Queen Mary University of London

Evaluation of Synthesised Sound Effects

by

David James Moffat

A thesis submitted in partial fulfilment for the
degree of Doctor of Philosophy

in the

Centre for Digital Music

September 2019

Declaration of Authorship

I, David James Moffat, confirm that the research included within this thesis is my own work or that where it has been carried out in collaboration with, or supported by others, that this is duly acknowledged below and my contribution indicated. Previously published material is also acknowledged below.

I attest that I have exercised reasonable care to ensure that the work is original, and does not to the best of my knowledge break any UK law, infringe any third party's copyright or other Intellectual Property Right, or contain any confidential material.

I accept that the College has the right to use plagiarism detection software to check the electronic version of the thesis.

I confirm that this thesis has not been previously submitted for the award of a degree by this or any other university.

Signed: _____

Date: _____

Abstract

The current field of sound synthesis research presents a range of methods and approaches for synthesising a given sound. Sounds are synthesised to facilitate interaction or control of a sound, to enable sound searching through parametric control of a sound or to allow for the creation of an artificial nonexistent sound. In all of these cases, the ability of a synthesis technique to reproduce a desired sound is integral.

This thesis uses an audio feature representation of audio to produce a sonically inspired taxonomy, based entirely on the sonic content of sound, which enables a user to search through a large set of sounds without the need for understanding of context. This provides an approach for using audio features to compare similarity between different audio effect samples in a sound effects library. This thesis then develops approaches for evaluation of synthesised sound effects.

A large scale methodic subjective evaluation of synthesised sound effects is performed, evaluating a range of different synthesis methods in a range of different sound classes or sonic contexts. It is then identified that there are cases where synthesised sound effects can be considered as realistic as a recorded sample. An objective evaluation approach is then presented. Audio feature vectors are used to measure the relative objective similarities between two samples, and this is correlated with a perceptual evaluation of sound similarity. These objective measures are then compared based on the perceptual evaluations. Both evaluation approaches are then demonstrated in a case study of aeroacoustic sound effects, where these subjective and objective evaluation techniques are demonstrated for a specific case.

There is no single best approach to synthesising sound effects. More consistent and rigorous evaluation methodologies will lead to a better understanding as to the advantages and disadvantages of each method. The outcome of this research suggests that further consistent perceptual and objective evaluation within the sound effect synthesis community will lead to a better understanding as to the successes and failings of existing work and thus facilitate an enhancement of current sound synthesis technologies.

Acknowledgements

I would like to express my thanks to my supervisor Josh Reiss. His effort, dedication and commitment to all of my work, and enthusiasm for project, has kept me going far long after I wanted to give up. C4DM is a wonderful welcoming place were everyone is always supportive and ready to tell you how impressed they are with your work, so thanks to every single C4DM member. A special thanks to Giulio Moro, Elio Quinton, Dave Ronan and Rod Selfridge.

To the numerous people around C4DM, past and present, who have picked me up when I was down: Will, Adan, Thomas, Parham, Maria, Antonella, Chris, Florian, Alo, Delia, Adib, Sebastian, Ken, Marco, Matthias, Sarah, Sophie, Eddie, Thomas and many many more.

Friends keep you motivated on your worst day and celebrate with you on your best. Neale Dutton, though we never see each other enough, your advice, friendship and support has been incredible. Mike and Lucy, putting the world to rights, regardless of which pub we are in. Tuomas, the random distractions have been the best chance to escape. Stuart went through the painful process of reading and reviewing this thesis, for which I am exceedingly grateful.

It would not have been possible to complete this thesis without the continual support from my parents. Their support and encouragement has kept me going through it all.

Words cannot express the thanks due to my partner and teammate, Rebekah Stackhouse, without whom, I would have never even applied for a PhD. She has been the constant encouragement and calming voice throughout the entire process and there is no doubt in my mind that I would have never completed this thesis without her.

This work was supported by the EPSRC grant EP/M506394/1.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Motivation

Sound effects are commonly defined as non-musical, non-speech sounds used in some artificial context, such as theatre, TV, film, video games or virtual reality. The purpose of a sound effect is typically to provide diegetic context to some event or action, that is, a sound that exists within the narrative. Sound effects can often be the linchpin of a soundscape, and the use of sounds and different styles of sound will vary drastically depending on the medium's design and format, among other factors.

Sound synthesis is the technique of generating sound through artificial means, either in analogue or digital or a combination of the two. Synthesis of sound effects is typically performed for one of three reasons:

- to create something that does not exist as a recorded file, such as creating new artificial sci-fi sounds;
- to repair damaged sound files, where a segment of the sound file has been lost;
- to facilitate some interaction with or control of a sound, whether for a performance or direct parameter-driven manipulation of an auditory effect; or
- to facilitate a sound designer searching for a suitable sound within a synthesis space rather than through a sound effects library.

Within the context of this thesis, sound synthesis is considered to be the process of computer generation of audio. Synthesised sound effects can be applied to a range of sound design fields including film, TV, video games, virtual reality and augmented reality [Merer *et al.*, 2013].

A clear overview of synthesis research is presented by Misra and Cook [2009a]. The aims of current sound synthesis research include producing realistic sounding or controllable systems for artificially replicating real world sounds. The primary focus is either on implementation efficiency [Horner and Wun, 2006], interfacing and interaction control [Nordahl *et al.*, 2010] or producing accurate models of the physical environment [Bilbao and Chick, 2013]. Evaluation is vital, as it helps us understand both how well our synthesis method performs, and how the community can improve these systems. Without a rigorous evaluation method, one cannot understand if a synthesis method performs as required or where it fails. There are many different methods of evaluating a sound synthesis system. Examples of various evaluation methods being employed in literature, include: evaluation of controls and control parameters [Merer *et al.*, 2013; Rocchesso *et al.*, 2003; Selfridge *et al.*, 2017b]; human perception of different timbre [Aramaki *et al.*, 2012; Merer *et al.*, 2011]; sound identification [Ballas, 1993; McDermott and Simoncelli, 2011]; sonic classification [Gabrielli *et al.*, 2011; Hoffman and Cook, 2006a; Moffat *et al.*, 2017] and perceived sonic realism [Moffat and Reiss, 2018b; Selfridge *et al.*, 2018a, 2017c].

Fundamentally, these evaluation methods can be broken down into one of two categories: evaluation of sonic qualities and evaluation of human control and interaction.

1.1.1 Evaluation of Sonic Qualities

One of the fundamental metrics by which a synthesis method can be evaluated is the sonic quality of the sound produced. Does the synthesis method produce the desired sound? If not, then no quantity of sonic interaction will make a synthesis model effective. Generally, this evaluation needs to be performed with human participants, where recorded samples of a given sound can be compared, by those participants, to samples

rendered from a synthesis method, in a multi-stimulus subjective evaluation experiment [Bech and Zacharov, 2007; Moffat and Reiss, 2018b]. This comparison method will evaluate synthesised sounds against recordings in the same contextual environment, with suitable isolation from external acoustic environments. This method of evaluation can be applied to a range of different sounds [Mengual *et al.*, 2016; Selfridge *et al.*, 2018a, 2017a,b,c,d].

It is important that similar sounds are compared, and that participants are asked suitable questions. Generally participants are asked to evaluate how real, plausible or believable a given sound is. This is important as, although participants may have a strong idea of what a sound is, this does not mean that their impression of that real sound is correct. It has often been the case that a participant will rate a synthetic sound as 'more realistic' than a real recording of the same sound [Moffat and Reiss, 2018b], especially in less common sounds. This is due to the hyper-realism effect. As people are generally expecting explosions and gunshot sounds to be 'larger than life', when they hear a real recording compared to a synthesised sound, the recording seems less exciting than the synthesised sound [Mengual *et al.*, 2016; Puronas, 2014].

1.1.2 Evaluation of Human Control and Interaction

The other popular evaluation method focuses on control and interaction. Evaluating the control and interaction of a synthesis engine is a vital aspect of understanding in which environment the sound can be used. Much in the same way as Foley is the performance of 'analog' sounds, synthesis is the performance of digital sounds, where the primary focus is on the control of the parameter interaction. However, in most cases, the physical interaction that creates the sound will not be suitable for directly driving the individual synthesis parameters and, as such, some mapping layer for parameters and physical properties will be required [Heinrichs and McPherson, 2014; Heinrichs *et al.*, 2014], such as a hardware sensor or a game engine physical parameter. There are numerous methods for evaluating these sonic interactions and, in many cases, the control evaluation has to be a bespoke design for the synthesis methods and parametric controls [Heinrichs and McPherson, 2014; Heinrichs *et al.*, 2014; Selfridge *et al.*, 2017b; Turchet *et al.*, 2016].

User listening tests, where participants are able to interact with the synthesis engine through some mapping layer, can be performed to evaluate a series of criteria. Key aspects of synthesis control systems to evaluate, as defined by Heinrichs and McPherson [2014], are whether it is:

- **intuitive** - Are the controls adjusting the sound in a way that a user would expect and understand?
- **perceptible** - To what extent can someone identify the change each control makes, at all times.
- **consistent** - Does a given control value always produce the same modification or reproduction of sound, or is there some control hysteresis?
- **reactive** - Do the controls immediately change the sound output, or is there a delay on control parameters, impacting the ease of usability? Typically 20ms of latency is acceptable, in most cases, so long as the latency is consistent [Jack *et al.*, 2018].

1.1.3 Sound Effect Searching

The ability to search through different perceivable dimensions of sound, such as pitch and timbre, is an advantage of using sound synthesis approaches, and is an important aspect for sound design. Sound design is the practice of constructing and controlling sonic elements creatively. This is usually done to tell a story, evoke a particular emotion or mood, or to emphasise a non-auditory element of the context or scene to happen simultaneously. A typical sound design process will involve combining a number of pre-recorded elements with bespoke recordings in an aesthetically pleasing manner. In the case of film, this could be a number of sound effects taken from a library, mixed with some specific recorded sound effects, vocal tracks, and Foley recordings [Farnell, 2010; Sonnenschein, 2001]. Sound effects can often be the linchpin of a sound scene, and different sounds and styles will vary drastically dependent on the style and design of the medium, among other factors [BBC, 1931; Tremblay *et al.*, 2000]. Sound designers will

regularly record their own specific sounds for a given project, but will often incorporate pre-recorded sounds from a library. As such, sound effects libraries typically consist of an individual's personal recordings and large commercial royalty free audio collections, such as the BBC sound effects library [BBC, 2011]. Searching through these large collections of sound recording can often be a challenging task, and typically the metadata associated can hold great importance in the ability to find and select appropriate audio recordings [Cano *et al.*, 2004]. Metadata could consist of sound effect description, or appropriate words to describe the sonic properties of the sound [Turnbull *et al.*, 2008]. There have also been numerous approaches in attempting to support the searching of sound effects without the use of supportive metadata [Black *et al.*, 2009; Pearce *et al.*, 2017; Rice and Bailey, 2004; Wold *et al.*, 1996].

The rest of this chapter will outline the principal aims and structure of this thesis. The primary contributions are outlined, and associated publications are identified.

1.2 Research Objectives

This thesis aims to develop the current state of evaluation within the field of sound effect synthesis. Within this thesis, the following research questions are addressed:

- What sort of sound effect taxonomy can be produced, based on the sonic content of sound samples? *(Chapter 4)*
- To what extent can unsupervised learning be used to produce an objective similarity measure of synthesised sound effects? *(Chapter 4)*
- Which synthesis methods are best able to synthesise a given sound? *(Chapter 5)*
- Which objective similarity metrics can be used to evaluate synthesised sound effects, through comparison to a reference audio sample? *(Chapter 6)*
- Does a physical approach of modelling aeolian tones produce a more plausible sound than other existing synthesis methods? *(Chapter 7)*

1.3 Contributions

The contributions of this thesis are the sonic comparison and evaluation of synthesised sound effects. Utilising a review of existing audio analysis tools, presented in Chapter 3, an unsupervised machine learning approach is used, in Chapter 4, to generate a sound effects taxonomy based purely on the sonic elements of the sound effect. It was identified that the repetitive nature and dynamic range of sounds are some of the best attributes to split sound effects into grouping, given the specific dataset used. This work proposes an approach for a fully sonically inspired hierarchical classification of sound effects, which may help with the ability to search for appropriate sounds. Further to this, the use of a new audio feature set for measuring the similarity between sounds was proposed, based on the loudness, pitch, periodicity, timbral, envelope and spectral contrast attributes of a sound. Chapter 5 presents a large scale investigation into the perception of synthesised sound effects. Six different synthesis approaches are used to produce eight different sound effects, of different sound classes. A subjective evaluation is performed to determine the perceived realism of each synthesis method, in the case of each sound effect. It is identified that there are cases where synthesised sound effects can be considered to be as realistic as a recorded sample. As such, it can be determined that real-time sound synthesis can be appropriate for replacing a number of sound samples in some sound design cases. There are also many cases where synthesis is not as realistic as a recorded sample. These cases are identified and potential justifications are proposed. As such, this new understanding as to the state-of-the-art of synthesis can be used to develop new synthesis approaches, through identifying where more work is required. Chapter 6 uses the sonic similarity measure proposed in Chapter 4, and evaluates this through comparison to a number of other similarity measures, utilising an evaluation method derived from Chapter 5. A set of six different objective evaluation metrics, used within the literature, are presented and evaluated through synthesis optimisation. This identifies that the Wichern objective similarity measure performed best, and is the only method to be statistically significant and correlate highly with human perception. The other five methods did not correlate with human perception. The Wichern similarity measure is therefore shown to provide an effective audio feature representation for sound

effect similarity. Chapter 7 demonstrates the range of this work, through performing a number of subjective evaluations on a range of different aeroacoustic sound effects. The plausibility of each of the sound effects is presented, and the ability to use a standard evaluation framework, with modifications to each of the specific requirements of the sound case is shown. Furthermore, the objective evaluation metrics presented in Chapter 6 are used to demonstrate their ability to correlate with human perception, and the cases where this method fails is identified.

1.4 Thesis Structure

Chapter 1 – Introduction

This chapter introduces the structure of the thesis, identifies the research questions addressed and states the contributions of this thesis.

Chapter 2 – Background

This chapter briefly introduced a range of sound synthesis approaches and presents a summary of existing synthesis methods. A review of existing sound synthesis evaluation approaches is performed. The background of evaluation within the field of sound effect synthesis is provided, and an overview of existing systems of evaluation, including subjective evaluation and objective evaluation. This systematic review demonstrates the lack of formalised and rigorous evaluation within the sound effect synthesis field.

Chapter 3 – Audio Feature Extraction Toolboxes

This chapter discusses a set of tools used within this thesis. An evaluation of existing audio feature extraction libraries was undertaken. Ten libraries and toolboxes were evaluated using the Cranfield Model for evaluation of information retrieval systems, to review the coverage, effort, presentation and time lag of a system. Comparisons between these tools were undertaken and example use cases are presented as to when toolboxes are most suitable.

Chapter 4 – Taxonomy of Sound Effects

This chapter produces a taxonomy of sound effects based on the sonic content of sounds, rather than the human interpretation of them, through the use of unsupervised machine learning. This approach further identifies the audio features most relevant for classifying sound effects.

Chapter 5 – Subjective Evaluation of Synthesised Sound Effects

This chapter presents a formal listening experiment . Existing research, as presented in Chapter 2 is reimplemented and evaluated, and subjective evaluation is used to assess the perceived 'realism' of different sound synthesis methods under various conditions. This detailed subjective evaluation provides insights which suggest the further research directions. It is found, among other things, that the performance of synthesis methods is highly dependent on the type of sound being synthesised. Subjective evaluation is used in this case to refer to a human subject performing an evaluation.

Chapter 6 – Objective Evaluation Metric for Synthesised Environmental Sounds

This chapter discusses a range of different objective methods (computational or automated evaluation processes) for comparing or measuring the similarity between environmental sound effects. These methods are currently used as objective evaluation techniques to measure the effectiveness of a sound synthesis method by assessing the similarity between synthesised and recorded samples. Evaluation is performed on a number of different synthesis objective evaluation metrics by using the different distance metrics as fitness functions within a resynthesis algorithm. The recorded samples are excerpts from a sound effects library, and the results are evaluated through a subjective listening test. Results show that one objective function performs significantly worse than several others. Only one method exhibited a significant and strong correlation between the user subjective similarity and the objective distance. A recommendation is made, for

a perceptually motivated evaluation function (that is, relating to our understanding of human perception) for measuring similarity between synthesised environmental sounds.

Chapter 7 – Case Study: Evaluation of Aeroacoustic Sound Effects

This chapter presents a case study that was undertaken in the field of synthesising aeroacoustic sounds. It builds on our understanding to demonstrate a rigorous evaluation of a new sound synthesis approach, and integrates the subjective and objective evaluation approaches into improved synthesis design. This further demonstrates the importance of evaluation, and how it benefits even physical modelling, where the model and its parameters are not derived directly from sample analysis or prior knowledge of perception.

Chapter 8 – Conclusion

This chapter draws conclusions across the field of sound effect synthesis and the importance of evaluation to the field. It identifies directions for new research and focus.

1.5 Related publications

This section presents work considered part of this thesis that has previously been published, with an indication to what chapter the work appears in. If the author is not the primary author of the paper, the author's contributions to the paper are detailed.

Chapter 2

Moffat, D., Selfridge, R., and Reiss, J. D. (2019). Sound effect synthesis. In Filimowicz, M., editor, *Foundations in Sound Design for Interactive Media: A Multidisciplinary Approach*. Routledge

Chapter 3

Moffat, D., Ronan, D., and Reiss, J. D. (2015). An evaluation of audio feature extraction toolboxes. In *Proc. 18th International Conference on Digital Audio Effects (DAFx-15)*

Chapter 4

Moffat, D., Ronan, D., and Reiss, J. D. (2017). Unsupervised taxonomy of sound effects. In *Proc. 20th International Conference on Digital Audio Effects (DAFx-17)*, Edinburgh, UK

Ronan, D., Moffat, D., Gunes, H., and Reiss, J. D. (2015). Automatic subgrouping of multitrack audio. In *Proc. 18th International Conference on Digital Audio Effects (DAFx-15)*. DAFx-15

The author supported initial concepts, aided in data analysis, edited text.

Chapter 5

Moffat, D. and Reiss, J. D. (2018b). Perceptual evaluation of synthesized sound effects. *ACM Transactions on Applied Perception (TAP)*, 15(2):19

Chapter 6

Moffat, D. and Reiss, J. D. (2018a). Objective evaluations of synthesised environmental sounds. In *Proc. 21th International Conference on Digital Audio Effects (DAFx-17)*, Aveiro, Portugal

Chapter 7

Selfridge, R., Moffat, D., Avital, E. J., and Reiss, J. D. (2018a). Creating real-time aeroacoustic sound effects using physically informed models. *Journal of the Audio Engineering Society*, 66(7/8):594–607

The author developed evaluation methodology, performed evaluations, analysed results and edited text.

Selfridge, R., Moffat, D., and Reiss, J. D. (2017c). Sound synthesis of objects swinging through air using physical models. *Applied Sciences*, 7(11)

The author developed evaluation methodology, performed evaluations, analysed results and edited text.

Selfridge, R., Moffat, D., and Reiss, J. D. (2017a). Physically inspired sound synthesis model of a propeller. In *ACM Audio Mostly Conference*, London, UK

The author developed evaluation methodology, performed evaluations, analysed results and edited text.

Selfridge, R., Moffat, D., and Reiss, J. D. (2017b). Real-time physical model for synthesis of sword swing sounds. In *International Conference on Sound and Music Computing (SMC)*, Espoo, Finland

The author developed evaluation methodology, performed evaluations, analysed results and edited text.

Selfridge, R., Moffat, D., Reiss, J. D., and Avital, E. J. (2017d). Real-time physical model for an aeolian harp. In *International Congress on Sound and Vibration*, London, UK

The author developed evaluation methodology, performed evaluations, analysed results and edited text.

Additional Papers

The following is a list of papers which do not directly relate to specific chapters, but generally contribute to the research performed throughout this thesis.

Jillings, N., Moffat, D., De Man, B., and Reiss, J. D. (2016). Web audio evaluation tool: A framework for subjective assessment of audio. In *Proc. 2nd Web Audio Conference*, Atlanta, Georgia, USA

The author produced initial set of evaluation frameworks, provided software implementations, edited text and supported users.

Jillings, N., De Man, B., Moffat, D., Reiss, J. D., and Stables, R. (2015). Web audio evaluation tool: A browser-based listening test environment. In *Proceedings of the International Sound and Music Computing Conference*, Maynooth, Ireland

The author produced initial set of evaluation frameworks, provided software implementations, edited text and supported users.

Turchet, L., Moffat, D., Tajadura-Jiménez, A., Reiss, J. D., and Stockman, T. (2016). What do your footsteps sound like? an investigation on interactive footstep sounds adjustment. *Applied Acoustics*, 111:77–85

The author aided in experimental design, evaluated, collated and analysed results and edited text.

Mengual, L., Moffat, D., and Reiss, J. D. (2016). Modal synthesis of weapon sounds. In *Proc. Audio Engineering Society Conference: 61st International Conference: Audio for Games*, London. Audio Engineering Society

The author created evaluation methodology, performed evaluations, analysed results, edited text and presented the work

Chapter 2

Background

2.1 Introduction

This chapter presents a summary of synthesis methods, particularly relating to work evaluated in Chapter 5. The breadth of evaluation methodologies that are often undertaken within sound effect synthesis research is presented, and the key failings within the field identified. It will be shown that there is limited consistency or rigour in the current methods for evaluating sound effect synthesis, therefore justifying the further research presented within this thesis. A review of existing sound effects is presented, which will be extended in Chapter 4.

2.2 Sound Effect Synthesis

There are many methods and techniques for synthesising different sound effects, and each one has varying advantages and disadvantages. There are almost as many sound synthesis classification methods, but the most prominent was produced by Smith [1991]. Sound synthesis can generally be separated into the following categories: sample based; signal modelling; abstract; and physical modelling synthesis [Smith, 1991].

2.2.1 Sample Based Synthesis

In sample based synthesis, audio recordings are cut and spliced together to produce new or similar sounds. This is effective for pulse-train or granular sound textures, based on a given sound timbre.

The most common example of this is granular synthesis. Granular synthesis is the method of analysing a sound file or set of sound files and extracting sonic 'grains'. A sound grain is generally a small element or component of a sound, typically between 10-200ms in length. Once a set of sound grains have been extracted, they can then be reconstructed and played back with components of the sound modified, such as selecting a subset of grains for a different timbre, or modifying the grain density, or rate, to change the pitched qualities of the sound.

2.2.1.1 Concatenative Synthesis

Concatenative synthesis is a subset of granular synthesis and a form of sample based synthesis. Segments or 'grains' are made from small segments of sound samples. Grains can range from 10ms to 1s samples of audio. Concatenative synthesis is the process of selecting and recombining the grains together, in such a manner that it does not create any perceivable discontinuities.

For the purpose of the work in Chapter 5, a library of 46ms audio grains was constructed, selected at 1.5ms intervals from the samples. Grain selection from the library was performed using a time domain probabilistic method. Given the current grain, a subset library of grains was selected based on the Spearman correlation distance of the time domain waveform signal, such that

$$d_t = 1 - \frac{(v_r - \bar{v}_r)(v_t - \bar{v}_t)'}{\sqrt{(v_r - \bar{v}_r)(v_r - \bar{v}_r)'}\sqrt{(v_t - \bar{v}_t)(v_t - \bar{v}_t)'}} \tag{2.1}$$

where v is a coordinate-wise vector of either the current grain r or query grain t, for which the distance is calculated. $\bar{v}$ denotes the mean of the vector, to normalise the vector around its current mean. The Spearman distance was used, as it considers the

sample vector in sequence, so small variations in sample do not result in a significant overall difference. The time domain vector to represent each grain is taken as the second half (23ms) of the current grain, and the first half of the grain within the grain library. The time domain waveform vectors of the current grain and all possible grains were selected.

From this calculated subset library of possible grains, one grain g_t is selected with probability

$$P(g_t) = \frac{1 - d_t}{\sum_{k=1}^{K} d_k} \tag{2.2}$$

where d_t is the Spearman distance from the current input grain and K is the number of selected nearest neighbours, in this case 10. The selected grain is then overlapped with the current audio grain, and the two audio samples crossfaded.

Due to the lack of available open source implementations of concatenate synthesis, this synthesis method was implemented by the authors, based on O'Leary and Robel [2014]. The implementation is available to download.[1]

2.2.2 Signal Modelling Synthesis

Signal modelling synthesis is the method where sounds are created based on some analysis of real world sounds, and then attempting to resynthesise the waveform sound, not the underlying physical system. The premise of signal modelling, is that through comparing and reproducing the actual sound components, one can extrapolate the control parameters and accurately model the synthesis system. The most common method of signal modelling synthesis is Spectral Modelling Synthesis (SMS) [Serra and Smith, 1990]. SMS assumes that sounds can be synthesised as a summation of sine waves and a filtered noise component. Spectral modelling is often performed by analysing the original audio file, selecting a series of sine waves to be used for resynthesis, and then creating some 'residual' noise shape, which can be summed together to produce the original sound [Amatriain *et al.*, 2002].

[1] `https://code.soundsoftware.ac.uk/projects/time-domain-probabilistic-concatenative-synthesis`

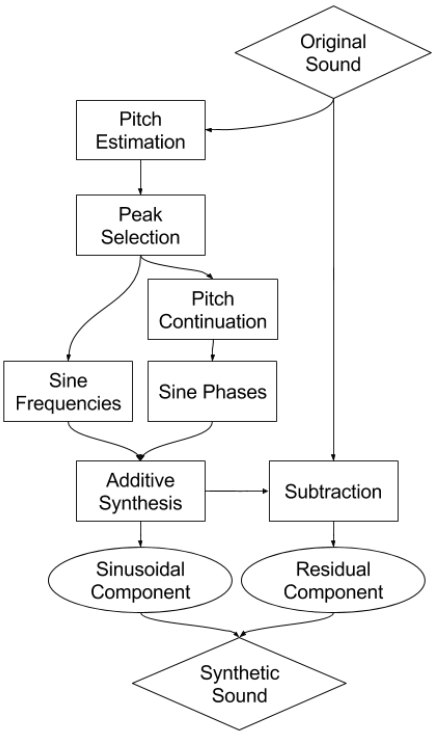

FIGURE 2.1: Flow diagram of sinusoidal modelling, based on Serra and Smith [1990].

2.2.2.1 Sinusoidal Modelling

Sinusoidal Modeling Synthesis or Spectral Modeling Synthesis (SMS) [Serra and Smith, 1990] is an example of signal modelling synthesis. Sinusoidal modelling assumes that sounds can be synthesised as a summation of sine waves and a filtered noise component such that any sound $x(t)$ can be represented as

$$x(t) = \sum_{r=1}^{R} A_r(t) \sin\left(\theta_r(t)\right) + e(t) \tag{2.3}$$

where $x(t)$ is a summation of R sinusoids, A_r and θ_r are the amplitude and phase, respectively, of a given sinusoid at time t and $e(t)$ is the noise component, referred to as the residual.

As presented in Figure 2.1, sinusoidal modelling is performed by peak selection from the frequency spectra. These peaks are resynthesised using sine waves. The output sine waves are summed together and the residual is calculated as the result of subtracting the summation of sine waves from the initial sound signal. The synthesis method evaluated was based on the documentation and implementation from Serra and Smith [1990]

and Amatriain *et al.* [2002]. SMS best performs on simple harmonic sounds, such as a metal impact sound [Mengual *et al.*, 2016; Van Den Doel *et al.*, 2001; Zheng and James, 2011].

2.2.2.2 Statistical Modelling and Marginal Statistics

Statistical modelling is a signal modelling synthesis technique where an input sound file is decomposed into a set of summary statistics. These statistics are used to shape an input noise signal, and resynthesise the input audio file. The extracted statistics are based on perceptual models of audio signals. Statistics of the sound are calculated from an auditory inspired cochlear filter bank representation of the signal.

There are two different use cases presented using this algorithm, one is described as Marginal Statistics and the other as Statistical Modelling. They both take the same form, but use a different set of statistics to represent the audio file. Marginal statistics are the mean, variance, skew and kurtosis of the subband envelope and modulation power, extracted from the filtered signal representation. Statistical Modelling includes all the statistics of the marginal statistics and includes the cross-sub-band envelope correlation and cross-sub-band modulation correlations. Full mathematical descriptions are presented in McDermott and Simoncelli [2011].

Sounds were resynthesised from the set of chosen statistics, through an iterative process of shaping Gaussian white noise, as can be seen in Figures 2.2 and 2.3. For the purposes of evaluation, the synthesis method, documentation and implementation were taken from McDermott and Simoncelli [2011].

2.2.2.3 Additive Synthesis

Traditionally additive synthesis was a form of signal based modelling where a series of sine waves were added together to produce complex waveform. This technique was further developed and became sinusoidal modelling, as discussed in Section 2.2.2.1. Additive synthesis has since become the process of modelling sounds as a summation of synthesised audio signals, such as noise signals, sinusoids and chirp sounds.

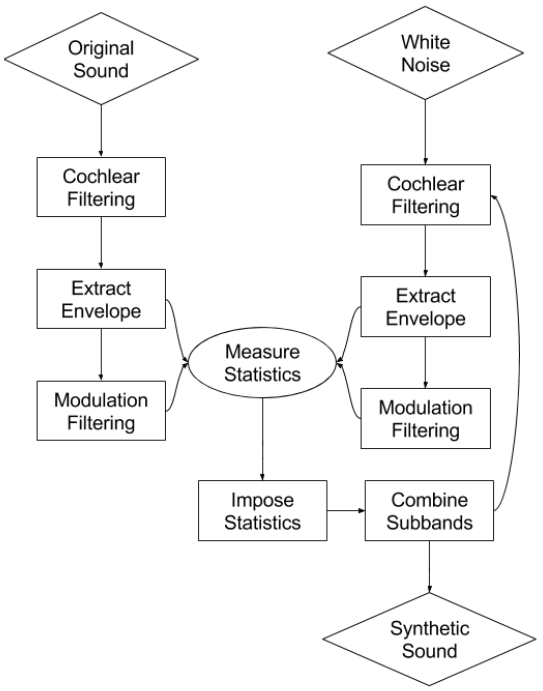

FIGURE 2.2: Flow diagram for Statistical Synthesis, based on McDermott and Simoncelli [2011].

1. Noise signal is filtered into cochlear subbands
2. Calculate Subband envelopes
3. Sub bands of envelopes are calculated, to represent detail
4. Down sample envelopes
5. Calculate envelope statistics
6. Calculate error by comparing statistics to the original signal
7. Envelopes are modified using gradient descent method
8. Envelopes are upsampled
9. Envelopes applied to input noise signal subbands
10. Sub bands combines to create synthesis signal

FIGURE 2.3: The iterative process undertaken to perform Statistical Synthesis

For the purposes of evaluation, performed within Chapter 5, the Spatialized Additive Synthesizer for Environmental Sounds (SPAD) from Verron *et al.* [2010a] was used. SPAD works on the principle of breaking every sound into one of five core sound elements, or atoms, and synthesising each sound as one of these core elements. Elements are synthesised as per Table 2.1. All synthesis of atoms occurs in the time domain, apart from band-limited noise, which is synthesised in the frequency domain.

2.2.3 Abstract Synthesis

Sounds are created from abstract methods and algorithms, typically to create entirely new sounds. A classic example of abstract synthesis is Frequency Modulation (FM) Synthesis [Chowning, 1973]. FM synthesis is a method derived from telecommunications. Two sine waves are multiplied together to create a much richer sound. These sounds can be controlled in real-time, as computation is low, to create a set of sounds that do not exist in the natural world. A lot of traditional video game sounds and 1980's keyboard sounds were based on FM synthesis.

TABLE 2.1: The five classes of atoms used for the sound synthesis models, with their respective synthesis equations and parameters, as taken from Verron *et al.* [2010a].

Atom	Equation	Parameters
Modal impact	$x_1(t) = \sum_{m=1}^{M} a_m \sin(2\pi f_m t) e^{-\alpha_m t}$	a_m initial amplitudes, α_m decays, f_m frequencies
Noisy impact	$x_2(t) = \sum_{n=1}^{N} a_n s_n(t) e^{-\alpha_n t}$	a_n subband amplitudes, α_n subband decays
Chirped impact	$x_3(t) = a \sin(2\pi \int_0^t f(\nu)\, d\nu) e^{-\alpha t}$	f_0 initial frequency, σ linear frequency shift, α decay
Band-limited noise	$X_4(f) = \begin{cases} A(t), & \text{if} \lvert f - F(t) \rvert < \frac{B(t)}{2} \\ A(t) e^{-\alpha(t)\left(\lvert f - F(t) \rvert - \frac{B(t)}{2}\right)}, & \text{otherwise} \end{cases}$	$F(t)$ center frequency, $B(t)$ bandwidth, $\alpha(t)$ filter slope, $A(t)$ amplitude
Equalized noise	$x_5(t) = \sum_{n=1}^{32} a_n(t) s_n(t)$	$[a_1(t)...a_{32}(t)]$ amplitudes

$s_n(t)$ represents subband filtered noise, band n at time step t. $x(t)$ represents a time domain signal, whereas $X(f)$ represents a frequency domain signal. $x_1...x_3$ and x_5 are calculated in the time domain, whereas X_4 is calculated in the frequency domain.

2.2.4 Physical Modelling Synthesis

Sounds are generated based on modelling of the physics of the system that created the sound. The more physics is incorporated into the system, the better the model is considered to be, however the models often end up very computationally intensive and can take a long time to run. Despite the computational nature of these approaches, with GPU and accelerated computing, physical models are beginning to be capable of running in real-time [Harrison-Harsley and Bilbao, 2018; Webb and Bilbao, 2015]. As such, physical models are based on fundamental physical properties of a system and solving partial differential equations at each step sample [Bilbao, 2009].

2.2.4.1 Physically Inspired Synthesis

Physically inspired synthesis [Cook, 2007] is derived from physical modelling. It is possible to construct synthesis systems by modelling the entire physical environment in which

the sound was created, but this can be incredibly complex to construct. Physically inspired or physically informed synthesis is considered as another form of signal based modelling or as a hybrid approach between signal based modelling and physical modelling, where the user controls represent the physics of the system, but the calculations are all approximations to allow the system to run in realtime. Sounds are constructed as a combination of base units, such as filtered noise, sine, triangle and square waves, envelope shapes and filters.

Producing a physically inspired synthesis simulation or model of a sonic context is considered a time consuming process. Each individual sound synthesis model needs to be manually constructed with knowledge of the physics, understanding of psychoacoustics and experience in sound synthesis model production and workflows. Despite the labor intensive nature, physically inspired synthesis is an effective and flexible method of sound synthesis, as once a context has been modelled, it is possible to vary a large range of parameters to create very different sounding environments, with physically and perceptually relevant interface controls.

For the purposes of evaluation performed within Chapter 5, a number of synthesis models were taken from Farnell [2010] and Peltola *et al.* [2007].

2.2.5 Synthesis Methods Summary

There are a range of different synthesis methods, that can produce a range of different sounds. From abstract synthesis techniques that are lightweight and can be implemented on old 80's hardware, to physical modelling techniques that require optimisation and GPU and even still, are only just able to operate in real-time. Each approach and methodology has its advantages and disadvantages. Misra and Cook [2009b] perform a rigorous survey of synthesis methods, and recommend different synthesis techniques for each type of sound to be produced. Abstract synthesis is highly effective for producing artificial sounds, 'retro' style synthetic sounds and some musical sounds. Signal modelling can produce excellent voiced sounds and environmental sounds. Physical models are great for impact or force driven sounds, such as the pluck of a string, whereas sound textures and environmental sounds are often best produced by sample based models.

Sound Type	**Synthesis Method**
Sci-Fi / Technology Sounds	Abstract Synthesis
Environmental Sounds	Sample Based Model / Signal Models
Impact Sounds	Physical Models / Signal Models
Voiced Sounds	Signal Models
Sound Textures / Soundscapes	Sample Based Models

TABLE 2.2: Recommendation of Synthesis Method for Each Sound Type

A summary of recommendations as to a method of synthesis that would work for each type of sound class can be found in Table 2.2. A large collection of these synthesised sound effects is presented on the FXive website,[2] developed by Queen Mary University of London [Bahadoran *et al.*, 2017, 2018a,b].

2.3 Synthesis Evaluation

The aims of sound synthesis, as noted above, are to produce realistic and controllable systems for artificially replicating real world sounds. Current research generally focuses on either implementation efficiency, interfacing control or physical modelling, and provides very limited evaluation. Subjective evaluation, performed with participants, is occasionally used in current sound synthesis research, and there is no consistency in objective evaluation metrics used.

Schwarz [2011] noted in a review of 94 published papers on sound texture synthesis that only seven contained any subjective evaluation of the synthesis method. Jaffe [1995] presents ten criteria for evaluating sound synthesis: five based on parameter control, three on computation of the synthesis method and two on the sonic qualities of the synthesis method. Tolonen *et al.* [1998] used this framework for evaluation to produce a rigorous review of a range of synthesis methods. Despite this work, these criteria and frameworks are not commonly used and there is no consistently used standard process for evaluating the subjective realism of sound synthesis.

The aim of this section is to identify and summarise the diversity of evaluation methods used for different synthesis approaches undertaken in literature.

[2] `http://fxive.com`

2.3.1 Subjective Evaluation

Bonebright *et al.* [2005] discussed three different methods for determining subjective qualities of audio: identification testing, context-based rating and attribute rating. Identification testing is an approach where sounds are rated based on whether a participant can identify the intended source of the sound. Context-based rating is where participants score how close a sound is to a provided context, such as a visual scene or text phrase. Attribute rating is where participants are asked to rate a set of given attributes of a sound, such as loudness, pitch, roughness or pleasantness. Bonebright *et al.* [2005] stated that context-based rating is appropriate for sound in video games and sound synthesis, whereas Merer *et al.* [2011], proposed an attribute rating is most appropriate for synthesised abstract sounds, as they believe abstract sounds are not easily associated with a source.

There are a range of different subjective test methodologies, and some work will use multiple different methodologies for evaluation. McDermott and Simoncelli [2011] performed an identification task where participants needed to pick the correct word describing the sound, from a set of five words. McDermott and Simoncelli [2011] then performed a context test, where an original recorded sound was played as a reference and participants had to select which of two different synthesised sounds were most similar to the reference. Participants were then asked to provide a rating for realism on a scale of one to seven for various different synthesised and recorded sounds. No formal anchors were identified as the lower bounds for sound quality. This is problematic, as Bech and Zacharov [2007] state that, with no lower reference against which to compare the effectiveness or realism of each sound effect, each participant's scores are based on an arbitrary scale. Given a fixed lower point, which is known to be a poor quality sample, the presentation of results makes it clear as to what should be considered reasonable or poor results.

