Modelling the Perception of Percussive Low Frequency Instruments in Rooms

Michael Howard

MSc By Research

University of Salford
School of Science, Environment & Engineering
2020

Abstract

Throughout the study of room acoustics, adverse effects and methods of controlling and mitigating modal behaviour are a well researched topic. However, despite this, a gap between objective metrics and subjective results is still prevalent, thus resulting in a limited understanding of perceived bass quality.

Previous work has suggested a group of perceptual attributes that are useful in describing the effect of room acoustics on the perceived bass quality, however an objective link to the perceptual attributes has not been quantified. Furthermore, the scope of previous work is mostly concerned with small listening rooms and rarely extends to other cases, such as that found in live sound reinforcement.

Hence, this work is focused on broadening the understanding of low frequency quality due to modal behaviour in rooms, through extending the scope of research to include larger listening environments and single instrument excitation of the room.

To investigate the characteristics of low frequency quality, various kick drums were auralised using an improvement to the modal decomposition model and were then rated in a subjective listening test using the descriptive bass quality attributes. From the results, the attributes were modelled through a novel approach using a Random Forest model, utilising a combination of acoustic and MIR features.

It was found that the perceptual attributes of both Resonance and Articulation were predicted effectively from signal features, however Bass Energy was unable to be modelled with any accuracy. Use of feature selection algorithms revealed that Resonance and Articulation attributes relied on temporal and decay based features, such as early decay time and temporal centroid. This result further suggests the importance of temporal modal behaviour when considering audible effects due to low frequency modes. The outcome of this work supports the growing body of work that the effects of modal density are not as important as traditionally thought and is therefore applicable to both small and large rooms.

Acknowledgements

The author would like to acknowledge Bruno Fazenda, Carlo Bolla, Alessandro Palladini and Jonathan Hargreaves for their support, mentoring and advice throughout the project.
Music Tribe UK for sponsoring the project.
The Machine Learning team at MIDAS for participating in the listening tests and for advice and support throughout the project.
Rita Campos for creating and implementing the listening test GUI.

Declaration

This work is my own. The work of others used in its completion has been duly acknowledged. Experimental or other investigative results have not been falsified. I have read and understood the University Policy on the Conduct of Assessed Work. By submitting this assessment, I am declaring that I am fit to do so. I understand that personal mitigating circumstances requests which relate to the standard I have achieved in this assessment may become null and void.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Background

The problems of modal behaviour in rooms are well reported across research into room acoustics with a plethora of methods of control. However a perceptual gap exists between current understandings of room modes on perceived quality of sound reinforcement.

Furthermore, this gap in perception is merely emphasised when moving the attention to larger room volumes - which extends beyond the scope of most low frequency room acoustic research. These larger rooms often have different use cases over the smaller listening room environments, one such example being live sound reinforcement, where the typical workflow involves correcting the mix of individual instruments to account for modal artefacts.

Another factor that must be taken into account when investigating live sound reinforcement is the time sensitive role of the Live Sound Engineer, where optimal correction can take a considerable amount of time. This issue is particularly problematic in band limited low frequency instruments such as bass guitars and kick drums. However, due to the highly contrasting natures of the two types of instruments, this work will only focus on kick drums, as there is far less variation across independent factors in kick drums than bass guitars.

On the topic of low frequency quality, recent advancements in machine learning will allow for the use of building complex models that can map signal features to perceptual attributes of low frequency quality. This may be used to understand the perceptual gap between room acoustic characteristics and perceived quality of bass reproduction.

1.2 Motivation of research

Due to the aforementioned issues, there exists a gap in research where single instrument excitation in large rooms has not been investigated when relating the perception of quality in low frequency reproduction. This is particularly of concern in workflows such as live sound engineering, where information about how the acoustic environment may affect the perceived quality of reproduction may aid the correction process. Therefore an informative model that can accurately predict the perceived attributes of low frequency quality in kick drums due to modal artefacts, may prove

a useful tool in aiding sound engineers to understand the potential problems that may arise.

1.3 Aims of the Thesis

The aim of this thesis is to therefore construct an interpretive model for descriptive bass quality attributes. This will help form a greater understanding into the perception of low frequency sound through interpretation of useful signal features and prediction of quality scores in novel rooms/kick drum combinations.

1.4 Objectives

The objectives of this thesis are as follows:

- Model the acoustic modal behaviour of rooms for the purpose of auralisation.
- Conduct a listening test using the auralisation process to obtain perceptually relevant ratings of bass quality attributes.
- Create an appropriate feature extraction pipeline for the bespoke nature of the research problem.
- Form interpretive perceptual predictors through modelling the bass quality attributes using informative signal features.
- Use of the perceptual predictors to form a greater understanding of low frequency perception.

1.5 Contribution to knowledge

The contributions to knowledge formed from this thesis are the following:

- Furthered the available knowledge and data of how perceptual descriptors: Articulation, Resonance and Bass Energy are rated, particularly for impulsive instruments and large rooms.
- Formed a perceptual model of signal features which can accurately describe the perceived resonance and articulation of an auralised kick drum in a variety of room sizes.

1.6 Outline

To outline the contents of this research, first a review of literature is conducted to form a relevant understanding in the work of low frequency sound reproduction and common methods of modal correction. A review on perceptual modelling is done to understand how previous work has formed perceptual predictors for different use cases and applications. A look into the perceptual nature of low frequency room acoustics is presented to investigate what may be useful when forming a perceptual

model. Finally, a look into how the scope of research changes when moving to the bespoke case of live sound reinforcements and single room excitation.

Then, relevant theory for low frequency room acoustics is presented, featuring an overview of modal resonances, their underlying characteristics, perception and how they may be modelled empirically using a source receiver example. Furthermore, a brief overview on some perceptual attributes of the low frequency are described which are used throughout this Thesis. Finally, an introduction to the machine learning approaches used throughout this work are described, as well as the underlying signal features that are used.

The Methodology is then outlined to describe the relevant methods that have been implemented to model and auralise room responses. Further to this, the machine learning model implementations and variables with their underlying signal features are then outlined.

To design and describe the subjective listening test, the experimental methodology outlines the variables in test, with justification for the control and scope is presented. Then an outline of the logistic elements of the test are described, as well as previous results from a preliminary test that helped guide the control of test conditions in the primary subjective test.

The results from the subjective listening test are then analysed and presented to first form an understanding between the acoustic spaces and subjective responses, where the subjective responses are then modelled using the signal features and subjective attribute ratings. The models are then put under a generalisation test to expand the understanding of the perceptual attributes.

Finally the results are discussed with relation to the suitability of attributes in the context of this research, interactions between the subjective attributes and how previous research relates to the findings of this study.

Chapter 2

Literature Review

The aim of this section is to provide relevant background to the research problem and to define both the scope of research and the research objectives, which will be guided by a review of literature and previous work.

To begin, background is provided by outlining typical issues that arise in bass sound reproduction in rooms and why lower frequencies are particularly problematic.

After outlining some of the complications that arise due to room acoustics, it is then key to understand how these effects are often mitigated to maximise perceived bass quality in sound reproduction.

Furthermore, exploration of current methods of modelling and measuring perceived quality are presented with the typical approaches that are used. The objective is to then understand what parameters may be influencing low frequency quality by turning the attention to previous work in perception of low frequency.

Finally, the scope is adjusted to investigate low frequency away from the typical environment that most literature is concerned with, such as the particular case of live sound, where single instruments are typically tackled and tuned through corrective methods to avoid problematic excitation of low frequency room modes.

2.1 Introduction to Low Frequency Room Acoustics

Negative effects of low frequency are well documented and reported as being one of the determining factors in control room acoustics (B. Fazenda & Davies, 2002). Further to this, difficulty in controlling sound reproduction systems for bass is a major focus of literature. (R. J. Wilson, 2006) provides a useful insight into where the current state of research and literature was focused at the time of writing, in the aptly named "Can we get the bass right?"

The fundamental aspect of why bass frequencies are so problematic, is due to the modal nature of the sound field at lower frequencies, where standing waves exist between the boundaries of the room, comprised of resonances where the wavelengths are comparative to the size of the room (Stephenson, 2012). Furthermore, this modal sound field interacts differently when compared with higher frequencies, which are more diffuse and are often more sensitive to the increasing absorption efficiency of materials, leading to an imbalance in reverberence time across the frequency spectrum (Angus, 1997). Consequently, this has a dominating effect on the reproduction

of low frequencies in rooms.

2.2 Modal Control

Early work in low frequency control is mostly concerned with modal spacing and density, where the low frequency is tackled from a perspective of individual modal resonances being ideally spaced through specific ratios of the room's Geometry. Examples of recommended approaches to modal spacing are Bolt's Area (Bolt, 1946) and Bonello's Criteria (Bonello, 1981) among other methods discussed in (T. J. Cox et al., 2004) and (Stephenson, 2012). However, there is much debate into the subjective validity of how modal spacing influences the low frequency, such as (T. S. Welti, 2009), (T. J. Cox et al., 2004) and (Wankling & Fazenda, 2009). One of the main limitations of these investigations into modal spacing, is that the ideal ratios only apply to perfectly rectangular rooms. Furthermore, these control methods only exist in the room design phase where geometry is at complete control of the Acoustic Engineer, which is not always the case.

(Toole, 2013) presents three methodologies that may be considered when tackling problematic room modes that can be implemented when design of the room is not at control; Loudspeaker position, listener positions and use of equalization. However, it must be noted that positional corrections are focused on small rectangular listening rooms, as the approach takes advantage of the regular shape of the room boundaries. Furthermore, the loudspeaker position approach is mostly concerned with reducing spatial variance of low frequency energy throughout the room (T. Welti & Devantier, 2003) (i.e. the difference in low frequency magnitude across the listening space). This is important to note, as this issue becomes far greater in smaller spaces, where modal resonances become higher in frequency and therefore have a shorter wavelength, leading to a greater variation across the listening environment (an example of this effect is shown later in Figure 3.1). Further to this, in a denser modal field (i.e. a larger acoustic space), the difference in peaks and troughs may be less noticeable, where there is a higher chance of the listener being positioned in a pressure null.

