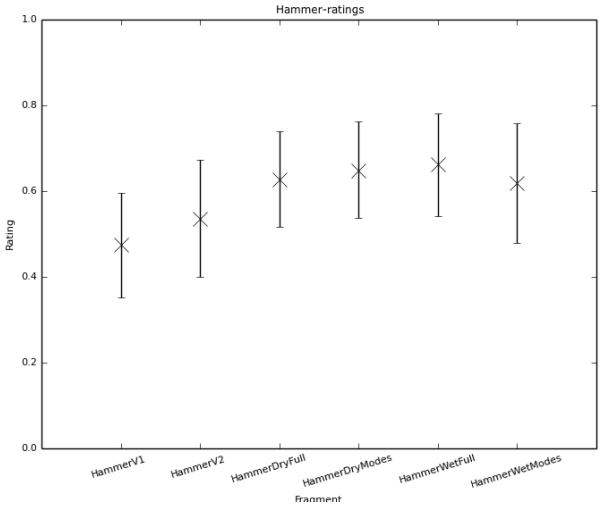

Figure 12: Ratings of Hammer

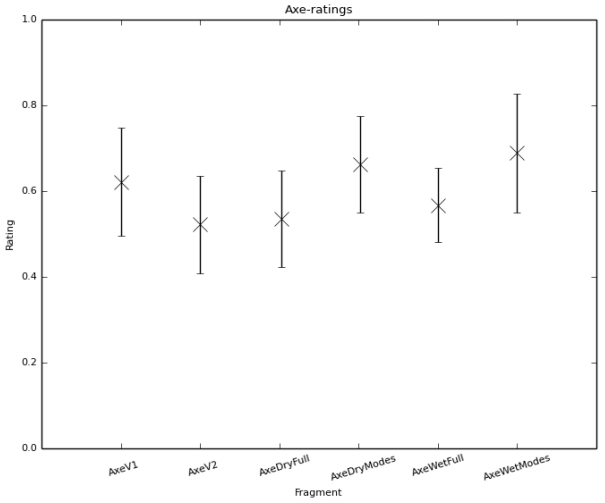

Figure 13: Ratings of Axe

In general, participants rated the wet (reverb and saturator processed effects) synthesized samples slightly higher than the dry versions. This is easily observed in the AK47, Rapier, and Axe plots (Fig. 7, 8, 11 and 13). This conveys two important points about perceptual preferences. First, we are more accustomed and tuned to hearing sounds in reverberant environments, and hence prefer some reverberation on any sound. Second, the two largest audio-visual industries where gunfire and combat weapon sounds can be found are in gaming and film, and in both of these cases, audio is entirely processed.

The art of audio post-production is key to a product´s (film and games) success, and it is no surprise that sound designers are aware of this. Therefore, even for sounds that we may only have heard in artificial settings, we expect it to be processed and may perceive the unprocessed dry version as unnatural and dull. This can also partly explain the low ratings in the gunfire tests, because the original sound samples were not complete dry signals, since they have a strong audible reverberation in the signal.

5 CONCLUSION

In conclusion, we presented a procedural audio model that synthesizes different impact-based sounds using spectral modelling synthesis. The SMS technique relied on a pre-processing analysis stage to collect the perceptual characteristics of sounds. The data from the analysis is mapped to different control parameters in our interface. Methods were taken to optimize in the CPU memory, to meet closely with gaming consoles and real-world performance requirements. Also, our model took incorporated post-processing audio effects such as reverb and distortion into a modal synthesis engine. The real-time synthesized output from both the stochastic and deterministic parts allows for an interactive performance system to be manageable by the user, and make manual tweaks or instantly play from stored presets. Subjective evaluation showed that in the case of metallic contact weapons, participants are not able to identify the difference between pre-recorded and synthesis generated sound samples.

Many of these control parameters required a manual setting. However, the interface model implements new variation techniques, like random frequency content in a sound sample, and time variation in the length of the sounds and between the frequency modes, with a similar approach to a granular synthesis implementation [16]. Another improvement in our model is the use of randomization in central frequency from the modes, borrowing the concept idea of the spectral peak-continuation algorithm [11].

It is, however, clear that resynthesis of a recording of a gunshot or impact sound is not enough. An overly processed sound with much more reverb that is generally plausible, is required to allow users to perceive the sound as realistic. This indicated the prevalence of hyperrealism, particularly within weapon sounds.

Further work suggests an automated analysis process, in particular for a frequency peak and high-quality amplitude envelope detection algorithms. Some other methods that are open for further research are the study of perceptual quality metrics to optimize the analysis process. Further study on the limits and impact of hyperrealism would also be beneficial to the understanding of the effects this has on sound design.

Full audio samples and data are available from https://code.soundsoftware.ac.uk/projects/modal-synthesis-of-weapon-sounds

REFERENCES

[1] A. Farnell, "Designing Sound: Procedural Audio Research," PhD, Escola des Artes, Universidade Catolica Portuguesa, 2012.

[2] N. Bottcher and S. Serafin, "Design and evaluation of physically inspired models of sound effects in computer games," in AES 35th International Conference: Audio for Games, London, 2009.

[3] Z. Ren, H. Yeh, and M. C. Lin, "Example-Guided Physically Based Modal Sound Synthesis," ACM Transactions on Graphics, vol. 32, Jan. 2013.

[4] C. Zheng and D. L. James, "Toward high-quality modal contact sound," ACM Transactions on Graphics - Proceedings of ACM SIGGRAPH vol. 30, 2011.

[5] N. Bonneel, G. Drettakis, et al, "Fast modal sounds with scalable frequency-domain synthesis," ACM Transactions on Graphics - Proceedings of ACM SIGGRAPH, vol. 27, 2008.

[6] P. Cook, Real sound synthesis for interactive applications: AK Peters Ltd., 2002.

[7] G. Durr, L. Peixoto, et al, "Implementation and evaluation of dynamic level of audio detail," in AES 56th International Conference, London, UK, 2015.

[8] S. Hendry and J. D. Reiss, "Physical Modeling and Synthesis of Motor Noise for Replication of a Sound Effects Library," in 129th AES Convention, San Francisco, 2010.

[9] G. Meurisse, P. Hanna, and S. Marchand, "A New Analysis Method for Sinusoids + Noise Spectral Models," in 9th Int. Conference on Digital Audio Effects, pp. 139-144.

[10] M. Klingbeil, "Software for Spectral Analysis, Editing and Synthesis," in International Computer Music Conference (ICMC), 2005.

[11] X. Serra and J. Smith, "A sound analysis/synthesis based on a deterministic plus stochastic decomposition," Computer Music Journal, vol. 14, pp. 12-24, 1990.

[12] J. A. Moorer, " About this reverberation business," Computer Music Journal, vol. 3, pp. 13-18, 1979.

[13] J. D. Reiss and A. McPherson, Audio Effects: Theory, Implementation and Application, CRC Press, 2014.

[14] N. Jillings, B. D. Man, et al, "Web Audio Evaluation Tool: A Browser-based Listening Test Environment," in 12th Sound and Music Computing Conference, 2015.

[15] B. D. Man and J. D. Reiss, "APE: Audio perceptual evaluation toolbox for MATLAB," 136th Convention of the Audio Engineering Society, April 2014.

[16] C. Picard, N. Tsingos, and F. Faure, "Retargetting Example Sounds to Interactive Physics-Driven Animations," in AES 35th International Conference, Audio in Games, London, 2009.