However, most subjective evaluation takes the form of a single test. An evaluation of concatenative synthesis methods was performed via an online MUSHRA (MUltiple Stimulus Hidden Reference and Anchor [ITU-R BS.1534-3, 2015]) style listening test, in which participants rated both the quality of samples and their similarity to a reference

sample [Schwarz and O'Leary, 2015; Schwarz *et al.*, 2016]. Their sample order was not randomised, so potential ordering bias may be an issue, and no recording of the participants' listening conditions was made. The authors concluded that all concatenative synthesis methods are indistinguishable from each other, in terms of both the perceived quality and realism of the sound produced, and thus no method was distinguishable from random. A similar evaluation methodology was used by Mengual *et al.* [2016], in evaluating gunshot and metal impact sounds, with order randomisation to remove bias, and under controlled listening conditions. Adami *et al.* [2017] used a MUSHRA test to rate density of applause sounds and the similarity of a sample to their synthesised sounds.

Bonneel *et al.* [2008] performed a MUS (MUltiple Stimulus) style test to evaluate the synthesis quality, though no references or anchors were included. Henter *et al.* [2014] asked participants to rate the 'naturalness', through a MUSHRA test and Fröjd and Horner [2009] instead used a fixed point scale to evaluate naturalness of a sound texture granular synthesis approach. Moss *et al.* [2010] evaluated a physical approach to producing liquid sounds using a ten point fixed scale, rating how realistic a sound is. Nordahl *et al.* [2010] performed evaluation of footstep sounds, where recognition of surface material and subjective realism were reported on a seven point scale. However, neither samples nor an alternative synthesis approach was used for comparison.

Selfridge *et al.* [2017b] evaluated synthesis of sword swing sounds using a similar structure: comparing to multiple different synthesis methods, recorded samples and a specific anchor. Murphy *et al.* [2008] performed an attribute test in which participants were asked to rate the quality of 'rollingness' of synthesised rolling sounds in a MUS style test, but no alternative synthesis methods, samples or hidden anchors were provided for comparison. In Rocchesso and Fontana [2003] participants were asked to browse through a range of synthesised sounds to find their preferred sound, and then asked to rate the perceived realism on a seven point Likert scale. 'Perceived realism' was also the evaluation criterion in Böttcher and Serafin [2009], on a five point scale and McDermott and Simoncelli [2011], on a seven point fixed scale.

Gabrielli *et al.* [2011] proposed an alternative form of evaluation of synthesis, an 'RS Test', where participants are played a single sound only once and had to determine if it was real or synthetic. It is then possible to use this approach to test a large number of audio samples, however it is important to include an anchor. Hahn [2015] evaluated musical instrument sounds using the RS test, and Hoskinson [2002] modified this evaluation framework using pairs of real and synthesised sounds. Hoskinson [2002] setup a paired test, where participants were played two sounds, one real and one synthesised, each sound was played only once and the participant had to identify which was the real sample. To evaluate instrument synthesis, Järveläinen *et al.* [2002] asked participants to match real and synthesised equivalents of samples together based on their harmonic components.

It is vital that the manner in which an individual interacts with a synthesis engine is understood, and there is considerable research on this. As part of the Sounding Object project, a large body of work was undertaken in sound effect synthesis, primarily focusing on interactions with sound synthesis models [Rocchesso *et al.*, 2003]. Böttcher and Serafin [2009] evaluated the perceived quality of an interaction with a synthesis engine, and this work was further developed in Böttcher *et al.* [2013]. However, this type of work often measures the parameter mapping more than the quality of the sound synthesis [Heinrichs and McPherson, 2014; Heinrichs *et al.*, 2014]. Hoffman and Cook [2006b] discussed the generalised process of synthesis parameter mapping to perceptual controls through feature vector mapping. Aramaki *et al.* [2012] presented other methods for mapping physical controls of a synthesis engine to perceptual parameters.

Scavone *et al.* [2001] created a program for presenting sound effects on a 2D plane using multi-dimensional scaling (MDS), while Lakatos *et al.* [1997] asked participants if they could identify the material dimensions of an impact sound in a two-alternatives forced choice experiment, and then employed MDS on the results. Aramaki *et al.* [2011] performed further synthesis evaluation through a forced choice ABCX test, in which participants are required to select one of three categories for a reference sound, and EEG (Electroencephalogram) recordings were taken. Kersten and Purwins [2010] evaluated sound textures through subjective sound categorisation, and Ma *et al.* [2010] played

samples to participants, who were asked to write free text response descriptions of the sound sample.

As can be seen from the above examples, whilst many studies use some form of evaluation, it is demonstrably diverse and intermittent, when it is employed at all.

2.3.2 No Evaluation

There are a number of reports of synthesis methods where no evaluation was performed. Concatenative synthesis is performed in An *et al.* [2012]; Schwarz [2000]; Schwarz *et al.* [2006]. Serra and Smith [1990] introduce the SMS method. Goodwin [1996] proposed a modification to include the residual element of the signal, and Pampin [2004] extend the model to include masking models, and no analysis of the output sounds were presented. Rath [2003]; Van Den Doel *et al.* [2001]; Van den Doel and Pai [2003] all performed modal synthesis approaches, where harmonic modes of materials are identified. This approach is often described as interpreting some physical properties of the material, and synthesising that based on harmonic resonances. Physically based modal synthesis approaches are presented by Castagné and Cadoz [2000]; Karjalainen *et al.* [1993], and the computational approaches by Peltola *et al.* [2007]; Zita [2003]. Lee *et al.* [2010],Kahrs and Avanzini [2001] and Hahn and Röbel [2013] all demonstrate further source filter synthesis approaches that relate filter coefficients to physical properties, but not the output result of each of the synthesis methods.

2.3.2.1 Evaluation by Visual Inspection

Hoffman and Cook [2006b] discussed control parameter mapping layers and evaluation is performed through visual comparison of spectrograms. Numerous other synthesis approaches have performed similar evaluation, through visual comparison of spectrograms [Bascou and Pottier, 2005; Bruna and Mallat, 2013; Doel, 1998; Dubnov *et al.*, 2002; O'Leary and Robel, 2014; O'Regan and Kokaram, 2007a; Ren *et al.*, 2013], comparison of time domain waveform plots [Bar-Joseph *et al.*, 1999; Smith and Serra, 1987;

Verma and Meng, 2000], or showing both spectrograms and time domain plots [Karjalainen *et al.*, 2002; Oksanen *et al.*, 2013; O'Regan and Kokaram, 2007b; Verron *et al.*, 2010b].

2.3.2.2 Informal Evaluation

Fontana and Bresin [2003] demonstrated a range of synthesis models, and informal listening tests, but no results were reported, instead comparing spectrogram plots. Similar mentions of informal subjective evaluation is present in Siddiq [2017] and Gaver [1993]. A five point rating scale for similarity was used by Miner and Caudell [2005], though no results were reported. Guggiana *et al.* [1995] discussed psychoacoustic tests for sound classification, without reporting results. O'Modhrain and Essl [2004] demonstrate a granular synthesis approach with a tactile interface and report performing evaluation, however no results are presented.

Often a lack of evaluation is due to the synthesis method being seen as a creative tool [Chowning, 1973; Kleimola, 2013; Sturm, 2004]. A lack of any baseline for comparison within any field of research, however, will inevitably stifle technological advances [Paul *et al.*, 2013; Rumsey *et al.*, 2005].

2.3.3 Objective Metrics

A summary of all sound synthesis papers that use objective evaluation is presented in Table 2.3. Objective evaluation is considered to be a computational or automated evaluation processes.

There are a range of methods for objective evaluation of synthesised sound effects, However there is little to no consistency on objective metrics to use. Horner and Wun [2006] objectively compared different wavetable synthesis methods using "Relative Spectral Error", with no comparison to samples. In contrast, Hendry and Reiss [2010] compared a synthesis method to reference samples, through visual comparison of spectrograms, and comparison of low level audio features, such as fundamental frequency, first 4 harmonic frequencies, spectral centroid and zero crossing rate, but no comparison with other

TABLE 2.3: Range of Objective Evaluation Metrics used in Current Sound Synthesis Research

Research paper	Objective Evaluation Methods
Sound Effects	
Hendry and Reiss [2010]	Fundamental and four harmonics
	Spectral centroid
	Zero crossing rate
Adami *et al.* [2017]	Spectral envelope
	Applause density
Raghuvanshi and Lin [2006]	Fundamental and harmonics
Horner and Wun [2006]	Fundamental and harmonics
Cook [1997]	Fundamental and harmonics frequency, amplitude and decay
Cook [2002]	Linear predictive coding (LPC) coefficients
Petrausch and Rabenstein [2003]	Number of harmonics
	Filter distortion from filterbank (dB)
Kersten and Purwins [2012]	Temporal centroid
	Signal to noise ratio
Musical Instruments	
Kreutzer *et al.* [2008]	Amplitude envelopes for frequency bands
Valimaki *et al.* [1999]	Envelope of fundamental frequency and 4 harmonics
Cantzos *et al.* [2005]	Ceptral distance in 3 frequency bands
Bensa *et al.* [2000]	Perceptual excitation model
	Perceptual brightness model (spectral centroid)
Hamadicharef and Ifeachor [2003]	PEAQ
Hamadicharef and Ifeachor [2005]	PEAQ
Garcia [2001a]	Least Square Error (LSE) in FD
	Simultaneous Frequency Masking (SFM)
Heise *et al.* [2009]	1D-DCT of each MFCC coefficient over time
	Fundamental frequency
	Spectral shape
	Attack and decay characteristics
	Overall duration
Speech	
Huang [2011]	FW-SNR
	Weighted spectral slope measure
	Itakura-Saito distance
	Log-Likelihood ratio
	PESQ
Valentini-Botinhao *et al.* [2011]	PESQ
Theobald and Matthews [2012]	Correlation
	Normalised RMS error
	Normalised peak RMS error
	Dynamic time warp cost
	Phone-based mahalonabis distance

synthesis methods was undertaken. Objective evaluation was also performed, through inspection and discussion of spectrogram plots. This work was extended to a range of other aeroacoustic sounds, with the same evaluation methodology [Selfridge *et al.*, 2018a, 2017a,c,d].

Hamadicharef and Ifeachor [2003] proposed evaluating sound using Perceptual Evaluation of Audio Quality (PEAQ [Thiede *et al.*, 2000]). PEAQ is an algorithm designed for determining the quality of audio compression codecs, which analyses the sound on a sample by sample basis to determine any perceptual artifacts. This work was further developed to select parameters for a piano synthesiser, to replicate an input audio signal [Hamadicharef and Ifeachor, 2005]. However, the notes will never be exactly the same if played with slightly different attack or at a different sample time - thus resulting in a perceptual difference that should not be attributed to the synthesis model.

Similarly, Heise *et al.* [2009] evaluated synthesis parameter selection using a range of low level audio features, such as fundamental frequency, spectral shape, envelope characteristics and overall duration. They also used the Discrete Cosine Transform (DCT) of the Mel-Frequency Cepstral Coefficients (MFCC's) as a measure of how similar the synthesised sound was to a recorded sample. Allamanche *et al.* [2001] evaluated a set of audio features for similarity on their ability to perform classification on a labeled data set, identifying different modifications of an audio file, such as whether equalisation or dynamic range compression had been applied, whether it had been encoded to MPEG or whether it had been processed through a loudspeaker and microphone chain. Moffat *et al.* [2017] used feature vectors to compare the sonic similarity of different sound effects.

2.3.4 Synthesis Evaluation Summary

In summary, although there is a large body of work on sound synthesis, many proposed methods have not performed evaluation of how effectively the desired sound is produced or the perceived sound quality. When evaluation has been performed, it is often objective, and it is even rarer for it to be comparative, where the proposed technique is compared against alternatives. Nor have standard methodologies been established. This

failing of the sound synthesis community to address evaluation is a clear contributing factor to the lack of understanding of the current state-of-the-art in sound synthesis.

Many fields have tried to address this issue within their respective research areas. This has resulted in the MIREX Competition [Downie, 2008], the BSS Eval toolbox [Vincent *et al.*, 2006] and the PEAQ algorithm [Câmpeanu and Câmpeanu, 2005], which are all attempts to standardise the evaluation approaches in their specific audio fields.

Evaluation of existing synthesis methods could potentially yield significant insight into the state-of-the-art in synthesis technology. Without understanding the benefits and weaknesses of current synthesis techniques, it is not possible to understand where current deficits exist. The lack of standardised evaluation methods and metrics is evident, and can potentially prohibit progress in this field.

As is evident from the literature, it is not expected that a single synthesis method is effectively able to produce all possible sounds. In every case, there may be a range of synthesis approaches that are appropriate. However, this simply highlights the importance of evaluation. Identification of suitable use cases and occasions where a particular sound synthesis method is applicable is vital to having a convincing synthesis process.

2.4 Sound Effects Taxonomy

There are numerous examples of work attempting to create a taxonomy of sound. In Schafer [1993], the author classified sounds by acoustic, psychoacoustic, semantic, aesthetic and referential properties. Russolo and Pratella [1967], classified "noise-sound" into six groups: roars, hisses, whispers, impactful noises, voiced sounds and screams. This is further discussed in Russolo [2004]. Production of a taxonomy of sounds heard in a cafe or restaurant were produced, basing the grouping on the sound source or context [Lindborg, 2016; Stevenson, 2016].

Gaver [1993] presented a classification scheme of sounds based on the state of the physical property of the material. The sound classifications were vibrating solids, liquids and aerodynamic sounds (gas). A series of sub-classifications based on hybrid sounds were

Reference	Type of Sound	Classification properties	Quantitive Analysis	Qualitative Analysis	Word Classification	Audio Feature Analysis	Hierarchical Cluster
Schafer [1993]	Environmental	Acoustics	N	N	N	Y	N
Schafer [1993]	Environmental	Aesthetics	N	N	N	N	N
Schafer [1993]	Environmental	Source/context	N	N	N	N	Y
Russolo [2004]	Environmental	Subjective	N	N	N	N	N
Russolo and Pratella [1967]	Environmental	Subjective	N	N	N	N	N
Stevenson [2016]	Cafe sounds	Source or context	N	N	N	N	Y
Lindborg [2016]	Restaurant	Subjective 'liking' score	N	Y	Y	N	N
Lindborg [2016]	Restaurant	Word occurrence	N	Y	Y	N	Y
Gaver [1993]	Environmental	Physical properties	Y	N	N	N	N
Houix *et al.* [2012]	Environmental	Subjective grouping	Y	N	N	N	Y
Ballas [1993]	Environmental	Subjective ratings	Y	N	N	Y	Y
Gygi *et al.* [2007]	Environmental	Subjective ratings	Y	N	N	N	Y
Aldrich *et al.* [2009]	Environmental	Subjective ratings	Y	N	Y	N	Y
Rychtáriková and Vermeir [2013]	Sound walks	Low level audio features	Y	Y	N	Y	N
Davies *et al.* [2013]	Sound walks	Semantic words	Y	N	Y	N	N
McGregor *et al.* [2006]	Soundscape	Subjective free text recurrence	N	Y	Y	N	N
Pedersen [2008]	Perceptual attribute	Definition of word	N	Y	Y	N	N
Woodcock *et al.* [2016]	Broadcast objects	Predefined word list	Y	Y	Y	N	Y
Salamon *et al.* [2014]	Urban sounds	Source	N	N	N	N	Y
Rocchesso and Fontana [2003]	Synthesised sounds	Control parameters	N	N	N	N	Y
Hemery and Aucouturier [2015]	Field recordings	Labels/audio features	Y	N	N	Y	N

Table 2.4: Summary of literature on sound classification

also produced along with a set of properties that would impact the perception of the sound. Houix *et al.* [2012] developed this further by attempting to understand how participants would arbitrarily categorise sounds. Ballas [1993] asked participants to identify how similar sounds are to each other along a series of different dimensions. They then performed hierarchical cluster analysis on the results, to produce a hierarchical linkage structure of the sounds. Furthermore, Gygi *et al.* [2007] performed a similar study where participants were asked how alike sets of sounds were. Audio features were then correlated to a likeness measure and a hierarchical cluster was produced on the set of selected features.

In Aldrich *et al.* [2009] participants were asked to rate the similarity of audio samples, and performed hierarchical cluster analysis to demonstrate the related similarity structure of the sounds. Rychtáriková and Vermeir [2013] captured the acoustic properties of sound walk recordings and performed unsupervised clustering. These clusters were identified and related back to some semantic terms. Similarly, Davies *et al.* [2013] used sound walks and interviews to identify appropriate words as sound descriptors. McGregor *et al.* [2006] performed classification of sound effects by asking individuals to identify suitable adjectives to differentiate sound samples and similarly in Pedersen [2008] where the authors define classes of sound descriptor words that can be used to relate the similarity of words. In an extension to this, Woodcock *et al.* [2016] asked participants to perform a sorting and labelling task on broadcast audio objects, again yielding a hierarchical cluster.

Salamon *et al.* [2014] produced a dataset of urban sounds, and a taxonomy for the dataset, where sounds are clustered based on the source of the audio, rather than the relative similarity of the audio sample themselves. This dataset is used for unsupervised learning classification [Salamon and Bello, 2015, 2017]. In the context of synthesised sounds, Rocchesso and Fontana [2003] grouped sounds by their control parameters.

There is no clear standard method for grouping sounds such as those found in a sound effects library. It becomes clear from the literature that there is limited work utilising audio features to produce a taxonomy of sound. Table 2.4 shows relevant work that performs classification of sound, and the approach taken in each case. In many cases,

classification is performed through subjective rating or word clustering. It is also apparent there is little work clustering the acoustic properties of individual samples. Schafer [1993] discussed sound classification based on the acoustic properties of samples, but only a high level discussion is presented and the full idea and taxonomy is not developed further.

Chapter 3

Audio Feature Extraction Toolboxes

3.1 Introduction

This chapter presents a series of audio feature extraction toolboxes that were evaluated as part of this work. Audio feature extraction is one of the cornerstones of current audio signal processing research and development. This chapter will perform a rigorous review of a range of toolboxes, supporting work later within this thesis.

An audio feature is a statistical or computational representation of an audio file. Audio features can range for low level measures, such as average frequency of a sound, to a high level interpretation, such as attack time or chord structure. Audio features are typically hand crafted statistical representations of audio, designed with a specific purpose, such as representing timbral context or the pitch of a piece of audio. Audio features are often used for sonic similarity [Aldrich *et al.*, 2009; Gygi *et al.*, 2007; Pachet and Aucouturier, 2004; Peeters, 2004; Virtanen and Helén, 2007]. This is typically performed by measuring distances between sets of audio features, which can then produce a relative distance or similarity measure between two different sounds. These audio features represent all of the important perceptual aspects of a sound, and allow for a set of summary statistics to represent the signal. This is particularly important in the area

of computational processing of audio. Computers have no ability to hear two different sounds for comparison, so instead can calculate some audio features and use this as a basis for comparison. Audio features are an important computational summary of a piece of audio, and are a prominent aspect of many audio based research fields.

Audio features are contextual information that can be extracted from an audio signal. Although these problems are somewhat dissimilar in nature, they rely heavily on a set of related audio features. Low level features are computed directly from the audio signal, often in a frame-by-frame basis such as zero-crossing rate, spectral centroid or signal energy, and generally have little perceptual relevance in comparison to higher level features, like chord or key of musical piece, which hold greater semantic meaning. In Music Information Retrieval (MIR), it is common to refer to audio features or descriptors, whereas, in psychology distinctions are made between dimensions and features, where dimensions are continuous and features are discrete. Audio features are relevant in a range of research fields including:

- feature extraction linked to audio effects [Stables *et al.*, 2014],
- statistical synthesis [McDermott and Simoncelli, 2011],
- feature-based synthesis [Hoffman and Cook, 2006a],
- evaluation of synthesis techniques [Hendry and Reiss, 2010],
- similarity measures [Gygi *et al.*, 2007],
- data classification [McKinney and Breebaart, 2003], and
- data mining [Li *et al.*, 2011].

These audio features can be broken down into groups, as presented in the Cuidado Project [Peeters, 2004], which includes definitions of a range of features. Descriptor is used as a general term that can refer to either continuous or discrete content [Peeters, 2004]. Seventeen low level descriptors (LLDs), or audio features, are defined in the MPEG-7 standard, with feature categorisation, for the purpose of performing audio similarity searching via associated metadata contained within the audio file [Lindsay

and Herre, 2001; Manjunath *et al.*, 2002]. The Cuidado project takes this work further to define 54 audio features for audio similarity and classification [Peeters, 2004]. This project provides definitions of a range of features, groups them, and identifies which are relevant as frame based features. These audio features can then be used to identify a particular aspect of an audio signal. A good overview of features for extraction is presented by Mitrović *et al.* [2010]

A range of audio feature extraction libraries and toolboxes have been constructed. Some are built as workflow tools, with pre-processing and batch operations, some are written for algorithmic efficiency or parallelisation, some for specific programming environments or platforms. Despite significant growth and research in the field of audio signal processing and feature extraction, there has been little research on evaluating and identifying suitable feature extraction tools and their appropriate applications.

It has been identified that MIR primarily focuses on precision and recall, which may be considered a limitation [Reiss and Sandler, 2002b, 2003]. Cleverdon and Keen [1966] developed a six point scale for measuring and evaluating information retrieval systems. This model is widely known as the Cranfield model of information retrieval evaluation. The Cranfield model properties are: Coverage; Time Lag; Effort; Presentation; Precision; Recall. This model is an appropriate platform for evaluation and benchmarking of MIR systems.

Reviews and evaluations of existing feature extraction libraries based on the Cranfield model are presented. The properties of the model can be suitably related to the MIR feature extraction tool evaluation [Reiss and Sandler, 2002a] and presents an evaluation based on the following criteria:

Coverage - The range of audio descriptor features presented by a toolkit, along with additional pre-processing or post-processing functionality.

Effort - User Interface, how easily one can create a new specific query or modify queries, and appropriate documentation.

Presentation - File Output format options and consistency.

Time Lag - Computational Efficiency of each tool.

Precision and recall are both included as part of the Cranfield Model. However, within the case of evaluating feature extraction toolboxes, precision and recall are not considered applicable to the task, and as such are not used. Existing work discusses the merits of using precision and recall within MIR application [Downie, 2004].

Ten audio feature extraction toolboxes are evaluated based on the Cranfield model, as proposed by Reiss and Sandler [2002b]. Section 3.3 compares the functionality of the tools with respect to the range of audio features that can be extracted and any further pre or post processing that the tool implements. The interface options of each toolboxes is presented and discussed in Section 3.4. The output format of data of each toolbox is presented in Section 3.5 and the computational time is presented in Section 3.6.

3.2 Existing Feature Extraction toolboxes

There are a large number of audio feature extraction toolboxes available, delivered to the community in differing formats, but usually as at least one of the following formats:

- stand alone applications,
- plug-ins for a host application, and
- software function library.

To allow for delivery of tools, some Application Programming Interfaces (API) have been constructed to allow for feature extraction plug-ins to be developed. Vamp [Cannam, 2009] is a C++ API specification which functions with standalone applications such as Sonic Visualiser, a content and feature visualiser with Graphical User Interface (GUI) and its command line interface (CLI) counterpart Sonic Annotator [Cannam *et al.*, 2010]. The Vamp Plugin API is an independent plugin development framework and as a result plugin libraries have been developed by numerous research labs and academic institutions. However due to the nature of the framework, it is not possible to create

plug-ins that depend on pre-existing plug-ins. This results in multiple implementations and instances of certain features being calculated, which causes potential system inefficiencies. Feature Extraction API is another plugin framework API in C and C++ [Lerch *et al.*, 2005], though it is less commonly used than the VAMP plugin format. There are also feature extraction libraries that provide their own plugin API for extending their stand alone system [McKay *et al.*, 2005], though this is less common. There has been a rise in MIR web services, such as the web based audio feature extraction API produced by Echo Nest,[1] where users submit files online and receive extensible markup language (XML) descriptions. These tools have resulted in large music feature datasets, such as the Million Song Dataset [Bertin-Mahieux *et al.*, 2011].

The feature extraction tools evaluated are:

Aubio A high level feature extraction library that extracts features such as onset detection, beat tracking, tempo, melody [Brossier, 2006].

Essentia Full function workflow environment for high and low level features, facilitating audio input, pre-processing and statistical analysis of output. Written in C++, with Matlab and Python binding and export data in YAML (YAML Ain't Markup Language) or JSON (JavaScript Object Notation) format [Bogdanov *et al.*, 2013].

jAudio Java based stand alone application with GUI and CLI. Designed for batch processing to output in XML or ARFF (Attribute-Relation File Format) format for loading into Weka [McKay *et al.*, 2005].

Librosa API for feature extraction, for processing data in Python [McFee *et al.*, 2015]

LibXtract Low level feature extraction tool written with the aim of efficient realtime feature extraction, originally in C but now ported to Max, Pure Data, Super Collider and Vamp formats [Bullock, 2007].

Marsyas Full real-time audio processing standalone framework for data flow audio processing with GUI and CLI. This programme includes a low level feature extraction tool built in C++, with ability to perform machine learning and synthesis within

[1] `http://static.echonest.com/enspex/`

the framework. The feature extraction aspects have also been translated to Vamp plugin format [Tzanetakis and Cook, 2000].

Meyda Web Audio API based low level feature extraction tool, written in Javascript. Designed for web browser based efficient real-time processing [Rawlinson *et al.*, 2015].

MIR Toolbox Audio processing API for offline extraction of high and low level audio features in Matlab. Includes pre-processing, classification and clustering functionality along with audio similarity and distance metrics as part of the toolbox functionality. Algorithms are fragmented allowing detailed control with simple syntax, but often suffers from standard Matlab memory management limitations [Lartillot and Toiviainen, 2007].

Timbre Toolbox A Matlab toolbox for offline high and low level feature extraction. A toolbox that provides different set of features to the MIR Toolbox, specifically made efficient for identifying timbre and to fulfil the Cuidado standards [Peeters *et al.*, 2011].

YAAFE Low level feature extraction library designed for computational efficiency and batch processing by utilising data flow graphs, written in C++ with a CLI and bindings for Python and Matlab [Mathieu *et al.*, 2010].

This list is not exhaustive, as there are many other feature extraction tools out there [Brent, 2010; Bryant, 2014; Deliege *et al.*, 2008; Eyben *et al.*, 2013]. However the list of tools selected was designed with popularity, programming environment range and how recently it has been updated all being taken into consideration.

3.3 Coverage

The coverage of an information retrieval system can be defined as the extent to which all relevant matters are covered by the system. Within the context of audio feature extraction tools, the coverage can be considered as the range of features a tool can

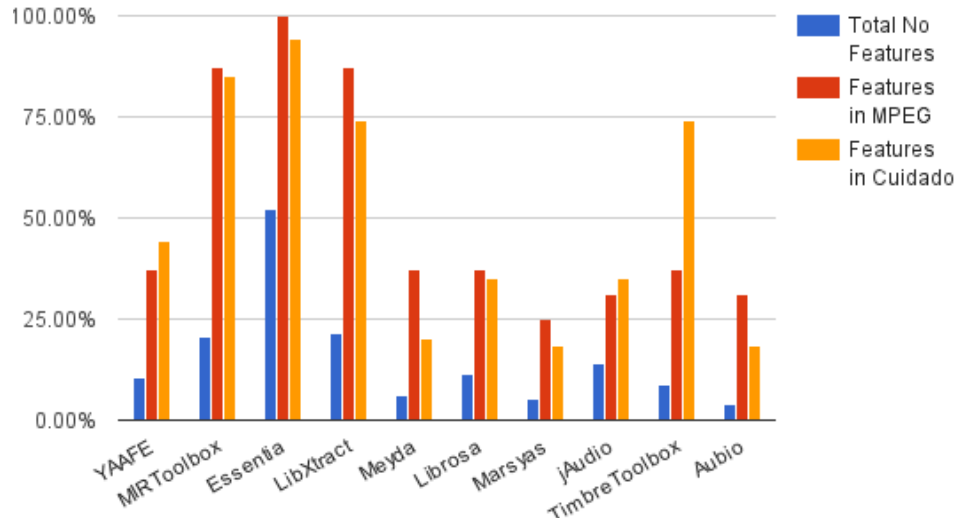

FIGURE 3.1: Percentage Coverage of Multiple Feature Sets

extract. This section presents the features provided by each toolbox, relative to the total number of unique features from all presented tool boxes and the features from the MPEG-7 and Cuidado standard sets of audio descriptors. The relative importance of audio features is heavily context based. To provide a meaningful measure of the relative importance of audio features within each toolbox, the toolboxes will be compared to their compliance with the MPEG-7 and Cuidado standards. Additional functionality, including pre-processing and post-processing available with each feature extraction tool will also be discussed. The accuracy of audio features or specific implementation detail is beyond the scope of this work, but is discussed by Raffel *et al.* [2014].

The features available within each tool are evaluated, and a list of unique features is created. Each tool is then compared to the total list of unique features. Each tool is also evaluated based on the feature coverage when compared to the MPEG-7 and Cuidado standard feature sets. The results of this can be seen in Figure 3.1. It can be seen that Essentia provides the largest range of features, and is the only toolbox to produce 100% coverage of the MPEG-7 audio descriptors. Following this the MIR Toolbox and LibXtract both fulfill over 85% of the MPEG-7 and provide 85% and 75% of the features contained within the Cuidado project, respectfully. The Timbre Toolbox provides nearly 75% of the Cuidado feature set, however this may be unsurprising as they were both written by the same principal author. YAAFE, jAudio and Librosa provide

fairly similar features sets, presenting between 30% and 38% of the MPEG-7 standard feature set, with YAAFE presenting some more perceptually motivated features than the others. Meyda and Aubio provide a relatively low number of features, however this is not without justification: Meyda is written for real-time processing in the browser, and as such is inherently limited to a set of frame based features; Aubio is designed to focus on more high level feature extraction tools, and though providing a number of low level features, this is only to facilitate the high level feature extraction. Marsyas performs the worst in terms of feature range, complying to just 25% of MPEG-7 standard and 20% of the Cuidado standard. Marsyas is designed as an audio processing platform, where clustering, classification and synthesis can all be performed within the entire workflow, so the range of available features may be limited, but the tool provides functionality beyond feature extraction. It is worth noting that only three features, spectral centroid, spectral rolloff and signal energy, are present in all the toolboxes and just 30 features are present in more than half of the toolboxes, and so attention must be paid if specific features are required.

Table 3.1 shows that LibXtract and Meyda do not provide any high level features. jAudio provides a limited range of high level features, such as beat histogram, and strongest beat. Aubio provides a limited range of low level features, such as spectral centroid, spread, skew and kurtosis. Aubio is designed specifically as a high level feature extraction tool, with particular focus on segmentation, whereas LibXtract, Meyda and jAudio are principally designed to extract low level features.

Additional functionality, such as pre-processing or post processing, is provided by a range of tools. Pre-processing is an important aspect of evaluating any audio processing system, as it allows the user to be confident of a classification in any low quality environment where the audio may be degraded Mauch and Ewert [2013]. The resample function allows a standardisation of sample rates within a toolbox, which can be used to ensure that expected results for spectral comparisons are within the same range. For example, if a sample rate of 96kHz is used, it would be possible to have a spectral centroid of 30kHz, which has no perceptual meaning, compared to a file sampled at 44.1kHz, where the maximum spectral centroid would be 22050Hz.

Table 3.1: Overview of Feature Extraction Tools

	Aubio	Essentia	jAudio	Librosa	LibXtract	Marsyas	Meyda	MIR	Timbre	YAAFE
High level Features	Y	Y	N[‡]	Y	N	Y	N	Y	Y	Y
Low Level Features	N*	Y	Y	Y	Y	Y	Y	Y	Y	Y
Resample	Y	Y	Y	Y	N	Y	Y[†]	Y[†]	Y[†]	Y
Filter	N	Y	N	N	N	N	Y[†]	Y[†]	Y[†]	N
Clustering	N	Y	N[§]	N	N	Y[§]	N	Y[§]	N	N
Similarity	N	N	N	N	N	N	N	Y	N	N
Real-Time	Y	Y			Y	Y	Y			
Vamp Plugin	Y	N	N	N	Y	Y	N	N	N	N
GUI	Y[+]	N	Y	N	Y[+]	Y[+]	N	N	N	N
CLI	Y	Y[E]	Y	N	Y[+]	Y	N	N	N	Y
APIs	C/C++ Python R PD	C/C++ Python Matlab[O] PD/Max-MSP	Java	Python	C/C++ Supercollider PD/Max-MSP Java	C/C++ Python Java Lua	JS	Matlab	Matlab	Matlab Python C/C++
Output	Vamp	YAML JSON	XML ARFF	CSV	Vamp XML	Vamp CSV ARFF		TSV ARFF	TSV	CSV HDF5

* = Except MFCC and FFT Statistics, [‡] = Some Mid-high level features but very limited, [†] = As part of environment, not toolbox, [+] = As result of being Vamp plugin, [§] = Can produce ARFF files, designed for being read directly into Weka.[E] = CLI is produced through C 'Extractor' files, with some examples provided. [O] = A project for calling Essentia from Matlab has been developed.

It can be seen from Table 3.1 that Aubio, Essentia, jAudio, Librosa, Marsyas and YAAFE all provide the user with some resample function as part of the toolbox, where as Meyda, MIR Toolbox and Timbre Toolbox all inherit a resample function from their native environments, as Web Audio API and Matlab both have resample functions built in. LibXtract does not provide a resample function, however, if used as a Vamp plugin, many Vamp hosts contain resample functions.

Clustering, as a post-processing tool, is also a useful functionality for many MIR processes. The post processing tools allow the user to directly analyse the output results as part of a single process. Essentia, Marsyas and MIR Toolbox all provide some form of clustering algorithm within them, and jAudio, Marsyas and MIR Toolbox can export files directly to ARFF format for loading directly into Weka, a data mining and clustering tool [Hall *et al.*, 2009].

Essentia, MIR Toolbox and LibXtract produce a strong range of feature coverage, and the Timbre Toolbox covers the Cuidado feature set well. In terms of feature range these tools seem to perform better than many other existing tools. Essentia and MIR Toolbox both provide a powerful range of additional pre and post processing tools to benefit the user.

3.4 Effort

Effort is used to define how challenging a user finds a system to use, and whether any user experience considerations have been made while developing a system. Within this section, effort is evaluated relative to the user interface that is provided, whether it is a GUI, CLI or an API. The existence and quality of documentation and suitable examples is evaluated. The purpose is to identify how intuitively a tool's interface is presented to a user.

Table 3.1 outlines the user interfaces presented by each of the feature extraction tools. It can be seen that jAudio is the only tool that comes with its own GUI, though Aubio, LibXtract and Marsyas all have GUI capabilities through virtue of being Vamp plugins, when paired with a visualisation tool such as Sonic Visualiser. CLI's are more

common, as Aubio, Marsyas, jAudio and YAAFE come with complete command line interfaces and Essentia comes with a series of precompiled C++ binary files that can be run from the command line. However, this limits control functionality, as all the control is included in the software implementation. LibXtract can also be controlled via command line, through the use of its Vamp plugin format and a Vamp CLI tool such as Sonic Annotator. All tools come with APIs, which means Librosa, Meyda, MIR Toolbox and Timbre Toolbox are all only presented as software APIs, and as such all require software implementation before feature extraction is possible.

There are five different APIs written for both C and Python. Four APIs are available for Matlab, including the Essentia Matlab project.[2] Java has three APIs and only a single API for Javascript, Pure Data, Max-MSP, Supercollider, Lua and R are provided. Although Python and C are common programming languages, it is believed that Matlab is one of the most common frameworks used within MIR [Page *et al.*, 2012]. Environments such as web audio, in Javascript, Pure Data and Max-MSP are much less common in the MIR and audio research field, but are advantageous as they are real-time audio environments where features are calculated in realtime and as such are excellent for prototyping.

Most toolboxes have clear documentation with examples, but there is limited documentation for LibXtract, Meyda and the Timbre toolbox. Though the documentation is not unclear, all the other tools provide a lot more information regarding basic access and software applications. Similarly, all toolboxes supply basic examples of implementation, however YAAFE, Essentia, Aubio and MIR Toolbox all have a strong range of examples that run straight away. Marsyas has clear documentation and a range of examples from which to draw inspiration, but required the user to learn a proprietary language for use as part of the system.

In conclusion, when looking for a standalone tool which covers a user flexibility and usability with a user interface, then the Vamp plugin route is a useful one to take. This provides a simple intuitive interface and any number of specific features can be loaded in as required. If batch processing is required, then either the GUI from jAudio or CLI

[2] `https://github.com/MTG/Matlab-c-tools/tree/master/essentia`

from YAAFE are intuitive, flexible and simple to use. If a user requires a programming API, then, depending on their environment, there are potentially a range of tools. C, C++ Python and Matlab APIs are provided by a range of tools, with often multiple being offered by each toolkit, as can be seen in Table 3.1.

3.5 Presentation

An important aspect of any information retrieval system is how the resulting information is presented back to the user. Within this section, the output format of data is discussed and the relative merits of each approach outlined.

Document output format is one of the most significant barriers in fully integrated workflow solutions within MIR [Casey *et al.*, 2008]. Output format is important, primarily as it impacts the ease and format of analysis someone can carry out on a dataset. Typically, a software API is used to store values in a relevant data structure within the given development language and as such, file output format becomes irrelevant in this case.

XML, YAML and JSON are all standard structured data formats, that allow the presentation of hierarchical structures of data. HDF5 (Hierarchical Data Format 5) is also a hierarchical data structure specifically designed for efficient storage and accuracy of contents and relating to the Semantic web - as such is well suited to big data tasks. CSV (Comma-separated values) and TSV (Tab-separated values) are table structures that allow users to view data in most spreadsheet applications, and ARFF is also a table structure with specific metadata about each column format and available options. ARFF is specifically designed for use with Weka, which is a powerful data mining tool.

CSV and TSV formats are considered to be suitable output formats if the resulting value can be considered as a table, however within the feature extraction, generally there is a much more complex data structure than simply two dimensions. Features carry varying levels of complexity, as some features are global for a signal where as some are based on windowed frames of a signal. Some features, such as MFCCs produce 13 numerical values per frame. As such it seems suitable that the data format used to output these results can represent these hierarchical feature formats. JSON and XML file formats

are well supported by almost all programming languages, so there should not be any issues with processing the data results. CSV is also well supported within programming languages, but the lack of complexity or data structure can lead to potential ambiguities or errors with data transfer. The benefits of producing ARFF files, which can be loaded direct into Weka allows the user a great range of data mining opportunities and should not be underestimated.