When looking to equalization as a means of room control, traditionally, electronic equalisation would be the main approach through means such as octave band filtering. (Groh, 1974) presents a method of equalising room modes through filtering at problematic frequencies using a 1/3rd octave band technique. This is reported as being a very limited solution, citing issues with setting ideal parameters for best results due to resonances rarely aligning with the centre frequency of the filter. Although this can be alleviated somewhat with parametric style equalisation that allows for fine control over the centre frequency and Q of the filter. More recent implementations of equalizers take the form of Digital Signal Processing to obtain greater control over the equalizer parameters. A comprehensive look at various implementations of advanced signal processing techniques for room correction are presented in (Cecchi et al., 2018). Furthermore, it must be noted that the typical method of applying room equalisation is to correct for a single (or average) of listener positions. Therefore, room equalisation cannot address issues in the spacial variance across the entire listening space, often only fixing a fixed listener position (Toole, 2013).

Moreover, there are active control methods which do not rely on equalization. A

prominent example of active low frequency control is a CABS system (Celestinos & Nielsen, 2008), where rear-wall acoustic reflections are cancelled by utilising loudspeakers on opposing boundaries. While CABS is a useful method of controlling the low frequency of a room, results are best found in cuboid rooms. Although results do show that there is some improvement when using CABS for non-rectangular room, it is not as effective as when utilised in rectangular rooms (Celestinos & Nielsen, 2011).

While modal control methods are not the primary aim of this Thesis, it is critical to understand typical approaches used, as their fundamental goal is to improve the subjective quality of bass reproduction through mitigating problematic excitation of room artifacts. Research into correction of small listening rooms has even suggested that some implementations performed worse than no correction at all (Olive et al., 2009), although this research may not reflect the most recent solutions available to date, it is still a salient result. To add further to the confounding factors of room control B. Fazenda et al., 2012 found that when testing control implementations using music, the sample in test was a highly important factor when assessing modal control methods.

Therefore, it is clear that in literature there is an abundance of implementations available for room correction, however, there still exists a gap where control methods do not increase the perceived quality of sound. Furthermore, many correction systems are focused on purely objective metrics ranging from modal spacing to flattening the room response through equalization, without taking subjective responses into consideration (Stephenson, 2012), thus leading to high variance among perceptual results and a gap between the objective metrics and subjective response.

2.3 Audio Quality and Perceptual Modelling

When defining perceived audio quality, it is important not to confuse *hedonic preference* and *technical quality*, where hedonic preference is closely linked to personal affective traits influenced by familiarity and technical quality relates to technical aspects of the signal (A. Wilson & Fazenda, 2016). This Thesis will primarily look at the subjective effects on technical quality due to room acoustic artifacts and when referring to audio quality, technical quality is of concern and *not* that of hedonic preference.

There currently exists a number of objective implementations of measuring audio quality, primarily split into two types of models; Single and Double Ended (sometimes referred to as non-intrusive and intrusive respectively). Where single ended metrics can predict quality from only the processed signal and double ended models require a reference (i.e. the original, unprocessed signal) (Akhtar & Falk, 2017). However, current implementations of these quality metrics are mostly used for bespoke cases; Such as signal degradation through codecs like the ITU PEAQ algorithm (Thiede et al., 2000), or the HAAQI quality metric used for hearing aids (Kates & Arehart, 2016). Both of which are comprised of an auditory model which is used to represent the human hearing mechanism. While evidence suggests these types of metrics can be applied and generalised to a variety of signal distortions (Kressner et al., 2013), there is little known evidence to suggest that these quality models will be useful when considering only low frequency artefacts, especially due to room acoustics.

Other methods of modelling the human hearing through an auditory model have been used to investigate the psychoacoustic validity of conventional room acoustic measurements (van Dorp Schuitman et al., 2013). However, these approaches can be highly sophisticated and building an auditory model is beyond the scope of this work, due to the high complexity that would be introduced into tackling the research problem.

While there does not currently exist a bespoke low frequency model for rooms, (Wankling et al., 2012) proposes 3 Bass Quality attributes for the purpose of defining the perceived quality due to low frequency modal behaviour. These Bass Quality Attributes are defined as Articulation, Resonance and Bass Energy. Where Articulation ranges from Muddy to Tight, describing the definition of notes, Resonance ranges from None to High, describing the long ringing of individual frequencies and Bass Energy, which is a composite of two closely related descriptors of Strength and Depth that relate to the low frequency loudness and extension respectively. These descriptors were found to be useful at describing the overall bass quality, where it was preferred to aim for low resonance and high articulation. While there is no current method of measuring these attributes directly from signal metrics, there do exist methods of defining descriptors through statistical modelling.

2.4 Statistical Models of Perception

Another approach to modelling specific perceptual phenomena, is to model listener responses through means of a statistical model, thus creating a perceptual model. Promising results have been observed when modelling perceived 'punchiness' (Fenton & Lee, 2019) through means of presenting a variety of signals with varying attack times. Linear regression is then used to model listener ratings for perceived punchiness, which furthered understanding by showing punchiness related strongly to band limited onset times (Fenton & Lee, 2015). This modelling approach is also similar to that used in (Olive et al., 2017), where a linear regression model is used to predict listener preference of headphones. While these approaches may produce simple interpretable models, a difficulty arises when considering the generalisation of the perceptual models. Careful consideration must be therefore taken in designing listening tests and collecting data, to ensure results aren't confounded with uncontrolled factors and biases such as those pointed out in (Zieliński et al., 2008).

One recent advancement of interest is MIR or Music Information Retrieval, which is the field of extracting signal features (i.e. numerical metrics) from audio in the context of music, such as those described in (Duncan et al., 2014). Coupled with advancements in Machine Learning, MIR has been used to tackle problems such as instrument classification (Bai & Chen, 2007), acoustic environment classification (Ma et al., 2006) and predicting emotional ratings from music (Eerola et al., 2009). While these results are promising, use of MIR features has come under question regarding their robustness for use in research. Wegener et al., 2008 suggests a proposed glass ceiling effect partially attributed to a "samantic gap" between features and responses/labels. Furthermore, Machine Learning techniques are far more complex than that of simple linear regression, where their black box approach reduces interpretation and does not often aid in understanding the problem (Rudin, 2019). Another consideration that comes from Data Mining is that of "Data Dredging", which can lead to false positives and mis-interpreted findings (Davey Smith &

Ebrahim, 2002) due to over-analysis of data.

To summarise, while there is no existing model of Audio Quality appropriate for this specific topic area of low frequency bass reproduction, there are approaches that may be applicable in solving a perceptual problem of this nature, such as modelling the quality attributes discussed in Section 2.3 through means of a statistical, Machine Learning approach.

2.5 Subjective Aspects of Room Modes

The aim of this section is to understand the previous work into low frequency perception and discern if it may aid in modelling the perception of bass quality. To obtain a full understanding of how we perceive modes, each characteristic of modal behaviour must be understood, such as Modal spacing which was introduced in Section 2.2. The underlying theory of each characteristic is explained further in Section 3.1.2.

Investigations into modal spacing suggests that there is an audible phenomenon when two modes (or single mode with a sufficiently periodic excitation signal) are close in frequency, where the sum of modes leads to "beating effects" akin to amplitude modulation. B. Fazenda and Wankling, 2008 suggest an optimal spacing of 25-40% of the modal bandwidth to avoid modulation beating. These effects are predominantly found in acoustically smaller rooms, where modal density will be sparser.

Moving the attention to modal density, research suggests little effect on the perceptual quality at low frequencies (Wankling & Fazenda, 2009), which found that increasing modal density was not a defining factor in low frequency quality and was not analogous to a flat magnitude response. Suggesting that the problem may not lie purely in the frequency domain, which is where many of these optimisations are focused, such as that found in (T. Cox & D'Antonio, 2001) and many of the EQ solutions in (Cecchi et al., 2018).

While most of the aforementioned literature is focused on the frequency domain, interesting results are found when turning the attention to temporal characteristics of modal behaviour. The primary factor of modal behaviour in the time domain is that of modal decay, where B. M. Fazenda et al., 2015 suggest a threshold for which modal resonances are audible. This leads to an interesting concept of modal audibility, where theoretically if the modal resonances are sufficiently damped, then there will be no perceptual effects due to modes. However, this is difficult in practice, where low frequency absorbers or "bass traps" are often very large due to the long wavelengths of low frequencies.

Therefore, when comparing the findings in modal decay to that of the frequency domain, it may be deduced that modal decay might play a more important role than the previously believed frequency domain characteristics.

2.6 Large Acoustic Environments and Single Instrument Excitation

Most, if not all of the previous work has been focused on small listening or control room acoustics, which although highly important, is a mere glimpse of many sound

reproduction environments. Furthermore, research in small room acoustics often look to increasing modal density/room volume as an means of increasing the low frequency quality due to a more diffuse low frequency field (T. Cox & D'Antonio, 2001). However, larger room volumes quickly start to move out of the aforementioned scope of small rectangular listening rooms. A typical example of a larger acoustic environment is that of a live sound venue, where a transition is made from small room acoustics to live sound reinforcement. (White, 2015) presents the key responsibilities of the role and common practises of the Live Sound Engineer, primarily that of mixing (setting the balance of the instruments) and performing sound checks to avoid feedback, typically working with single instruments at a time.

Furthermore, live sound venues vary widely both geometrically and acoustically; (Adelman-Larsen, 2014) provides acoustical measurements from Live Sound Performance venues from across the globe that range from small venues such as The Cavern Club (capacity 350), to Stadium venues like the MEN Arena (capacity 21,000). These venues are a far cry away from the typical modal behaviour found in the aforementioned acoustically small listening rooms, where the arena or stadium type large performance spaces most likely come under 'acoustically large' spaces where the critical frequency is below the lowest frequency produced in the room (Angus, 1997).

To understand the limits of single instrument excitation, (Teret et al., 2017) found that perceived reverberation time varied heavily on the signal interacting in the room, although a variety of musical instruments and signals were tested, percussive sounds were not. (Hill et al., 2011) emphasises the importance of the robustness of the characteristics of kick drum sounds in a live sound environment. Furthermore, many of the perceptual investigations discussed in Section 2.5 are either concerned with single excitation sinusoids or full range music. Hence, it is unclear how the use of single instruments and high modal density may effect tolerances to perceptual metrics such as the Modal Decay Thresholds (B. M. Fazenda et al., 2015), Bonello Criteria (Bonello, 1981) or even the aforementioned bass quality attributes. Although, contrary to this concern, the modal decay thresholds have been used in a live sound application in (Bolla et al., 2019), which describes a perceptually weighted frequency response that accounts for the perceptual thresholds of modal decay.

One key issue when moving the scope of research to live sound environments, is that the typical area of research deviates away from the concerns of small room acoustics and moves towards unique issues such as even coverage across audiences (Hill, 2018). Furthermore, considering previous work in room acoustics, these large acoustic environments (with high modal density) must therefore be devoid of modal low frequency problems (see Section 2.5). However, issues in low frequency reproduction are not solved in these environments, the control over low frequency can still be detrimental to the perceived quality of the live sound experience (Rumsey, 2011).

Therefore, the understanding into the perceptual characteristics of low frequency needs to be expanded when shifting perspective to a live sound environment where larger rooms and single instrument excitation is often of concern.

2.7 Summary

To conclude, low frequency reproduction in rooms is problematic and current methods of modal control vary widely in attempting to increase perceived low frequency quality. Furthermore, there is no bespoke method of measuring bass quality due to room acoustic effects other than perceptual descriptors. Current auditory methods may not be sufficient in modelling the niche scope of this research and therefore may be built upon the perceptual descriptors and statistical modelling approach. A look into low frequency perception suggests that the problem may not lie purely in the frequency domain and may be influenced more by modal decay. Moreover, a review of previous work in low frequency room acoustics has highlighted a limited scope, where perception of low frequency in rooms is typically only investigated in small listening or control rooms, with full range music excitation. However, since larger scale environments differ acoustically from small rooms and auditioning of single instrument excitation is a common in live sound engineering, it is unclear how these perceptual effects may change.

Henceforth, the research objective of this project will be focused on creating a perceptual model of bass quality via the Bass attribute descriptors defined in (Wankling et al., 2012). Therefore, the secondary objective will use the model to further the understanding the perception of low frequency room acoustics through use of signal features. The scope will be made specifically for the case of a variety of rooms that scale from listening rooms to live sound venues in the particular case of single instrument excitation.

Chapter 3

Theory

3.1 Low Frequency Room Acoustics

3.1.1 Introduction of Modal Theory

As an introduction to modal behaviour in rooms, it is important to first understand the basic theory of single mode resonances. These resonances are standing waves that occur at integer multiples of half wavelengths of the distance between room boundaries. Therefore, modes are often represented in terms of their order, which are the integer multiples of half wavelengths across the dimension of the room. The frequency of resonances (or eigenfrequency) is defined below in Equation 3.1 (adapted from (Toole, 2013)).

It should also be noted that most of the theory outlined in this section is discussing modes in cuboid shaped rooms as a means of reducing the complexity of the presented theory, thus the scope of what is described is somewhat limited.

$$f(n_x n_y n_z) = \frac{c}{2}\sqrt{\left(\frac{n_x}{L_x}\right)^2 + \left(\frac{n_x}{L_x}\right)^2 + \left(\frac{n_x}{L_x}\right)^2} \tag{3.1}$$

Where c is the speed of sound in air, n is the modal order across dimension (x, y, z) and L is the length of the room in the respective dimension. While predicting potential problematic frequencies is useful, it does not provide a full understanding of the modal effects in a room. A key defining factor of modal effects results from the coupling between a source and receiver to each mode, where the listening positions across a room dimension will determine whether the listener is placed on a node or anti-node. (Kutruff, 2009) defines the the pressure at a single point due to a modal resonance in Equation 3.2.

$$P(n_x n_y n_z)(x, y, z) = C \cdot \cos\left(\frac{n_x \pi x}{L_x}\right) \cdot \cos\left(\frac{n_y \pi y}{L_y}\right) \cdot \cos\left(\frac{n_z \pi z}{L_z}\right) \tag{3.2}$$

Where n is the modal order, x, y, z are the positions in the room respective to length L of the room geometry and C is an arbitrary constant. Figure 3.1 illustrates this effect, where the pressure distribution across a single room axis is shown for the first three orders of modes.

While these low modal orders are simple to conceptualise across a single dimension, when concerned with a 3 dimensional enclosure (i.e. a room), the complexity

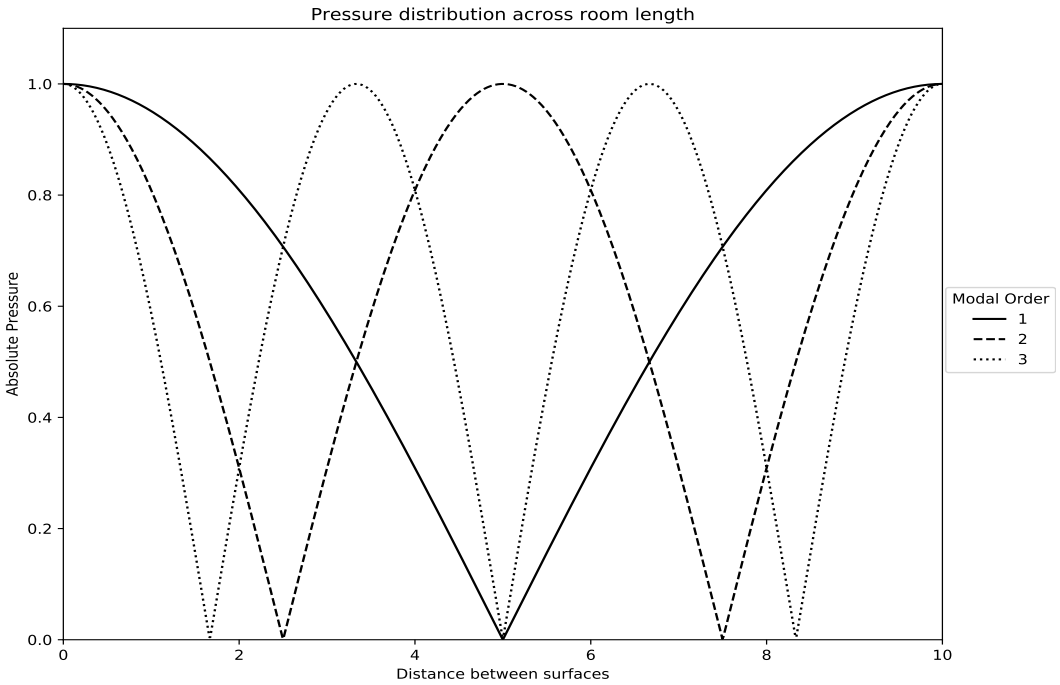


Figure 3.1: Modal pressure distribution across a single dimension illustrating the first 3 orders of resonance.

is greatly increased. The corresponding dimensions that a mode acts across can be described by the type of mode - Axial, Tangential and Oblique; where axial interacts between 2 surfaces, tangential interacts between 4 and oblique interacts with all surfaces in a rectangular enclosure. Therefore, using the modal order system in Equations 3.1 and 3.2, a first order Axial mode may defined as $(1, 0, 0)$, tangential as $(1, 1, 0)$ and oblique as $(1, 1, 1)$. To illustrate this further, Figure 3.2 shows the pressure distribution of the first 3 orders of a tangential mode across an idealised rectangular room (note: z axis is not taken into account for this example). Where the modal order of each room is $(n, n, 0)$ for (x, y, z) respectively (i.e. the same order in both the x and y axis).

Finally, it must be noted that both Equations 3.1, 3.2 are only valid for the case of rectangular rooms and do not take into account any geometrical complexity that is found in most real rooms (such as windows and alcoves).

3.1.2 Modal Characteristics

After understanding how single modes are formed and the basis of the resonances in rooms, there are further characteristics of modal behaviour that aid in describing the effects of modes in rooms. Therefore, this section aims to provide an understanding of these characteristics, their contributing factors and any effects on reproduced sound.