Although it is clear that the file format is reliant on further applications, any feature extraction library should be able to present its data output in a data structure to suitably represent the hierarchical nature of the data it intends to represent. YAAFE, Essentia, jAudio and LibXtract all provide some from of suitable data structure, however only jAudio can also pass files direct into Weka. Marsays and MIR Toolbox both allow for unstructured data, but can produce ARFF files for easy data mining, and Librosa and Timbre Toolbox will only allow users an unstructured data in a table format. YAAFE and jAudio are the only two applications that allow users the choice of structured or tabular data.

3.6 Time Lag

Time lag is the measure of how long a given task will take to complete. Understanding the time necessary to perform a task, and comparing the relative speed of systems will give users an informed choice as to what system to use, particularly when they want to analyse large data sets. This section will discuss the computational complexity of the ten feature extractions tools and identify whether they are implemented in real-time or if they are offline methods.

Meyda and the various LibXtract ports to Pure Data, Supercollider and Max-MSP are all designed to run in realtime. Each of these real-time environments are provided with suitable feature extraction, which provides a user with powerful visualisation and real-time interactivity but is less useful for users wishing to focus on offline approaches.

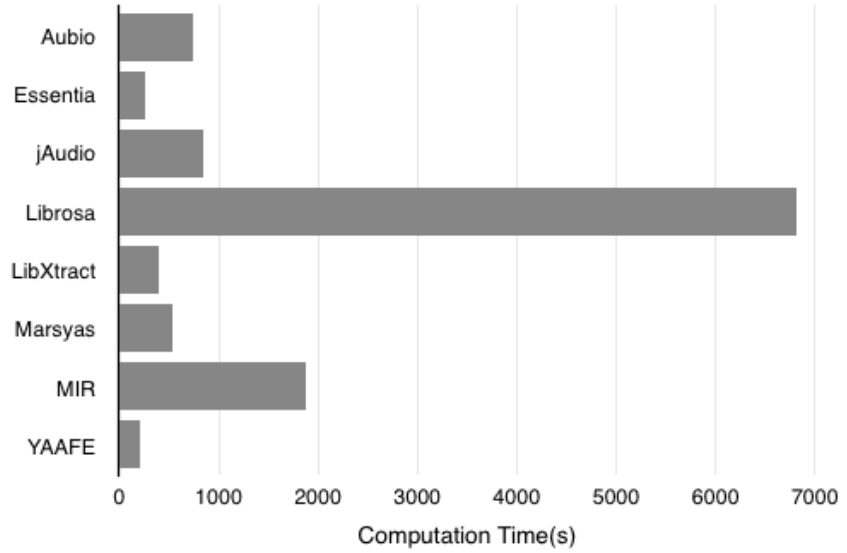

FIGURE 3.2: Graph of Computational Time of Feature Extraction Tools

The offline approaches were all evaluated for computational efficiency. A dataset for evaluation is a subset of the Cambridge Multitrack Data Set.[3] Thirty two different songs were used. The dataset consists of 561 tracks with an average duration of 106s, which totalled over 16.5 hours of audio and 8.79Gb of data. Each toolbox is used to calculate the MFCC's from this data set, with a 512 sample window size and 256 sample hop size. The input audio is at a variety of different sample rates and bit depths to ensure that variable input file formats is allowable. This test is run on a MacBook Pro 2.9GHz i7 processor and 8Gb of RAM. The results are presented in Figure 3.2. The MFCCs were used, as they are a computational method, that exists within nine of the ten given tool boxes, and so should provide a good basis for comparison of computational efficiency. MFCCs are not computed by the Timbre Toolbox and Meyda will only run in real-time.

As can be seen from Figure 3.2, Yaafe is the fastest toolbox, processing over 16.5 hours of audio in just over 3 minutes 30s, with Essentia coming in as a close second place at 4 minutes 12s. LibXtract and Marsyas both completed in under 10 minutes, and both Aubio and jAudio ran in under 15 minutes. The MIR toolbox took over 31 minutes to

[3] `http://www.cambridge-mt.com/ms-mtk.htm`

run and Librosa took 1hour 53 minutes. It is evident that tools written in C or C++ run faster than tools written in Python or java.

3.7 Discussion

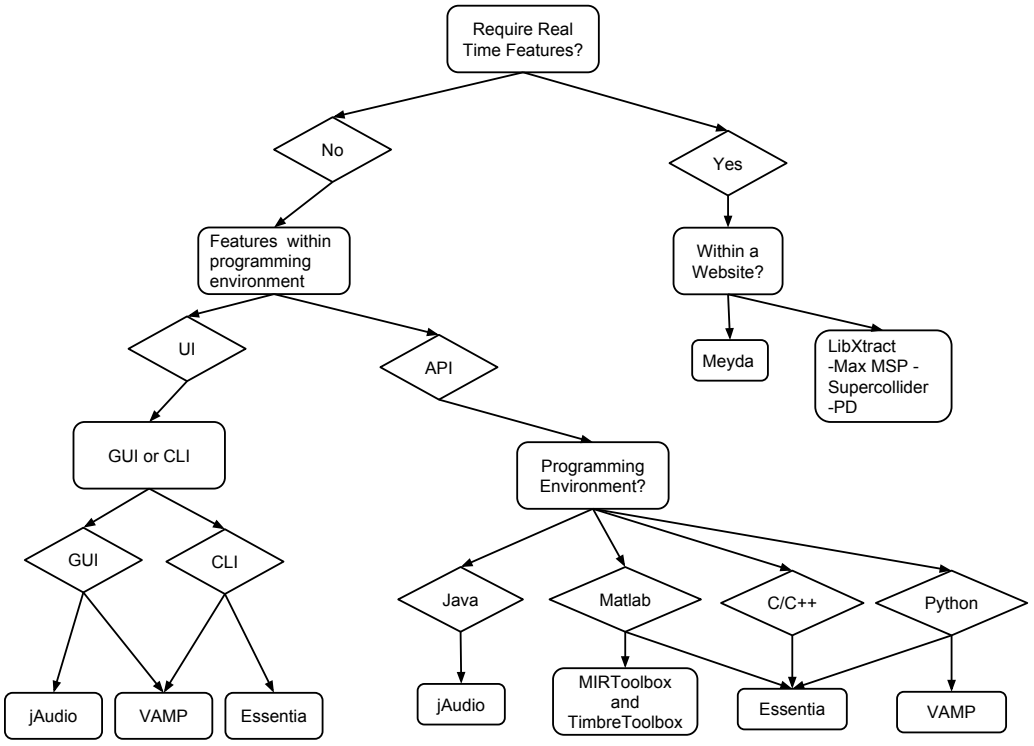

FIGURE 3.3: Flowchart to recommend what tool to use

N.B. Vamp = LibXtract and Marsyas Vamp Packages

Ten audio feature extraction toolboxes are discussed and evaluated relative to four of the six criteria of the Cranfield Model.

Meyda and LibXtract provide excellent real-time feature extraction tools in various programming environments. When high level features and segmentation is required, Aubio provides a simple and intuitive tool. It also provides a Vamp plugin format. When visualisation is required, for annotation or basic exploration of features, using the Vamp plugin format is very powerful, and the combination of LibXtract and Marsysas as Vamp plug-ins provide excellent coverage of audio features. Research based in MATLAB should use the MIR Toolbox combined with the Timbre Toolbox for maximum feature

coverage, or Essentia where computational efficiency is important with little sacrifice of feature range. Essentia performs the best with regards to computation, feature coverage and output presentation, with a range of APIs.

As the suitable feature extraction toolbox is entirely application dependent, there is no single case where a certain tool is better or more powerful than another. However suggestions for suitable toolboxes to use can be identified from Figure 3.3. When working on real-time applications, either Meyda or LibXtract will be most suitable for applications, whereas when working in an offline fashion, there is an option for either user interfaces or APIs. If a user interface is required then Vamp plug-ins, of LibXtract and Marsyas, are very powerful and advantageous tools, that can be hosted in either graphic interfaces or command line interfaces. jAudio provides a strong user interface with batch processing tool, but its range of features is limited. Essentia provides a strong CLI with large range of features, but low level of control, so implementation is required for accurate control of the features. If an API is required, then a range of example suggestions are proposed for some commonly used programming languages. A Java API is provided by jAudio, which is powerful and efficient, but performs on a reduced feature set. Strong Matlab APIs are provided by either a combination of MIR Toolbox and Timbre Toolbox or Essentia, with the 'Matlab-C-tools'.[4] As such, the Essentia tool box will be used for feature extraction throughout the rest of this thesis.

[4] `https://github.com/MTG/matlab-c-tools`

Chapter 4

Taxonomy of Sound Effects

4.1 Introduction

This chapter demonstrates the production of a hierarchical taxonomy of sound effects, based entirely on the sonic properties of the audio samples, through the use of unsupervised machine learning. Unsupervised machine learning is a machine learning approach that can be performed on unlabelled data. This will provide a better understanding of the relative structure and similarity of different sound effects. An unsupervised, sonically inspired taxonomy offers an alternative to standard categorisation, in the hope that it will aid the search process by alleviating dependence on manually annotated labels and inconsistent grouping of sounds.

Sound designers regularly use sound effects libraries to design audio scenes, layering different sounds in order to realise a design aesthetic. For example, numerous explosion audio samples are often combined to create an effect with the desired weight of impact. A large part of this work involves the use of Foley, where an artist will perform sound with a range of props. A key aspect of Foley is that the prop being used may not match the object in the visual scene, but is capable of mimicking its sonic properties. An example would be the use of a mechanical whisk, which becomes a convincing gun rattle sound effect when combined in a scene with explosions and shouting.

Sound designers are less interested in the physical properties or causes of a sound, and more interested in their sonic properties. Despite this, many sound effects libraries are organised into location based or physical categories. It is common for sound effects libraries to use location categories such as 'Forest', 'Urban', 'Boat' or 'Kitchen'. Physical properties of the object such as 'Telephone' or 'Glass' are often used. These two grouping approaches can be problematic as often a sound derived from a specific property or location may be appropriate for use in other cases, such as in the field of Foley sound. In Wold *et al.* [1996] a sound search tool based on sonic properties is proposed, considering loudness, pitch and timbral attributes. A similar tool for semantic browsing of a small library of urban environmental sounds has also been proposed by Lafay *et al.* [2016]. No other known classification methods for sound effects based on their sonic attributes exist, though Heller and Wolf [2002] discuss the importance of acoustical features for associating sound effects to target events. Most previous work focuses either on subjective similarity or the context and source of the sound.

Given that the practical use for a sound sample is often abstracted from its original intention, source, or semantic label, categorisation based on this information is not always desirable. Furthermore, no standard exists for the labelling of recorded sound, and the metadata within a sound effects library can be highly inconsistent. This can make the task of searching and identifying useful sounds extremely challenging, and the neuances desired are often missing. For this reason, along with many others, sound designers may resort to recording new sound effects for each new project.

Different approaches to developing taxonomies of sound are discussed in Section 2.4. A review of different feature extraction toolboxes was presented in Section 3.1. Section 4.2 presents the dataset, feature selection technique and unsupervised learning method undertaken to produce a hierarchy within a sound effects library. The taxonomy produced is presented in Section 4.3. The evaluation of the presented taxonomy is undertaken in Section 4.4 and discussed in Section 4.5. Finally, the validity of the taxonomy is discussed in Section 4.6.

4.2 Methodology

An unsupervised machine learning technique is used to develop an inherent taxonomy of sound effects. This section will detail the various development stages of the taxonomy, as presented in Figure 4.1. The Adobe sound effects library was used. A set of audio features were extracted, feature selection was performed using Random Forests (see Section 4.2.3), and a Gaussian Mixture Model (see Section 4.2.3) was used to predict the optimal number of clusters in the final taxonomy. From the reduced feature set, unsupervised hierarchical clustering was performed to produce the number of clusters as predicted using the Gaussian Mixture Model. Finally the hierarchical clustering results are interpreted. All software is available online.[1]

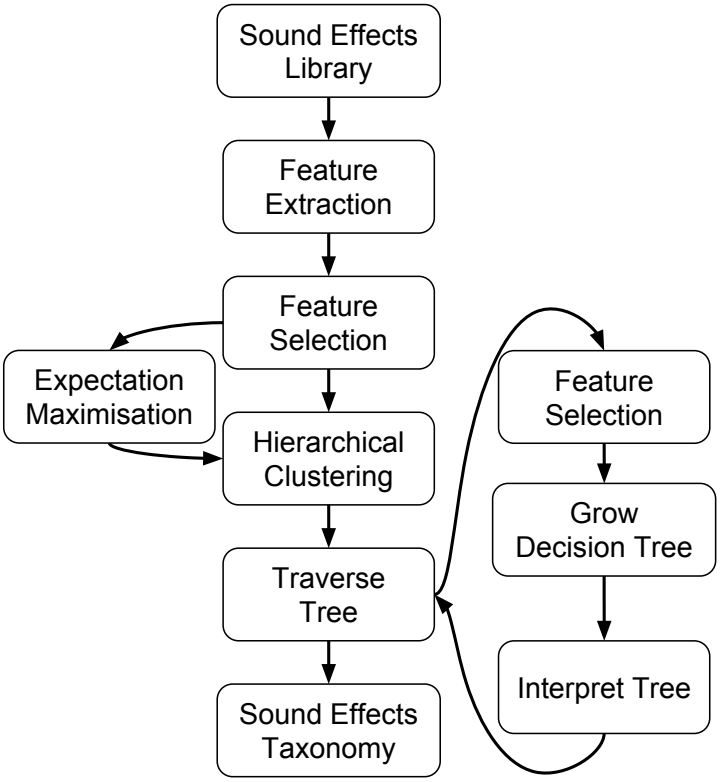

FIGURE 4.1: Flow Diagram of unsupervised sound effects taxonomy system.

[1] `https://goo.gl/9aWhTX`

4.2.1 Dataset

A dataset containing around 9,000 audio samples from the Adobe sound effects library is used.[2] This sound effects library contains a range of audio samples, between 0.1s and 4 minutes in length. The average sample length was 6 seconds with a standard deviation of 14.7. All input audio signals were downmixed to mono, downsampled to 44.1 kHz if required, and had the initial and final silence removed. All audio samples were loudness normalised using ReplayGain [Robinson and Hawksfords, 2000]. Each sound effect was placed in a different folder, describing the context of the original sound effect. The original labels from the sound effects library can be found in Table 4.1, along with the number of samples found in each folder.

TABLE 4.1: Original label classification of the Adobe Sound Effects Dataset.

Class Name	Quantity of Samples	Class Name	Quantity of Samples
Ambience	92	Animals	173
Cartoon	261	Crashes	266
DC	6	DTMF	26
Drones	75	Emergency Effects	158
Fire and Explosions	106	Foley	702
Foley Footsteps	56	Horror	221
Household	556	Human Elements	506
Impacts	575	Industry	378
Liquid-Water	254	Multichannel	98
Multimedia	1223	Noise	43
Production Elements	1308	Science Fiction	312
Sports	319	Technology	219
Tones	33	Transportation	460
Underwater	73	Weapons	424
Weather	54		

DC are offset component audio signals, typically used for system tests. DTMF is Dual Tone Multi Frequency - a set of old telephone tones.

4.2.2 Feature Extraction

The dataset described in Section 4.2.1 was used. The Essentia Freesound Extractor was used to extract audio features [Bogdanov *et al.*, 2013]; as Essentia allows for extraction of a large number of audio features, is easy to use in a number of different systems and

[2] `https://goo.gl/TzQgsB`

produced the data in a highly usable format, as discussed in Section 3.1. A range of 180 different audio features were extracted, which are the essentia freesound extractor set.[3] This extractor set was used, as it is purposely designed for computations on large non-musical sound collections, and it was felt that this best applies to the data being investigated. All frame based features were calculated using a frame size of 46ms with a hop size of 23ms, with the exception of pitch based features, which used a frame size of 92ms and the hop size 46ms. The statistics of these audio features were then calculated, to summarise frame based features over the audio file. The statistics used are the mean, variance, skewness, kurtosis, median, mean of the derivative, the mean of the second derivative, the variance of the derivative, the variance of the second derivative, the maximum and minimum values. Non-frame based features were computed directly, so no statistics were required. This produced a set of 1450 features, extracted from each file. The original feature set was reduced to 1364 features, as features were removed if they provided no variance over the dataset. All features were then normalised to the range $[0, 1]$.

4.2.3 Feature Selection

Feature selection was performed using a similar method to the one described in Ronan *et al.* [2015], where a Random Forest classifier is used to determine audio feature importance. Random forests are an unsupervised classification technique where a series of decision trees are created, each with a random subset of features. As such, the clusters were intended to separate out the data in the most natural manner presented from the data. The out-of-bag (OOB) error is then calculated, as a measure of the random forests classification accuracy. OOB error is a measure of the predicted error for each feature, if that feature were to be removed from the feature set, for each given tree. From this OOB error, it is possible to allocate each feature with a Feature Importance Index (FII), which ranks all audio features in terms of importance by evaluating the OOB error for each tree grown with a given feature, to the overall OOB error [Breiman, 2001].

[3] `https://essentia.upf.edu/documentation/freesound_extractor.html`

Ronan *et al.* [2015] eliminated the audio features from a Random Forest that had an FII less than the average FII and then grew a new Random Forest with the reduced audio feature set. This elimination process would repeat until the OOB error for a newly grown Random Forest started to increase.

In this work, the 1% worst performing audio features were eliminated, on each step of growing a Random Forest, in a similar to but more conservative than the approach taken by Genuer *et al.* [2010]. In order to select the correct set of audio features that fitted the dataset, the feature set that provided the lowest mean OOB error over all the feature selection iterations was chosen.

On each step of the audio feature selection process, the data was clustered using a Gaussian Mixture Model (GMM). GMM's are an unsupervised method for clustering data, on the assumption that data points can be modelled by a gaussian. In this method, the number of clusters is specified and get a measure of GMM quality using the Akaike Information Criterion (AIC). The AIC is a measure of the relative quality of a statistical model for a given dataset. The number of clusters used to create each GMM was continuously increased, while performing 10-fold cross-validation until the AIC measure stops decreasing. This provides the optimal number of clusters to fit the dataset.

4.2.4 Hierarchical Clustering

There are two main methods for hierarchical clustering. Agglomerative clustering is a bottom up approach, where the algorithm starts with singular element clusters and recursively merges two or more of the most similar clusters. Divisive clustering is a top down approach, where the data is one large cluster, and is recursively separated out into a fixed number of smaller clusters.

Agglomerative clustering was used in this chapter, as it is frequently applied to problems within this field [Aldrich *et al.*, 2009; Ballas, 1993; Gygi *et al.*, 2007; Ronan *et al.*, 2015; Rychtáriková and Vermeir, 2013; Woodcock *et al.*, 2016]. It also provides the benefit of providing cophonetic distances between different clusters, so that the relative distances between nodes of the hierarchy are clear. Agglomerative clustering was performed,

on the feature reduced dataset, by assigning each individual sample in the dataset as a cluster. The distance was then calculated for every cluster pair based on Ward's method [Ward Jr, 1963],

$$d(c_i, c_j) = \sqrt{\frac{2n_{c_i}n_{c_j}}{n_{c_i} + n_{c_j}}} euc(x_{c_i}, x_{c_j}) \tag{4.1}$$

where for clusters c_i and c_j, x_c is the centroid of a cluster c, n_c is the number of elements in a cluster c and $euc(x_{c_i}, x_{c_j})$ is the euclidean distance between the centroids of clusters c_i and c_j. This introduces a penalty for clusters that are too large, which reduces the chances of a single cluster containing the majority of the dataset and that an even distribution across a hierarchical structure is produced. The distance is calculated for all pairs of clusters, and the two clusters with the minimum distance d are merged into a single cluster. This is performed iteratively until there is only a single cluster. This provides us with a full structure of the data, which can be visualised from the whole dataset, down to each individual component sample.

4.2.5 Node Semantic Context

In order to interpret the dendrogram produced from the previous step, it is important to have an understanding of what is causing the separation at each of the node points within the dendrogram. Visualising the results of machine learning algorithms is a challenging task. According to Baehrens *et al.* [2010], decision trees are the only classification method which provides a semantic explanation of the classification. This is because a decision tree faciliates inspection of individual features and threshold values, allowing interpretation of the separation of different clusters. This is not possible with any other classification methods. As such, feature selection was undertaken and then a decision tree is grown to provide some semantic meaning to the results.

Each node point can be addressed as a binary classification problem. For each node point, every cluster that falls underneath one side is put into a single cluster, and everything that falls on the other side of the node is placed in another separate cluster. Everything that does not fall underneath the node is ignored. This produces two clusters, which represent the binary selection problem at that node point. From this node point,

a random forest is grown to perform the binary classification between the two sets and feature selection is then performed as described in Section 4.2.3. The main difference here is that only the five most relevant features, based on the FII are selected at each stage. Five features were selected, so as to allow for a low number of features in order to allow some interpretability to the separation. This is performed for interpretability of the separation at this node point.

A decision tree is grown with this reduced set of five audio features, to allow manual visualisation of the separation of data at each node point within the hierarchical structure. The decision tree is constructed by minimising the Gini Diversity Index (GDI), at each node point within the decision tree, which is calculated as:

$$GDI = 1 - \sum_{i} p(i)^2 \tag{4.2}$$

where i is the class and $p(i)$ is the fraction of objects within class i following the branch. The decision trees are grown using the CART algorithm [Breiman, 1984]. To allow for a more meaningful visualisation of the proposed taxonomy, the audio features and values were translated to a semantically meaningful context based on the audio interpretation of the audio feature. The definitions of the particular audio features were investigated and the perceptual context of these features are identified, providing relevant semantic terms in order to describe the classification of sounds at each node point.

4.3 Results and Evaluation

4.3.1 Feature Extraction Results

Figure 4.2 plots the mean OOB error for each Random Forest that is grown for each iteration of the audio feature selection process. In total there were 325 iterations of the feature selection process, where the lowest OOB error occurred at iteration 203 with a value of 0.3242. This reduced the number of audio features from 1450 to 193. 193 features were remaining, as the bottom 1% of features were removed in each one of the 203 iterations, as discussed in Section 4.2.3

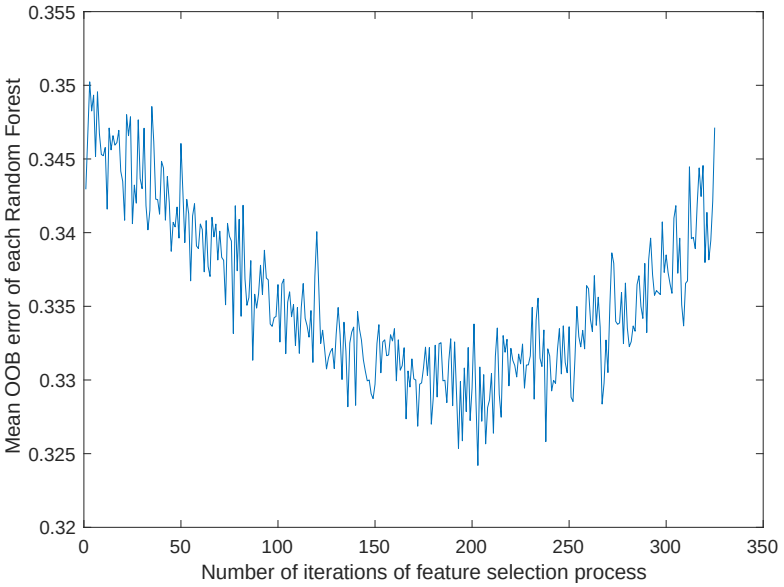

FIGURE 4.2: Mean OOB Error for each Random Forest grown plotted against number of feature selection iterations

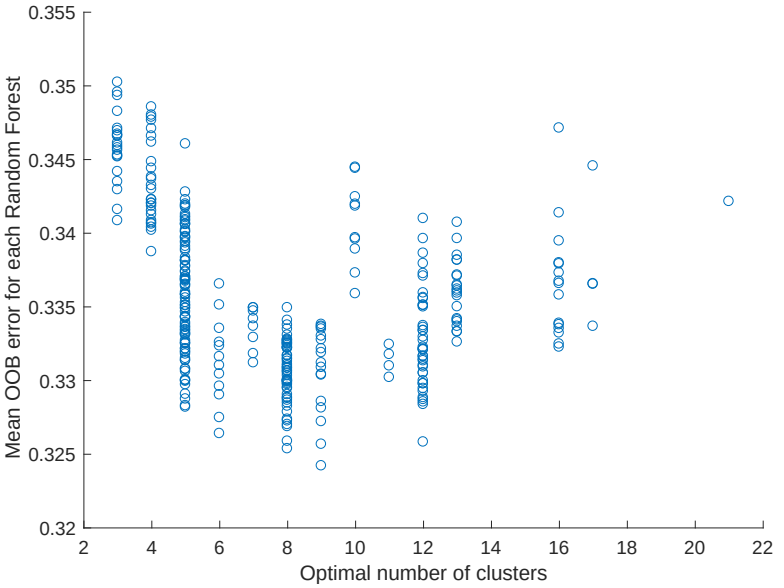

FIGURE 4.3: Mean OOB Error for each Random Forest grown plotted against optimal number of clusters for each feature selection iteration

Figure 4.3 depicts the mean OOB error for each Random Forest feature selection iteration against the optimal amount of clusters, where the optimal amount of clusters was calculated using the AIC for each GMM created. The optimal amount of clusters was found to be 9, as this coincides with the minimum mean OOB error in Figure 4.3.

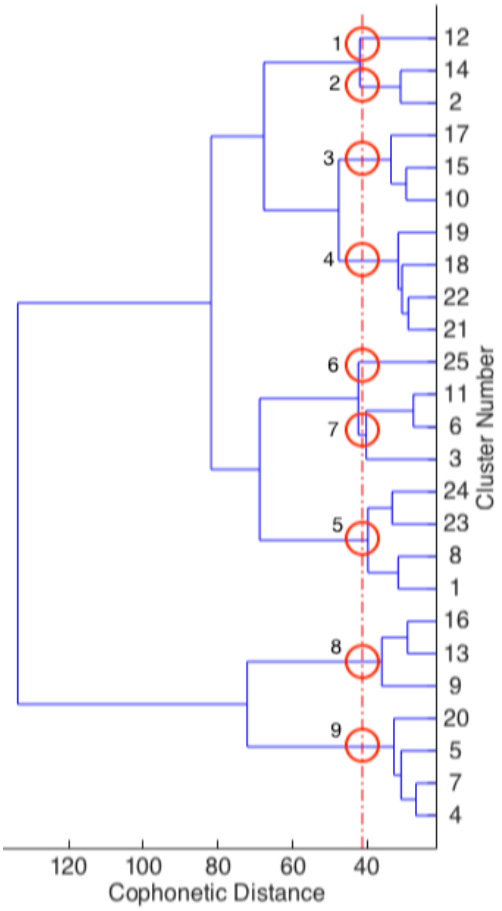

FIGURE 4.4: Dendrogram of arbitrary clusters

The dotted line represents the cut-off for the depth of analysis (9 clusters)

4.3.2 Hierarchical Clustering Results

Having applied agglomerative hierarchical clustering to the reduced dataset, the resultant dendrogram can be seen in Figure 4.4. The dotted line represents the cut-off for depth analysis, chosen based on the result that the optimal choice of clusters is 9.

The results of the pruned decision trees are presented in Figure 4.5. Each node point identified the specific audio feature that provides the best split in the data, to create the structure as presented in Figure 4.4.

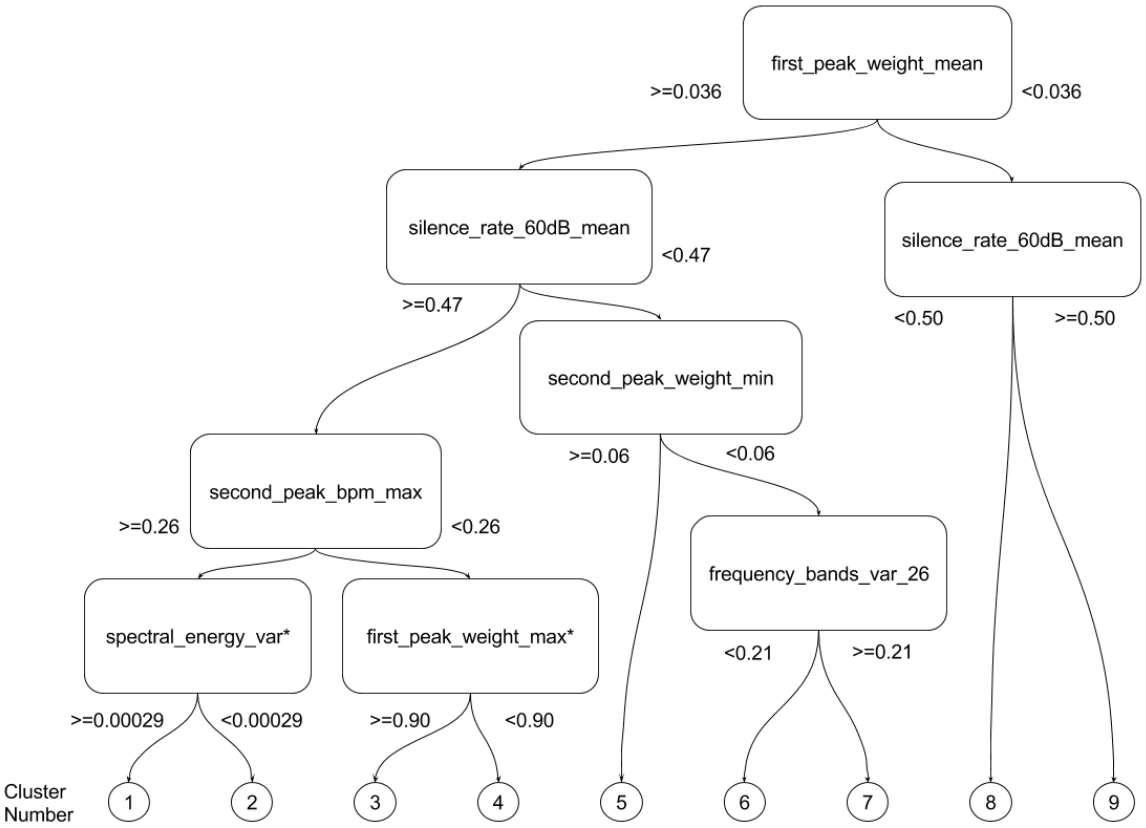

FIGURE 4.5: Machine learned structure of sound effects library, where clusters are hierarchical clusters.

The single audio feature contributing to the separation is used as the node point,with normalised audio feature values down each branch to understand the impact the audio feature has on the sound classification. The $*$ represents a feature separation where the classification accuracy is less than 80%, never less than 75%.

4.3.3 Sound Effects Taxonomy Result

The audio features used for classification were related to their semantic meanings by manual inspection of the audio features used and the feature definitions. This is presented in Figure 4.6. As can be seen, the two key factors that make a difference to the clustering are periodicity and dynamic range.

Periodicity is calculated as the relative weight of the tallest peak in the beat histogram. Therefore strongly periodic signals have a much higher relative peak weight than random signals, which is expected to have near-flat beat histograms. Dynamic range is represented by the ratio of analysis frames under 60dB to the number over 60dB as all audio samples were loudness normalised and all leading and trailing silence was removed, as discussed in Section 4.2.2. Further down the taxonomy, it is clear that periodicity

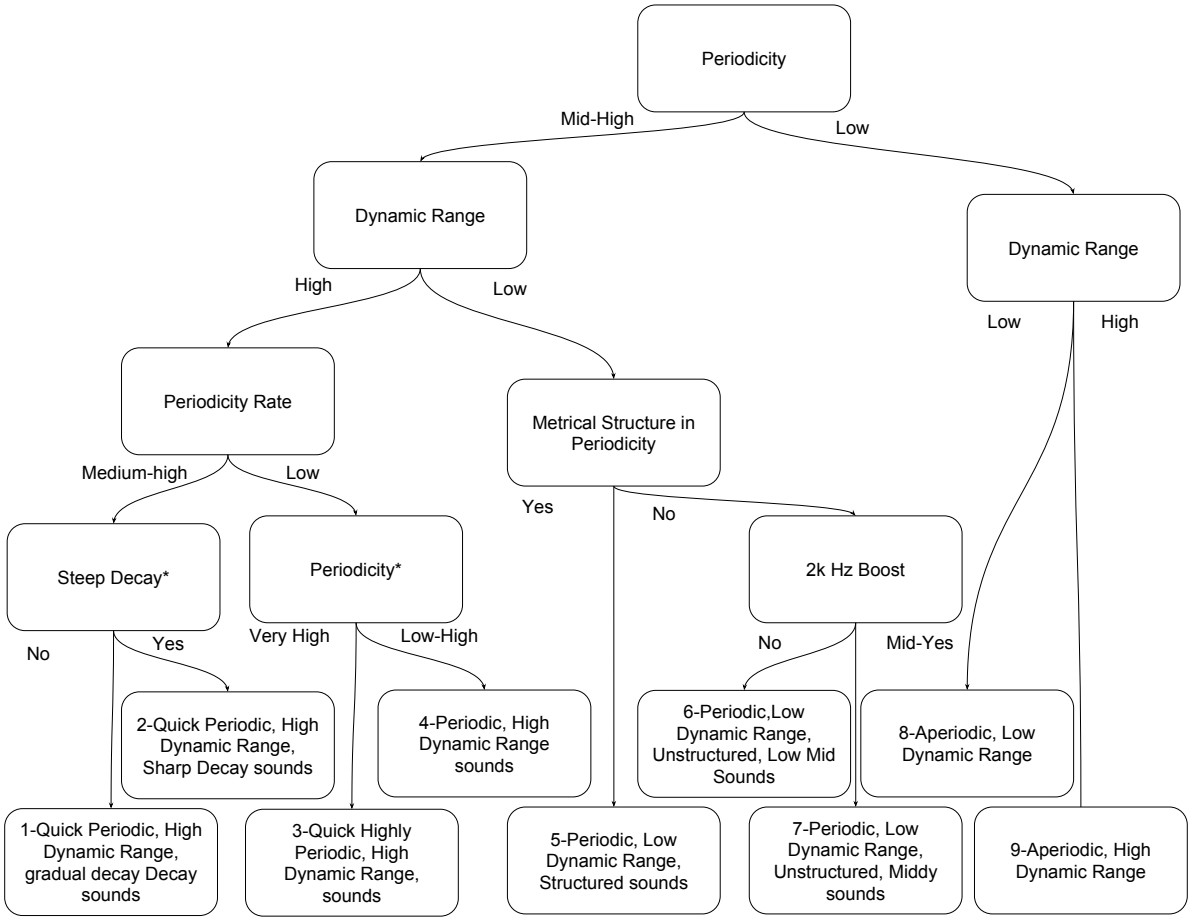

FIGURE 4.6: Interpretable machine learned taxonomy

Each node separation point is determined by hierarchical clustering and text within each node is an semantic interpretation of the most contributing audio feature for classification. Each final cluster is given a cluster number and a brief semantic description. The $*$ represents a feature separation where the classification accuracy is less than 80%, never less than 75%.

stands out as a key factor, in many different locations, along with the metric structure of periodicity, calculated as the weight of the second most prominent peak in the beat histogram. Structured music with beats and bars will have a high metrical structure, whereas single impulse beats or ticks will have a high beat histogram at one point but the rest of the histogram should look flat.

4.4 Evaluation

To evaluate the results of the produced sound effects taxonomy, as presented in Figure 4.6, the generated taxonomy was compared to the original sound effects library

classification scheme, as presented in Section 4.2.1. The purpose of this is to produce a better understanding of the resulting classifications, and how it compares to more traditional sound effects library classifications. It is not expected that these clusters will appropriately represent any pre-existing data clusters, but that it may give us a better insight into the representation of the data.

Each of the 9 clusters identified in Figures 4.5 and 4.6 were evaluated by comparing the original classification labels found in Table 4.1 to the new classification structure. This is presented in Figures 4.7–4.15, where each cluster has a pie chart representing the distribution of original labels from the dataset. Only labels that make up more than 5% of the dataset were plotted. Each of the legend items are plotted alphabetically, clockwise from the centre top of the figure.

In cluster 1 (Figure 4.7), which has quick, periodic, high dynamic range sounds with a gradual decay, the majority of the results are from a range of production elements which are highly reverberant repetitive sounds, such as slide transition sounds. Many of these sounds are artificial or reverberant in nature, which follows the intuition of the cluster identification.

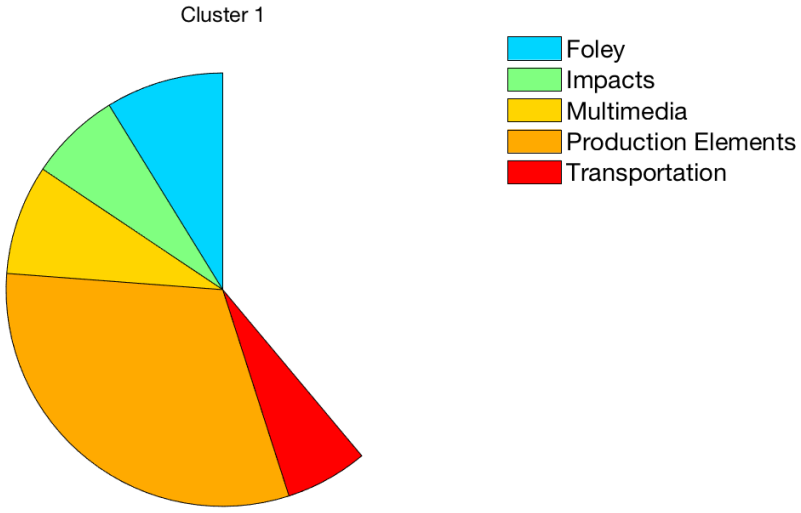

FIGURE 4.7: Dataset labels of cluster 1

Cluster 2 (Figure 4.8), with quick periodic highly dynamic sharp sounds contains a combination of Foley sounds and water-splashing sounds. These sounds are somewhat periodic, such as lapping water, but do not have the same decay as in cluster 1.