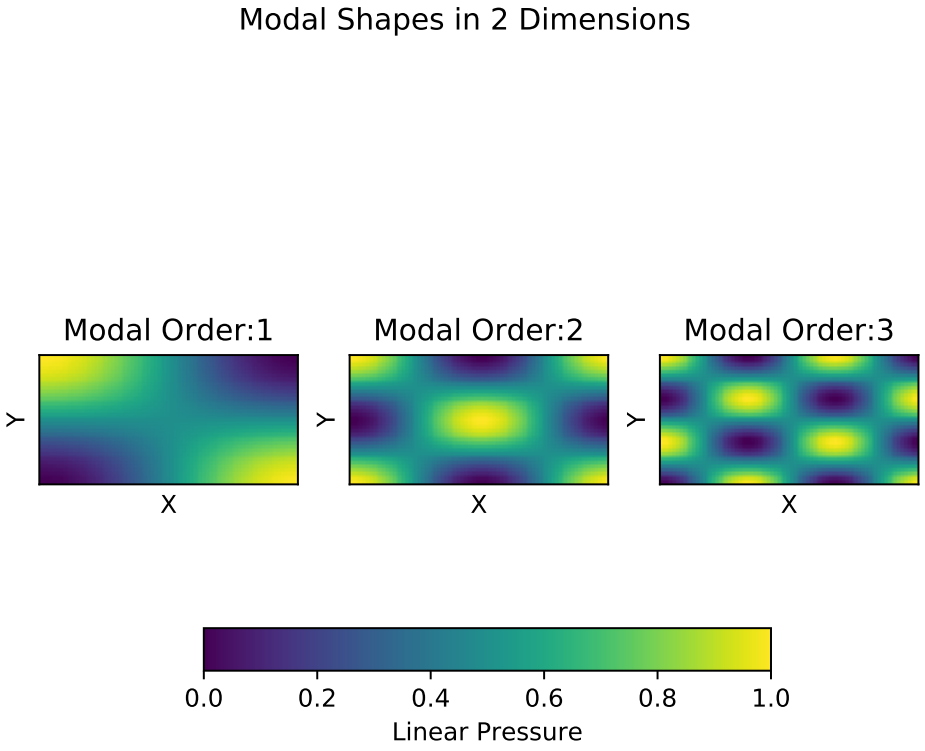


Figure 3.2: First 3 orders of a tangential mode across a 2:1 ratio rectangle demonstrating the pressure distribution across the horizontal axis - where the modal pattern is (1,1,0), (2,2,0) & (3,3,0).

Modal Decay

Modal decay is the main temporal factor in describing modes. Modal decay time is somewhat defined similarly to reverberation time, where both decays are also represented by the time taken to decay to a relative -60dB (Karjalainen et al., 2002). However, a key difference is that reverberation is only applicable in a diffuse field, which is not the case of the aforementioned modal sound field (Angus, 1997). Therefore, there lies a strict difference between the two sound fields, where modal decay is defined in equation 3.3.

$$T_{\text{modal}} = \frac{2.2Q}{f} \tag{3.3}$$

Where T_{Modal} is the modal decay time defined by the relative -60dB point, Q is the Q factor due to the modal bandwidth and f is the modal frequency.

Similarly to reverberation time, the modal decay time is primarily determined by two factors; The damping of the room boundaries and time taken between losses (see Equation 3.11). Therefore, the absorption of the room and volume are important factors when considering the suitability of listening spaces.

Modal decay is a key factor in the audibility of room modes, where proposed thresholds of modal decay partly determine whether effects of room modes are audible (B. M. Fazenda et al., 2015) (explored in more detail in Section 3.1.4).

Modal Coupling

While the main application of discussion has been that of single resonance excitation, a more realistic context describes excitation of multiple modes through a source receiver model. Modal coupling describes the effect of modal excitation due to the position of source receiver position.

The typical way of describing modal coupling is weak and strong coupling to particular modes. Where weak coupling occurs when the source is placed on a node null (minima) and strong coupling occurs on a node maxima, such as the room boundaries. This effect can be imagined when considering a moving receiver across a room dimension in figure 3.1, where the pressure at the receiver fluctuates from strong to weak modal coupling.

Ideal modal coupling is obtained when a point source and receiver are in opposing corners of the room (B. M. Fazenda et al., 2005). Assuming even excitation (i.e. no losses due to the quality of the source or receiver), opposing corners allows for strong coupling of all modes. However, this ideal case is not practically viable in a listening environment.

Modal coupling is often used as a control method in listening rooms, such as the techniques described in (Toole, 2013), where the suggested subwoofer and listener position configurations describe placement of source in the 1/4 position into the room and listeners in the centre. Broadly speaking, modal coupling is a difficult variable to control as subtle shifts in source and receiver position vary the coupling across all modes (assuming even excitation) and across all dimensions. Therefore, when using modal coupling as a control method, typically only the strong first order axial modes are factored. It must be noted that many examples of idealistic receiver positions, are typically focused on rectangular rooms and may not be applicable to more complex room geometries.

Modal Spacing

Modal Spacing is the consideration on the of distance in frequency between two resonances. This is a particularly important characteristic in where two overlapping modes (that may occur in a square-cuboid like room with identical lengths of dimensions), where there is no spacing between modes will compound the effect of the resonances. However, in rooms with more complex geometry or where room dimensions are not equal, the spacing of modes is still of concern due to potential 'beating' effects due to closely spaced modes (B. Fazenda & Wankling, 2008). Effects due to modal spacing are therefore similar to that of perception of two tones, where regions of identical resonances, beating, roughness and separate resonances exists depending on the distance in frequency between the two tones.

The primary contributing factor to modal spacing is the room volume and geometry. This is due to the length of room dimensions dictating the wavelength that the resonances are comprised from. Therefore, there has been considerable work to suggest ideal room ratios to avoid adverse effects of modal spacing (T. Cox & D'Antonio, 2001), (Bolt, 1946) and (Bonello, 1979).

Modal Density

Closely linked to modal spacing is modal density. Where modal density describes the number of modes across a given frequency range. A sparse modal density describes the condition of room modes that are spaced far apart in frequency and high density occurs where there are many overlapping modes. Again, the major contributing factor to modal density is the room volume due to the room dimension lengths.

(Stephenson, 2012) defines the average number of modes up to a frequency f in Equation 3.4. Note that this is an idealised case for modal density that assumes a linearly increasing density due to the room dimensions.

$$N_f = \frac{4\pi}{3} V \left(\frac{f}{c}\right)^3 + \frac{\pi}{4} S \left(\frac{f}{c}\right)^2 + \frac{L}{8}\frac{f}{c} \tag{3.4}$$

Where S and L are defined below.

$$S = 2(L_x L_y + L_x L_z + L_y L_z) \tag{3.5}$$

$$L = 4(L_x + L_y + L_z) \tag{3.6}$$

An important factor of Modal density is the determining factor of the transition between a modal and diffuse sound field, where a sufficiently high modal density can be assumed to be a diffuse sound field. The commonly referred to Schroeder or critical frequency (equation 3.7) is an attempt to describe the transition frequency between the modal and diffuse sound field.

$$f_c = 2000 * \sqrt{\frac{RT_{60}}{V}} \tag{3.7}$$

Therefore, modal density is largely linked to the volume of the room. (Angus, 1997) describes the critical frequency of the room as a determining factor to whether a room is acoustically "large" or "small" Where a critical frequency for a large room lies below the lowest audible frequency of the room and small rooms are determined by a critical frequency higher than the lowest audible frequency of the room, which will therefore reside in the audible range.

The consequence for small rooms is a modal sound field, which modal effects are audible and are therefore susceptible to the adverse effects discussed throughout this section. However, in larger spaces, the modal sound field is below the audible threshold or lowest excitable frequency, resulting in a diffuse field, where modal characteristics described throughout this section are assumed to no longer be present (Kutruff, 2009).

3.1.3 Modal Decomposition Model

Room acoustic modelling is a large field of study of where many approaches to modelling exist, with varying degrees of accuracy and complexities for different modelling applications. For example, implementations vary from high complexity such as FDTD and FEM, to more simple approaches such as the image source method. These models are typically based on broad band modelling of the room, where the approaches are defined by their method of discretising the wave equation.

However, when concerned with only the modal sound field, a rectangular room can be described as a sum of all possible modes due to the room geometry. This approach is referred to as the modal decomposition model, which has been used to model low frequency room acoustics, such examples include (Stephenson, 2012), (Walker, 1992) and (Kutruff, 2009).

Formal derivation of the wave equation can be found in (Kutruff, 2009), which is not part of the concern for this thesis. Furthermore, the room transfer function at a point receiver due to the summation of modes is described in equation 3.8, which is adapted from (Stephenson, 2012).

$$P_\omega(r) = iQc^2\omega\rho_0 \sum_n \frac{p_n(r)p_n(r_0)}{K_n(\omega^2 - \omega_n^2 - 2i\delta_n\omega_n)} \tag{3.8}$$

Where: Q is the source strength; ω is the driving frequency of the source; n is the n'th order mode; $p_n(r)p_n(r_0)$ are the respective source and receiver pressures as defined by equation 3.2; K_n is a constant scaling factor and δ_n is the damping coefficient of n'th mode defined in equation 3.9.

$$\delta_0 = \frac{6.91}{RT_{60}} \tag{3.9}$$

Equation 3.9 describes the weighted average damping which assumes a rigid walled room (i.e. no transmission loss is accounted for) with a uniform damping coefficient and sound field. This assumption for modal damping is a somewhat inaccurate model for how modes take losses through their path. For example, an axial mode will convey less losses due to fewer interactions with surfaces than a tangential mode, which will take losses due to each boundary in a room (assuming a rigid, rectangular room). Therefore, a more robust method of modelling the damping of room modes must be used.

Proposed in (Rindel, 2015), a different approach for factoring modal damping of the nth order mode due to the modal decay time (rather than the overall reverberent time) is shown in Equation 3.10.

$$\delta_n = \frac{3}{\log_e T_n} \cong \frac{2.2\pi}{T_n} \tag{3.10}$$

Where modal decay time is defined in Equation 3.11.

$$\begin{aligned} T_n = {} & 13.8 \cdot \sqrt{(l_x n_x)^2 + (l_y n_y)^2 + (l_z n_z)^2} \\ & \cdot [-c \ln([(1-\alpha_{x1})(1-\alpha_{x2})]^{nx} \\ & \cdot [(1-\alpha_{y1})(1-\alpha_{y2})]^{ny} \\ & \cdot [(1-\alpha_{z1})(1-\alpha_{z2})]^{nz})]^{-1} \end{aligned} \tag{3.11}$$

Where α denotes the acoustic absorption coefficient of dimensions x, y, z, which consider 2 opposing surfaces for a rectangular room. Use of the absorption coefficients allows for a much more intuitive approach in defining the damping coefficient of the room when compared to more traditional methods such as using wall admittances. For full derivation, refer to (Rindel, 2015). Therefore, this approach of defining individual damping tied to the mode characteristics rather than assuming a one size fits all approach across all modes will be used to model the rooms.