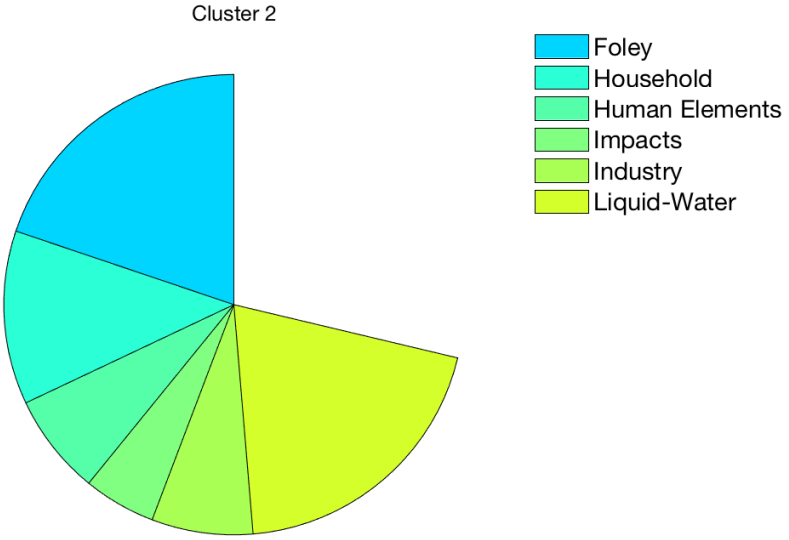

FIGURE 4.8: Dataset labels of cluster 2

Cluster 3 (Figure 4.9) is very mixed. The cluster is made up of quick, highly periodic, high dynamic range sounds. Impacts, household sounds and Foley make up the largest parts of the dataset, but there is also contribution from crashes, production elements and weapon sounds. It is clear from the distribution of sounds that this cluster contains mostly impactful sounds. It is also evident that a range of impactful sounds from across the sound effects library have been grouped together.

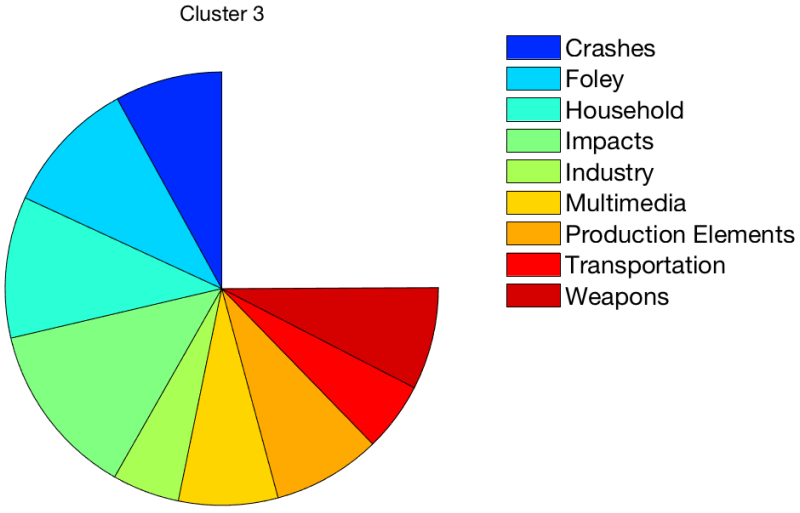

FIGURE 4.9: Dataset labels of cluster 4

In cluster 4 (Figure 4.10), with highly periodic sounds with a high dynamic range, most of the samples are from the production elements label. These elements are moderately

periodic at a high rate, such as clicking and whooshing elements, which are also similar to the next category of multimedia.

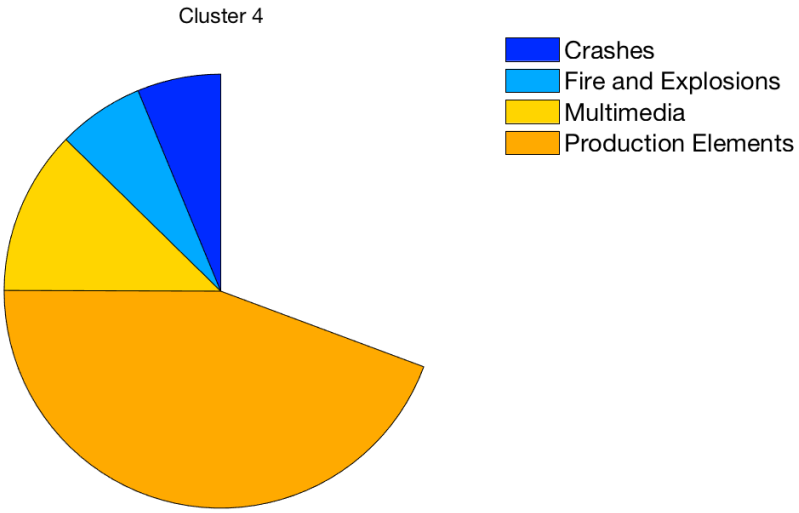

FIGURE 4.10: Dataset labels of cluster 4

Cluster 5 (Figure 4.11) contains sounds that are periodic, structured sounds with a low dynamic range. A spread of sound labels is included in this cluster, which includes transport and production elements as the two largest components. In particular, the transport sounds will be a periodic repetition of engine noises or vehicles passing, while remaining at a consistent volume.

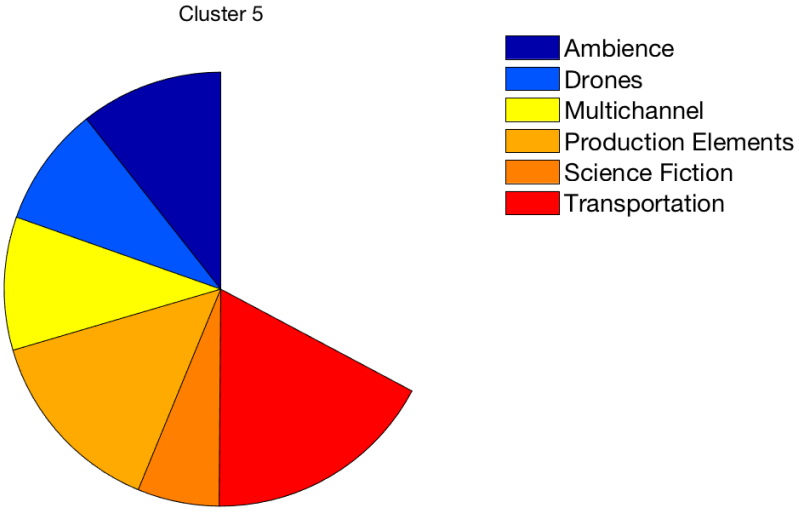

FIGURE 4.11: Dataset labels of cluster 5

Cluster 6 (Figure 4.12) contains sounds which have a low dynamic range, periodic, unstructured sounds. There is a large range of labels within cluster 6. The three

most prominent are human, multimedia and production elements, though cartoon and emergency sounds also contribute to this cluster. Human elements are primarily speech sounds, so the idea that periodic sounds that do not have a lot of high mid seems suitable, as the human voice fundamental frequency is usually between 90Hz and 300Hz.

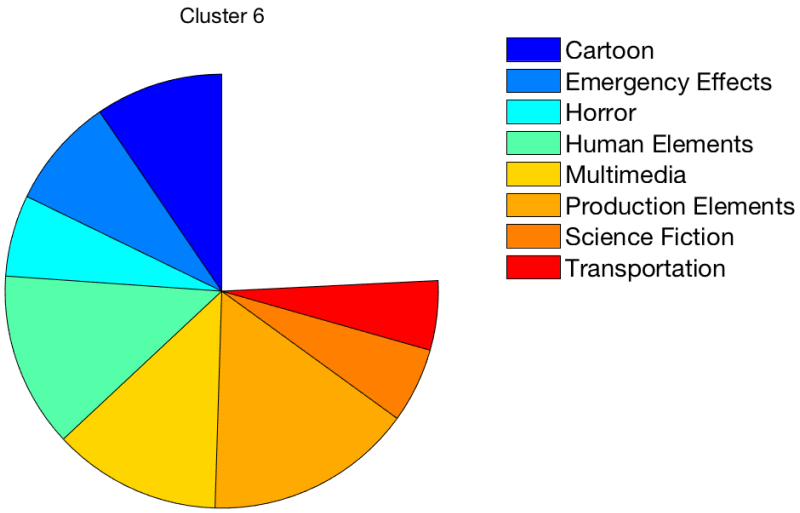

FIGURE 4.12: Dataset labels of cluster 6

Cluster 7 (Figure 4.13) is entirely represented by the science fiction label. This is made up of periodic, unstructured sounds with a low dynamic range and a large emphasis on the high mid frequency range. These fairly repetitive, constant volume sounds have an unnaturally large amount of high mid frequency around the 2kHz range.

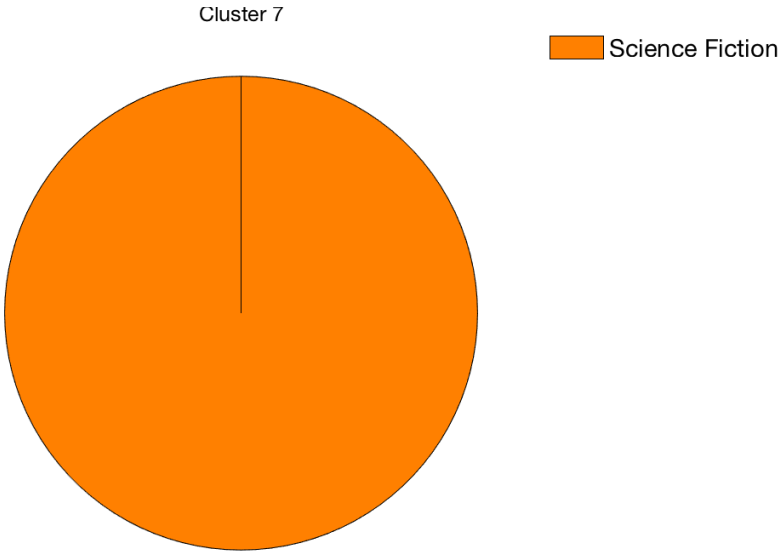

FIGURE 4.13: Dataset labels of cluster 7

Within cluster 8 (Figure 4.14) there are aperiodic low dynamic range sounds. The largest group of samples in this cluster is multimedia, which consists of whooshes and swipe sounds. These are aperiodic, and the artificial nature suggests a long reverb tail or echo. A low dynamic range suggests that the samples are consistent in loudness, with very few transients.

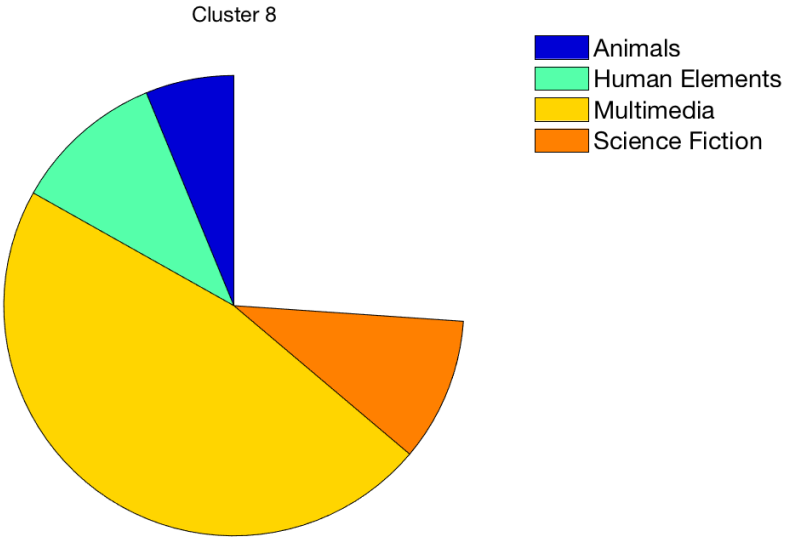

FIGURE 4.14: Dataset labels of cluster 8

Finally, cluster 9 (Figure 4.15) consists of a range of aperiodic impactful sounds with a high dynamic range from the impact, Foley, multimedia and weapon categories.

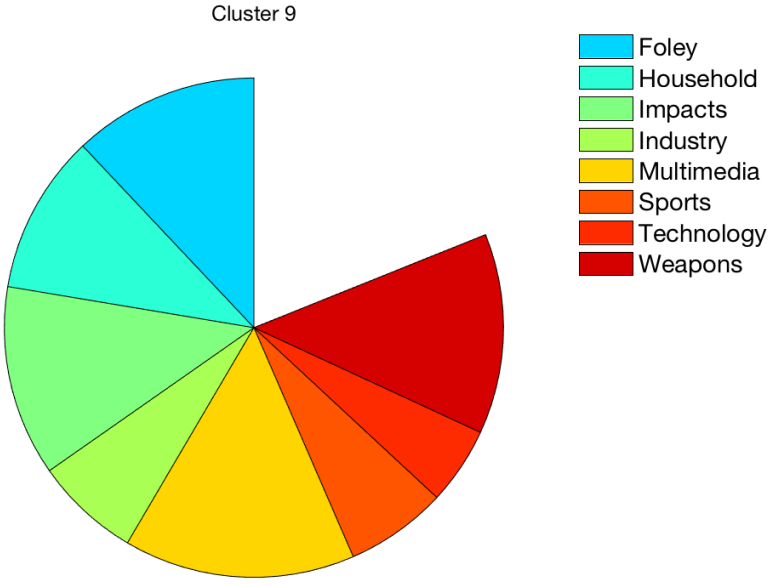

FIGURE 4.15: Dataset labels of cluster 9

4.5 Discussion

The nine inferred clusters were compared to the 29 original source based labels. It is clear that some clusters relate to intuition, and that this structure may aid a sound designer and present a suitable method for finding sounds, such as impactful sounds in cluster 9. Despite this, there are some clusters that do not make intuitive sense, or are difficult to fully interpret. It is suspected that this is due to the depth of analysis on the dataset. Despite the GMM predicting 9 clusters within the data, it is believed that a greater depth of analysis and clustering could aid in providing more meaningful, interpretable results, as many of the clusters are currently too large. The target of nine clusters was used within this work, as that was the optimal number of clusters suggested by the GMM approach, and any other depth level of study would have been an arbitrary selected number of clusters.

As can be seen from Figure 4.6 and discussed in Section 4.3, dynamic range and periodic structure are the key factors that create separations in this dataset. It is surprising that no timbral attributes and only one spectral attribute appears in the top features for classification within the dataset, and that seven of the eight features are time domain features.

Cluster 7 was described entirely as 'Science Fiction' in Section 4.4. This set of sound effects is entirely artificial, created using synthesisers and audio production. The grouping using this audio feature is most likely an artefact of the artificial nature of the samples and the fact they all come from a single source. This is also caused by the analysis and evaluation of a single produced sound effects library. This artefact may be avoided with a large range of sound effects from different sources.

Section 4.4 shows that the current classification system for sound effects may not be ideal, especially since expert sound designers often know what sonic attributes they wish to obtain. This is one of the reasons that audio search tools have become so prominent, yet many audio search tools only work using tag metadata and not the sonic attributes of the audio files. As such, by considering the sonic elements of the audio files, it has

been possible to form a taxonomy that separates audio into clusters based entirely on their sonic properties.

The produced taxonomy is very different from current work. As presented in Section 2.4, most literature bases a taxonomy on either audio source, environmental context or subjective ratings. This alternative taxonomy can provide a new approach for searching though groups of sounds that are related only to the sonic properties of the sounds. The hope is that this could provide variety in searching for specific sound effects by focusing on sonic attributes of the sound sample, rather than associated tags.

4.6 Conclusion

Given a commercial sound effects library, a taxonomy of sound effects has been learned using unsupervised learning techniques.

At the first level, a hierarchical structure of the data was extracted and presented in Figure 4.4. Following from this, a decision tree was created and pruned, to allow for visualisation of the data, as in Figure 4.5. Finally a semantically relevant context was applied to data, to produce a meaningful taxonomy of sound effects which is presented in Figure 4.6. A semantic relationship between different sonic clusters was identified.

The hierarchical clusters of the data provide deeper understanding of the separating attributes of sound effects, and gives us an insight into relevant audio features for sound effects classification. The importance of the periodicity, dynamic range and spectral features for classification is demonstrated. It should be noted that although the entire classification was performed in an unsupervised manner, there is still a perceptual relevance to the results and there is a level of intuition provided by the decision tree and the presented semantic descriptors. Furthermore, the clustering and structure will be heavily reliant on the sound effects library used.

It has also been demonstrated that current sound effects classification and taxonomies may not be ideal for their purpose. They are both non-standard and often place sonically similar sounds in very different categories, as demonstrated in Section 4.3. This

new approach could potentially afford new approaches for a sound designer to find an appropriate sound. This work proposes a new direction for producing new sound effects taxonomies based purely on the sonic content of the samples, rather than source or context metadata. As such, a natural extension of this work would be to scale this investigation up to larger sound effects libraries.

This work requires some user experience led development to produce an interactive system to allow users to search through large systems of sound effects. This approach has the potential to be a disruptive technology [Bower and Christensen, 1995]. Regardless of whether the current sound design practitioners recognise a direct need for new technology, there is the potential for developments to have considerable impact on the field. The uptake of technology in sound design may well grow as technological innovation and advancements develop further, and the technology has been demonstrated to be effective.

Chapter 5

Subjective Evaluation of Synthesised Sound Effects

5.1 Introduction

This chapter proposes to evaluate sounds produced by a synthesis system and compare them against recorded samples, in the same contextual environment. This facilitates direct comparison and helps establish if a particular synthesis method can be considered indistinguishable from a recording of the original sound. In this context, there may be instances where a synthesis method would be beneficial for use in a professional capacity, since there are typically more direct ways to control the sonic properties of a synthesis method than of a sample. The synthesis methods used will inevitably rely heavily on the sonic properties of the sound effect being represented. This builds on the previous chapter on investigating sonic properties and their ability to group sounds together, and investigates the extent to which realistic sound effects can be computationally generated.

The ability to produce realistic, real-time synthesised sounds is considered a challenging [Miner and Caudell, 2005] and unsolved problem [Caramiaux *et al.*, 2014]. This work aims to highlight the shortcomings of current research and to provide insight into which synthesis methods are most effective given a specific context. Through better understanding of the subjective realism of a range of synthesis methods, on a range of

different sounds, the intention is to highlight particular sound classes or contexts that would benefit from further work. Section 2.2 presented the range of synthesis methods to be evaluated. The listening test set-up is presented in Section 5.2 and the results presented in Section 5.3. An evaluation of the results and discussion of the impact of these results is presented in Section 5.4. Section 5.5 will present conclusions and discussion of this work.

5.2 Experimental Method

5.2.1 Participants

Eighteen participants between the ages of 18 and 40 took part in the experiment, of which 11 were male and 7 female. The procedure was approved by the local ethics committee. The average test duration was 17.5 minutes, so fatigue was not an issue Schatz *et al.* [2012].

5.2.2 Experimental Setup

The experiment took place in a dedicated, professionally acoustically treated listening room at Queen Mary University of London. The audio was played back over a pair of PMC AML2 loudspeakers, where the participant could adjust the volume of the audio to a comfortable level. Participants were asked to set the volume during the first test, and then refrain from adjusting it during the remainder of the test. No participant moved the volume more than 3dB from its starting position, so this effect is considered negligible. Background noise was measured to be below 40dB SPL (Sound Pressure Level) for all frequencies below 100Hz on a callibrated SPL meter.

5.2.3 Materials

Six different synthesis methods were used to synthesise a range of different sound effects. The synthesis methods were selected to represent a large range of published work in the

field, as presented in Section 2.2 (additive, concatenative, sinusoidal modelling, physically inspired synthesis and statistical modelling and marginal statistics). Evaluation of physical models is a difficult task and is specific to the type of physical modelling. The complex nature and detail of some physical models makes it challenging to compare these to more general sample based or signal based synthesis methods. As such, evaluation of physical models is beyond the scope of this work, but physically inspired synthesis will be used for evaluation purposes.

Participants were asked to evaluate sound textures for eight categories (applause, babble, bees, fire, rain, stream, waves, wind). These textures comprise a large range of sounds that have been used for sound synthesis evaluation in existing work [McDermott and Simoncelli, 2011; Schwarz *et al.*, 2016]. They represent composite scenes containing a range of different timbres of sounds. The long term evolution and structure of the sound are as important contributing factors as the timbre of each individual sonic element within the complex scenes. Thus any synthesis method should model the temporal development of the sound along with the instantaneous qualities. In particular, the applause and babble sounds were selected as they are known to be challenging sounds to reproduce, and may test sound synthesis methods to the limit of their capabilities.

In every category between six and eleven samples were provided. Sixty-six samples were evaluated in total. All samples were 44.1kHz audio files, that were loudness normalised in accordance with ITU-R BS.1387-1 [1998]. Each category had at least one anchor and at least one recorded sample. The recorded samples were all selected by a group of five experienced critical listeners as being realistic samples, given at least 5 different sample options all selected from a professional sound effects library. Each anchor was constructed from a trivial additive synthesis model, produced by deconstructing either the additive or physically inspired model to the point that it was barely perceivable as the intended sound. Anchors were not used from the MUSHRA standard (MUltiple Stimuli Hidden Reference and Anchor [ITU-R BS. 1534-1, 2001]), as the standard MUSHRA anchors are downsampled versions of the original recorded sample. MUSHRA was originally designed to evaluate audio encoding algorithms, but in the context of measuring realism, it was a concern that a downsampled real recording would potentially be more

TABLE 5.1: Synthesis method used to created each sound sample

Synthesis Method	Applause	Babble	Bees	Fire	Rain	Stream	Waves	Wind
Physically Inspired	N	N	Y	Y	Y	Y	N	Y
Marginal Statistics	Y	Y	N	Y	Y	Y	N	Y
Sinusoidal Modelling	Y	Y	Y	Y	Y	Y	N	N
Additive	N	N	N	Y	Y	N	Y	Y
Statistical Modelling	Y	Y	Y	Y	Y	Y	Y	Y
Concatenative	Y	Y	Y	Y	Y	Y	Y	Y

realistic than some of the synthesis methods. As such, a poor quality synthesis method was generated, to act as a suitable anchor.

The references and anchors were important within this test to encourage participants to use the entire evaluation scale, and to review how samples were distributed within that scale, in accordance with ITU-R BS.1534-3 [2015]. The reference samples allowed evaluation of how synthesis methods compared to the genuine sound, and to allow for identification as to whether the samples are distinguishable from the real sample. The purpose of the anchor sample was to support evaluation of how synthesis methods compared to each other. If every synthesis method was highly realistic, participants may decide to use the entire evaluation scale, to identify micro-differences between samples, or may decide to group all samples together at the high end of the scale. The anchor ensures that there is a lower limit sample to compare against. It also performs as a confirmation that a participant has fully understood the requirements for the experiment. If a participant rated the anchor as higher than the sample, then it would be inferred that the participant may not have fully understood the requirements, or may have some hearing defect.

A list of synthesis methods used within each sound class is presented in Table 5.1. To demonstrate the full range of reference sound samples, audio features were extracted from the samples using the Essentia toolbox [Bogdanov *et al.*, 2013] based on recommendations from Chapter 3, and in Moffat *et al.* [2015], and summarised attributes are presented in Table 5.2. All sound samples used, software implementations and parameter settings, are available online.[1]

[1] `https://code.soundsoftware.ac.uk/projects/perceptual-evaluation-of-sound-synthesis`

TABLE 5.2: Summary of attributes of different sounds classes used for evaluation

	Environmental	Animal/Human	Synchronised	Noisy	Harmonic	Granular
Applause	N	Y	Y	Y	N	Y
Babble	N	Y	N	N	Y	Y
Bees	N	Y	N	N	Y	N
Fire	Y	N	N	Y	Y	Y
Rain	Y	N	N	Y	Y	Y
Stream	Y	N	Y	N	Y	Y
Waves	Y	N	Y	Y	N	N
Wind	Y	N	Y	Y	N	N

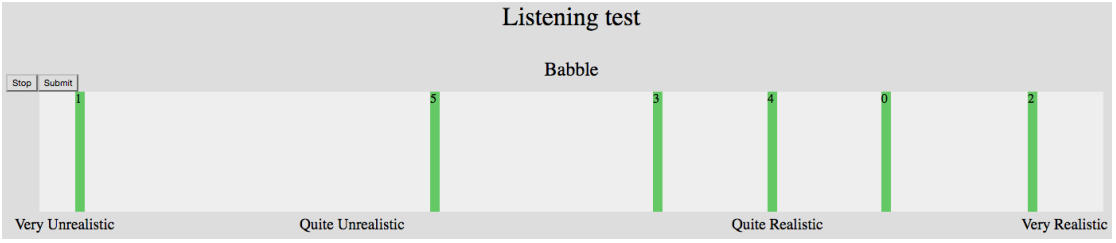

FIGURE 5.1: A screenshot of the user interface used by participants for inter-comparison of sound samples.

5.2.4 Web Audio Evaluation Tool

The listening test was set up using the Web Audio Evaluation Tool [Jillings *et al.*, 2015]. A screenshot of the user interface used for this experiment is presented in Figure 5.1. An online version of the listening test is available, with the same user interface and set of samples that were used by participants.[2]

5.2.5 Procedure

Participants were provided with instructions as to the experiment they were to undertake, and were asked whether they had previous experience of listening tests and whether they would consider themselves as accomplished musicians or audio engineers.

Participants were then asked to rate how realistic they perceived the samples within a given category, relative to all the other samples within that category. Participants were provided with a continuous linear scale on which to rate all sounds, labeled from "very unrealistic" to "very realistic". All sounds were rated on a single horizontal scale, to encourage inter-sample comparison. Participants were provided with the sound category

[2] `https://goo.gl/789eWl`

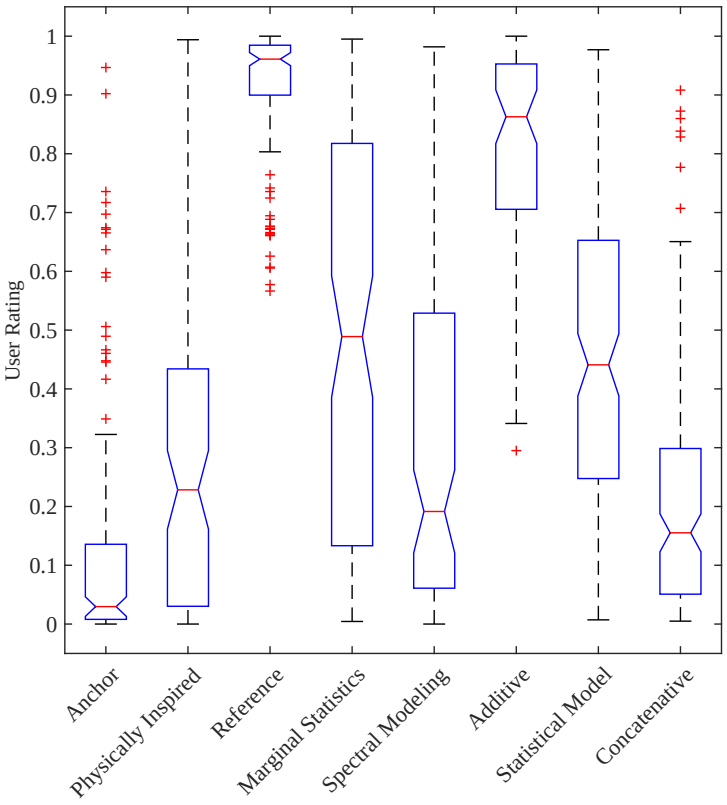

FIGURE 5.2: Plot of the median, standard deviation and 95% confidence intervals of all synthesis results.

TABLE 5.3: Mean and standard deviation of each sound class

Sound Class	Mean	Standard Deviation
Applause	0.40	0.37
Babble	0.35	0.33
Bees	0.44	0.35
Fire	0.38	0.32
Rain	0.40	0.37
Stream	0.39	0.36
Waves	0.49	0.37
Wind	0.57	0.32

name, but other than that, did not have any information regarding the samples. Both the ordering of categories and the initial ordering of samples within a category were randomised.

5.3 Results

The overall results for the experiment are presented in Figure 5.2 using a notched box plot. In all plots the red line represents the median. The end of the notches, where

the angled lines become parallel within the box plot, represents the 95% confidence intervals, and the end of the boxes represent the 1st and 3rd quartiles. The end of the whiskers represent the data range not considered as an outlier. Red crosses are outliers. The anchor and reference have very low and very high median values, respectively, with small confidence ranges. This informs us that the anchor and reference function as intended.

The null hypothesis is that the subjective evaluation scores are from the same distribution. A one way ANalysis Of VAriance (ANOVA), with Bonferroni correction, shows that for all sound classes, the effect of each synthesis method on user perception was statistically significant $F(7,946) = 176.51$, $p < 0.0001$. Table 5.4 shows the statistical significance of the difference in ratings between synthesis methods, for all sound samples. A post-hoc Tukey pairwise comparison, with Bonferroni correction to reduce the chance of type I errors, was used. It can be seen, for example, that concatenative synthesis is significantly different from the reference sample, marginal statistics, additive synthesis and statistical modelling all with a $p < 0.0001$. However, concatenative synthesis is not significantly different from the anchor, physically inspired synthesis or the sinusoidal modelling. These results are presented in more detail, broken down by sound class in Table 5.5.

Table 5.4: Results of pairwise comparison of synthesis method on subjective realism rating, with Bonferroni Correction

	Anch	PhISM	Ref	Marg	SMS	Add	Stat	Concat
Anch	.	**	****	****	***	****	****	o
PhISM	**	.	****	****	o	****	***	o
Ref	****	****	.	****	****	o	****	****
Marg	****	****	****	.	****	****	o	****
SMS	***	o	****	****	.	****	***	o
Add	****	****	o	****	****	.	****	****
Stat	****	***	****	o	***	****	.	****
Concat	o	o	****	****	o	****	****	.

o > 0.05, * < 0.05, ** < 0.01, *** < 0.001, **** < 0.0001, . = no comparison made.
Anch = Anchor, Ref = Reference, PhISM = Physically Inspired, Marg = Marginal Statistics, SMS = Sinusoidal Modelling, Add = Additive, Concat = Concatenative, Stat = Statistical Modelling.

The additive method performed best overall, and was the only synthesis method where the results were not significantly different from the reference. It was also significantly

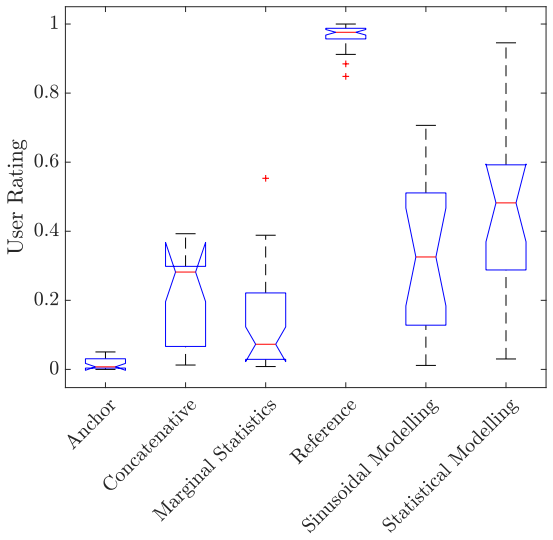

FIGURE 5.3: Applause result distribution

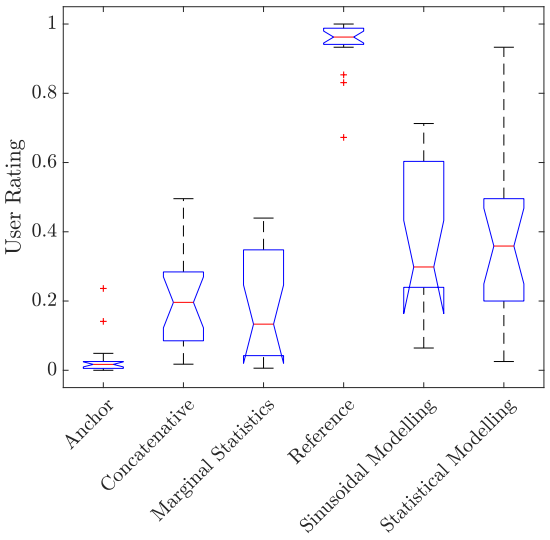

FIGURE 5.4: Babble result distribution

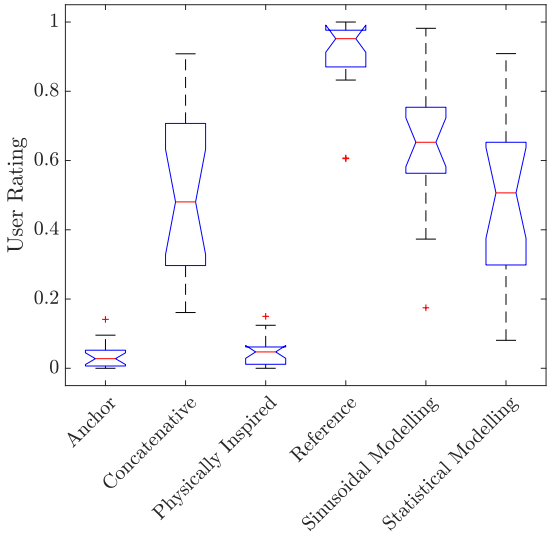

FIGURE 5.5: Bees result distribution

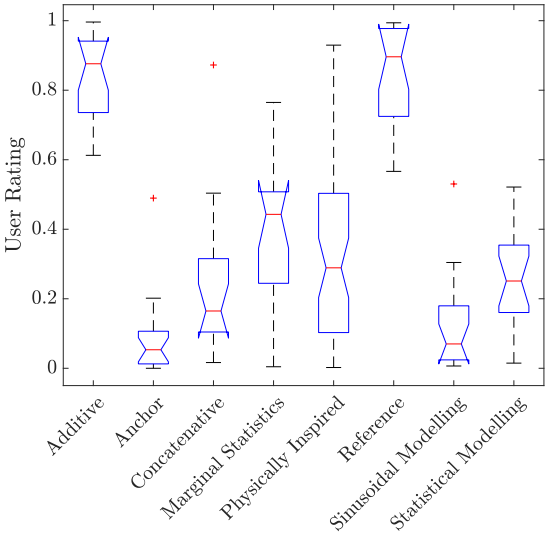

FIGURE 5.6: Fire result distribution

different from all other synthesis methods. However, this method was not used in all tests, as only a subset of sounds (fire, rain, wind and waves) could be synthesised using additive synthesis. Table 5.3 shows the mean and standard deviation of each sound class. With the exception of wind, there is little variation between the means of each sound class. This suggests that the superior performance of additive synthesis was not due to higher ratings for these sound classes, but instead, the synthesis method itself must have performed well.

Concatenative synthesis is the only method not significantly different from the provided anchor sounds. Table 5.4 shows that the different synthesis techniques can be broken

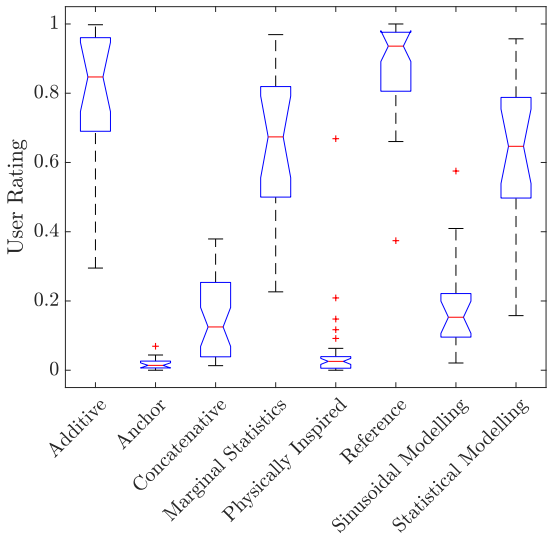

FIGURE 5.7: Rain result distribution

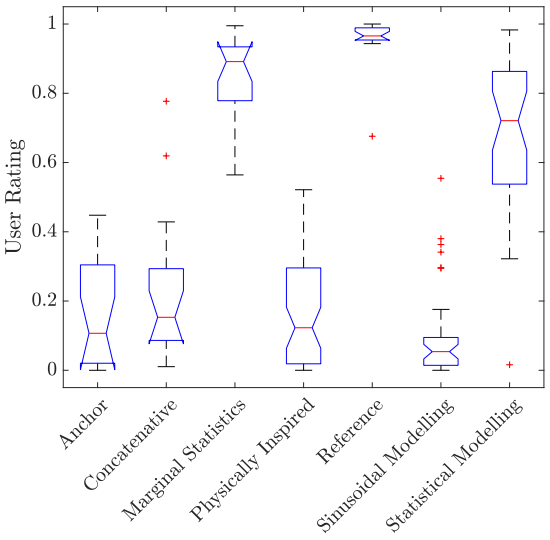

FIGURE 5.8: Stream result distribution

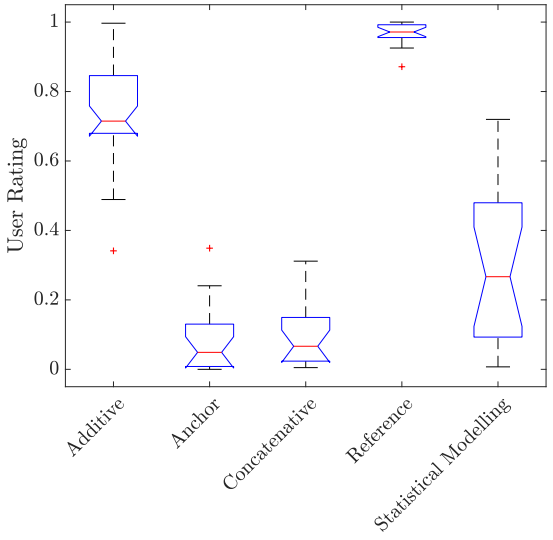

FIGURE 5.9: Waves result distribution

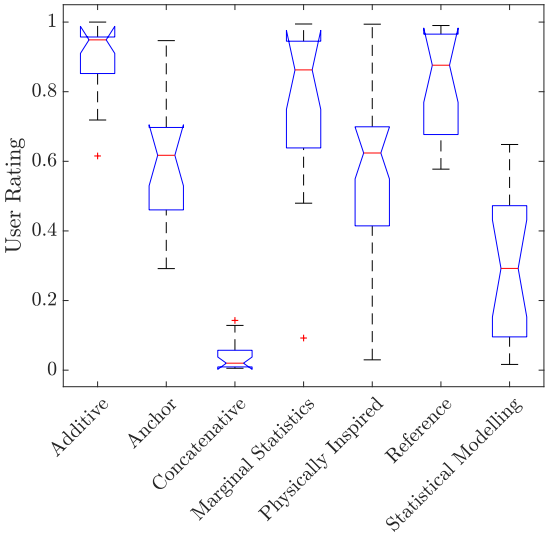

FIGURE 5.10: Wind result distribution

down into three subjective groupings, where Sinusoidal Modelling, Physically Inspired and Concatenative are all grouped together with the Anchor.