3.1.4 Perception of Low Frequency Room Acoustics

This section aims at describing the underlying theory of the perceptual nature of low frequency room acoustics that are used throughout the test methodology.

Perceptual Modal Thresholds

The perceptual modal thresholds describe the auditory threshold of which modes become audible due to their decay time. Described in (B. M. Fazenda et al., 2015), the modal decay thresholds are defined at 85dB SPL (as the thresholds are level dependent) for both music and single frequency excitation. Where music (general broadband excitation of the low frequency) is less sensitive and therefore has a higher modal decay threshold.

Further to these perceptual decay thresholds, (Heddle, 2016) presents a means of modelling the perceptual modal threshold (PMT), where the adapted equation from (Bolla et al., 2019) is modelled in Equation 3.12.

$$PMT(f) = 0.15 + \frac{755}{f^2} \tag{3.12}$$

Modal Density Function (MDF)

Further to the perceptual modal thresholds, (Bolla et al., 2019) presents a means of using the thresholds to weight a room response by the relative exceedance of modal decay times to the perceptual decay thresholds. This weighted response is presented as the MDF or Modal Density Function and is shown in Equation 3.13.

$$MDF(f) = \sum_{n=0}^{N} a_{f,n} \left\{ \frac{L_{f,n}}{PMT_{f,n}} \right\} \tag{3.13}$$

Where N is the number of ratios between the PMT and decay profile at levels L, f denotes the modal frequency and α corresponds to a weighting coefficient. Full detail of the Modal Density Function is outlined with recommended values for N, α, L_f are found in (Bolla et al., 2019).

Perceptual Descriptor Attributes

To encapsulate the subjective descriptive response of a listener, (Wankling et al., 2012) presents low frequency subjective attributes under Articulation, Resonance, Strength and Depth. The perceptual attributes are defined by a panel to represent different characteristics of how room modes effect the perceived quality of reproduced sound. Where Articulation represents the modal effects on the attack and definition of a sound, Resonance describes the the audible resonant effects due to the modal resonances and Strength and Depth refer to the low frequency loudness and extension respectively. The full definitions and scale extremes are shown in Table 3.1.

It should be noted however, that the research found that both Strength and Depth attributes were highly correlated and were advised to form a new metric under Bass Energy, where Bass Energy was the combination of the two attributes of Strength and Depth.

Descriptor	Scale Values	Definition
Articulation	Muddy	Each sound (or note) has a lack of definition and could sometimes be described as "smeared".
	Tight	Each sound (or note) is distinct, well defined and precise.
Resonance	None	A resonant sample has some notes which sound louder, ring and last longer.
	High	
Strength	Weak	Relates to the loudness of the low frequency when compared to the rest of the frequency range in the sample.
	Strong	
Depth	Shallow	Lacks notes that extend down lower in frequency.
	Deep	Has notes that extend down lower in frequency.

Table 3.1: Original bass quality perceptual attribute descriptors for Articulation, Resonance, Bass Strength & Depth as defined in (Wankling et al., 2012).

3.2 Signal Features and The Machine Learning Pipeline

As the research objective is to further the understanding of low frequency sound quality through perceptual modelling, this section describes the approaches that may be considered when creating a statistical model from an audio signal.

To begin, methods of how features can be used to describe audio signals are discussed. Then an overview of statistical modelling techniques are investigated with a particular focus on machine learning.

3.2.1 Audio Features

One of the disadvantages of working with audio signals, is that raw data is incredibly challenging to use for statistical modelling due to large amounts of data (e.g. 44100 samples for one second of audio). While raw audio can be used in some machine learning tasks such as end to end learning (where the input and output of a model is raw audio), these applications often require large scale neural networks and highly complex models such as the one described in (Stoller et al., 2018). Although a powerful and sophisticated tool, large scale neural networks greatly reduce the interperetability of the relationship between the input features and outputs, creating a black box approach to solving the problem (Rudin, 2019).

Therefore, Audio features are used as an abstraction from the audio to provide a more meaningful and condensed form of information over raw sample data. Hence, audio features increase the ability to interpret the relationship between characteristics of a signal and a response. An example of an audio feature may be loudness, where a single numerical value allows for a meaningful description of a section of audio, without finding patterns in hundreds of thousands of samples and creating a complex black box approach.

Currently in literature, the most common form of audio features is that of Music Information Retrieval (referred to as MIR). These features describe characteristics of

audio content through different categories of timbral, spectral and signal properties (International Organisation for Standardisation, 2004).

However, it must be noted that while there are typically used features for the application of music, an audio feature is a generic term for any numeric abstraction of an audio signal. Therefore, bespoke or 'hand-crafted' audio features may be used to provide features that are more relevant to the problem-space, although the validity and ability to generalise cannot be ensured.

3.2.2 Introduction to Machine Learning

This section describes a brief introduction into the basic principles behind machine learning, guided by information from (Mehta et al., 2019), a comprehensive introduction to a wide variety of machine learning information and techniques. Hence, the machine learning techniques discussed throughout this section are focused on those used throughout the methodology and only look at a mere glimpse of machine learning approaches and models available.

Machine learning describes a statistical model that is built on an automatic learning process. Most machine learning applications are suited to an input/output pipeline, where a vector of numerical input features are mapped through some system of weighting to produce a continuous numerical output (regression) or categorical output in the form of classification. It must be noted that the only use case observed throughout this work is the application of regression, where the aim is to predict an output score of perceptual attributes.

Furthermore, machine learning applications can be split into two categories, supervised and unsupervised. Supervised methods are used to train the model on "ground truths", which are typically human or expert labelled data provided to the model as an objective criterion. While this is useful for smaller scale learning applications, difficulties arise when considering the scalability of modelling, where labelling data becomes a increasingly difficult as humans are often required to manually label data.

Unsupervised methods rely on a large machine learning model, often through deep learning to form understandings of the data space. While accuracy does not often differ between the two applications (Love, 2002), there is a trade-off where collection of labelled data becomes impractical or suitable features are not known or applicable for the chosen problem. Therefore, smaller scale problems such as the research objective for this work, where labelled data can be obtained through listening tests, is deemed to be more suitable to make use of supervised learning techniques. This is especially useful as direct relationships between signal features and perceptual attributes are required to be interpreted to provide an insight into the problem, which is far more complex in large scale machine learning models (Rudin, 2019).

3.2.3 Linear Multivariate Regression

Linear regression is a simplistic and powerful statistical modeling approach that imposes simple linear weights onto input features to provide direct mapping from said features to an output criterion (Dawes, 1979). It must be noted that the type of linear regression referred to throughout this thesis is that of multivariate regression

(i.e. use of multiple variables) over simple linear regression which commonly refers to one single feature.

A linear regression for n features can be summarised by Equation 3.14. Where b_0 describes the y-intercept and b denotes the weighting for input feature x.

$$y = b_0 + b_1x_1 + b_2x_2 + ... + b_nx_n \tag{3.14}$$

For a typical machine learning approach these weighting coefficients are tuned using a gradient descent method to find optimal weightings to reduce the residual error between input features and the output variables.

3.2.4 Random Forest

Random forest differs from typical regression models as it is a Ensemble learning approach (Liaw & Wiener, 2002), which describes aggregating multiple simple models together to form one, large complex machine learning model. Ensemble approaches make use of many poor/weak models that come together to form a more stable and accurate model that is less prone to overfitting (Sagi & Rokach, 2018). Random forest achieves this by aggregating many decision trees together to form a "forest" of trees (Breiman, 2001).

Decision trees are inherently prone to overfitting, however due to the converging nature of an ensemble learning approach random forests become robust to overfitting (Liaw & Wiener, 2002). Random Forests are also useful in creating simple and interpretive models due to the way trees are formed using an embedded feature selection process (Saeys et al., 2007). Furthermore, individual decision trees can be investigated to understand how the data is being split and common features that are used, albeit a narrow interpretation of a larger forest.

To form a greater understanding of how a random forest functions, it is important to understand how a random forest model is formed and how individual trees are created. Individual decision trees are formed by binary splits of the feature space, which creates "sub-sections" of the original feature space that through inequality checking, can direct new input feature vectors to similar splits of the data. Decision trees are formed of two types of nodes - decision nodes, which form the inequality checking of the input feature and leaf nodes which indicate the predicted regression or classification output. Trees are then constructed using a network of true/false paths that lead to a leaf node indicating the predicted score or classification. A simple example of a decision tree of arbitrary X features regressing output Y is shown in figure 3.3.

To form an ensemble of trees and make a forest, many methods can be used, however two of the most popular methods are boosting and bagging (Breiman, 2001). Both boosting and bagging describe some form of random sampling and replacement of the data space and aggregation of many estimators/predictors together to form one model. The main difference between bagging and boosting is as follows; bagging describes individual trees trained on sub-samples of the data-space and boosting describes a sequential approach where trees are trained on instances which were poorly modelled by the previously trained model (Sagi & Rokach, 2018). Finally, the trees are aggregated together which can be done in a variety of ways, such as a linear average, online bagging or weighted aggregation (Sagi & Rokach, 2018).