Statistical Modelling and Marginal Statistics can also be grouped together. This is to be expected, as they are based on the same implementation with different sets of synthesis statistics.

TABLE 5.5: Results of pairwise comparison of synthesis method on subjective realism rating for each class of sound, with Bonferroni Correction

Group 1	Group 2	Applause	Babble	Bees	Fire	Rain	Stream	Waves	Wind
Anch	Ref	****	****	****	****	****	****	****	o
Anch	PhISM	.	.	o	*	o	o	.	o
Anch	Marg	o	o	.	*	****	****	.	o
Anch	SMS	****	****	****	o	o	o	.	.
Anch	Add	.	.	.	****	****	.	****	**
Anch	Stat	****	***	****	o	****	****	*	***
Anch	Concat	o	o	****	o	o	o	o	****
Ref	PhISM	.	.	****	****	****	****	.	***
Ref	Marg	****	****	.	****	****	o	.	o
Ref	SMS	****	****	*	****	****	****	.	.
Ref	Add	.	.	.	o	o	.	**	o
Ref	Stat	****	****	****	****	***	***	****	****
Ref	Concat	****	****	****	****	****	****	****	****
PhISM	Marg	.	.	.	o	****	****	.	*
PhISM	SMS	.	.	****	o	o	o	.	.
PhISM	Add	.	.	.	****	****	.	.	****
PhISM	Stat	.	.	****	o	****	****	.	***
PhISM	Concat	.	.	****	o	o	o	.	****
Marg	SMS	***	o	.	o	****	****	.	.
Marg	Add	.	.	.	****	o	.	.	o
Marg	Stat	****	o	.	o	o	o	.	****
Marg	Concat	o	o	.	o	****	****	.	****
SMS	Add	.	.	.	****	****	****	.	.
SMS	Stat	o	o	o	o	****	.	.	.
SMS	Concat	o	o	o	o	o	o	.	.
Add	Stat	.	.	.	****	o	.	****	****
Add	Concat	.	.	.	****	****	.	****	****
Stat	Concat	***	o	o	o	****	****	o	o

o > 0.05, * < 0.05, ** < 0.01, *** < 0.001, **** < 0.0001, . = no comparison made.
Anch = Anchor, Ref = Reference, PhISM = Physically Inspired, Marg = Marginal Statistics, SMS = Sinusoidal Modelling, Add = Additive, Concat = Concatenative, Stat = Statistical Modelling.

5.3.1 Results Per Sound Class

An ANOVA with Bonferroni correction showed that for a given sound class the effect of each synthesis method on user perception was significant and in all cases $p < 0.0001$. A post-hoc Tukey pairwise comparisons shows the statistical significance of the differences between each synthesis methods, given each sound class, seen in Table 5.5. For all sound classes the anchor, the physically inspired model, the sinusoidal modelling, the statistical modelling and the concatenative synthesis all had subjective rating distributions that

were significantly different from the reference. Marginal statistics and additive were the only two synthesis methods under which there is sometimes no clear difference between their subjective rating distributions, given a specific sound class.

The median, standard deviation and 95% confidence intervals for each synthesis method, for each sound class, are reported in Figures 5.3-5.10. For all sounds except wind and stream sounds, the anchor had the lowest median rating, with small confidence intervals. For wind sounds, though sinusoidal modelling had a lower median rating, there is statistically no discernible difference in their distributions, and as such, it can be said they are equally poor. Wind is the only case where the anchor is not one of the worst samples selected. This suggests that the anchor may not have been ideal. However, the concatenative synthesis method produced a very low subjective rating with small confidence intervals, so the concatenative synthesis method can be considered as the anchor in this case. This is confirmed by the fact there is no significant difference between the anchor and the reference for wind.

In the case of synthesising wind, additive performed better than the reference sound. This is the only case where a synthesis method outperformed the reference recorded signal. The difference in distributions between additive and the reference is not significant. The null hypothesis was not rejected, and thus additive might be considered as realistic as a recording of wind, and possibly more realistic. In the case of fire and rain synthesis, additive could also be considered as realistic as a recorded reference sample, since the null hypothesis could not be rejected, and the confidence intervals are significantly overlapping.

A summary of the results are presented in Table 5.6, including summaries of the effectiveness of each synthesis method at producing the relevant sounds.

5.4 Discussion

The results of this subjective evaluation suggest that additive synthesis is an effective approach for environmental sounds such as fire, water and wind sounds. These sounds can be considered as sounds constructed from band-pass filtered noise. Marginal

TABLE 5.6: Rating of Synthesis Method per Sound Class

Synthesis Method	Applause	Babble	Bees	Fire	Rain	Stream	Waves	Wind
Physically Inspired	.	.	5	3	5	4	.	3
Marginal Statistics	4	5	.	3	3	**1**	.	**1**
Sinusoidal Modelling	3	3	**3**	5	4	5	.	.
Additive	.	.	.	**1**	**1**	.	**2**	**1**
Statistical Modelling	**3**	**3**	4	4	3	2	3	4
Concatenative	3	3	4	4	4	4	4	5

1 = Best Method, Comparable with Reference, 5 = Worst Method, Comparable with Anchor, . = No Comparison Made. The best score for each sound class in highlighted in bold.

Statistics are effective for synthesising wind and stream type sounds. For applause and babble sounds, which are more dynamic and impulsive, the statistical modelling synthesis proved to be the most effective approach in synthesising these types of sounds. As can be seen in Table 5.5, wind and stream sound synthesis can effectively be produced with marginal statistical synthesis, in such a manner that the realism rating distribution is not significantly different from that of the reference sounds. Despite this, it is noted that marginal statistics are "sufficient to produce compelling synthetic examples of many water textures (rain, streams etc.), but not much else" [McDermott *et al.*, 2009]. McDermott *et al.* [2009] suggests this applies to all sounds that are based around filtered noise signals, where sounds are primarily made up of noisy audio signals with little harmonic component. As such, water and wind sounds are all effectively synthesised using the Marginal Statistical method for synthesis.

In the case of the wind sounds, the additive synthesis method performed better than the reference sample. However, the difference was not considered to be significant, so this may be a statistical abnormality. This may also be an indication of hyper-realism. The idea of hyper-realism is simply that an unreal sound can sound 'more real' than a real sound. This is particularly prominent in weapon and explosion sounds [Mengual *et al.*, 2016], where a listener may never have heard a real gunshot sound, but will have a strong opinion of a gun sound based on TV, film and video games [Puronas, 2014].

Concatenative synthesis created noticeable artifacts in some of the samples. The artefacts seem to be caused by non-smooth transitions between frames, but are only perceivable in a small number of sound contexts. This caused this synthesis method to under-perform in certain cases, particularly rain and fire sounds and to a lesser extent,

babble. These are impulsive sounds, where the individual sonic elements may be smaller than an individual grain of sound, with variable size.

The sinusoidal modelling method also caused some audible artefacts, particularly in the fire and babble sounds. It is suspected that this was caused by spectral peaks being modelled as harmonic components, when they are actually noisy spectral peaks. There also appears to have been issues with phase recognition, which again is due to noisy signal components being modelled as harmonic components resulting in an audible vocoder-like effect.

Many of the physically inspired models were taken from Farnell [2010], which is designed as a textbook for teaching the principles of procedural audio. Thus, these sound synthesis models were designed for their sonic interaction capabilities rather than the exact replication of realistic sounds. These sound synthesis models did not produce convincing sounds. Despite this, it is considered important to evaluate this range of algorithms, as they are popular, well known synthesis methods.

The results show that additive synthesis is an effective synthesis method for both slow moving and impulsive sounds. Despite this, additive synthesis allows for a very large range of possible parameters, and the individual parameter ranges were manually selected by the original authors. As such, this sound synthesis method cannot easily be generalised to a large number of sounds.

Particularly for slow changing sounds, statistical synthesis is effective, either using a reduced feature set, or the full feature set. It is speculated that a granular synthesis method may be most effective for impulse sounds, due to the fact that these sound textures are generally made up of a large number of small sound atoms eg. individual plosives in babble, claps within applause or raindrops in rain.

No synthesis technique was capable of producing convincing applause or babble sound. This was expected, as these sounds are known for being challenging to synthesise. However the sinusoidal modelling and statistical modelling performed relatively well on these sounds. This suggests that noise components are important in the reproduction of a realistic applause or babble sound, since statistical modelling and sinusoidal modelling

involve careful noise shaping. Additive synthesis produced realistic sounding examples of fire and rain sounds. This may be because the method focuses on synthesising individual sonic elements separately and then constructing a composite scene from these elements, rather than alternative methods, such as statistical modelling, which models the statistics of the entire sounds. In particularly composite scenes, such as fire and rain, the individual sonic element synthesis is more important than overall sonic structure.

5.5 Conclusions

An experiment in which participants were asked to rate 66 examples of synthesised sounds from eight different sound classes and five different synthesis methods, in terms of their perceived 'realism' was presented. The results demonstrate that, in some cases, sound synthesis methods can be as convincing as a recorded audio sample. However, in the case of wind, the users consistently rated the sound as more realistic than the recorded sample. In five of the eight sound classes tested, there exist synthesis techniques where synthesised sounds were indistinguishable, in terms of realism, from recorded samples.

This experiment presents a method for evaluation of synthesised sounds in a range of different sound classes and provides recommendations for synthesising different types of sounds. It is clear that although sound synthesis can effectively synthesise a range of realistic sounds, there are many potential future directions for development of plausible sound synthesis across the full sonic range.

Despite this, there are limitations of the work presented. Only a relatively small number of sound synthesis methods were evaluated, and as such substantial claims cannot be made about entire areas of synthesis research. There is a requirement for further work in comparing and evaluating more synthesis techniques.

The need for evaluation and further development of sound synthesis was clearly identified. Evaluation of sound synthesis can assist in improving upon the state-of-the-art and developing future sound synthesis. A clear and rigorous method for evaluation of sound synthesis was presented, through a double blind multiple comparison evaluation

test. This test methodology can be used to evaluate any sounds synthesis method, to determine the perceived realism of the synthesised sound, given a single word or phrase context.

Chapter 6

Objective Evaluation Metric for Synthesised Environmental Sounds

6.1 Introduction

Following on from the subjective evaluation technique presented in the previous chapter, this chapter considers ways to improve synthesised sound effect objective evaluation methods. Taking the audio feature representation and analysis presented in Chapter 4, this chapter will compare a set of different audio feature sets to determine which approach correlates the best with human perception. The perception of these sound effects will be compared in a method similar to the work presented in Chapter 5. The aim of this is to provide a consistent validated objective evaluation metric. Through improved standardised objective evaluation, a greater consistency within the evaluation methods for sound synthesis can be produced, without the necessity for intensive and difficult to perform subjective evaluations. A comparison of sound similarity measures, through resynthesis, is proposed. The aim is to identify an objective measure that can encapsulate the perceptual similarity of sounds. Optimisation of this measure would then select appropriate parameters for a synthesis engine to match a given sound. Optimisation

of synthesis parameters to evaluation of sound perception has been previously demonstrated [McDermott and Simoncelli, 2011]. Parameter selection can be viewed as an optimisation problem in which synthesis parameters are dimensions through a fitness landscape. In many cases, search spaces are highly nonlinear, and thus evolutionary optimisation functions are effective methods to use [Garcia, 2001b; McDermott *et al.*, 2008; Yee-King and Roth, 2011].

The objective metrics and evaluation framework will be presented in Section 6.3. The subjective listening test is presented in Section 6.4. Results of the subjective and objective measures are given in Section 6.5. Recommendations for objective synthesis evaluation metrics are presented in Section 6.6, and final comments and outline of impact in the community are presented in Section 6.7.

6.2 Parameter Optimisation Background

There have been a number of approaches to searching audio parameter spaces, within a synthesised environment. An iterative process to control parameters and minimise a set of perceptually motivated audio features was developed by McDermott *et al.* [2011] and McDermott and Simoncelli [2011]. The results were subjectively evaluated based on participants identification and synthesis realism. Further approaches using genetic algorithms that have attempted to modify musical parameters based on varying fitness functions were used. No other method performed any formal evaluation of the synthesis results, typically reporting their final distance measure. Fitness function methods are typically calculated as distances features such as between Mel Frequency Cepstrum Coefficients (MFCCs) [Yee-King and Roth, 2011], the Discrete Cosine Transform of the MFCCs [Heise *et al.*, 2009]. The Perceptual Evaluation of Audio Quality (PEAQ Thiede *et al.* [2000]) distances were measured for piano string synthesis [Hamadicharef and Ifeachor, 2003, 2005], where as the distance between Least Square Error(LSE) of time domain waveform, LSE of spectrograms and LSE of spectrograms with some masking weighting were all used as distance measures [Garcia, 2001a,b]. McDermott *et al.* [2008] used sets of different audio features to measure distances.

6.3 Objective Measure Through Synthesis

In this section, the methodology of evaluating a range of objective measures will be presented. The principal is that evaluation of different objective measures can be compared through resynthesis. By using the objective measure as a fitness function in an iterative synthesis process, the most effective measure that best encapsulates aspects of the perception of the sounds can be identified. The aim of the optimisation is to interpret which objective similarity measure correlates with the perceptual similarity, evaluated through a listening test. Essentially, this will be used for comparison of different objective similarity measures, by producing audio samples. Through subjective listening tests, the results of the objective measures can be validated and compared.

6.3.1 Parameter Optimisation

Given a specific synthesis method and a recording of a sound, the parameters of that synthesis model were selected to best reflect an original audio sample. This parameter selection was performed using a particle swarm optimisation (PSO) technique. PSO is an evolutionary inspired, population based, optimisation technique in which a swarm of particles iteratively propagate in a search space, where a weighting between individual and global preferences is modelled. Each particle is evaluated with a 'fitness' function, which, in this case, is a computational objective similarity measure. This fitness function was used to compare each objective function presented in Section 6.3.3, given a range of audio samples.

Given a single input audio file, and a specific synthesis method, an iterative approach was undertaken. The aim of this iterative approach was to find the set of synthesis parameters that match the closest possible sound the synthesis method can make to the original input audio file.

This is demonstrated in Figure 6.1. An initial input audio file is analysed, and this is used to measure the similarity between a synthesised example and the input audio file. The synthesis engine parameters were initialised at random values, so the sample is likely to be very different. The PSO algorithm then makes suggestions for how to modify the

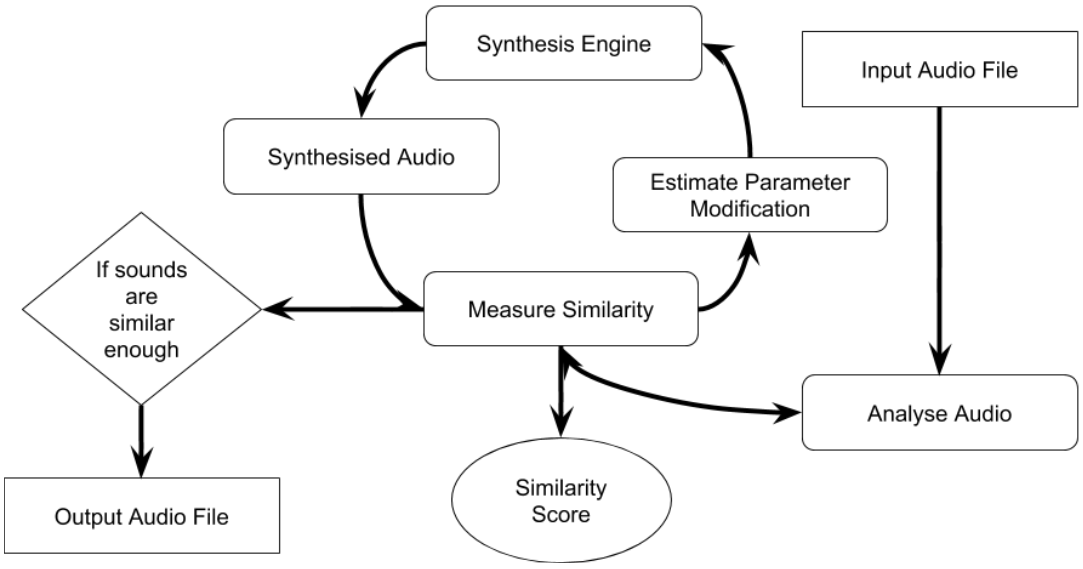

FIGURE 6.1: Flow Diagram of Synthesis Optimisation Approach

synthesis parameters, which are passed to the synthesis engine. The new audio sample is then compared, using the objective similarity measure or fitness function, to compare how similar the sounds are. Once the PSO algorithm believes that a global minima has been found, the algorithm stops, and reports the best synthesis parameters and the output audio example.

PSO is a method of optimisation, inspired by evolutionary biology and social behaviour of a flock of birds or swarm of insects, and is an effective optimisation method for highly nonlinear search spaces. There are many examples of evolutionary algorithms applied to audio research [Garcia, 2001b; Mäkinen *et al.*, 2012; McDermott *et al.*, 2008; Ronan *et al.*, 2018; Yee-King and Roth, 2011]. PSO works on the basis of a large number of particles, each of which is a potential solution, which can sample the search space. Each particle can then evaluate its effectiveness and both the local and global results are considered when directing the particle in which direction to move. A comprehensive overview of PSO is presented by Marini and Walczak [2015].

6.3.2 Sound Synthesis Methods

Four different sound effects were used for evaluation purposes. All of them are available and hosted online as part of the FXive synthesis platform [Bahadoran *et al.*, 2017,

2018a].[1] The FXive platform is a free online hosted web service for sound synthesis, created at Queen Mary University. This system was used, as it contains a large number of synthesis methods which are all openly accessible through a simple URL (Uniform Resource Locator) interface, which allowed for a simple and standard method for interaction. Each of the selected synthesis methods have a limited number of control parameters, and are all designed to produce a sound of a specific type. All approaches used within this chapter were originally derived from work by Farnell [2010] and are all examples of physically inspired synthesis.

Fire The fire synthesis model[2] is a noise shaping synthesis method. Individual sonic components of a fire, the hiss, crackle and lapping, are all modelled through filtered and envelope shaped noise signals. Three control parameters are exposed to the user, which are *lapping*, *hissing* and *crackling*.

Rain In the rain model,[3] components of rain are broken into a number of categories. Ambience, which is modelled as constant shaped noise, droplets, rumble and drips. Three control parameters are exposed to the user, which are *density*, *rumble* and *ambience*.

Stream The stream[4] is modelled entirely on the bubbling sounds that are made as water runs over substances, based on control of filtered chirp sounds. Three control parameters are exposed to the user, which are *bubbles*, *frequency* and *filter Q*.

Wind The wind model[5] uses a varying filtered noise approach, where wind parameters control the overall envelope of the sound. Different wind hitting materials, such as door or branches/wires, select the timesteps over which the wind envelope shaping will occur. Ten parameters are exposed to the user: *Wind Speed*, *Gustiness*, *Squall*, *Buildings*, *Doorways*, *Branches*, *Leaves*, *Pan*, *Directionality* and *Gain*. The parameters *Pan*, *Directionality* and *Gain* were all left constant at their default values, as discussed in Section 6.3.2.

[1] https://fxive.com/
[2] https://fxive.com/app/main-panel/fire.html
[3] https://fxive.com/app/main-panel/rain.html
[4] https://fxive.com/app/main-panel/stream.html
[5] https://fxive.com/app/main-panel/wind.html

Parameters Not Changed

Several parameters were not used, to limit the search space and as these parameters were considered to make no immediate impact to the synthesis of the sound. During analysis, all samples were loudness normalised, so output gain controls were redundant. As no evaluation metric used spatial aspects to evaluate synthesis, pan controls were also not considered. With each sound effect, there was the ability to apply a range of audio effects, including equalisation, distortion, delay, convolution reverb and HRTF spatialisation. However, because all of these controls can be added to every single synthesised sample, it was felt this would significantly grow the search space without significant improvements in the synthesis. The impact of individual audio effects on the perceived realism of a synthesised sound is out of the scope of this work.

6.3.3 Objective Function

The set of fitness functions, or objective functions, were taken from literature. Each of the objective measures describe a set of audio features that are used to summarise the audio sample, as a set of audio statistics. The distance measure between sets of audio samples was constructed as the euclidian distance between these audio feature sets. This measure was then used as the fitness function for the PSO, described in Section 6.3.1. The audio features that make up each objective measure are described in Table 6.1. To standardise implementations, all audio features were extracted using Essentia [Bogdanov *et al.*, 2013], based on recommendations from Chapter 3, and in Moffat *et al.* [2015].

The Allamanche *et al.* [2001] distance measure was first designed for measuring similarity of musical content, and is included as part of the MPEG-7 standard list of low level descriptors. Gygi *et al.* [2007] produced a set of audio features to apply to environmental sounds, both for measuring similarity and performing categorisation. Moffat *et al.* [2017] produced a set of reduced audio features to perform hierarchical classification of an audio effects library, where the audio features were selected through a machine learning feature selection approach, as discussed in Chapter 4. The Wichern *et al.* [2007] feature set was produced for indexing and segmenting natural sound environments. MFCCs

TABLE 6.1: Attributes of Each Objective Function

Objective Function	Features and Attributes
Allamanche [Allamanche *et al.*, 2001]	Loudness Spectral Flatness Spectral Crest Factor
Gygi [Gygi *et al.*, 2007]	Envelope Statistics Pitch Autocorrelation Waveform Peaks Spectral Centroid Spectral Moments Frequency Band Energy Modulation Statistics Subband Correlation Spectral Flux
MFCC [Yee-King and Roth, 2011]	MFCC
Moffat [Moffat *et al.*, 2017]	Loudness Pitch MFCC Envelope Statistics Spectral Contrast Spectral Flux
PEAQ [Thiede *et al.*, 2000]	Signal Bandwidth Masking Content Modulation Difference Distortion Harmonic Structure
Wichern [Wichern *et al.*, 2007]	Loudness Spectral Centroid Spectral Sparsity Harmonicity Temporal Sparsity Transient Index (ΔMFCC)

were used for parameter optimisation of musical tones, and are generally considered as a good measure of timbre [Pachet and Aucouturier, 2004; Yee-King and Roth, 2011]. This method was intended to be a baseline measure, through which to compare each synthesis method. PEAQ [Thiede *et al.*, 2000] (Perceptual Evaluation of Audio Quality) is an evaluation method specifically designed to measure the sample by sample difference in audio samples. This method was first designed to measure the impact different audio compression algorithms will have, such as comparing MP3 to WAV.

6.4 Synthesis Evaluation - Listening Test

6.4.1 Participants

Nineteen participants took part in the experiment, of which 7 were female and 12 were male. Their average age was 29 with a standard deviation of 3 years. The average test duration was 23 minutes, so fatigue was not an issue [Schatz *et al.*, 2012]. The procedure was approved by the local ethics committee.

6.4.2 Experimental Setup

The experiment was performed in a similar methodology as presented in Section 5.2. The listening test was set up using the Web Audio Evaluation Tool [Jillings *et al.*, 2015]. A pair of high quality calibrated PMC AML-2 loudspeakers were used in the Queen Mary Studio [Morrell *et al.*, 2011]. The volume was adjusted by participant to a comfortable level at the beginning of the test. The listening test is available with the same user interface and set of samples that were used by participants.[6]

6.4.3 Materials

Participants were asked to evaluate sound samples for four categories (fire, rain, stream and wind). In each category there were two different reference samples compared. For each sound example, six synthesised samples were provided and compared to a recorded sample reference. Each synthesised sample was produced using a different objective function. The reference samples were all selected from a professionally available sound effects library.[7] All samples were 48kHz wav files, and loudness normalised in accordance with ITU-R BS.1387-1 [1998]. Each category had one anchor, where random parameter values were used to generate a sample.

The reference samples were: Big Fire; Candle; Wind Gusts; Medium Wind; Fast Flowing Stream; Gentle Bubbling Stream; Rain on Umbrella; Rain Storm. There were two

[6] `https://goo.gl/2Bp3Ju`
[7] `https://www.prosoundeffects.com/hybrid-library/`

reference samples selected for each category, and they were chosen to be distinct and different from each other, to cover a range of sonic properties. On each stage of the listening test, there were eight audio samples: one hidden reference; one random parameter anchor; and six synthesised samples, each using a different objective metric.

The anchors were included to encourage participants to use the entire evaluation scale, and to review how samples were distributed within that scale, in accordance with ITU-R BS.1534-3 [2015]. This ensures that there is a lower limit sample to compare against. It also performs as a confirmation that a participant has fully understood the requirements for the experiment. If a participant rated the anchor as higher than the sample, then it could be inferred that the participant may not have fully understood the requirements, or may have some hearing defect.

6.4.4 Procedure

Participants were then asked to rate how similar they perceived a set of given samples to a provided reference, on a continuous linear scale on which to rate all sounds, labeled from "most similar" to "very different". All other aspects of the procedure are as described in Section 5.2.5.

6.5 Results

One participant's results was identified as an outlier as over 30% of their answers was more than three scaled median absolute deviations from the median result. As such all results presented are of the remaining 18 participants. User similarity ratings are presented in Figure 6.3, where the distributions of the results can be seen.

A Shapiro-Wilk normality test showed that the data is not-normally distributed ($W = 0.95208$, $p < 2.2e\text{-}16$). A Kruskal Wallis test was performed to evaluate the impact of each objective function. A significant difference between the objective evaluation methods was found ($H=18.2$, $p=0.0057$). A post-hoc multiple comparison was performed, with results presented in Table 6.2.

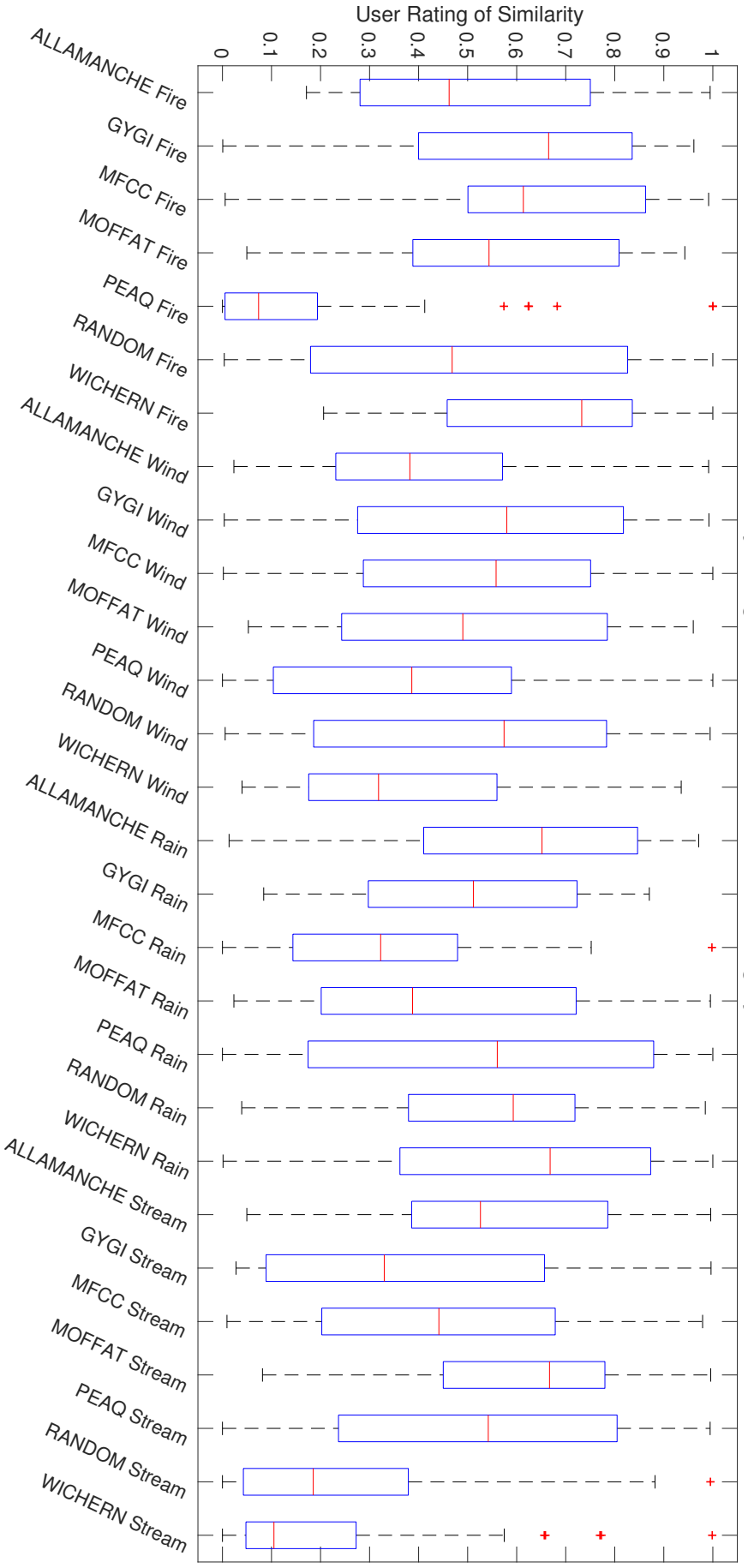

FIGURE 6.2: Distribution of User Similarity Ratings over Objective Function and Synthesis Model

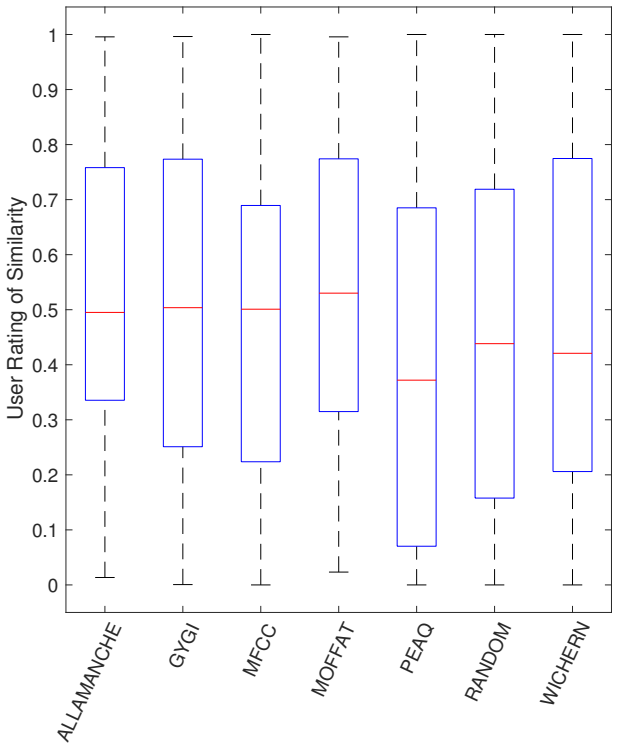

FIGURE 6.3: Distribution of User Similarity Ratings per Objective Function

6.5.1 Results per Synthesis Method

Table 6.2 shows that across all sound synthesis models, there is limited consistent variation. The PEAQ objective function is significantly worse than both Allamanche and Moffat. There are no further significant results at this level. To analyse the data further, the results per synthesis method are investigated, as shown in Figure 6.2. Kruskal Wallis tests were performed to identify the impact of each objective function for each synthesis method. The results show that there are significantly different groupings in three of the four sound synthesis methods. These results are presented in Tables 6.3-6.5. Within the wind synthesis method, no significant differences in subjective similarity to the reference sample were found between different objective synthesis methods (H=11.72, p=0.069).

As seen in Table 6.3, the PEAQ method is significantly worse than every other objective evaluation function with regards to fire sounds. Other than the rain sounds, in Table 6.4 MFCCs are significantly worse than Allamanche, PEAQ, random and Wichern. For stream sounds, Table 6.5 shows that Allamanche, Moffat and PEAQ are all significantly

TABLE 6.2: Multiple Comparisons Test Significance Results for All Synthesis Models, Kruskal Wallis Results (H=18.2, p=0.0057)

All Methods	Allamanche	Gygi	MFCC	Moffat	PEAQ	Random	Wichern
Allamanche	.	o	o	o	**	o	o
Gygi	o	.	o	o	o	o	o
MFCC	o	o	.	o	o	o	o
Moffat	o	o	o	.	*	o	o
PEAQ	**	o	o	*	.	o	o
Random	o	o	o	o	o	.	o
Wichern	o	o	o	o	o	o	.

o >0.05, * <0.05, ** <0.01, *** <0.001, **** <0.0001, . = no comparison made

TABLE 6.3: Multiple Comparisons Test Significance Results for Fire Synthesis Method, Kruskal Wallis Results (H=53.19, p=1.08e-9)

Fire	Allamanche	Gygi	MFCC	Moffat	PEAQ	Random	Wichern
Allamanche	.	o	o	o	***	o	o
Gygi	o	.	o	o	****	o	o
MFCC	o	o	.	o	****	o	o
Moffat	o	o	o	.	****	o	o
PEAQ	***	****	****	****	.	***	****
Random	o	o	o	o	***	.	o
Wichern	o	o	o	o	****	o	.

o >0.05, * <0.05, ** <0.01, *** <0.001, **** <0.0001, . = no comparison made

TABLE 6.4: Multiple Comparisons Test Significance Results for Rain Synthesis Method, Kruskal Wallis Results (H=26.81, p=1.57e-4)

Rain	Allamanche	Gygi	MFCC	Moffat	PEAQ	Random	Wichern
Allamanche	.	o	***	o	o	o	o
Gygi	o	.	o	o	o	o	o
MFCC	***	o	.	o	*	*	***
Moffat	o	o	o	.	o	o	o
PEAQ	o	o	*	o	.	o	o
Random	o	o	*	o	o	.	o
Wichern	o	o	***	o	o	o	.

o >0.05, * <0.05, ** <0.01, *** <0.001, **** <0.0001, . = no comparison made

better than both random and Wichern. MFCC is also significantly better than Wichern, and Moffat is significantly better than Gygi.

TABLE 6.5: Multiple Comparisons Test Significance Results for Stream Synthesis Method, Kruskal Wallis Results (H=54.91, p=4.84e-10)

Stream	Allamanche	Gygi	MFCC	Moffat	PEAQ	Random	Wichern
Allamanche	.	o	o	o	o	***	****
Gygi	o	.	o	*	o	o	o
MFCC	o	o	.	o	o	o	*
Moffat	o	*	o	.	o	****	****
PEAQ	o	o	o	o	.	**	**
Random	***	o	o	****	**	.	o
Wichern	****	o	*	****	**	o	.

o >0.05, * <0.05, ** <0.01, *** <0.001, **** <0.0001, . = no comparison made

TABLE 6.6: Correlations of Objective Function Distance Measure with Mean User Similarity Rating

Objective Function	Correlations ρ	P-Value p
Allamanche	-0.3095	0.4618
Gygi	-0.0952	0.8401
MFCC	0.0238	0.9768
Moffat	-0.3095	0.4618
PEAQ	-0.4059	0.3155
Wichern	**0.7857**	**0.0279***

o >0.05, * <0.05, ** <0.01, *** <0.001, **** <0.0001, . = no comparison made

6.5.2 Comparison with Objective Function Results

Each of the objective functions also produced a distance measure, which is the value that was minimised as part of the synthesis. These distances indicate how successful the synthesis method believes it has performed in each case. The objective distances are compared with the subjective distances, and are plotted in Figure 6.4, along with linear regression lines of best fit. The user similarity ratings were inverted to make the graphical representation easier to interpret, and correlations more clear. Each of the objective and subjective results were correlated, using a Spearman correlation, for non-parametric data, and the results presented in Table 6.6. Only the Wichern result is statistically significant, with a strong positive correlation.

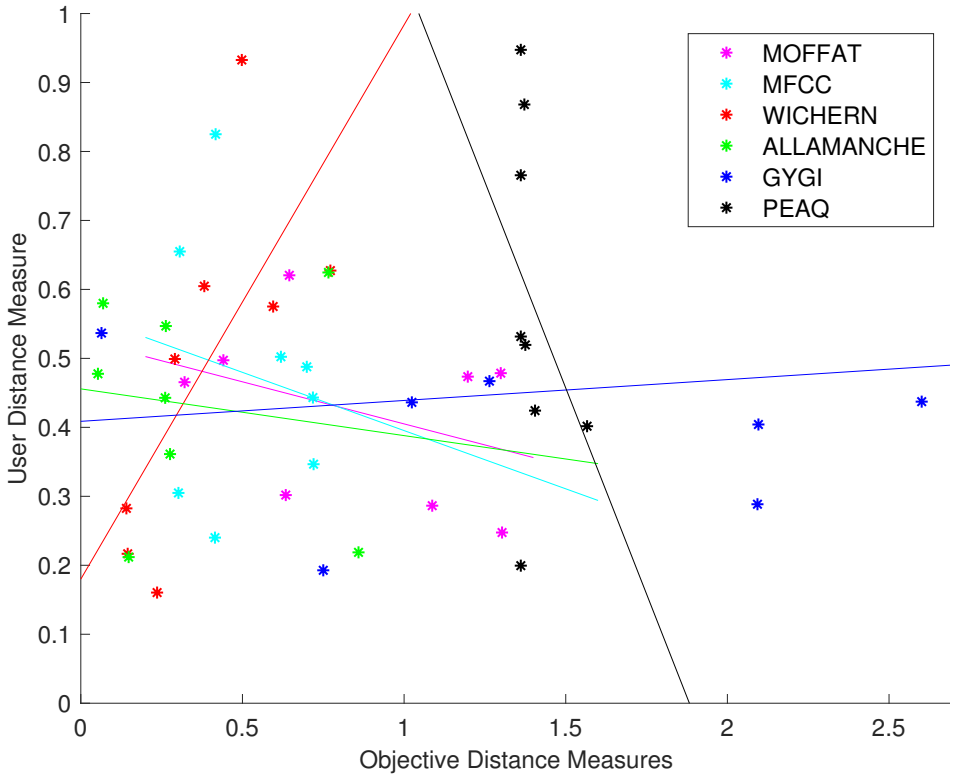

FIGURE 6.4: Inverse User Similarity Compared Against Objective Distance Metric for Each Objective Function, with Linear Best Fit Lines

6.6 Discussion

Figure 6.3 and Table 6.2 shows minimal significant variation in the distributions of similarity ratings. Overall Moffat performs as the best objective evaluation method, whereas Allamanche shows positive results with a lower variance in the data. PEAQ performs the worst, and is significantly worse than both Allamanche and Moffat, which is the only significant generalised result.

For further analysis, the breakdown per synthesis method was investigated. Within the fire sound, every objective function was significantly better than PEAQ. PEAQ is the only method that models distortion and bandwidth, and it is believed that these components of the objective function caused it to perform poorly for fire. A large portion of a fire sound is crackling and popping, and broadband noise. As PEAQ is designed for evaluating the quality of audio compression algorithms, it is designed to be sensitive to cracking and distortion artefacts. However, this is principally what makes up a fire sound. As such, it is expected that PEAQ failed to appropriately model fire due to the

wide-band, impulsive nature of the sound, which PEAQ is often identifies as a flaw. It is suspected that PEAQ will also fail to accurately model other sounds that are broadband and highly impulsive, such as applause [Adami *et al.*, 2017] or gunshot [Mengual *et al.*, 2016] sounds.

Within the rain sounds, the MFCC evaluation metric performed significantly worse than Allamanche, PEAQ, Wichern and random. MFCCs are often used in music information retrieval as a descriptor for timbre. However, the variation in rain sounds are less timbral and more related to the ambient noise versus individual impulsive tones. The separation between constant noise tones and impulsive tones will not be identified by MFCCs. As MFCCs are no better than the random parameters, it is clear that MFCCs are not a good measure for parameter estimation within rain sounds. There is no other significant variation in objective evaluation functions. Wichern was the only method to perform better than random parameter selection, though this was not significantly better. This could be due to the random parameters being very good parameters selected by chance, or that there is limited variation within the synthesis method.

Regarding stream sounds, Figure 6.2 shows that Wichern and random both perform poorly, and are significantly worse than Allamanche, Moffat and PEAQ methods, and Alllamanche is significantly worse than MFCCs. It is suspected that this is due to Wichern primarily looking at harmonic content and transient sounds, where less attention was paid to broadband sound similarities. Within the stream model, most water noises will be highly broadband signals, and Wichern will most likely tend to produce more harmonic tuned sounds, than those present in a real signal. Wichern and random are not significantly worse than Gygi, which is most likely due to the large variation in the distribution of the Gygi results. This suggests that individuals were undecided or opinions were split on the result. Moffat was the best performing result and is significantly better than Gygi, along with random and Wichern. It is suspected that this is due to the inclusion of the spectral contrast feature. Spectral contrast is an audio feature that identifies the peaks and valleys in the magnitude spectrum, and performs dimensionality reduction on the result. Spectral contrast is often considered an effective method for evaluating audio masking and for identifying high contrast variations in

TABLE 6.7: Ratings of Success of each Objective Evaluation Method

	Overall	Fire	Rain	Stream	Wind
Allamanche	2	4	5	1	1
Gygi	2	1	1	3	4
MFCC	2	1	2	5	3
Moffat	1	3	3	4	1
PEAQ	5	5	5	3	2
Wichern	4	1	5	1	5

1 = Best, 5 = Worse. Ratings were created manually, based on ranking and clustering of results

frequency spectra.

The wind model failed to produce any significant difference between any objective metrics. Gygi performed the best, closely followed by random parameter allocation, but all methods are fairly similar to each other. This could be a failing of the synthesis model, as there were highly harmonic artefacts within the synthesis model, that no parameters could be removed. Further investigation of the synthesis model shows that a number of filter center frequencies are hard-coded into the model, which most likely lead to inconsistent and inconclusive results. It is also possible that the number of parameters may also have influenced the results. Wind had more than twice the parameters to optimise compared to any other synthesis model, which the PSO algorithm may have had challenges optimising. The larger search space may have lead to issues in finding appropriate minima.

Each of the objective functions were compared and grouped in terms of their effectiveness on a 1-5 rating scale, as presented in Table 6.7. It can be seen that the Gygi method performs best for both fire and wind sounds and fairly well for rain sounds, but is one of the worse sounds for stream. Gygi contains a large set of parameters relating to subband correlations and modulation statistics, which have been tied to the human auditory system [McDermott and Simoncelli, 2011]. As such, Gygi method seems to be the best overall performer, as consistently produced reasonable results in all cases, and between that and Moffat, it never produced the worst results. Moffat performed best overall, and was best for wind sounds, which it is suspected is due to the spectral contrast feature. It also performed reasonably well for fire and rain sounds, as the spectral contrast and

spectral flux sounds will perform well for granular impulsive sounds. The Allamanche method performs best for rain sounds and reasonably well for stream sounds, but is one of the worse methods for wind and fire sounds. This suggests that the spectral characteristics are more complex for wind and fire sounds, as Allamanche only uses a spectral flatness and spectral crest factor as the evaluation, as all samples were loudness normalised before analysis. PEAQ performed worse overall, through performing worse in both fire and rain sounds, however performed reasonably well for stream and wind sounds. This demonstrates that PEAQ represents broadband noisy signals fairly well, however the low level textual and highly impulsive sounds are not effectively modelled by this method. The Wichern method is highly inconsistent as it performs best for fire and stream however is the worse for rain and wind sounds.

Wichern was the only objective evaluation method where the objective distance significantly correlated with the subjective distance ratings. The correlations of the objective distance are a vital aspect of any objective evaluation function, where it is possible to predict how well the objective function performs and how effective the synthesised sound is.

6.7 Conclusion

A set of six different objective evaluation functions, for measuring similarity between environmental sounds, were tested and compared, for their ability to direct a resynthesis algorithm towards an appropriate parameter setting. The Wichern method results correlate significantly and strongly to subjective distance measures. This suggests that the Wichern method can be used as an effective objective metric, comparing similarity between different sets of sounds. Further evaluation with different synthesis methods is required to verify these results and to identify whether the synthesis methods themselves impacted the results. The use of further different sounds samples and sound classes would also provide further data points, which would aid in correlating the objective results with the subjective ratings. This would ensure that the results can be applied to a range of different sound types. It has been clearly identified that, in the

context of environmental sounds, the Wichern set of audio features can be used to measure the similarity of two audio samples, and these results correlate with the subjective similarity measure reported from a listening test.

The PEAQ method performed the worst, performing significantly worse than both Moffat and Allamanche. This demonstrates that PEAQ is not suitable for evaluating sound similarity in a range of different cases, though it was effective for comparing broadband noisy signals, such as wind.

The limited scope of this work, and the focus on environmental sounds could be a limitation. Further work will be needed to verify if this work can be generalised to other areas of sounds. Another potential issue could be the synthesis models used. The limitation for each method to produce a wide range of sounds, could result in many different samples being challenging to synthesise, and thus cause all methods to underperform. Furthermore, there could be a large set of other audio features that could also be reviewed, such as the work in McDermott and Simoncelli [2011].

Chapter 7

Case Study: Evaluation of Aeroacoustic Sound Effects

7.1 Introduction

This chapter develops the evaluation approaches presented in Chapters 5 and 6, and applies it to a set of aeroacoustic sound effect synthesis models developed at Queen Mary University. A set of evaluations for each of the different types of sound effects is presented, to identify the benefit and contribution to understanding that was provided by the evaluation format for each sound effect. The modifications made to the standardised evaluation approach will be outlined, and the different context of each sound effect will have some impact on the evaluation methodology. Based on the work presented in Chapter 5, evaluations are presented, and then the experimental results and modifications for each of the types of sound effects evaluation performed are discussed, identifying specific experimental design choices. The results of this subjective evaluation will then be compared to the objective evaluation metrics discussed in Chapter 6. The benefit of these evaluation methodologies will then be discussed and conclusion drawn as to the appropriateness of this evaluation framework.

This chapter performs evaluation of aeroacoustic sound effects. A range of real-time sound effects synthesis models were developed, based on semi-empirical calculations

of computational fluid dynamics. An aeroacoustic sound effect, is any sound that is created by air moving over object. It produces a distinctive noisy sound, dependant on the type of object that is being passed by air. This could be a sword or golf club object being swung through the air [Selfridge *et al.*, 2017b], a propellor swinging round in the air [Selfridge *et al.*, 2017a] or wind passing over vibrating strings, such as in an Aeolian Harp [Selfridge *et al.*, 2017d]. In each of these cases, the core sound component modelled is an Aeolian Tone [Selfridge *et al.*, 2018a, 2017c, 2016].

A series of sound synthesis models will be evaluated, for the sounds of objects swinging through the air, propellors and aeolian harps. In each case, a sound is produced as a function of the distance and direction of the object and object diameter, length and air flow speed. The sounds produced are the result of sampling individual points along the simulated object, which allowed the creation of a range of different objects with varying profiles. Full description of the synthesis method can be found in Selfridge *et al.* [2018a, 2017a,b,c,d,e, 2016].

7.2 Generalised Experimental Method

The subjective evaluations were carried out for a series of different sound synthesis methods. The general experimental method, based on Moffat and Reiss [2018b], is described within this section, and specific variations of the experimental method will be presented in the relevant section for the synthesis method.

A double-blind listening test was carried out to evaluate the effectiveness of the synthesis model, where participants were asked to evaluate a range of samples in terms of 'plausibly', or how realistic they perceived the sounds to be. Participants were provided with a continuous linear scale to rate each sample based on how plausible the sound represented the given sonic category. Rating the plausibility of sound from a physical model was the preferred judgement in Castagné and Cadoz [2003], stating a plausible sound as one that listeners thought "was produced in some physical manner". The Web Audio Evaluation Tool [Jillings *et al.*, 2015, 2016] was used to build and run listening tests in the browser. This allowed test page order and samples on each page to be randomised

and loudness normalised in accordance with ITU-R BS.1534-3 [2015]. Headphones were used to administer the sounds to participants. Headphones were used for simplicity and convenience, and it has been shown that high quality headphones will provide a comparable listening test experience to a configured loudspeaker setup [Koehl *et al.*, 2011]. All experimental procedure was approved by local ethics committee and tests were all designed to take less than 30 minutes, to ensure fatigue was not an issue. High quality real world recordings were used to compare each synthesis method to real world examples, to allow for an understanding of whether the sound is indistinguishable from a recording. Alternative synthesis methods were used as a benchmark, to understand how well within a scale the synthesis method performed, or whether another standard analysis-synthesis method can produce better results. Anchors were used to standardise the bottom level of the scale, and to allow for comparison as to how badly a synthesis method performed. The knowledge that a method performed better or worse than an example designed to be unrealistic provides a better understanding of the participants' accuracy in rating the other samples.

The aeolian approaches for synthesis will be referred to as the physical approach, and all other approach will be referred to as the alt or alternative approach.

7.3 Aeolian Harp

A synthesis model for an aeolian harp was developed from aeolian tones in Selfridge *et al.* [2017d]. A full software implementation of the aeolian harp is available.[1] To evaluate how close the model sounds like an Aeolian harp, a subjective listening test was undertaken. Since Aeolian harps are not common, participants were given training to assist identification of an Aeolian harp. Prior to the start participants were invited to watch a short video explaining the Aeolian harp and how it sounds.[2] Participants then undertook three steps of pre-training, where, at each step, they were presented with a real examples of Aeolian Harps, and then provided a list of six audio samples and asked to identify if these samples were aeolian harps. In total there were 18 training samples

[1] `https://code.soundsoftware.ac.uk/projects/aeolianharp`
[2] `https://www.youtube.com/watch?v=d6c6-u3MQDk`

TABLE 7.1: Post-hoc multiple comparison test results for aeolian harp, for different synthesis models with subjective ratings

Harp	Physical	Alt	Sample
Physical	.	o	****
Alt	o	.	****
Sample	****	****	.

o > 0.05, * < 0.05, ** < 0.01, *** < 0.001, **** < 0.0001, . = no comparison made.

that participants were asked to identify. Nine were genuine recordings of aeolian harps, and nine were constructed with digital synthesisers, to be noticeably different from any aeolian harp samples. Thereafter participants were asked to rate four groups of Aeolian harp sounds, each having six sound clips, two from the physical model, two created by SMS [Zölzer and Amatriain, 2002], and two recordings of actual harps giving 24 clips overall. The SMS clips are produced by analysis of real recordings, not used in the test, and then synthesised from these, therefore wind generation should not be an issue with these. Recorded samples were taken from the Windsongs CD [Winfield, 1993]. The test is available online.[3] There were 32 participants in the tests, 22 male and 10 female, aged from 16 years old to 77 years old with an average age of 36. Eight participants had previously heard an Aeolian harp. No anchor was included within this test, as it was known that participants were most likely not familiar with an Aeolian Harp, and as such, wanted to make the listening test as simple as possible for them. This allowed for them to select one of three options as their preferred sound. The possibility of excluding participants based on their score in the pre-training was reviewed, however, no participant scored below 70% accuracy and setting a higher threshold did not meaningfully change the distribution of the results.

The results, as presented in Figure 7.1 show that the recorded sample was considered better than both synthesis methods. However, there seems to be little difference between the two synthesis methods. A Kruskal Wallis test shows that the results for each method are not drawn from the same distributions (H = 169.84, p = 1.32e-37), and a post-hoc multiple comparison test, as presented in Table 7.1 shows that both the physical and the alternative synthesis methods are significantly different from the recorded sample, but there is no significant difference between them. This demonstrates that although the

[3] `https://goo.gl/cHp49J`

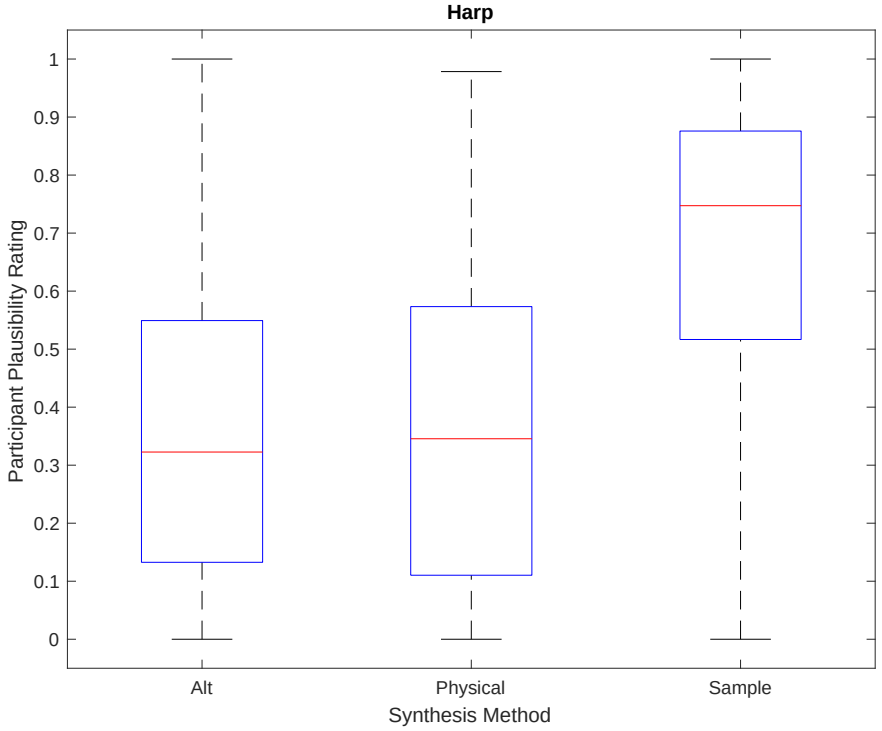

FIGURE 7.1: Plausibility rating of Aeolian Harp.

Alt = Alternative Synthesis Model (SMS)

physical approach is indistinguishable from an alternative synthesis method; the physical approach provides the participant with a much greater level of detailed control over the sound produced. The model gives users control over up to 13 strings, controlling length, diameter, tension, mass and damping.

7.4 Propellor

A synthesis model for a propellor was developed in Selfridge *et al.* [2017a]. Propellor controls include blade length, number of blades, RPM, number of propellors and a series of presets for pre measured planes. In total six planes were modelled, they are the Hercules C13, Boeing B17, Tiger Moth, Yakovlev Yak-52, Cessna 340 and P51 Mustang. A full software implementation of the propellor, with engine noise, is available.[4] A listening test was carried out to evaluate the effectiveness of the propellor model. The test is available online.[5] 20 participants, 6 female and 14 male, aged between 17 and 70,

[4] `https://code.soundsoftware.ac.uk/projects/propeller-model`
[5] `https://goo.gl/36bXQm`

with a median age of 39, were asked to rate a number of sounds for authenticity. Each participant was presented with 4 test pages, in which each page contained 2 real samples, 2 samples from our model, 2 samples generated by SMS [Amatriain *et al.*, 2002] of a recording, and an anchor, which was created by downgrading the quality of our model. The anchors were created from the downgraded synthesis signal, to allow a thorough comparison of how plausible the synthesis method is compared to the recorded sample. It was expected that a low pass filtered sample, as used in the MUSHRA standard, would still be considered plausible, whereas a low quality downgraded anchor would encourage the full use of the scale and allow for better understanding as to effectiveness of the synthesis method.

The physical approach is producing the sound of a propellor alone, and typically a propellor will never be heard without the sound of the associated airplane engine. As such, for a fair evaluation, an engine noise would be added to all propellor sounds, which is designed to match all the audio samples found. It is outside the scope of this work to design and replicate the motor component of a plane, so a model was adapted from the helicopter sound effect in Farnell [2010].

As presented in Figure 7.2, the sample again performed as the most plausible result, and the anchor performed poorest. This was entirely to be expected. There is also a small difference between the median of the alternative synthesis approach and the physical one, and the physical approach also has a more concentrated distribution. This shows that it may be a slight favourite, and that there is more consistence towards the positive plausibility rating of the physical approach, than that of the alternative synthesis approach. A Kruskal Wallis test demonstrated that the distributions of each method are significantly different ($H = 292.37$, $p = 4.45\text{e-}63$), and the results of a post-hoc multiple comparison test is presented in Table 7.2. It is shown that there are significant differences between every set of stimuli in Table 7.2, with the exception of the two different synthesis methods. As such, there are no significant differences in the subjective rating of the physical approach and the alternative synthesis approach. Despite this, both synthesised approaches are significantly worse than a recorded sample.

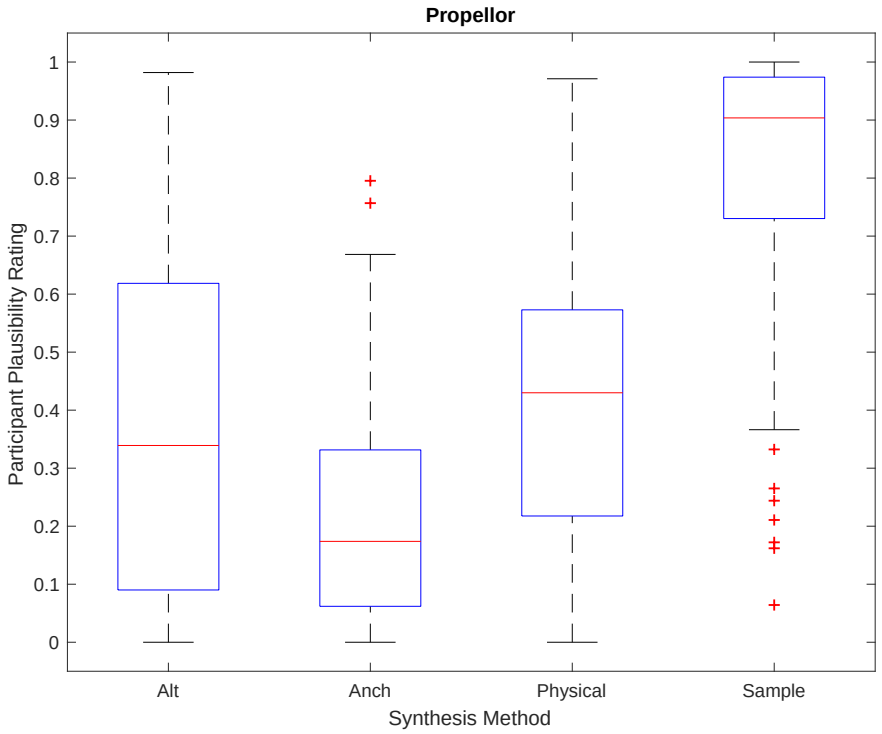

FIGURE 7.2: Plausibility rating of Propellor.

Alt = Alternative Synthesis Model (SMS), Anch = Anchor

TABLE 7.2: Post-hoc multiple comparison test results for propellor, for different synthesis models with subjective ratings

	Anch	Sample	Physical	Alt
Anch	.	****	****	**
Sample	****	.	****	****
Physical	****	****	.	o
Alt	**	****	o	.

o > 0.05, * < 0.05, ** < 0.01, *** < 0.001, **** < 0.0001, . = no comparison made.

7.5 Swinging Object

A real-time physical model was produced of objects swinging through the air, such as a baseball bat or a golf club. This work was originally presented in Selfridge *et al.* [2017b] and then developed further in Selfridge *et al.* [2017c]. Controls are presented for the length, hilt, profile/thickness along its length and details on any cavities [Selfridge *et al.*, 2017e] within the object. The full implementation of is available online.[6]

[6] `https://code.soundsoftware.ac.uk/projects/physicallyderivedswingingobjects`

The experimental evaluation was split into two different tests, a plausibility rating test and an object recognition test. A total of 26 participants undertook the test. The order of the plausibility rating test and object recognition was split to examine if the order had any influence on the results.

The author of this thesis collected five objects, measured their physical dimensions, and recorded the sound of each of them swinging through the air to create a set of reference samples for the plausibility rating component of the listening test. These objects were synthesised through the approach described in Selfridge *et al.* [2017c]. A metal sword, a wooden sword, a baseball bat a 3-wood golf club and a broom handle. Samples of each object being swung were recorded. The objects were then measured to allow for synthesis. All the sampled recordings were captured in the Listening Room of Queen Mary University of London [Morrell *et al.*, 2011]. They were recorded on a Neumann U87 microphone placed approximately 20 cm from the midpoint of the swing and at 90 degrees to the plane of the swing. The impulse response of the room was captured and applied to all other sounds in the listening test so that the natural reverb of the room would not influence the results, except samples from Böttcher and Serafin [2009] and Dobashi *et al.* [2003].

Five categories of sounds were presented to the participant, on 5 different test pages. The wooden sword, baseball bat, golf club and broom handle pages contained two real samples, two samples from the physical model, two samples generated by an alternative synthesis approach (SMS [Amatriain *et al.*, 2002]) from a recording and an anchor. The metal sword page included two real samples, one synthesis sample from Böttcher and Serafin [2009], in which a granular synthesis approach is taken, one synthesis sample from Dobashi *et al.* [2003], where a computational fluid dynamics approach is taken, one SMS sample, one sample from the physical aeolain model and a sample from the physical model with cavity tone compact sound sources added. The anchors were created from a real-time browser-based synthesis effect [Bahadoran *et al.*, 2017, 2018a], to allow a thorough comparison of how plausible the synthesis method is compared to the recorded sample.

During the object recognition test, participants were provided with a Wii Controller, and asked to swing the controller that was directly controlling physical parameters of the synthesis engine. The five preset objects were presented in a pseudorandom order and the user asked to identify which object they were swinging from the list of presets. Fourteen participants completed the object recognition test prior to the listening test, and 12 completed it after the listening test. Each preset was presented twice giving 10 individual tests in total. The listening test is available online.[7]

7.5.1 Plausibility Rating

Box plots for all five objects are shown in Figures 7.3-7.7. The physical model outperforms the alternative synthesis methods on all of the objects except the metal sword. The metal sword performed poorly for plausibility in this test.

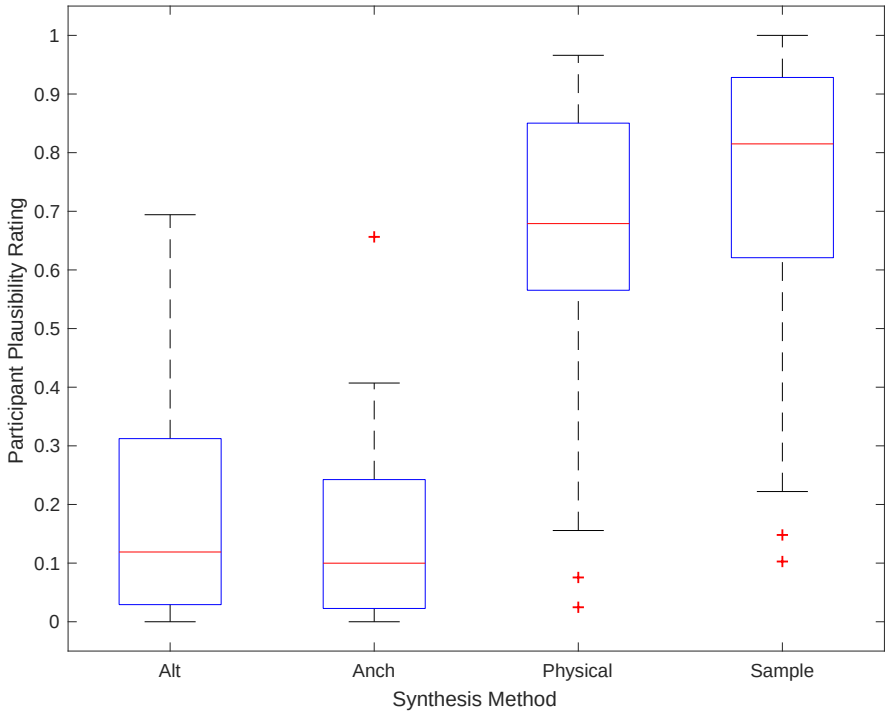

FIGURE 7.3: Broom Handle Plausibility Rating

Alt = Alternative Synthesis Model (SMS), Anch = Anchor

A Kruskal Wallis test was performed for each sound type, which demonstrated that there is significant difference between the sound method distributions for each of the sound

[7] `https://goo.gl/v63D7m`

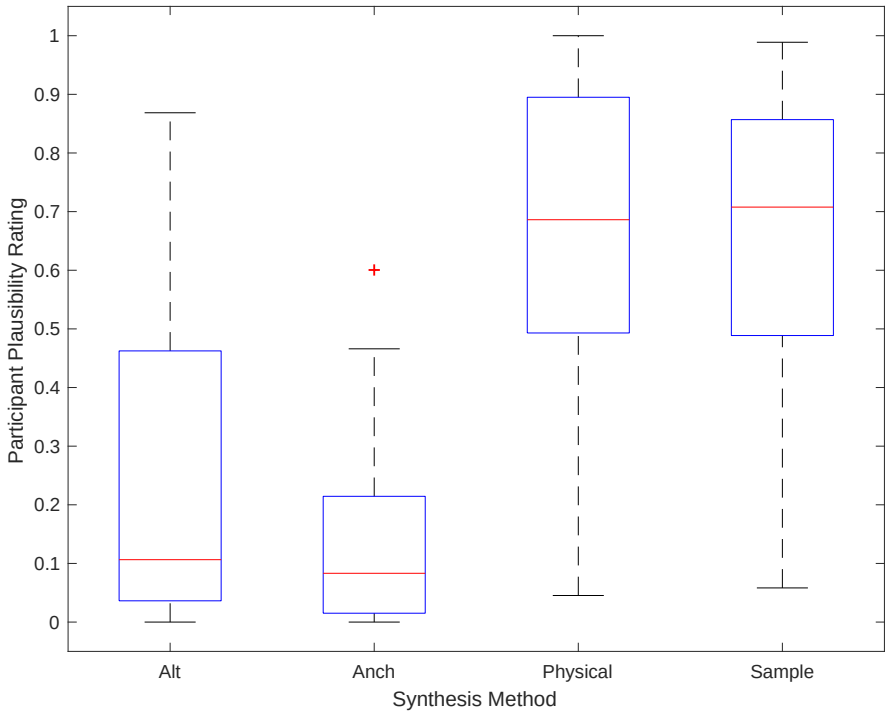

FIGURE 7.4: Baseball Bat Plausibility Rating

Alt = Alternative Synthesis Model (SMS), Anch = Anchor

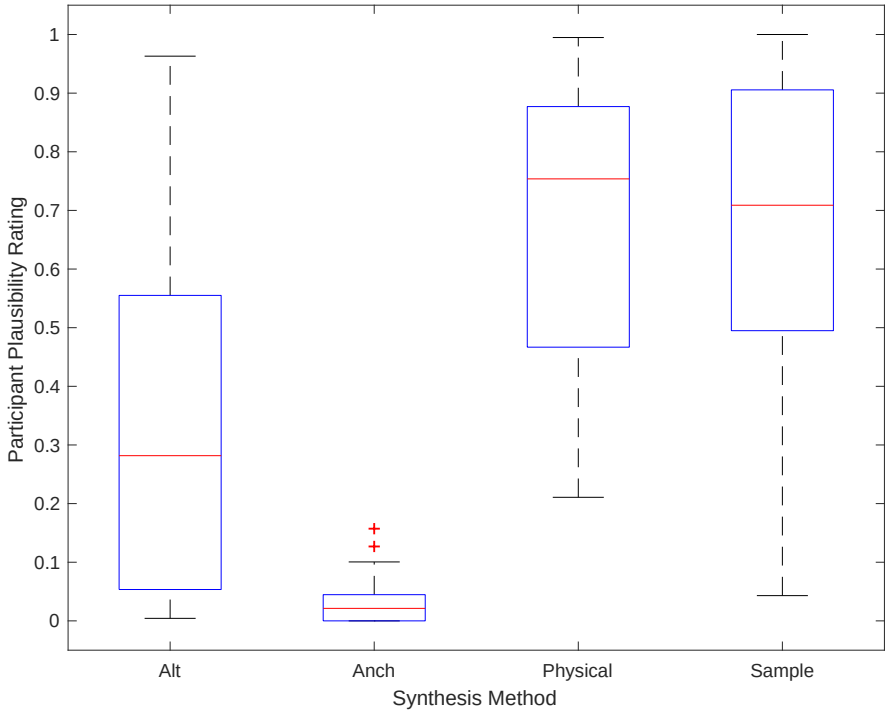

FIGURE 7.5: Golf Club Plausibility Rating

Alt = Alternative Synthesis Model (SMS), Anch = Anchor

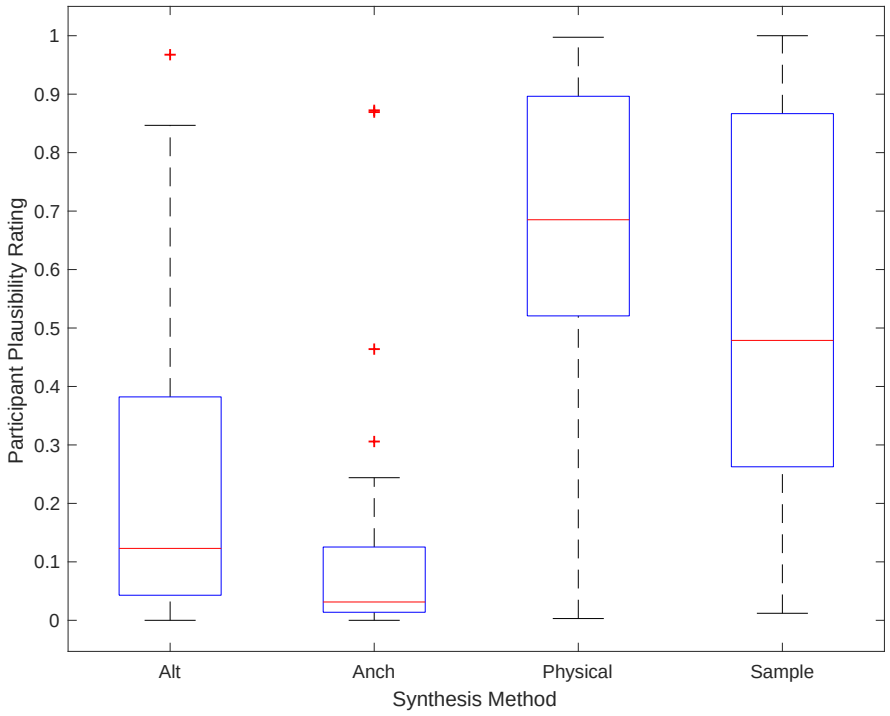

FIGURE 7.6: Wooden Sword Plausibility Rating

Alt = Alternative Synthesis Model (SMS), Anch = Anchor

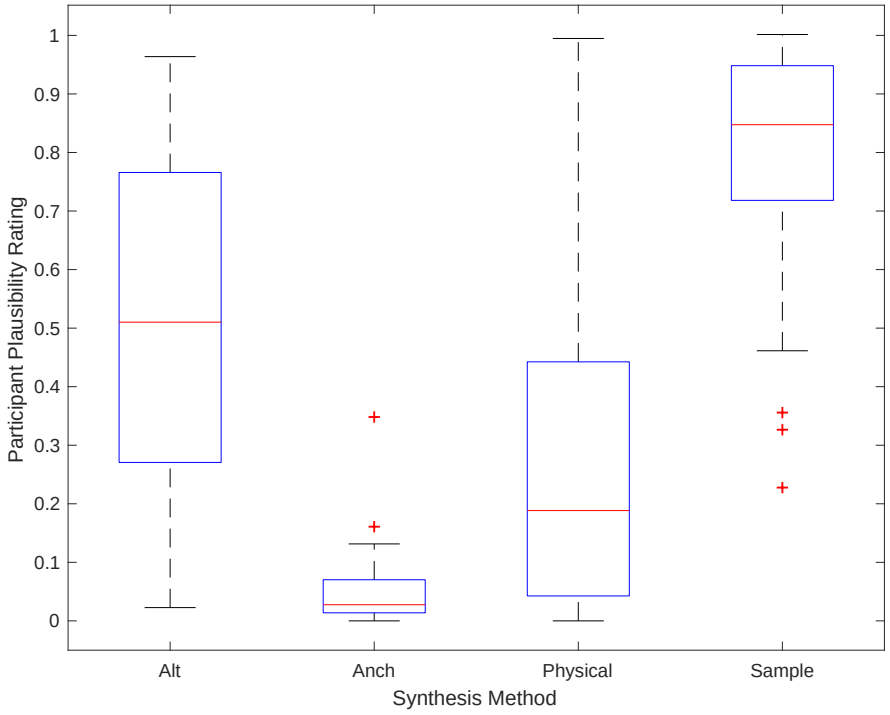

FIGURE 7.7: Metal Sword Plausibility Rating

Alt = Alternative Synthesis Model, Anch = Anchor

types, as shown in Table 7.3, and a post-hoc multiple comparisons test was performed, and presented in Table 7.4

TABLE 7.3: Kruskal wallis test results

Sound Type	H	p
Broom Handle	106.27	6.97e-23
Baseball Bat	82.61	8.46e-18
Golf Club	89.74	2.49e-19
Wooden Sword	68.70	8.10e-15
Metal Sword	102.31	4.96e-22

TABLE 7.4: Post-hoc multiple comparison test results for swinging objects, with different synthesis models subjective ratings

Broom Handle	Anchor	Physical	Alternative Synthesis	Sample
Anchor	.	****	o	****
Physical	****	.	****	o
Alternative Synthesis	o	****	.	****
Sample	****	o	****	.

Baseball Bat	Anchor	Physical	Alternative Synthesis	Sample
Anchor	.	****	o	****
Physical	****	.	****	o
Alternative Synthesis	o	****	.	****
Sample	****	o	****	.

Golf Club	Anchor	Physical	Alternative Synthesis	Sample
Anchor	.	****	***	****
Physical	****	.	****	o
Alternative Synthesis	***	****	.	****
Sample	****	o	****	.

Wooden Sword	Anchor	Physical	Alternative Synthesis	Sample
Anchor	.	****	o	****
Physical	****	.	****	o
Alternative Synthesis	o	****	.	****
Sample	****	o	****	.

Metal Sword	Anchor	Physical	Alternative Synthesis	Sample
Anchor	.	*	***	****
Physical	*	.	o	****
Alternative Synthesis	***	o	.	**
Sample	****	****	**	.

o $>$ 0.05, * $<$ 0.05, ** $<$ 0.01, *** $<$ 0.001, **** $<$ 0.0001, . = no comparison made.

As can be seen in Table 7.4, in every single case except the metal sword, the physical synthesis approach is indistinguishable from the real recording. In all these cases, both the physical synthesis and the sample are significantly different from an alternative synthesis approach. Only in the case of the golf club, is the alternative synthesis approach significantly different from the anchor. This demonstrates that under four different conditions, the physical model is as plausible as a recording. In the case of the metal sword, there were a lot more variations in the evaluation process. As the alternative synthesis approaches were taken from existing sword swing synthesis work, and there were variations within the physical approach used, with a cavity model included, there is a requirement to investigate the results in more detail.

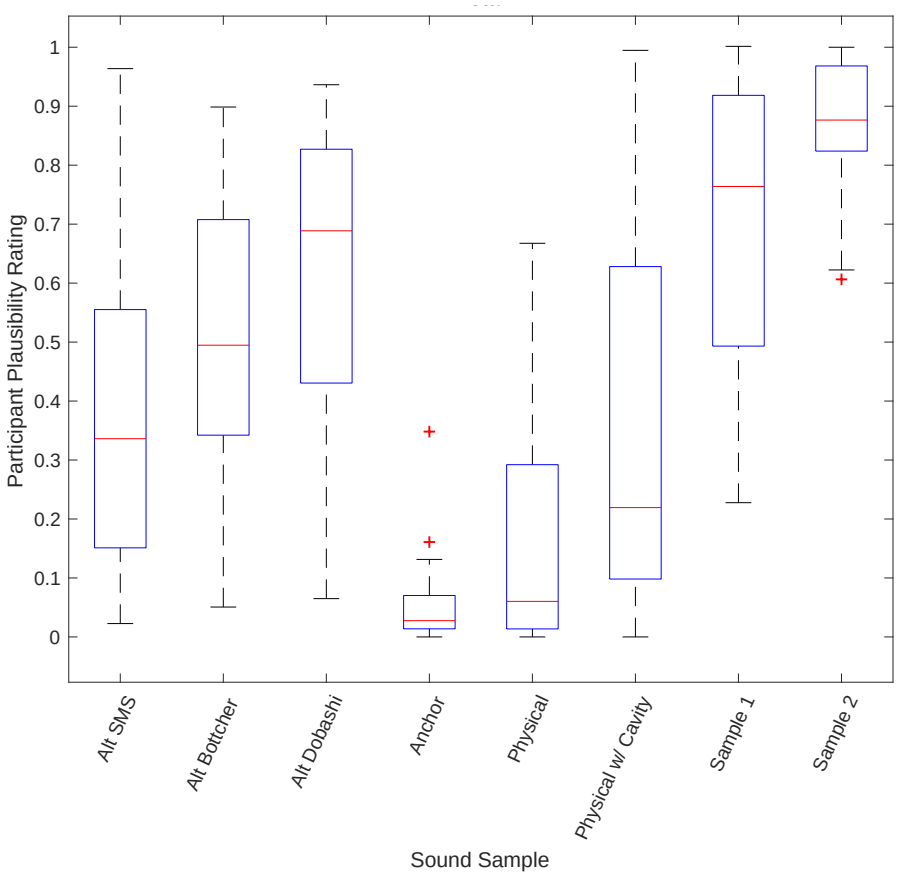

FIGURE 7.8: Metal Sword Plausibility Rating, per sound sample

The two sword samples are not significantly different from each other and the Dobashi sample, as can be seen in Figure 7.8 and Table 7.5. Bottcher is significantly different from one of the two samples. Both of the physical approaches, and the SMS model are significantly different from all the samples. The physical model with no cavity is

TABLE 7.5: Post-hoc multiple comparison test results for metal swords, for subjective ratings of each sample

	Bottcher	Dobashi	Anchor	Physical	Phys+Cavity	SMS	Sample1	Sample2
Bottcher	.	o	****	**	o	o	o	**
Dobashi	o	.	****	***	o	o	o	o
Anchor	****	****	.	o	*	**	****	****
Physical	**	***	o	.	o	o	****	****
Phys+Cavity	o	o	*	o	.	o	**	****
SMS	o	o	**	o	o	.	*	****
Sample1	o	o	****	****	**	*	.	o
Sample2	**	o	****	****	****	****	o	.

o > 0.05, * < 0.05, ** < 0.01, *** < 0.001, **** < 0.0001, . = no comparison made.

not significantly different from the Anchor, and is significantly worse than Bottcher, Dobashi and the recorded samples. The physical model with cavity is indistinguishable from the other three alternative synthesis methods, SMS, Bottcher and Dobashi. It is clear that the Dobashi synthesis model performed best, followed by the Bottcher method. One possible reason for the poorer performance of the metal sword physical model was that all the other modelled objects were thicker than the metal sword. Thicker objects produce much noisier signals, with lower filter Q values, whereas thinner objects produce sounds much closer to pure tones. SMS analyses and extracts sinusoidal components, much closer to pure tones, and so are more appropriate for modelling thinner objects. Results given in Selfridge *et al.* [2017b] indicated that the lower quality physical model sounds were rated as more plausible. These sounds had a fixed Q value that gave the impression of a thicker object. The diameter used to generate sounds by Dobashi *et al.* [2003] was 10mm, more than twice the thickness of the sword modelled with the physical approach, which was 4.6mm at the widest and 1.3mm at the tip. It may be the case that listeners perceive a thicker sound as more plausible even if not physically accurate. Aeolian tones make the assumption that the object air is flowing over is a cylinder, however in cases of a thin sword, this assumption is not valid. The assumption of cylindricality is more valid for baseball bats or broom handles.