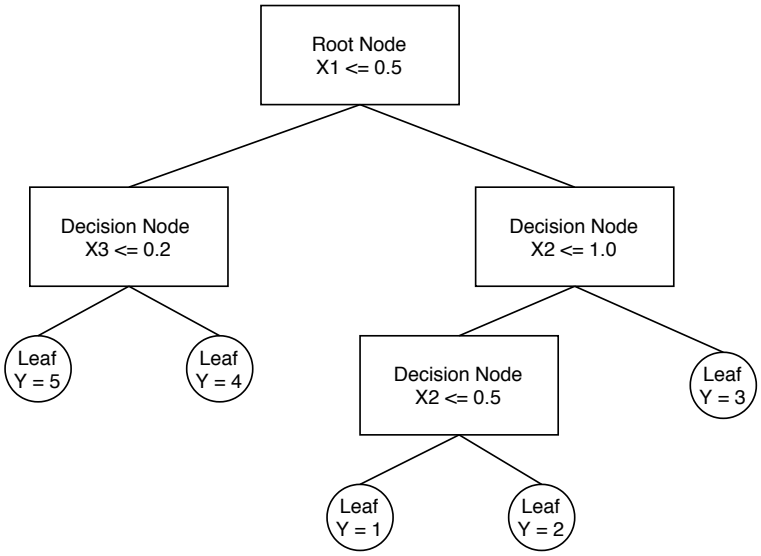


Figure 3.3: Simplified example of an arbitrary decision tree of depth 2. Where binary feature splitting is performed at inequalities in the decision nodes and predicted outputs are shown in the leaf nodes.

3.2.5 Hyper-parameter tuning

Hyper parameters describe the parameters that are not tuned in the learning process, which act as intermediate parameters that influence the created model. For example linear regression does not have hyper parameters as there are no variables to control for other than the weights applied to the input feature vector. However, random forest has many variables that are used to control the learning process, such as the maximum tree depth, the minimum numbers of samples that may be split at nodes and even the number of trees used (estimators) (Luo, 2016). Choice of ideal hyper parameters differs between learning applications and hence cannot apply a "one size fits all" approach applicable to selecting hyper parameters.

However, due to the potential size and complexity in choosing ideal hyper parameters, automated methods are often used, where the standard approach is a simple grid search of hyper parameter values (Hutter et al., 2015). Two primary methods of grid searches for hyper-parameter optimisation are commonly used, random search and grid search. Where random search involves a stochastic search across a given hyper-parameter grid and a grid search is a purely iterative search across the entire grid. While a full grid search would yield the optimal set of hyper parameters in the given grid, random search is an effective method of selecting hyper parameters without an exhaustive search across the entire grid (Luo, 2016). This more efficient approach is highly useful when the range of hyper-parameters is not well understood and large grids are required.

3.2.6 Feature Selection

Feature selection is a crucial part of streamlining the machine learning process, where redundant and useless features are removed, this is useful as large amounts of features results in overfitting (Mao, 2004) and also vastly improves the ability to interpret useful features in relation to the research problem. Feature selection methods

come under three main approaches; wrapper, embedded and filter/threshold based methods - (Saeys et al., 2007) provides a comprehensive review of the relevant advantages and disadvantages of each approach and is used to inform the trade-offs of the corresponding algorithms discussed throughout this section.

Threshold methods are either uni-variate or multi-variate, where the correlation between input and output variables is assessed directly. These methods are useful approaches for speed purposes, particularly in the case of large feature sets. Threshold methods also allow for assessment between features and criterion independently without influence of the classifier or regressor. However, the independence of the model can find discontinuities between the feature selection and modelling process.

Wrapper based methods encompass finding the best model that obtains a reduced feature set. This simplistic approach is useful, as direct interactions between the features and model result in fewer discontinuities between the feature selection and modelling process. Wrapper feature selection methods also model feature dependencies, where feature interactions and dependencies are factored into the selection process, unlike uni-variate methods. However, again this results in a highly model dependent feature selection result, where different features will be selected dependent on the chosen model. Wrapper based methods are also prone to overfitting due to the dependency of the machine learning model used for the feature selection process.

Finally, embedded methods are feature selection algorithms which are "in-built" to the modelling process such as the previously mentioned Decision Tree (see Section 3.2.4). As these methods are directly tied to the modelling process, they incur many of the same trade-offs of being dependent on the model used.

Chapter 4

Methodology

4.1 Overview of Methodology

The methodology for this work is comprised of three key areas, auralisation of rooms, feature extraction and finally, modelling using machine learning. This section outlines the chosen methodology and provides an overview of the implementations utilised. For exact control and justification over the variables in test, refer to section 5.1.

To accompany the diagrams throughout this methodology, refer to the legend shown in Figure 4.1, where a function is defined as any abstracted process; Auralisation is the process outlined in Section 4.2.2; A data set is a subset of a Database used for the train and test sets; A split of data describes further folds of the data sets and a Database is defined as the feature/rating space in its entirety.

To provide an overview of the methodology, a high level diagram of the room auralisation process and feature extraction pipeline is shown in Figure 4.2. The two form the combined feature and rating space, which is then used to power the machine learning pipeline (introduced in Section 4.4).

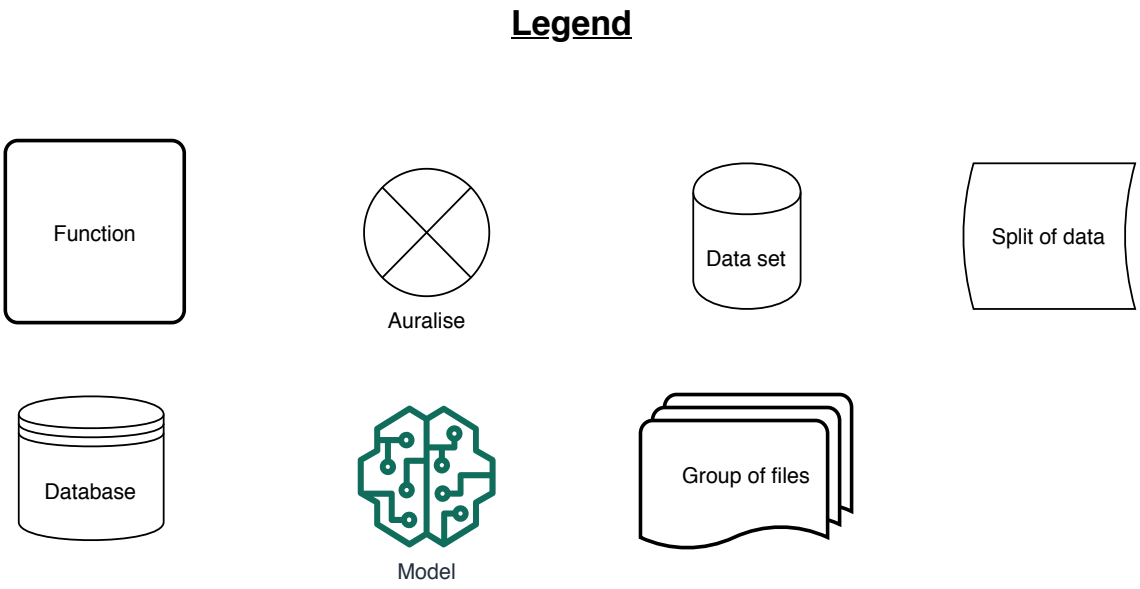


Figure 4.1: Legend for diagrams used throughout the methodology.

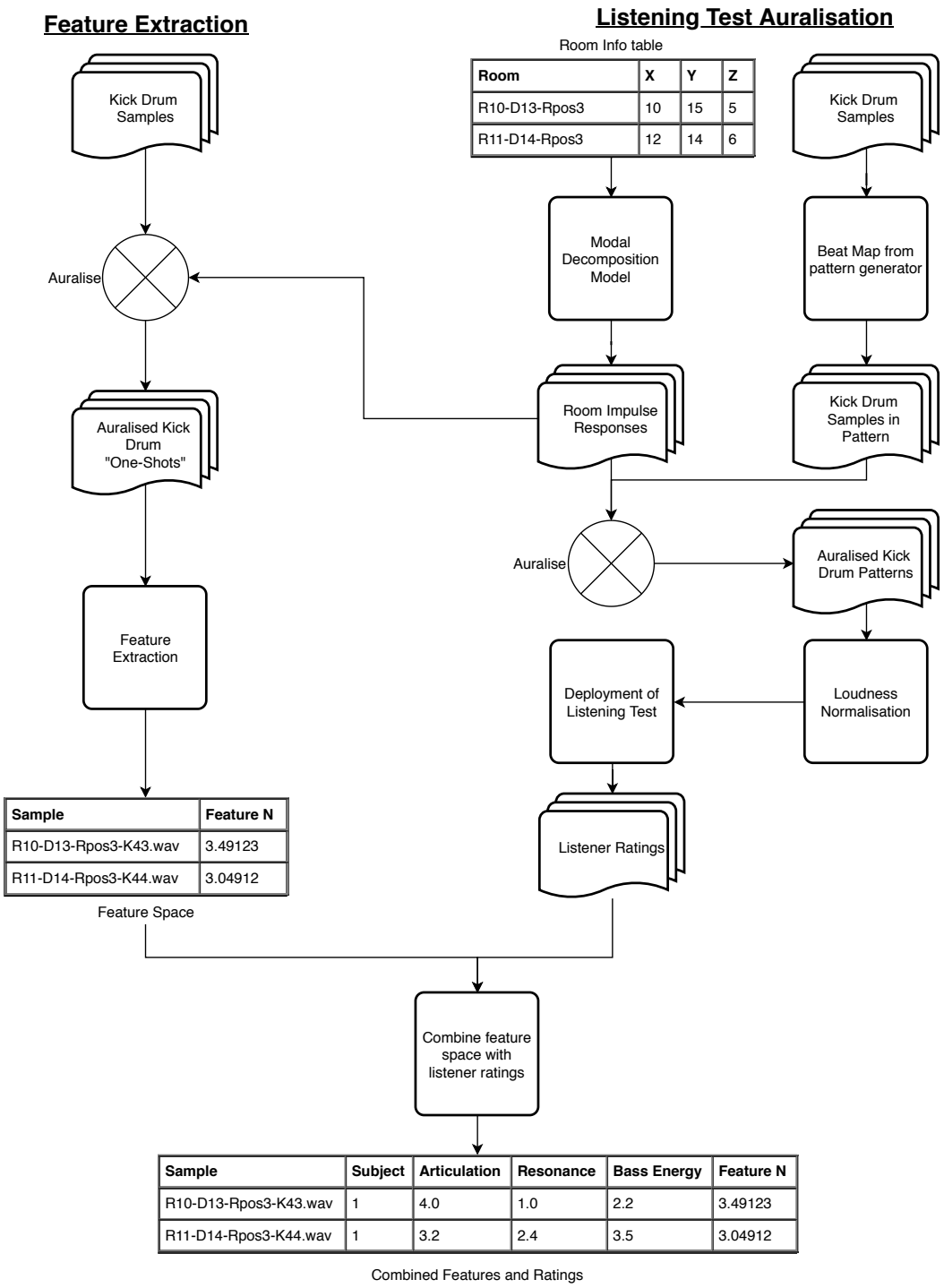


Figure 4.2: High level overview of the feature extraction process on kick drums auralised from room responses. Where the input is a group of kick drum samples and room information table to be modelled. The output is a combined feature and perceptual response data space to be used as part of a modelling procedure.