Another possible reason for the poor rating of the metal sword object compared to the other objects is that the number of participants who have swung a real sword and heard the sound may well be less than those who have perhaps swung a golf club and the other objects. Memory plays an important role in perception [Gaver and Norman,

1988]. As such, participants may be revealing the effects of hyper-realism [Moffat *et al.*, 2019; Puronas, 2014], as they are far more likely to have heard a Foley sound effect or artificial sword, than to have swung a sword themselves in real life. This may influence their perception of the physical model, though does not explain why the real recording was rated higher than the synthesis model.

7.5.2 Object Recognition

Table 7.6 give the results of how often participants correctly identified the object being represented by the physical model. This shows a clear difference can be seen between participants who completed the object recognition test prior to the listening test compared to those who completed the object recognition after. It is reasonable to conclude that completing the listening test first provides some level of training for the sound object recognition.

TABLE 7.6: Objects correctly identified from the Wii Controller test

Object	Correctly Guessed Before Listening Test(%)	Correctly Guessed After Listening Test(%)
Wooden Sword	0	38
Metal Sword	36	63
Broom Handle	7	42
Baseball Bat	11	46
Golf Club	21	38

Results presented in Table 7.6 show that participants were far less able to identify the object being modelled by our synthesis model when having to choose before the listening test. In fact, it was more common to choose one of the other objects being modelled rather than the correct one. The wooden sword model was never correctly identified, while the metal sword object was correctly identified more than any other object, but still less than 50% of the time.

On examination of those who completed the object recognition test after the listening test, shown in Table 7.6, it can be seen that there was an increase for all objects being correctly identified. The metal sword object was correctly identified more often than the other objects and on this occasion, more often than not. Although the results for the other objects are higher than those presented, it was still more common for participants

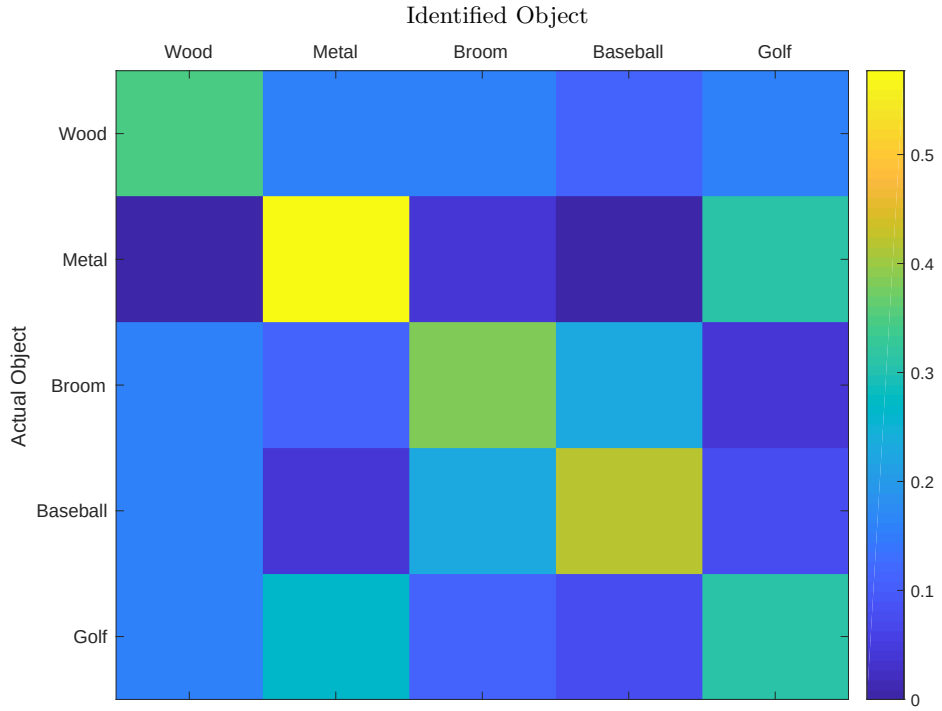

FIGURE 7.9: Confusion Matrix for Object Recognition

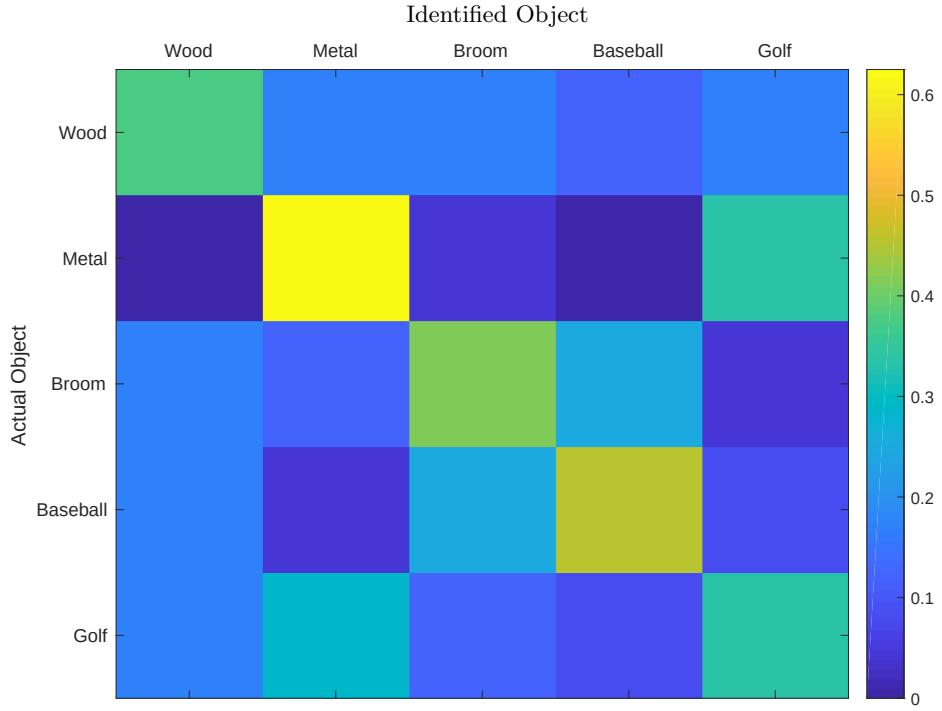

FIGURE 7.10: Confusion matrix for object recognition for group with pre-training

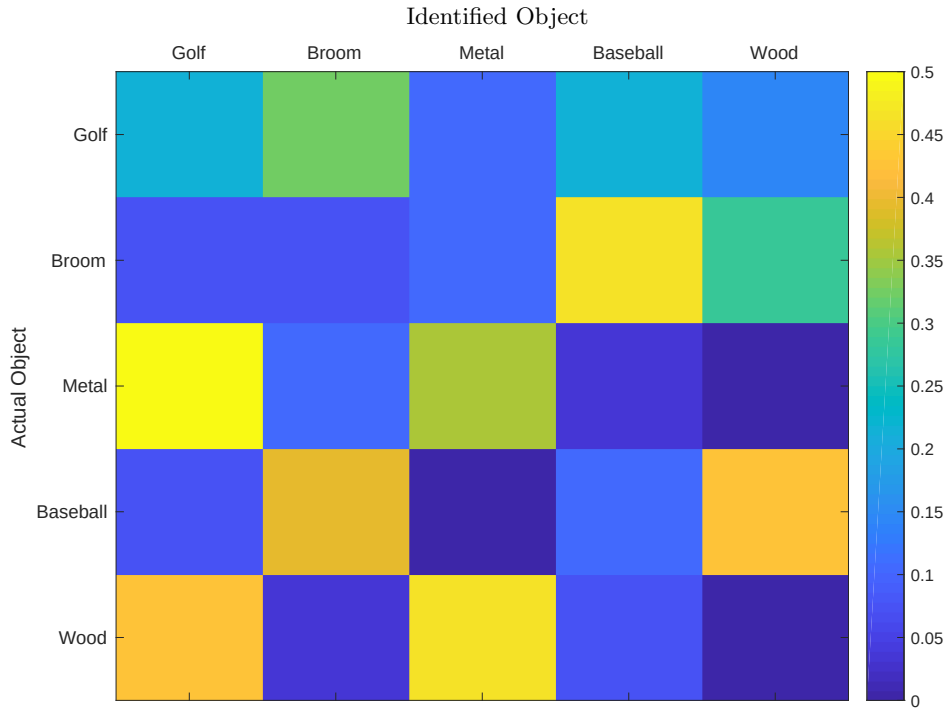

FIGURE 7.11: Confusion matrix for object recognition for group with no training

to choose one of the other objects being modelled rather than the one being replicated by our synthesis model.

Figures 7.10 and 7.11 show the confusion matrices for the object recognition part of the test. It is particularly interesting the difficulties participants had in recognising a set of fairly similar sounds into the discrete categories, particularly as recognition and categorisation tasks are commonplace for evaluating sound synthesis work, as discussed in Chapter 2

In all cases the metal sword and the golf club are one of the most commonly confused objects. It is believed that this is due to these being the two thinnest objects, and thus are likely to sound very similar. However, the golf club makes a louder, more powerful sound, due to the large golf club head and being one of the longest objects tested. The sonic properties of this object may be more associated with a weapon, such as a sword, than a golf club.

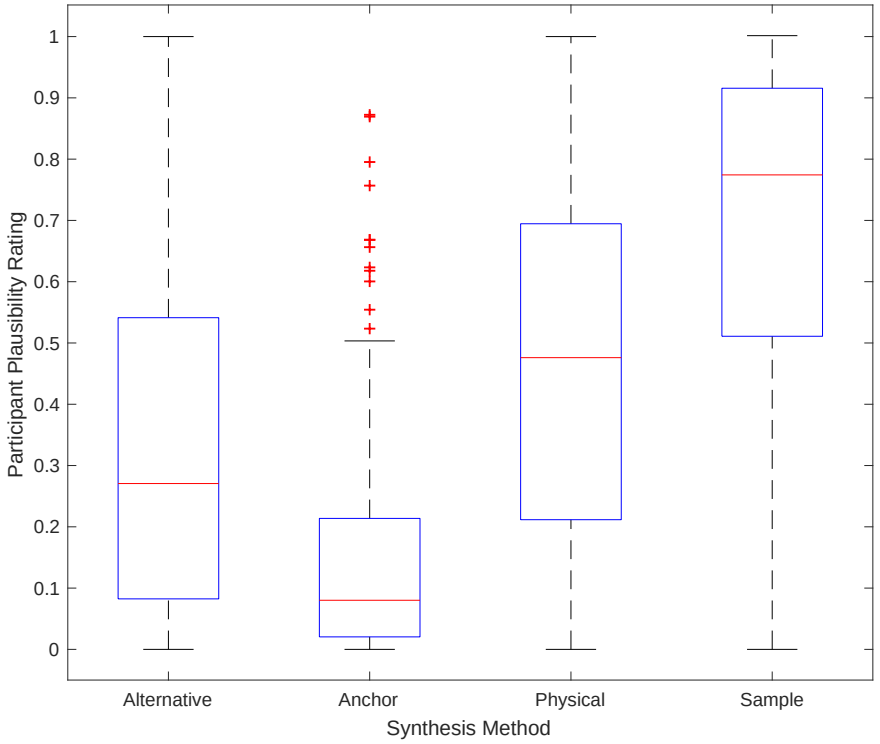

FIGURE 7.12: Plausibility Rating for All Sounds

7.6 Overall Evaluation

A series of different approaches for modelling aeroacoustic sound effects, have been subjectively evaluated, through the use of a range of listening tests. Some specific modifications have been made for each specific case. The overall results for all listening tests and synthesis models are presented in Figure 7.12. It can be seen that the physical approach overall outperformed the alternative synthesis approaches and anchors, though not as plausible as recorded samples. A Kruskal Wallis test showed that these distributions are significantly different (H = 752.9624, p = 6.8739e-163), and there are significant differences between each approach ($p < 0.0001$), as presented in Table 7.7. From this, it can be concluded that the physical approach can, in general, improve the perceived plausibility of sound synthesis methods, though this is dependant heavily on specific developments for each specific physical case.

TABLE 7.7: Post-hoc multiple comparison test results for all sounds, for different synthesis models with subjective ratings

	Sample	Alternative	Physical	Anchor
Sample	.	****	****	****
Alternative	****	.	****	****
Physical	****	****	.	****
Anchor	****	****	****	.

o > 0.05, * < 0.05, ** < 0.01, *** < 0.001, **** < 0.0001, . = no comparison made.

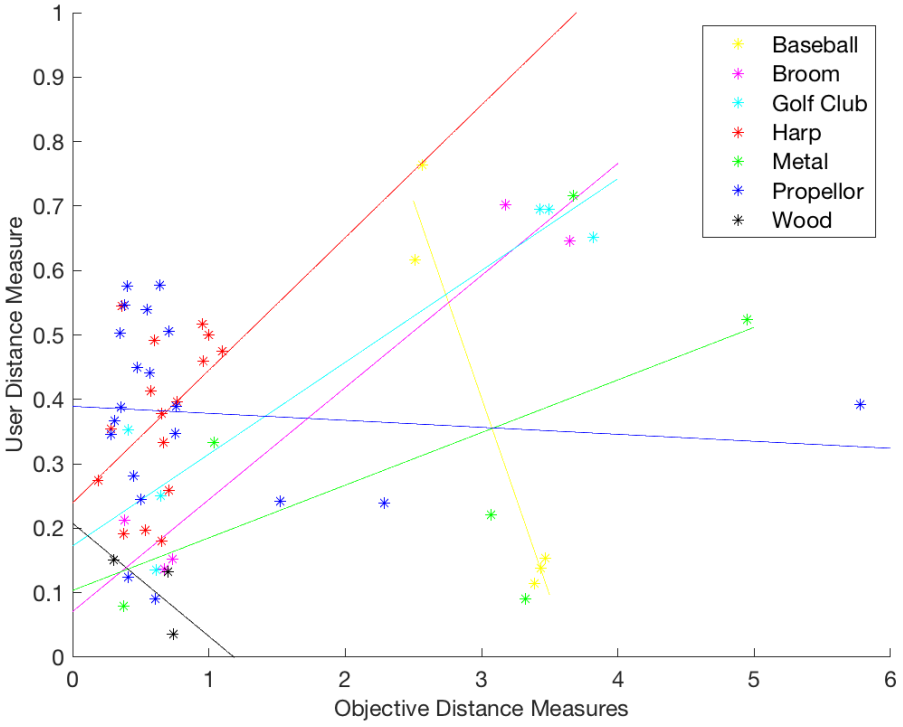

FIGURE 7.13: Comparison between Similarity of Subjective Ratings and Objective Distance Measure for Each Sound Category

7.7 Objective Evaluation

Following the approach of objective evaluation discussed in Chapter 6, the Wichern objective function was used to evaluate each of the different sound classes identified in this chapter. To validate the results, the objective distances were correlated with the subjective distance metrics. In each case, the distance was measured from the recorded sample to each of the other recordings in that group. Figure 7.13 shows the mean user distance rating with the Wichern objective distance measure. The lines, of each colour,

represent a regression best fit line, which is to ease the visualisation of the data and correlations.

TABLE 7.8: Spearman Correlation between Objective Distance Measure and User Distance Measure

Sample	Correlation (r)	Probability Value (p)
All	**0.25**	**0.048***
Baseball	-0.50	0.45
Broom	0.60	0.35
Golf Club	0.64	0.14
Harp	0.36	0.18
Metal	0.71	0.14
Propellor	-0.14	0.56
Wood	-1.0	0.33

Table 7.8 shows the correlation between user distance rating and the objective distance measures taken. 'All' represents the correlation with every single point shown, and then each other sonic class, and the correlations and probability of those correlations occurring by chance. The condition with all data points is the only one which correlates the subjective and objective evaluations, with a $p < 0.05$. However the correlation is a very weak one. There are other cases, such as with the golf club and metal sword, where the correlations are stronger, however these results are not statistically significant.

This demonstrates that the objective evaluation metric does perform reasonably, in some cases. This approach does extend past the original domain of environmental sounds, to aeroacoustic sounds, however different statistical significance of results were found, as compared to Chapter 6.

These results may also be difficult to interpret. Participants were not asked to measure the similarity between the samples, but were asked to compare which are more plausible. This does mean that, in some cases, the recorded sample and synthesised sample were never intended to be identical replications of the exact same world environments. In the case of the harps and propellor sounds, the exact physical constructions of the original recordings was not known, so a number of estimations had to be made about original recording environments.

7.8 Conclusion

A series of subjective evaluation experiments were presented, utilising a generalised evaluation framework to identify how plausible the synthesised sounds are perceived to be. The specific use cases of each sound effect were discussed and modifications to the subjective evaluation process were implemented to facilitate evaluations appropriate to the types of sounds being tested. In every case, the synthesis method was compared to at least one other synthesis method and to recorded samples, and in all cases, except the aeolian harp, an anchor was used, to standardise test scores within given range. This was able to appropriately identify the effectiveness of each given synthesis method, and how it compared to other such synthesis methods.

Objective evaluations were then performed, comparing recorded samples to synthesised samples. The objective metric performs best for physically narrower objects, such as metal swords, golf clubs, and broom handles. The results are worse for thicker objects, such as a wooden sword, baseball bat or propellor. This suggests that the harmonic elements of the sonic model perform well, however the more broadband sounds, are not effectively measured or compared in this objective evaluation metric. The overall results were statistically significant, however the correlations were very weak. As such, objective evaluations can be utilised for comparison between audio samples, but will not replace subjective listening tests and human perception.

Overall, the results demonstrate the importance of pre-training of participants before any identification or classification tasks can be undertaken, agreeing with existing work within other fields of audio technology [Reiss, 2016]. This suggests that the environmental context and individuals' experience will significantly impact their perception of a specific sound. Memory plays an important role in perception [Gaver and Norman, 1988]. Many participants may be revealing the effects of hyper-realism [Moffat *et al.*, 2019; Puronas, 2014], as they are far more likely to have heard a Foley sound effect or artificial sword, than to have swung a sword themselves in real life. This may influence their perception of the physical model. This could be revealed in future subjective evaluations. This may suggest that synthesised sounds do not need to be indistinguishable

from real recordings, in order to be believable in the context of a virtual world.

Chapter 8

Conclusion

The evaluation of sound effect synthesis has been reviewed throughout this thesis and a number of developments and directions for improvements have been proposed. Developments for methods of evaluation, through objective and subjective approaches are presented, and a better understanding of sound effect structures is identified.

Chapter 4 presented a sonically motivated sound effects taxonomy, produced using unsupervised learning techniques applied to a commercial sound effects library. The purpose of the research presented in this chapter is to identify an alternative approach to structuring and searching sound effects libraries and ensure that sound groupings are made which are more sonically relevant. Hierarchical clustering within an audio feature space was used to produce sonic structures and hierarchical grouping of sounds.

It is also demonstrated that current sound effects classification and taxonomies may not be ideal for their purpose. It has been shown that existing classification approaches differ from the sonic similarity approaches used within Chapter 4. An approach for producing new sound effect taxonomies based on the sonic content of the samples has been proposed. This approach has also provided a set of audio features to represent sound similarity.

However, in the current format the taxonomy will be difficult for people to search for individual samples without training and regular use. Each of the sound effects categories identified are not necessarily intuitive, and the branches are inconsistent. A simple

structure taxonomy, with consistent separating attributes would be more advantageous and intuitive to use, or a suitable sound browsing tool. Any resulting taxonomy produced through this method, will be highly dependant on the dataset used, and as such a small dataset was used, the result cannot be generalised to larger datasets. Artefacts from the Adobe Sound Effects Library may well result in some specific categorisation that would not be present within other datasets. This could negatively impact the details of the findings presented, or justify further validation requirements. Further to this, more intuitive and consistent data sets could be produced with a better semantic understanding and relationship between audio features and human perception.

There are many approaches to search for sound, and any successful approach will need to be a hybrid approach of sonic similarity and sound context or source labelling to provide an intuitive approach to explore a sonic space.

The work on audio features for evaluation of sound effects is contrasted with the importance of human subjective perspectives. A subjective listening test, presented in Chapter 5, identifies a method for subjective evaluation of the effectiveness of sound effects synthesis algorithms, based on how well the intended sound was synthesised. There is a lack of consistency for evaluation across the field of synthesis research, as identified in Chapter 2, and this subjective evaluation approach has addressed this failing. The failing of consistent evaluation across the field is not unique to sound synthesis, however, will have an impact on the focus of future research. Through the inter-comparison of different synthesis methods, it is possible to gain a better understanding as to the successes and failings of each synthesis approach and thus inform approaches for improving synthesis.

It is unsurprising to find that there is no single best approach to sound synthesis. One synthesis approach, additive synthesis, did consistently perform well, though this was the result of handcrafting specific parameters and can only be applied to a small number of applications - as such this approach is highly labour intensive. Additive synthesis performs well for granular type sounds, such as rain and fire sounds, where many small sonic elements combine together to produce a sound texture, whereas envelope shaped noise approaches performs well for slower moving sound types, such as wind and flowing

water. This seems intuitive, based on the structures of the sound and the approaches used to perform the synthesis.

The restricted number of sound examples evaluated limits the extent to which claims can be made regarding the entire field of synthesis. However, it is clear that better evaluation and understanding of synthesis will identify and develop the quality of synthesis models produced. It was also identified that some synthesis approaches are good enough that participants will find them indistinguishable from recorded samples, particularly environmental sounds. More human sounds, such as babble and applause are, however, consistently difficult to synthesise. Evaluation of sound synthesis can assist in improving upon the state-of-the-art and developing future sound synthesis. The subjective evaluation approaches suggested here, would support effective comparison of sound synthesis methods and thus support improvements in the field. It has also been shown that there are cases where synthesised sound effects can be considered as realistic as recorded sound effects. This is important, as there are many cases where a synthesised sound effect could be used in place of a recorded sample, which can further justify the use of computationally generated sounds.

Chapter 6 utilised the audio feature comparison measures developed in Chapter 4, and evaluated them using an analysis and synthesis approach with subjective evaluation inspired from Chapter 5. Chapter 6 presented an approach for evaluating and comparing objective evaluation metrics for synthesised sound effects. Through analysis and synthesis, where an objective function is used as a fitness function for a parameter optimisation problem, any objective function that correlates highly with a subjective perception of similarity must encapsulate a perceptual representation of similarity. A series of existing objective measures for sound effects were compared. It was shown that the Wichern *et al.* [2007] approach has a strong and statistically significant correlation with subjective similarity measures. The Wichern method defines a set of audio features, described as "loudness, spectral centroid, spectral sparsity, harmonicity, temporal sparsity, and transient index" [Wichern *et al.*, 2007]. This work presented these audio features, as purpose selected for indexing of natural environmental sounds. This agrees with the work presented in this chapter, which had a particular focus on environmental sounds.

It was also uncovered that the objective similarity measure produced in Chapter 4, and evaluated in Chapter 6, while an effective method in some conditions, particularly wind sounds, did not correlate with any subjective ratings of similarity.

It is surprising that only one of the tested objective evaluation functions presented had a high correlation with subjective results, as each of the objective functions had been used in published research for comparing sound effects. Due to the low number of data points being used, the impact of a single outlier will change the results significantly, so viewing the results with caution is essential. Further to this, random parameter settings were compared to the objective functions, and it was found that no method was ever consistently better than random parameter assignment. This demonstrates that there must be an issue either with the random parameter allocation, or every single method performed poorly. To fully explore the implications, in future tests the random parameters should be recalculated for each listening experiment in order to achieve a reasonable distribution, or hand engineered parameters to ensure that the random parameters were not accidentally a good fit. It would also be worthwhile to further explore the appropriateness of this evaluation method outside of the environmental sound context.

Further failings of the synthesis methods may have impacted the results. There were occasions that the synthesis method was unable to produce the sounds represented in the recorded samples, particularly in the context of wind sounds. A larger range of synthesis models, including non-environmental sounds, of better quality, and a further range of sound samples would all be required to confirm the results. The set of objective evaluation metrics used was based primarily on audio feature vector distances, however there are a range of objective functions used in other studies that implement different similarity measures, such as LSE of spectrograms or auditory models of similarity that are not explored here, such as that used by McDermott and Simoncelli [2011]. The method presented, whilst not covering the full range of measures available, still shows a strong correlation with subjective similarity measures for some objective functions, which could have considerable impact in the field of sound synthesis. Either using this objective function for parameter tuning on a range of synthesis methods or for

synthesis method evaluation and validation would aid the field of synthesis research. A standardised objective evaluation methodology with standardised sound samples can transform the field of synthesis research. This has the potential to identify current sound synthesis failings and push the boundary of the perceptual quality of synthesis.

The case study presented in Chapter 7 employed generalised subjective and objective evaluation methodologies of aeroacoustic sound effects. Modifications were made for different test cases, such as an aeolian harp, a propellor and a series of objects swinging through the air. This demonstrated the practical application of a standard evaluation technique to the field, and presents some examples of modifications that can be made for specific use cases. The results show that a physical modelling approach to synthesising these range of sounds perform as least as well as an alternative synthesis approach, with the exception of the metal sword. In the case of object swinging sounds, four of the five different objects were significantly better than SMS. When the results are combined, the physical model approach is significantly better than the alternative synthesis approach, though not as good as a sample. The object swinging results clearly impacted these results, and with just the harp and propellor sounds, it is expected that there would be no discernible difference between the two synthesis approaches.

It should also be noted that the physical modelling approach focuses on modelling aeolian tones, which are produced by the interaction of cylindrical objects and air. In many cases, including a thin metal sword and a propellor blade, this assumption is broken. This is being corrected in work modelling edge tones [Selfridge *et al.*, 2018b]. One of the key advantages of a physically modelled synthesis approach is the interaction and control that can be achieved, so the advantage of a synthesis approach, with some physical interaction parameters, cannot be understated. During the object recognition evaluation, the importance of participant pre-training before an experiment was highlighted. There are many cases in literature where evaluation was performed through recognition or classification [Aldrich, 2005; Gygi *et al.*, 2007; Houix *et al.*, 2012; Kersten and Purwins, 2010; McDermott *et al.*, 2009; Woodcock *et al.*, 2016], however it has been clearly demonstrated that this can be highly dependant on the pre-training and does

not provide us with a greater insight into the effectiveness of our synthesis approach, though it does demonstrate the value of consistent evaluation approaches.

8.1 Future Perspectives

There are clear opportunities to develop a better understanding of the current state-of-the-art within sound synthesis. The following recommendations in this thesis could certainly lead to significant improvements of evaluation within the field of sound effects synthesis. A single process for evaluating synthesis will never be able to encapsulate everything that is required to evaluate such a multidimensional problem as sound synthesis, however, it can help to provide a systematic approach to analysis. It is also the case that measuring the effectiveness of a synthesis method designed to synthesise a real world sound is only one of a range of important evaluation metrics. This type of evaluation does not negate the need for other evaluation forms, but merely adds to the understanding of the utility of existing work. Globally comparing sound synthesis methods and looking within sound groupings can both yield meaningful results. Identification of suitable sonic groupings, in which groups of sounds can be produced and inter-evaluated would be highly beneficial, and would encourage bespoke grouping-based synthesis research, rather than global synthesis approaches.

Developments of a biologically-inspired model for sound similarity, where a psychoacoustic model could be produced to represent similarity in the context of two sounds, would aid in production of objective metrics for system evaluation, and would even facilitate the opportunities of machine learned approaches to synthesising sound effects. The single biggest constraint to high quality intractable synthesis models is the labor intensive time taken to produce effective and realistic sound synthesis. A biologically inspired perceptual model for similarity, combined with an unsupervised learning hierarchical clustering approach, could remove some of these challenges, and potentially allow for an intelligent approach to synthesis.

The potential for implementation of sound similarity measures could be highly advantageous to the sound design community. The creation of sound replacement tools, or the

ability to find synthesis parameters based on sound effect searching, could significantly change the field. The idea of searching for sounds based on similarity would allow a sound designer to quickly limit the search for the correct audio sample. In this case, there could be a list of different sounds, each similar in some different sonic dimension. This would provide a new approach to searching for sounds effects. Furthermore, the synthesis and alignment of previously recorded sounds could be advantageous to the work in sound design communities. For example, footsteps sounds often have to be re-created after the scene is filmed, in order to accurately reflect the sound designer's creative intention. An approach to automatically synchronise and control parameters of a sound from the recording could potentially afford a much greater sense of expressive control within the restrictive time constraints many sound designers experience.

Acronyms

AIC Akaike information criterion

ANOVA ANalysis of VAriance

API Application Program Interface

ARFF Attribute-Relation File Format

BS British Standard

CLI Command Line Interface

CSV Comma-separated values

DC Direct Current

DCT Discrete Cosine Transform

FII Feature Importance Index

FM Frequency Modulation

GMM Gaussian Mixture Model

GPU Graphics Processing Unit

GUI Graphical User Interface

HDF Hierarchical Data Format

HRTF Head Related Transfer Function

ITU-R International Telecommunication Union - Radiocommunication Sector

JSON JavaScript Object Notation

LLD Low Level Descriptors

LSE Least Square Error

MDS Multi-Dimensional Scaling

MFCC Mel Frequency Cepstrum Coefficient

MIR Music Information Retrevial

ML Machine Learning

MPEG Moving Picture Experts Group

MUS MUltiple Stimulus

MUSHRA MUltiple Stimulus Hidden Reference and Anchor

OOB Out-Of-Bag

PEAQ Perceptual Evaluation of Audio Quality

PSO Particle Swarm Optimisation

RS Real Synthetic

SMS Sinusoidal Modelling Synthesis / Spectral Modelling Synthesis

SPAD SPatialized ADditive synthesiser for environmental sounds

SPL Sound Pressure Level

TSV Tab-separated values

URL Uniform Resource Locator

WAET Web Audio Evaluation Tool

XML eXtensible Markup Language

YAML YAML Ain't Markup Language

Glossary

Aeroacoustic A branch of acoustics relating to the sound produced by turbulent fluid motion. In this context of this thesis, the fluid in question is air.

Audio Feature An attribute or aspect description of a piece of audio. Can be a low level attribute, eg. mean frequency, or a high level attribute eg. chord sequence.

Diegetic An aspect of story telling, where the element is within the narration of the story, but not within the world of the story.

Foley The creative practice of performing sound effects.

Gaussian Mixture Model An approach where a feature space is modelled with a number of multivariate gaussians [Bilmes, 1998].

Multivariate A variable which contains more than a single dimension.

Objective Evaluation a computational process for automated evaluation.

OOB Error Out-of-bag error. A process for measuring the error introduced by a single attribute or feature. This is calculated as the difference in error with a given feature both included and excluded.

Perceptual a method that relates to our understanding of human perception.

Random Forest A machine learning approach, where random subsets of features are selected, and a number of decision trees are grown using the CART algorithm [Breiman, 1984, 2001].

Sound Effect A non-musical non-speech sound which can be used to create a specific effect.

Sound Synthesis Artificially generated audio signal.

Subjective Evaluation A human subject performing an evaluation.

Taxonomy A scheme of classification, typically with some form of hierarchical structure.

Unsupervised Learning Shorthand for unsupervised machine learning.

Unsupervised Machine Learning A machine learning approach where the data is unlabelled - often used for data exploration or determining inherent structures of data.

VAMP A plugin format for audio processing and analysis.

Bibliography

Adami, A., Taghipour, A., and Herre, J. (2017). On similarity and density of applause sounds. *Journal of the Audio Engineering Society*, 65(11):897–913.

Aldrich, K. M. (2005). *Perceived similarity between complex sounds: The contribution of acoustic, descriptive and categorical features.* PhD thesis, University of Plymouth.

Aldrich, K. M., Hellier, E. J., and Edworthy, J. (2009). What determines auditory similarity? the effect of stimulus group and methodology. *The Quarterly Journal of Experimental Psychology*, 62(1):63–83.

Allamanche, E., Herre, J., Hellmuth, O., Fröba, B., Kastner, T., and Cremer, M. (2001). Content-based identification of audio material using MPEG-7 low level description. In *Proc. 2nd International Symposium on Music Information Retrieval (ISMIR)*, Indiana, USA.

Amatriain, X., Bonada, J., Loscos, A., and Serra, X. (2002). Spectral processing. In Zölzer, U., editor, *DAFx: Digital Audio Effects*, chapter 10, pages 373–438. John Wiley and Sons, Ltd., Chichester, UK.

An, S. S., James, D. L., and Marschner, S. (2012). Motion-driven concatenative synthesis of cloth sounds. *ACM Transactions on Graphics (TOG)*, 31(4):102.

Aramaki, M., Besson, M., Kronland-Martinet, R., and Ystad, S. (2011). Controlling the perceived material in an impact sound synthesizer. *Audio, Speech, and Language Processing, IEEE Transactions on*, 19(2):301–314.

Aramaki, M., Kronland-Martinet, R., and Ystad, S. (2012). Perceptual control of environmental sound synthesis. In *Speech, Sound and Music Processing: Embracing Research in India*, pages 172–186. Springer Berlin Heidelberg.

Athineos, M. and Ellis, D. P. (2007). Autoregressive modeling of temporal envelopes. *IEEE Transactions on Signal Processing*, 55(11):5237–5245.

Baehrens, D., Schroeter, T., Harmeling, S., Kawanabe, M., Hansen, K., and Müller, K.-R. (2010). How to explain individual classification decisions. *Journal of Machine Learning Research*, 11(Jun):1803–1831.

Bahadoran, P., Benito, A., Vassallo, T., and Reiss, J. D. (2017). A system for online sound effect synthesis. UK Patent 1719854.0 (Pending).

Bahadoran, P., Benito, A., Vassallo, T., and Reiss, J. D. (2018a). FXive: A web platform for procedural sound synthesis. In *Audio Engineering Society Convention 144*, Milan, Italy. Audio Engineering Society.

Bahadoran, P., Benito, A. L., Buchanan, W., and Reiss, J. D. (2018b). FXive: Investigation and implementation of a sound effect synthesis service. In *International Broadcast Convetion.*

Ballas, J. A. (1993). Common factors in the identification of an assortment of brief everyday sounds. *Journal of experimental psychology: human perception and performance*, 19(2):250.

Bar-Joseph, Z., Lischinski, D., Werman, M., Dubnov, S., and El-Yaniv, R. (1999). Granular synthesis of sound textures using statistical learning. In *International Computer Music Conference (ICMC).*

Bascou, C. and Pottier, L. (2005). New sound decomposition method applied to granular synthesis. In *International Computer Music Conference (ICMC)*, Barcelona.

BBC (1931). "the use of sound effects". In *The British Broadcasting Corporation Year Book 1931*, pages 194–197.

BBC (2011). The BBC sound effects library. CD.

Bech, S. and Zacharov, N. (2007). *Perceptual audio evaluation-Theory, method and application*. John Wiley & Sons.

Bensa, J., Jensen, K., Kronland-Martinet, R., and Ystad, S. (2000). Perceptual and analytical analysis of the effect of the hammer impact on the piano tones. In *Proc. International Computer Music Conference (ICMC)*, Berlin, Germany.

Bertin-Mahieux, T., Ellis, D. P., Whitman, B., and Lamere, P. (2011). The million song dataset. In *ISMIR 2011: Proceedings of the 12th International Society for Music Information Retrieval Conference, October 24-28, 2011, Miami, Florida*, pages 591–596. University of Miami.

Bilbao, S. (2009). *Numerical Sound Synthesis: Finite Difference Schemes and Simulations in Musical Acoustics*. Wiley Online Library.

Bilbao, S. and Chick, J. (2013). Finite difference time domain simulation for the brass instrument bore. *The Journal of the Acoustical Society of America*, 134(5):3860–3871.

Bilmes, J. A. (1998). A gentle tutorial of the em algorithm and its application to parameter estimation for gaussian mixture and hidden markov models. *International Computer Science Institute*, 4(510):126.

Black, D., Heise, S., and Loviscach, J. (2009). Generic sound effects to aid in audio retrieval. In *Audio Engineering Society Convention 126*. Audio Engineering Society.

Bogdanov, D., Wack, N., Gómez, E., Gulati, S., Herrera, P., Mayor, O., Roma, G., Salamon, J., Zapata, J. R., and Serra, X. (2013). Essentia: An audio analysis library for music information retrieval. In *International Symposium on Music Information Retrieval (ISMIR)*, pages 493–498.

Bonebright, T. L., Miner, N. E., Goldsmith, T. E., and Caudell, T. P. (2005). Data collection and analysis techniques for evaluating the perceptual qualities of auditory stimuli. *ACM Transactions on Applied Perception (TAP)*, 2(4):505–516.