All of the described methodology throughout this section is implemented in Python unless stated otherwise.

4.2 Auralising and Room Modelling

An overview of the auralisation process is shown in Figure 4.4, which is adapted from (B. M. Fazenda et al., 2015).

4.2.1 Generating the Room Impulse Response

The aim of the auralisation process is to synthesise the low frequency modal sound field of a room, where different kick drums are auditioned to provide perceptual ratings for the quality attributes. While there are a plethora of room modelling techniques (such as those discussed in (Savioja & Svensson, 2015)), many of these tackle the mid-high frequency range where modal behaviour is not taken into account. Furthermore, techniques such as FEM, BEM and FDTD, require a more involved modelling approach, where the key benefit is to gain accuracy in complex geometries and boundary conditions. However, this work is limited to only modelling rectangular spaces for the purpose of simplicity, hence the advantages in boundary complexity are not required. Therefore, the chosen method of room modelling will be that of the Modal Decomposition model, introduced in Section 3.1.3. The Modal Decomposition Model has been successfully used for auralisation in previous work, such as (Stephenson, 2012), (B. Fazenda et al., 2012) and (Wankling et al., 2012). Although this approach is only being used to model simple rectangular spaces, it is thought that since the objective is to regress ratings from signal features, the scope of different room volumes used is sufficient enough where more complex geometries may not aid the pool of signal features significantly.

To model a room using modal decomposition, 3 aspects must be considered; Source, receiver and room. Where the source and receiver are assumed infinitesimal points placed within the room boundaries and the room boundaries are defined by their dimensions and acoustic damping. While the use of a point source and receiver will result in an onmidirection polar pattern and may not be a true reflection in what may be implemented in a ideal sound reinforcement setup, it is used to create a simple room model, where an assumption can be made for both the source and receiver to be mostly omnidirectional at lower frequencies.

For this implementation, the damping coefficient is defined by the acoustic absorption coefficient, which is used to calculate the modal decay time in Equation 3.11 and substituted into Equation 3.10. While previous studies have made use of modal decay times directly to calculate the damping of the room (Wankling et al., 2012), this assumption assumes a flat damping coefficient for all modes as in Equation 3.9. However, this study makes use of the more realistic case for modes acting across different dimensions taking different losses, (i.e. axial over tangential) using Equation 3.11. However, use of a fixed decay time throughout frequency using Equation 3.11 would involve an inaccurately defined absorption coefficient, as the modal decay time assumes less absorption due to fewer boundary interactions. A comparison of the two approaches as expressed in modal decay time is shown in Figure 4.3.

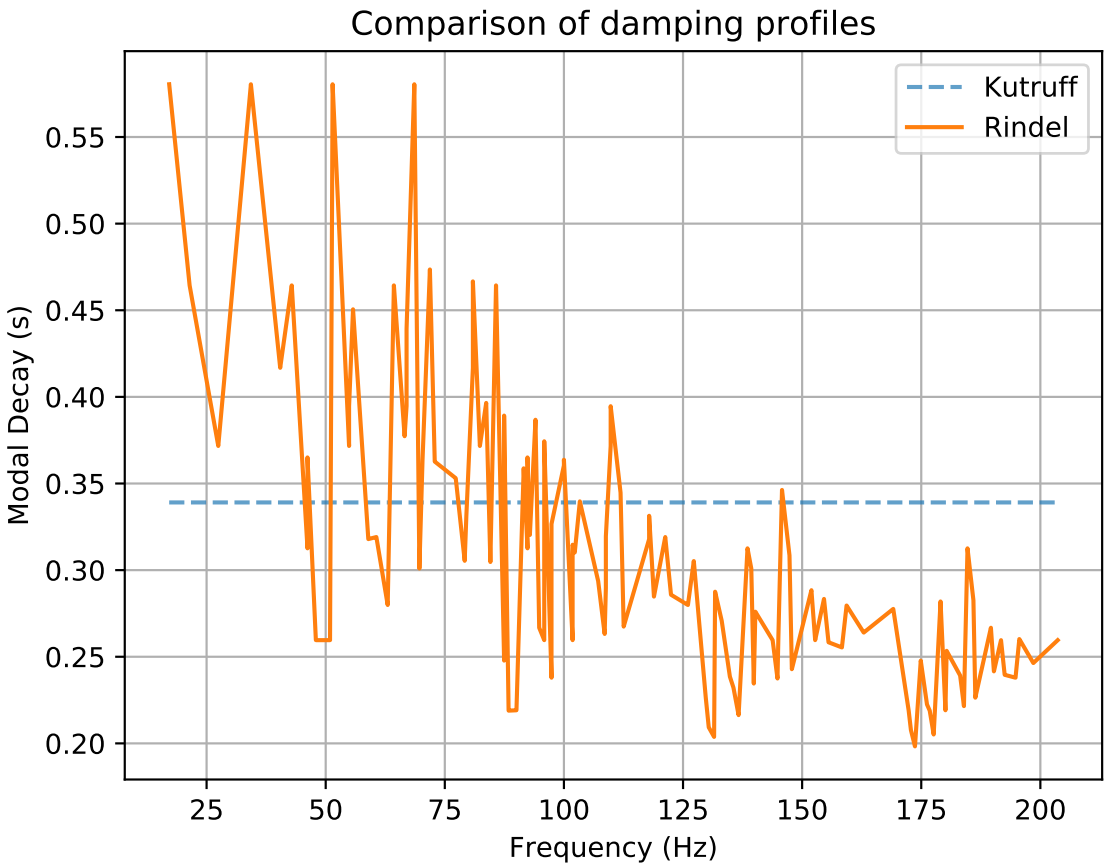


Figure 4.3: Comparison expressed in modal decay time between use of Equation 3.3 (Kutruff Method) and 3.11 (Rindel Method) - Assumed constant acoustic absorption coefficient of 0.5 across frequency range.

The modal decomposition model describes the steady state room transfer function (defined in Equation 3.8). From this, the room impulse response (RIR) can be obtained via an inverse Fourier transform. However, since the model can compute modes below the audible limit of hearing (thus reducing the available headroom of the signal) and above the frequency range of interest, two steps are used to reduce noise and unwanted artifacts of the RIR generation process; filtering and trimming of the impulse response.

First, the impulse response is filtered using a low pass filter (shown in figure 4.4 to avoid excitation of the room response above 250Hz (a commonly used upper limit of low frequency region in room acoustics). Then the impulse response is filtered using a 4th order zero-phase IIR high pass filter with a cutoff frequency of 20Hz to avoid any unwanted modes below the threshold that would not be audible to humans or reproducible through most sound reinforcement equipment. Furthermore, a zero phase is used to avoid any unwanted phase artifacts that may potentially affect the auralisation process. It must be further clarified that although perception does extend below 20Hz, due to the auralisation through headphones (to avoid room in room effects), it is beyond the scope of this Thesis to account for perceptual effects of infra-sound.

Trimming of the impulse response is achieved via the method described in (Lundeby et al., 1995), which is intended to trim the impulse response to obtain more

reliable acoustic metrics, accounting for potential shortcomings in acoustic measurement procedures. For this use case, trimming was necessary in order to avoid systematic effects of long ringing modes below the audible threshold that were still present in the signal, causing long periods of silence after the initial IR.

To summarise, the impulse response is sectioned into 50ms chunks, where the RMS level is calculated for each chunk. For this case, the final 10% of the impulse response was considered noise as suggested in (Lundeby et al., 1995). Finally, to ensure there are no issues due to early trimming of the impulse response or any effect due to sample discontinuities (causing audible clicking), a half window was applied from the +10dB point above the noise floor to the assumed -120dB linear decay point to ensure it was well below the audible limit.

4.2.2 Low Frequency Auralisation

The aim of the auralisation process is to auralise only the low frequency region of the room and have no auralisation above 250Hz. There are two main arguments for using this process, the first is to only present the low frequency so the listener is assessing the low frequency and modelled modal sound field without any bias due to higher frequency content. The second argument is to avoid any assumption in modelling and to avoid any biases in high frequency auralisation, where the reverberent sound field is no longer modal, as the modelled rooms assume no geometric complexity either at the room boundaries or throughout the room. This effect is not significant in a modal sound field, but plays a far more significant role when assuming high frequency scattering and complexities when modelling such behaviour. This in turn greatly extends the complexity in auralisation, where the high frequency may become the focus and cause a bias in response from the listener. Therefore, it is assumed that the trade-off for a somewhat unrealistic representation of a broadband room response is valid over increasing the complexity and potentially introducing confounding factors into the auralisation process.