Bonneel, N., Drettakis, G., Tsingos, N., Viaud-Delmon, I., and James, D. (2008). Fast modal sounds with scalable frequency-domain synthesis. In *ACM Transactions on Graphics (TOG)*, volume 27(3), page 24. ACM.

Böttcher, N., Martínez, H. P., and Serafin, S. (2013). Procedural audio in computer games using motion controllers: an evaluation on the effect and perception. *International Journal of Computer Games Technology*, 2013:6.

Böttcher, N. and Serafin, S. (2009). Design and evaluation of physically inspired models of sound effects in computer games. In *Audio Engineering Society Conference: 35th International Conference: Audio for Games*, London. AES.

Bower, J. L. and Christensen, C. M. (1995). Disruptive technologies: catching the wave. *Harvard Business Review*.

Breiman, L. (1984). *Classification and regression trees*. CRC press.

Breiman, L. (2001). Random forests. *Machine learning*, 45(1):5–32.

Brent, W. (2010). *A timbre analysis and classification toolkit for pure data*. Ann Arbor, MI: MPublishing, University of Michigan Library.

Brossier, P. M. (2006). The aubio library at MIREX 2006. *MIREX 2006*, page 1.

Bruna, J. and Mallat, S. (2013). Audio texture synthesis with scattering moments. *arXiv preprint arXiv:1311.0407*.

Bryant, D. L. (2014). Scalable audio feature extraction. Master's thesis, University of Colorado Colorado Springs.

Bullock, J. (2007). Libxtract: A lightweight library for audio feature extraction. In *Proceedings of the International Computer Music Conference*, volume 43.

Câmpeanu, D. and Câmpeanu, A. (2005). PEAQ–an objective method to assess the perceptual quality of audio compressed files. In *Proceedings of International Symposium on System Theory, SINTES*, volume 12, pages 487–492.

Cannam, C. (2009). The vamp audio analysis plugin API: A programmer's guide. *Available online: http://vamp-plugins. org/guide. pdf*.

Cannam, C., Landone, C., and Sandler, M. (2010). Sonic visualiser: An open source application for viewing, analysing, and annotating music audio files. In *Proceedings of the ACM Multimedia 2010 International Conference*, pages 1467–1468, Firenze, Italy.

Cano, P., Koppenberger, M., Herrera, P., Celma, O., and Tarasov, V. (2004). Sound effects taxonomy management in production environments. In *Proc. AES 25th Int. Conf.*

Cantzos, D., Mouchtaris, A., and Kyriakakis, C. (2005). Multichannel audio resynthesis based on a generalized gaussian mixture model and cepstral smoothing. In *Applications of Signal Processing to Audio and Acoustics, 2005. IEEE Workshop on*, pages 215–218. IEEE.

Caramiaux, B., Bevilacqua, F., Bianco, T., Schnell, N., Houix, O., and Susini, P. (2014). The role of sound source perception in gestural sound description. *ACM Transactions on Applied Perception (TAP)*, 11(1):1:1–1:19.

Casey, M. A., Veltkamp, R., Goto, M., Leman, M., Rhodes, C., and Slaney, M. (2008). Content-based music information retrieval: Current directions and future challenges. *Proceedings of the IEEE*, 96(4):668–696.

Castagné, N. and Cadoz, C. (2000). Physical modeling synthesis: balance between realism and computing speed. In *Proc. International Conference on Digital Audio Effects (DAFx-00)*, page 6.

Castagné, N. and Cadoz, C. (2003). 10 criteria for evaluating physical modelling schemes for music creation. In *Proc. 6th International Conference on Digital Audio Effects (DAFx03)*.

Chowning, J. M. (1973). The synthesis of complex audio spectra by means of frequency modulation. *Journal of the Audio Engineering Society*, 21(7):526–534.

Cleverdon, C. W. and Keen, M. (1966). ASLIB cranfield research project-factors determining the performance of indexing systems; volume 2, test results. Technical report, Cranfield University.

Cook, P. R. (1997). Physically informed sonic modeling (PhISM): Synthesis of percussive sounds. *Computer Music Journal*.

Cook, P. R. (2002). Modeling bill's gait: Analysis and parametric synthesis of walking sounds. In *Audio Engineering Society Conference: 22nd International Conference: Virtual, Synthetic, and Entertainment Audio*, Espoo, Finland. AES.

Cook, P. R. (2007). *Real Sound Synthesis for Interactive Applications (Book & CD-ROM)*. AK Peters, Ltd.

Davies, W. J., Adams, M. D., Bruce, N. S., Cain, R., Carlyle, A., Cusack, P., Hall, D. A., Hume, K. I., Irwin, A., Jennings, P., Marselle, M., Plack, C. J., and Poxon, J. (2013). Perception of soundscapes: An interdisciplinary approach. *Applied Acoustics*, 74(2):224–231.

Deliege, F., Chua, B. Y., and Pedersen, T. B. (2008). High-level audio features: Distributed extraction and similarity search. In *Ninth International Conference on Music Information Retrieval*, pages 565–570.

Dobashi, Y., Yamamoto, T., and Nishita, T. (2003). Real-time rendering of aerodynamic sound using sound textures based on computational fluid dynamics. In *ACM Transactions on Graphics (TOG)*, volume 22, pages 732–740. ACM.

Doel, C. P. v. d. (1998). *Sound synthesis for virtual reality and computer games*. PhD thesis, University of British Columbia.

Downie, J. S. (2004). The scientific evaluation of music information retrieval systems: Foundations and future. *Computer Music Journal*, 28(2):12–23.

Downie, J. S. (2008). The music information retrieval evaluation exchange (2005–2007): A window into music information retrieval research. *Acoustical Science and Technology*, 29(4):247–255.

Dubnov, S., Bar-Joseph, Z., El-Yaniv, R., and Werman, D. L. M. (2002). Synthesizing sound textures through wavelet tree learning. In *Proceedings of the IEEE Conference on Computer Graphics and Applications*.

Eyben, F., Weninger, F., Groß, F., and Schuller, B. (2013). Recent developments in openSMILE, the munich open-source multimedia feature extractor. In *Proceedings of the 21st ACM international conference on Multimedia*, pages 835–838. ACM.

Farnell, A. (2010). *Designing sound*. MIT Press Cambridge, UK.

Fontana, F. and Bresin, R. (2003). Physics-based sound synthesis and control: crushing, walking and running by crumpling sounds. In *Proc. Colloquium on Musical Informatics*, pages 109–114.

Fröjd, M. and Horner, A. (2009). Sound texture synthesis using an overlap–add/granular synthesis approach. *Journal of the Audio Engineering Society*, 57(1/2):29–37.

Gabrielli, L., Squartini, S., and Välimäki, V. (2011). A subjective validation method for musical instrument emulation. In *131st Audio Engineering Society Convention*, New York, USA.

Garcia, R. A. (2001a). Automatic generation of sound synthesis techniques. Master's thesis, Massachusetts Institute of Technology.

Garcia, R. A. (2001b). Automating the design of sound synthesis techniques using evolutionary methods. In *Proc. 4th International Conference on Digital Audio Effects (DAFx)*, Limerick, Ireland.

Gaver, W. W. (1993). What in the world do we hear?: An ecological approach to auditory event perception. *Ecological psychology*, 5(1):1–29.

Gaver, W. W. and Norman, D. A. (1988). *Everyday listening and auditory icons*. PhD thesis, University of California, San Diego, Department of Cognitive Science and Psychology.

Genuer, R., Poggi, J.-M., and Tuleau-Malot, C. (2010). Variable selection using random forests. *Pattern Recognition Letters*, 31(14):2225–2236.

Goodwin, M. (1996). Residual modeling in music analysis-synthesis. In *Acoustics, Speech, and Signal Processing, 1996. ICASSP-96. Conference Proceedings., 1996 IEEE International Conference on*, volume 2, pages 1005–1008. IEEE.

Guggiana, V., Darvishi, A., Munteanu, E., Schauer, H., Motavalli, M., and Rauterberg, M. (1995). Analysis and synthesis of environmental sounds. Technical report, University of Zürich.

Gygi, B., Kidd, G. R., and Watson, C. S. (2007). Similarity and categorization of environmental sounds. *Perception & psychophysics*, 69(6):839–855.

Hahn, H. (2015). *Expressive Sampling Synthesis-Learning Extended Source–Filter Models from Instrument Sound Databases for Expressive Sample Manipulations.* PhD thesis, UPMC Université Paris VI.

Hahn, H. and Röbel, A. (2013). Extended source-filter model for harmonic instruments for expressive control of sound synthesis and transformation. In *Proc. of the 16th International Conference on Digital Audio Effects (DAFx-13).*

Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., and Witten, I. H. (2009). The weka data mining software: an update. *ACM Special Interest Group on Knowledge Discovery and Data Mining (SIGKDD) Explorations Newsletter*, 11(1):10–18.

Hamadicharef, B. and Ifeachor, E. (2003). Objective prediction of sound synthesis quality. In *115th Audio Engineering Society Convention*, New York, USA.

Hamadicharef, B. and Ifeachor, E. (2005). Perceptual modeling of piano tones. In *Audio Engineering Society Convention 119*, Barcelona, Spain.

Harrison-Harsley, R. and Bilbao, S. (2018). Separability of wave solutions in nonlinear brass instrument modelling. *The Journal of the Acoustical Society of America*, 143(6):3654–3657.

Heinrichs, C. and McPherson, A. (2014). Mapping and interaction strategies for performing environmental sound. In *1st Workshop on Sonic Interactions for Virtual Environments at IEEE VR 2014.*

Heinrichs, C., McPherson, A., and Farnell, A. (2014). Human performance of computational sound models for immersive environments. *The New Soundtrack*, 4(2):139–155.

Heise, S., Hlatky, M., and Loviscach, J. (2009). Automatic cloning of recorded sounds by software synthesizers. In *Audio Engineering Society Convention 127*, New York, USA. AES.

Heller, L. M. and Wolf, L. (2002). When sound effects are better than the real thing. *The Journal of the Acoustical Society of America*, 111(5):2339–2339.

Hemery, E. and Aucouturier, J.-J. (2015). One hundred ways to process time, frequency, rate and scale in the central auditory system: a pattern-recognition meta-analysis. *Frontiers in computational neuroscience*, 9.

Hendry, S. and Reiss, J. D. (2010). Physical modeling and synthesis of motor noise for replication of a sound effects library. In *Audio Engineering Society Convention 129*, Los Angeles, CA, USA.

Henter, G. E., Merritt, T., Shannon, M., Mayo, C., and King, S. (2014). Measuring the perceptual effects of modelling assumptions in speech synthesis using stimuli constructed from repeated natural speech. In *Proceedings of the 15th Annual Conference of the International Speech Communication Association, Interspeech*, pages 1504–1508.

Hoffman, M. D. and Cook, P. R. (2006a). Feature-based synthesis: A tool for evaluating, designing, and interacting with music ir systems. In *International Symposium on Music Information Retrieval (ISMIR)*, pages 361–362.

Hoffman, M. D. and Cook, P. R. (2006b). Feature-based synthesis: Mapping acoustic and perceptual features onto synthesis parameters. In *International Computer Music Conference (ICMC)*.

Horner, A. and Wun, S. (2006). Evaluation of iterative matching for scalable wavetable synthesis. In *Audio Engineering Society Conference: 29th International Conference: Audio for Mobile and Handheld Devices*, Seoul, Korea.

Hoskinson, R. (2002). *Manipulation and Resynthesis of Environmental Sounds with Natural Wavelet Grains*. Phd, University of British Columbia.

Houix, O., Lemaitre, G., Misdariis, N., Susini, P., and Urdapilleta, I. (2012). A lexical analysis of environmental sound categories. *Journal of Experimental Psychology: Applied*, 18(1):52.

Huang, D.-Y. (2011). Prediction of perceived sound quality of synthetic speech. In *Proceedings. Asia Pacific Signal and Information processing Association (APSIPA) Annual Summit and Conference*.

ITU-R BS. 1534-1 (2001). Method for the subjective assessment of intermediate sound quality (mushra). *International Telecommunications Union, Geneva*.

ITU-R BS.1387-1 (1998). BS. 1387, method for objective measurements of perceived audio quality. Technical report, ITU-R.

ITU-R BS.1534-3 (2015). BS. 1534, method for subjective assessment of intermediate quality level of audio systems. Technical report, ITU-R.

Jack, R. H., Mehrabi, A., Stockman, T., and McPherson, A. (2018). Action-sound latency and the perceived quality of digital musical instruments: Comparing professional percussionists and amateur musicians. *Music Perception: An Interdisciplinary Journal*, 36(1):109–128.

Jaffe, D. A. (1995). Ten criteria for evaluating synthesis techniques. *Computer Music Journal*, 19(1):76–87.

Järveläinen, H., Verma, T., and Välimäki, V. (2002). Perception and adjustment of pitch in inharmonic string instrument tones. *Journal of New Music Research*, 31(4):311–319.

Jillings, N., De Man, B., Moffat, D., Reiss, J. D., and Stables, R. (2015). Web audio evaluation tool: A browser-based listening test environment. In *Proceedings of the International Sound and Music Computing Conference*, Maynooth, Ireland.

Jillings, N., Moffat, D., De Man, B., and Reiss, J. D. (2016). Web audio evaluation tool: A framework for subjective assessment of audio. In *Proc. 2nd Web Audio Conference*, Atlanta, Georgia, USA.

Kahrs, M. and Avanzini, F. (2001). Computer synthesis of bird songs and calls. In *Proc. 4th International Conference on Digital Audio Effects (DAFx)*, pages 23–27, Limerick, Ireland.

Karjalainen, M., Välimäki, V., and Esquef, P. A. (2002). Efficient modeling and synthesis of bell-like sounds. In *the 5th International Conference on Digital Audio Effects,(Hamburg, Germany)*, pages 181–186.

Karjalainen, M., Välimäki, V., and Jánosy, Z. (1993). Towards high-quality sound synthesis of the guitar and string instruments. In *Proceedings of the International Computer Music Conference*, pages 56–56.

Kersten, S. and Purwins, H. (2010). Sound texture synthesis with hidden markov tree models in the wavelet domain. In *7th Sound and Music Computing Conference (SMC2010)*, Barcelona, Spain.

Kersten, S. and Purwins, H. (2012). Fire texture sound re-synthesis using sparse decomposition and noise modelling. In *Proc. 15th International Conference on Digital Audio Effects (DAFx-12)*, York, UK.

Kleimola, J. (2013). *Nonlinear abstract sound synthesis algorithms*. PhD thesis, Aalto University.

Koehl, V., Paquier, M., and Delikaris-Manias, S. (2011). Comparison of subjective assessments obtained from listening tests through headphones and loudspeaker setups. In *Audio Engineering Society Convention 131*. Audio Engineering Society.

Kreutzer, C., Walker, J., and O'Neill, M. (2008). A parametric model for spectral sound synthesis of musical sounds. In *Audio, Language and Image Processing, 2008. ICALIP 2008. International Conference on*, pages 633–637. IEEE.

Lafay, G., Misdariis, N., Lagrange, M., and Rossignol, M. (2016). Semantic browsing of sound databases without keywords. *Journal of the Audio Engineering Society*, 64(9):628–635.

Lakatos, S., McAdams, S., and Caussé, R. (1997). The representation of auditory source characteristics: Simple geometric form. *Attention, Perception, & Psychophysics*, 59(8):1180–1190.

Lartillot, O. and Toiviainen, P. (2007). A matlab toolbox for musical feature extraction from audio. In *International Conference on Digital Audio Effects*, pages 237–244.

Lee, J. S., Depalle, P., and Scavone, G. (2010). Analysis/ synthesis of rolling sounds using a source filter approach. In *Proc. 13th International Conference on Digital Audio Effects (DAFx-10)*, Graz, Austria.

Lerch, A., Eisenberg, G., and Tanghe, K. (2005). FEAPI: A low level feature extraction plugin api. In *Proceedings of the International Conference on Digital Audio Effects*.

Li, T., Ogihara, M., and Tzanetakis, G. (2011). *Music data mining*. CRC Press.

Lindborg, P. (2016). A taxonomy of sound sources in restaurants. *Applied Acoustics*, 110:297–310.

Lindsay, A. T. and Herre, J. (2001). MPEG-7 and MPEG-7 audio – an overview. *Journal of the Audio Engineering Society*, 49(7/8):589–594.

Ma, X., Fellbaum, C., and Cook, P. R. (2010). Soundnet: investigating a language composed of environmental sounds. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 1945–1954. ACM.

Mäkinen, T., Kiranyaz, S., Pulkkinen, J., and Gabbouj, M. (2012). Evolutionary feature generation for content-based audio classification and retrieval. In *Signal Processing Conference (EUSIPCO), 2012 Proceedings of the 20th European*, pages 1474–1478. IEEE.

Manjunath, B. S., Salembier, P., and Sikora, T. (2002). *Introduction to MPEG-7: multimedia content description interface*, volume 1. John Wiley & Sons.

Marini, F. and Walczak, B. (2015). Particle swarm optimization (PSO). a tutorial. *Chemometrics and Intelligent Laboratory Systems*, 149:153–165.

Mathieu, B., Essid, S., Fillon, T., Prado, J., and Richard, G. (2010). YAAFE, an easy to use and efficient audio feature extraction software. In *ISMIR*, pages 441–446.

Mauch, M. and Ewert, S. (2013). The audio degradation toolbox and its application to robustness evaluation. In *Proc. 14th International Society for Music Information Retrieval Conference (ISMIR13)*, pages 83–88, Curitiba, Brazil.

McDermott, J., Griffith, N. J., and O'Neill, M. (2008). Evolutionary computation applied to sound synthesis. In *The Art of Artificial Evolution*, pages 81–101. Springer.

McDermott, J. H., Oxenham, A. J., and Simoncelli, E. P. (2009). Sound texture synthesis via filter statistics. In *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, pages 297–300, New Paltz, NY, USA.

McDermott, J. H. and Simoncelli, E. P. (2011). Sound texture perception via statistics of the auditory periphery: evidence from sound synthesis. *Neuron*, 71(5):926–940.

McDermott, J. H., Wrobleski, D., and Oxenham, A. J. (2011). Recovering sound sources from embedded repetition. *Proceedings of the National Academy of Sciences*, 108(3):1188–1193.

McFee, B., Raffel, C., Liang, D., Ellis, D. P., McVicar, M., Battenberg, E., and Nieto, O. (2015). librosa: Audio and music signal analysis in python. In *Proceedings of the 14th python in science conference*, pages 18–25.

McGregor, I., Leplâtre, G., Crerar, A., and Benyon, D. (2006). Sound and soundscape classification: establishing key auditory dimensions and their relative importance. In *12th International Conference on Auditory Display*, London, UK.

McKay, C., Fujinaga, I., and Depalle, P. (2005). jAudio: A feature extraction library. In *Proc. 6th International Conference on Music Information Retrieval*, pages 600–3.

McKinney, M. F. and Breebaart, J. (2003). Features for audio and music classification. In *Proc. 4th International Conference on Music Information Retrieval(ISMIR)*, pages 151–158, Baltimore, Maryland, USA.

Mengual, L., Moffat, D., and Reiss, J. D. (2016). Modal synthesis of weapon sounds. In *Proc. Audio Engineering Society Conference: 61st International Conference: Audio for Games*, London. Audio Engineering Society.

Merer, A., Aramaki, M., Ystad, S., and Kronland-Martinet, R. (2013). Perceptual characterization of motion evoked by sounds for synthesis control purposes. *ACM Transactions on Applied Perception (TAP)*, 10(1):1–24.

Merer, A., Ystad, S., Kronland-Martinet, R., and Aramaki, M. (2011). Abstract sounds and their applications in audio and perception research. *Exploring music contents*, pages 176–187.

Miner, N. E. and Caudell, T. P. (2005). Using wavelets to synthesize stochastic-based sounds for immersive virtual environments. *ACM Transactions on Applied Perception (TAP)*, 2(4):521–528.

Misra, A. and Cook, P. R. (2009a). Toward synthesized environments: A survey of analysis and synthesis methods for sound designers and composers. In *International Computer Music Conference (ICMC)*.

Misra, A. and Cook, P. R. (2009b). *Toward synthesized environments: A survey of analysis and synthesis methods for sound designers and composers*. Ann Arbor, MI: MPublishing, University of Michigan Library.

Mitrović, D., Zeppelzauer, M., and Breiteneder, C. (2010). Features for content-based audio retrieval. *Advances in computers*, 78:71–150.

Moffat, D. and Reiss, J. D. (2018a). Objective evaluations of synthesised environmental sounds. In *Proc. 21th International Conference on Digital Audio Effects (DAFx-17)*, Aveiro, Portugal.

Moffat, D. and Reiss, J. D. (2018b). Perceptual evaluation of synthesized sound effects. *ACM Transactions on Applied Perception (TAP)*, 15(2):19.

Moffat, D., Ronan, D., and Reiss, J. D. (2015). An evaluation of audio feature extraction toolboxes. In *Proc. 18th International Conference on Digital Audio Effects (DAFx-15)*.

Moffat, D., Ronan, D., and Reiss, J. D. (2017). Unsupervised taxonomy of sound effects. In *Proc. 20th International Conference on Digital Audio Effects (DAFx-17)*, Edinburgh, UK.

Moffat, D., Selfridge, R., and Reiss, J. D. (2019). Sound effect synthesis. In Filimowicz, M., editor, *Foundations in Sound Design for Interactive Media: A Multidisciplinary Approach*. Routledge.

Morrell, M. J., Harte, C. A., and Reiss, J. D. (2011). Queen Mary's "Media and Arts Technology studios" audio system design. In *130th Audio Engineering Society Convention*. Audio Engineering Society.

Moss, W., Yeh, H., Hong, J.-M., Lin, M. C., and Manocha, D. (2010). Sounding liquids: Automatic sound synthesis from fluid simulation. *ACM Transactions on Graphics (TOG)*, 29(3):21.

Murphy, E., Lagrange, M., Scavone, G., Depalle, P., and Guastavino, C. (2008). Perceptual evaluation of a real-time synthesis technique for rolling sounds. In *Conference on Enactive Interfaces*, Pisa, Italy. Interactive Design Foundation.

Nordahl, R., Serafin, S., and Turchet, L. (2010). Sound synthesis and evaluation of interactive footsteps for virtual reality applications. In *IEEE Virtual Reality Conference*, pages 147–153, Waltham, MA, USA. IEEE.

Oksanen, S., Parker, J., and Välimäki, V. (2013). Physically informed synthesis of jackhammer tool impact sounds. In *Proc. of the 16th Int. Conference on Digital Audio Effects (DAFx-13)*, Maynooth, Ireland.

O'Leary, S. and Robel, A. (2014). A montage approach to sound texture synthesis. In *22nd European Signal Processing Conference (EUSIPCO)*, pages 939–943, Lisbon, Portugal. IEEE.

O'Modhrain, S. and Essl, G. (2004). Pebblebox and crumblebag: tactile interfaces for granular synthesis. In *New Interfaces for Musical Expression (NIME)*, Singapore.

O'Regan, D. and Kokaram, A. (2007a). Multi-resolution sound texture synthesis using the dual-tree complex wavelet transform. In *15th European Signal Processing Conference EUSIPCO*, pages 350–354, Poznan, Poland. IEEE.

O'Regan, D. and Kokaram, A. (2007b). Wavelet based high resolution sound texture synthesis. In *Proc. Audio Engineering Society 31st International Conference: New Directions in High Resolution Audio*.

Pachet, F. and Aucouturier, J.-J. (2004). Improving timbre similarity: How high is the sky. *Journal of negative results in speech and audio sciences*, 1(1):1–13.

Page, K. R., Fields, B., De Roure, D., Crawford, T., and Downie, J. S. (2012). Reuse, remix, repeat: the workflows of MIR. In *Proc. 13th International Society for Music Information Retrieval Conference (ISMIR-12)*, pages 409–414.

Pampin, J. (2004). ATS: A system for sound analysis transformation and synthesis based on a sinusoidal plus critical-band noise model and psychoacoustics. In *Proceedings of the International Computer Music Conference.*, pages 402–405, Miami, FL.

Paul, S., McCulloch, P., and Sedrakyan, A. (2013). Robotic surgery: revisiting "no innovation without evaluation". *BMJ: British Medical Journal*, 346(f1573).

Pearce, A., Brookes, T., and Mason, R. (2017). Timbral attributes for sound effect library searching. In *Audio Engineering Society Conference: 2017 AES International Conference on Semantic Audio.*

Pedersen, T. H. (2008). *The Semantic Space of Sounds.* Delta.

Peeters, G. (2004). A large set of audio features for sound description (similarity and classification) in the cuidado project. Technical report, IRCAM.

Peeters, G., Giordano, B. L., Susini, P., Misdariis, N., and McAdams, S. (2011). The timbre toolbox: Extracting audio descriptors from musical signals. *The Journal of the Acoustical Society of America*, 130(5):2902–2916.

Peltola, L., Erkut, C., Cook, P. R., and Valimaki, V. (2007). Synthesis of hand clapping sounds. *IEEE Transactions on Audio, Speech, and Language Processing*, 15(3):1021–1029.

Petrausch, S. and Rabenstein, R. (2003). Sound synthesis by physical modeling using the functional transformation method: Efficient implementations with polyphase-filterbanks. In *Proc. International Conference on Digital Audio Effects.*

Puronas, V. (2014). Sonic hyperrealism: illusions of a non-existent aural reality. *The New Soundtrack*, 4(2):181–194.

Raffel, C., McFee, B., Humphrey, E. J., Salamon, J., Nieto, O., Liang, D., and Ellis, D. P. (2014). mir_eval: A Transparent Implementation of Common MIR Metrics. In

Proc. of the 15th International Society for Music Information Retrieval Conference, Taipei, Taiwan.

Raghuvanshi, N. and Lin, M. C. (2006). Interactive sound synthesis for large scale environments. In *Proceedings of the 2006 symposium on Interactive 3D graphics and games*, pages 101–108. ACM.

Rath, M. (2003). An expressive real-time sound model of rolling. In *6th Int. Conference on Digital Audio Effects (DAFx-03)*, London.

Rawlinson, H., Segal, N., and Fiala, J. (2015). Meyda: an audio feature extraction library for the web audio api. In *Web Audio Conference*. Web Audio Conference.

Reiss, J. D. (2016). A meta-analysis of high resolution audio perceptual evaluation. *Journal of the Audio Engineering Society*, 64(6).

Reiss, J. D. and Sandler, M. (2002a). Benchmarking music information retrieval systems. *JCDL Workshop on the Creation of Standardized Test Collections, Tasks, and Metrics for Music Information Retreival (MIR) and Music Digital Library (MDL) Evaluation*, pages 37–42.

Reiss, J. D. and Sandler, M. (2002b). Beyond recall and precision: A full framework for MIR system evaluation. In *3rd Annual International Symposium on Music Information Retrieval*, Paris, France.

Reiss, J. D. and Sandler, M. (2003). MIR benchmarking: Lessons learned from the multimedia community. *The MIR/MDL Evaluation Project White Paper Collection*, 3:114–120.

Ren, Z., Yeh, H., and Lin, M. C. (2013). Example-guided physically based modal sound synthesis. *ACM Transactions on Graphics (TOG)*, 32(1):1.

Rice, S. V. and Bailey, S. M. (2004). Searching for sounds: A demonstration of findsounds. com and findsounds palette. In *ICMC*.

Robinson, D. J. M. and Hawksfords, M. J. (2000). Psychoacoustic models and non-linear human hearing. In *109th Audio Engineering Society Convention*, Los Angeles, CA, USA.

Rocchesso, D., Bresin, R., and Fernstrom, M. (2003). Sounding objects. *IEEE MultiMedia*, 10(2):42–52.

Rocchesso, D. and Fontana, F. (2003). *The sounding object*. Mondo estremo.

Ronan, D., Ma, Z., Mc Namara, P., Gunes, H., and Reiss, J. D. (2018). Automatic minimisation of masking in multitrack audio using subgroups. *ArXiv e-prints*.

Ronan, D., Moffat, D., Gunes, H., and Reiss, J. D. (2015). Automatic subgrouping of multitrack audio. In *Proc. 18th International Conference on Digital Audio Effects (DAFx-15)*. DAFx-15.

Rumsey, F., Zieliński, S., Kassier, R., and Bech, S. (2005). On the relative importance of spatial and timbral fidelities in judgments of degraded multichannel audio quality. *The Journal of the Acoustical Society of America*, 118(2):968–976.

Russolo, L. (2004). The art of noises: Futurist manifesto. *Audio culture: Readings in modern music*, pages 10–14.

Russolo, L. and Pratella, F. B. (1967). *The art of noise:(futurist manifesto, 1913)*. Something Else Press.

Rychtáriková, M. and Vermeir, G. (2013). Soundscape categorization on the basis of objective acoustical parameters. *Applied Acoustics*, 74(2):240–247.

Salamon, J. and Bello, J. P. (2015). Unsupervised feature learning for urban sound classification. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 171–175.

Salamon, J. and Bello, J. P. (2017). Deep convolutional neural networks and data augmentation for environmental sound classification. *IEEE Signal Process. Lett.*, 24(3):279–283.

Salamon, J., Jacoby, C., and Bello, J. P. (2014). A dataset and taxonomy for urban sound research. In *Proceedings of the 22nd ACM international conference on Multimedia*, pages 1041–1044. ACM.

Scavone, G. P., Lakatos, S., Cook, P. R., and Harbke, C. (2001). Perceptual spaces for sound effects obtained with an interactive similarity rating program. In *Proceedings of International Symposium on Musical Acoustics*.

Schafer, R. M. (1993). *The soundscape: Our sonic environment and the tuning of the world*. Inner Traditions/Bear & Co.

Schatz, R., Egger, S., and Masuch, K. (2012). The impact of test duration on user fatigue and reliability of subjective quality ratings. *Journal of the Audio Engineering Society*, 60(1/2):63–73.

Schwarz, D. (2000). Data-driven concatenative sound synthesis. *DAFx*.

Schwarz, D. (2011). State of the art in sound texture synthesis. In *14th International Conference Digital Audio Effects (DAFx)*, pages 221–231, Paris, France.

Schwarz, D., Beller, G., Verbrugghe, B., and Britton, S. (2006). Real-time corpus-based concatenative synthesis with catart. In *Proceedings of the International Conference on Digital Audio Effects (DAFx-06)*, pages 279–282, Montreal, Canada.

Schwarz, D. and O'Leary, S. (2015). Smooth granular sound texture synthesis by control of timbral similarity. In *Proc. Sound and Music Computing (SMC)*, page 6.

Schwarz, D., Roebel, A., Yeh, H., and La Burthe, A. (2016). Concatenative sound texture synthesis methods and evaluation. In *19th International Conference on Digital Audio Effects (DAFx)*, Brno, Czech Republic.

Selfridge, R., Moffat, D., Avital, E. J., and Reiss, J. D. (2018a). Creating real-time aeroacoustic sound effects using physically informed models. *Journal of the Audio Engineering Society*, 66(7/8):594–607.

Selfridge, R., Moffat, D., and Reiss, J. D. (2017a). Physically inspired sound synthesis model of a propeller. In *ACM Audio Mostly Conference*, London, UK.

Selfridge, R., Moffat, D., and Reiss, J. D. (2017b). Real-time physical model for synthesis of sword swing sounds. In *International Conference on Sound and Music Computing (SMC)*, Espoo, Finland.

Selfridge, R., Moffat, D., and Reiss, J. D. (2017c). Sound synthesis of objects swinging through air using physical models. *Applied Sciences*, 7(11).

Selfridge, R., Moffat, D., Reiss, J. D., and Avital, E. J. (2017d). Real-time physical model for an aeolian harp. In *International Congress on Sound and Vibration*, London, UK.

Selfridge, R., Reiss, J. D., and Avital, E. (2017e). Physically derived synthesis model of a cavity tone. In *Proc. 20th International Conference on Digital Audio Effects (DAFx-17)*.

Selfridge, R., Reiss, J. D., and Avital, E. J. (2018b). Physically derived synthesis model of an edge tone. In *Proc. 144th International Convention of the Audio Engineering Society*, Milan, Italy.

Selfridge, R., Reiss, J. D., Avital, E. J., and Xiaolong, T. (2016). Physically derived synthesis model of an aeolian tone. In *141th Audio Engineering Society Convention*, Los Angeles, CA, USA.

Serra, X. and Smith, J. (1990). Spectral modeling synthesis: A sound analysis/synthesis system based on a deterministic plus stochastic decomposition. *Computer Music Journal*, 14(4):12–24.

Siddiq, S. (2017). Data-driven granular synthesis. In *Audio Engineering Society Convention 142*.

Smith, J. O. (1991). Viewpoints on the history of digital synthesis. In *Proceedings of the International Computer Music Conference*.

Smith, J. O. and Serra, X. (1987). *PARSHL: An analysis/synthesis program for non-harmonic sounds based on a sinusoidal representation*. CCRMA, Department of Music, Stanford University.

Sonnenschein, D. (2001). *Sound design: The expressive power of music, voice, and sound effects in cinema*. Michael Wiese Productions Studio City.

Stables, R., Enderby, S., De Man, B., Fazekas, G., and Reiss, J. D. (2014). SAFE: A system for the extraction and retrieval of semantic audio descriptors. In *15th International Society for Music Information Retrieval Conference (ISMIR 2014)*.

Stevenson, I. (2016). Soundscape analysis for effective sound design in commercial environments. In *Sonic Environments Australasian Computer Music Conference*. Australasian Computer Music Association.

Sturm, B. L. (2004). Matconcat: an application for exploring concatenative sound synthesis using matlab. In *Proc. 7th International Conference on Digital Audio Effects (DAFx-04)*, volume 2, Naples, Italy.

Theobald, B.-J. and Matthews, I. (2012). Relating objective and subjective performance measures for aam-based visual speech synthesis. *Audio, Speech, and Language Processing, IEEE Transactions on*, 20(8):2378–2387.

Thiede, T., Treurniet, W. C., Bitto, R., Schmidmer, C., Sporer, T., Beerends, J. G., and Colomes, C. (2000). PEAQ-The ITU standard for objective measurement of perceived audio quality. *Journal of the Audio Engineering Society*, 48(1/2):3–29.

Tolonen, T., Välimäki, V., and Karjalainen, M. (1998). Evaluation of modern sound synthesis methods. Technical report, Helsinki University of Technology.

Tremblay, S., Nicholls, A. P., Alford, D., and Jones, D. M. (2000). The irrelevant sound effect: Does speech play a special role? *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 26(6):1750.

Turchet, L., Moffat, D., Tajadura-Jiménez, A., Reiss, J. D., and Stockman, T. (2016). What do your footsteps sound like? an investigation on interactive footstep sounds adjustment. *Applied Acoustics*, 111:77–85.

Turnbull, D., Barrington, L., Torres, D., and Lanckriet, G. (2008). Semantic annotation and retrieval of music and sound effects. *Audio, Speech, and Language Processing, IEEE Transactions on*, 16(2):467–476.

Tzanetakis, G. and Cook, P. (2000). Marsyas: A framework for audio analysis. *Organised sound*, 4(03):169–175.

Valentini-Botinhao, C., Yamagishi, J., and King, S. (2011). Evaluation of objective measures for intelligibility prediction of hmm-based synthetic speech in noise. In *Acoustics, Speech and Signal Processing (ICASSP), 2011 IEEE International Conference on*, pages 5112–5115. IEEE.

Valimaki, V., Tolonen, T., and Karjalainen, M. (1999). Plucked-string synthesis algorithms with tension modulation nonlinearity. In *Acoustics, Speech, and Signal Processing, 1999. Proceedings., 1999 IEEE International Conference on*, volume 2, pages 977–980. IEEE.

Van Den Doel, K., Kry, P. G., and Pai, D. K. (2001). FoleyAutomatic: physically-based sound effects for interactive simulation and animation. In *28th Annual conference on Computer graphics and interactive techniques (SIGGRAPH)*, pages 537–544, Los Angeles, CA, USA. ACM.

Van den Doel, K. and Pai, D. K. (2003). Modal synthesis for vibrating objects. *Audio Anectodes. AK Peter, Natick, MA*, page 8.

Verma, T. S. and Meng, T. H. (2000). Extending spectral modeling synthesis with transient modeling synthesis. *Computer Music Journal*, 24(2):47–59.

Verron, C., Aramaki, M., Kronland-Martinet, R., and Pallone, G. (2010a). A 3D immersive synthesizer for environmental sounds. *IEEE Transactions on Audio, Speech, and Language Processing*, 18(6):1550–1561.

Verron, C., Aramaki, M., Kronland-Martinet, R., and Pallone, G. (2010b). Spatialized synthesis of noisy environmental sounds. In *International Conference on Auditory Display*, pages 392–407, Atlanta, Georgia, USA.

Vincent, E., Gribonval, R., and Févotte, C. (2006). Performance measurement in blind audio source separation. *IEEE transactions on audio, speech, and language processing*, 14(4):1462–1469.

Virtanen, T. and Helén, M. (2007). Probabilistic model based similarity measures for audio query-by-example. In *Applications of Signal Processing to Audio and Acoustics, 2007 IEEE Workshop on*, pages 82–85. IEEE.

Ward Jr, J. H. (1963). Hierarchical grouping to optimize an objective function. *Journal of the American statistical association*, 58(301):236–244.

Webb, C. J. and Bilbao, S. (2015). On the limits of real-time physical modelling synthesis with a modular environment. In *Proceedings of the International Conference on Digital Audio Effects*, page 65.

Wichern, G., Thornburg, H., Mechtley, B., Fink, A., Tu, K., and Spanias, A. (2007). Robust multi-features segmentation and indexing for natural sound environments. In *International Workshop on Content-Based Multimedia Indexing (CBMI)*, pages 69–76. IEEE.

Winfield, R. (1993). Windsongs: The sound of aeolian harps. Saydisc Records.

Wold, E., Blum, T., Keislar, D., and Wheaten, J. (1996). Content-based classification, search, and retrieval of audio. *IEEE multimedia*, 3(3):27–36.

Woodcock, J., Davies, W. J., Cox, T. J., and Melchior, F. (2016). Categorization of broadcast audio objects in complex auditory scenes. *Journal of the Audio Engineering Society*, 64(6):380–394.

Yee-King, M. and Roth, M. (2011). A comparison of parametric optimization techniques for musical instrument tone matching. In *Audio Engineering Society Convention 130*.

Zheng, C. and James, D. L. (2011). Toward high-quality modal contact sound. In *ACM Transactions on Graphics (TOG)*, volume 30.

Zita, A. (2003). Computational real-time sound synthesis of rain. Master's thesis, Institutionen för teknik och naturvetenskap.

Zölzer, U. and Amatriain, X. (2002). *DAFX: digital audio effects*, volume 1. Wiley Online Library.