The auralisation process described in this methodology is largely adopted from (B. M. Fazenda et al., 2015), however there are some subtle changes to streamline the methodology. While the resultant auralisation of the process is identical, the primary differences made in this work is the removal of downsampling and upsampling of the RIR and no equalisation of the transducer/headphones. The streamlined auralisation is shown in Figure 4.4.

The low frequency gain matching is achieved via calculating the RMS of both the low frequency dry signal and the low frequency convolved signal. The convolved signal gain is then adjusted to achieve the same RMS as the original dry signal. Finally, the filtered high pass dry signal is then summed with the low frequency convolved signal to form the full auralised signal.

The output of the auralisation process is therefore a sample (in this case a kick drum) which the low frequency ($20 <= f <= 250$Hz) is auralised with a room impulse response described in Section 4.2.1.

4.2.3 Pattern Generation

The main focus of this work is to investigate the specific case of low frequency percussive instruments, however when considering the presented stimuli to listeners

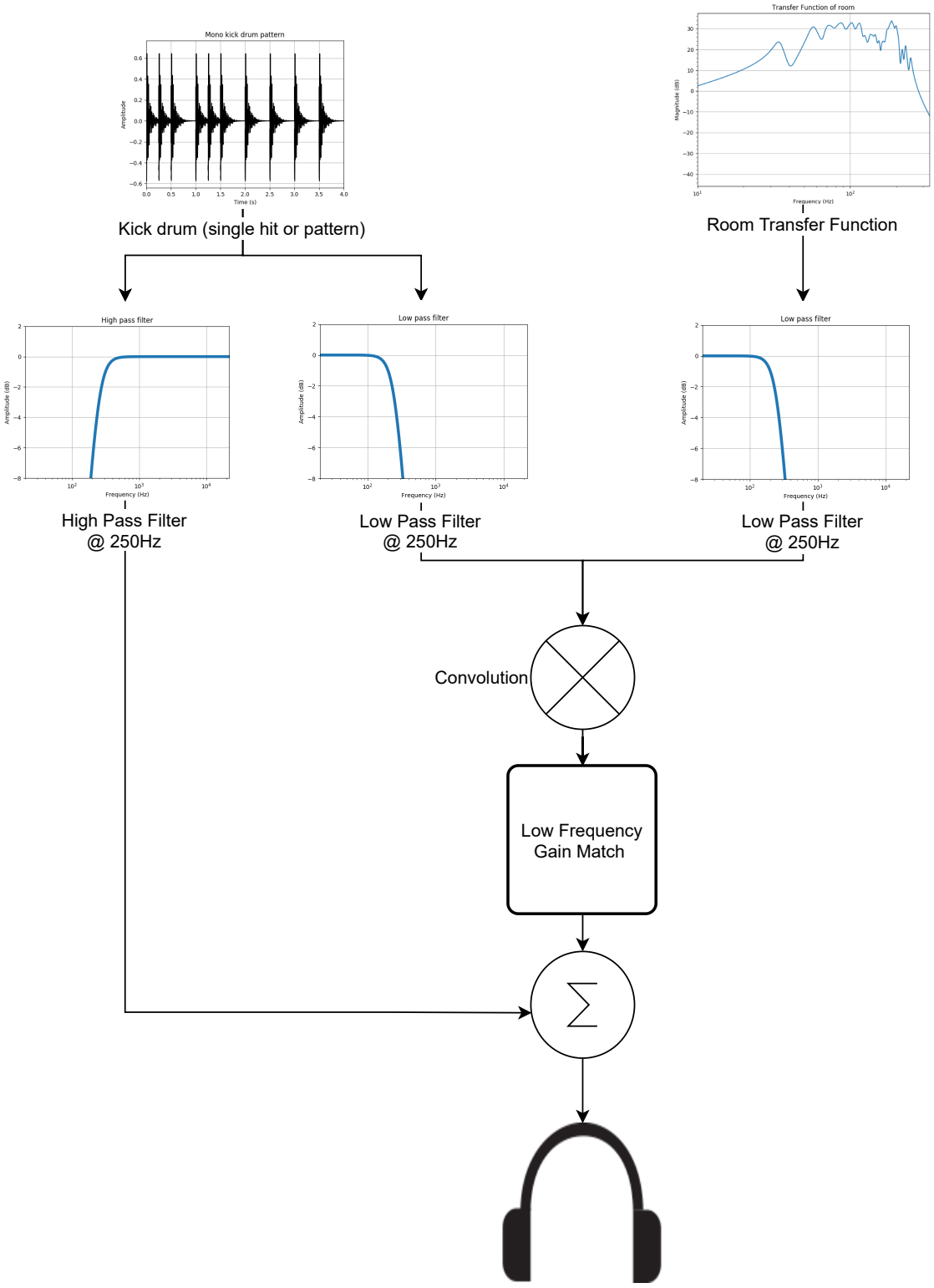


Figure 4.4: Diagram of the auralisation process adapted from (B. M. Fazenda et al., 2015) where output is an auralised kick drum between 20Hz and 250Hz with an assumed anechoic high frequency.

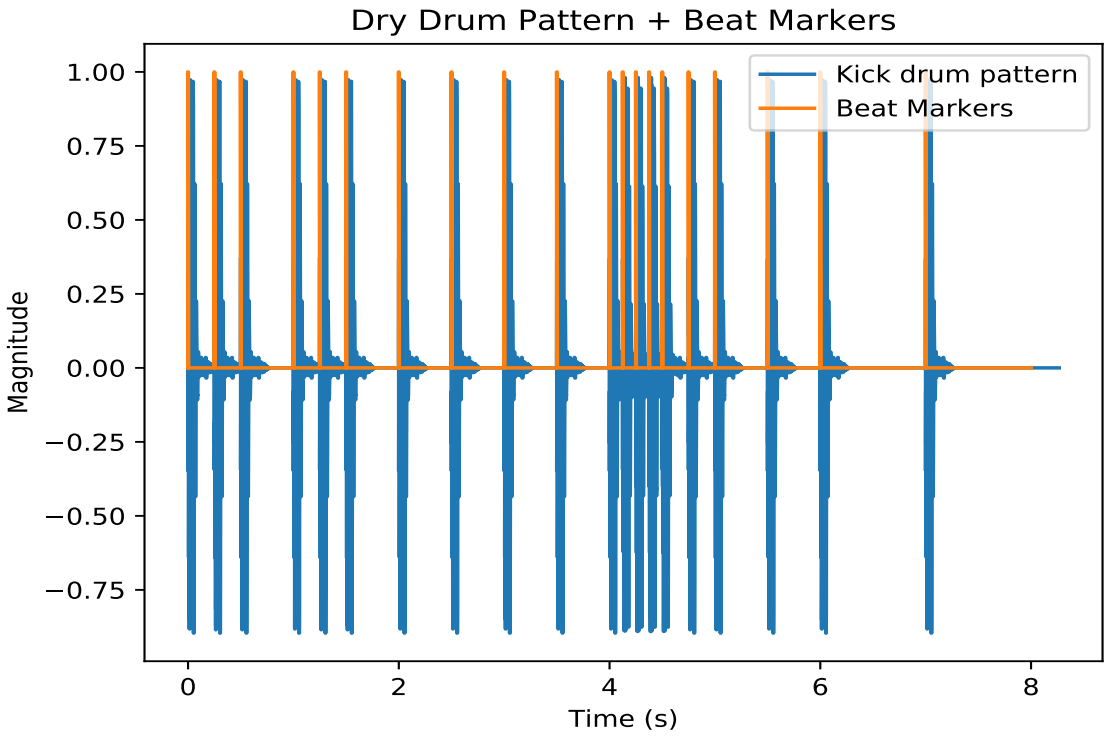


Figure 4.5: An example output of the pattern generation algorithm, where a kick drum is placed at user defined beat markers (shown in orange) to form an output pattern (shown in blue).

for rating a kick drum - room interaction, a problem arises. The issue comes with presenting a single kick-room combination or "one-shot", where the kick drum is played once in isolation. This method may prove to obtain unreliable results due to the potential short duration for auralisation and not allowing for natural build up of sound energy in the room (B. M. Fazenda et al., 2015). An alternative method may be to loop the auralised sample, however this may be biased by the shortest and longest impulse responses, as the whole IR must be audible leading to varying time lengths between loops with potential exaggeration of the decay time. Therefore, a drum pattern may be used as a method of presenting a realistic interaction with the room and remove biases due to presentation. However, there are caveats to consider when presenting a pattern, which is discussed further in Section 5.2.2.

A difficulty arises when introducing kick drum patterns into the stimuli, as there were little to no anechoic recordings of kick drums in a varied and controlled pattern for this use case. Therefore, a bespoke pattern generation tool is included in this methodology to have full control over the presented stimuli.

The overarching method consists of input beat placement markers through positions of bars and beats at a given tempo (expressed in beats per minute). From here, beat markers are placed throughout the pattern timeline, where a kick drum sample is convolved with the markers (pulse train) to form a completed pattern. An example of this process is illustrated in Figure 4.5.

Furthermore, one observation must be made as a limitation of this approach. A limitation arises in the lack of dynamics expressed in the pattern, where each kick drum hit is at an identical level. While this is a subtle sacrifice that leads to a less

"human"/realistic feel, it was thought that this limitation was suitable as adding dynamic control may lead to more variables and potential confounding factor to control in test and was therefore not accounted for.

4.3 Feature Extraction

The feature libraries used throughout this methodology were from various resources covering different purposes. MIR features comprise of standard Music Information Retrieval features; Room acoustic features describe mostly impulse response based metrics; Features From Literature describe various features that have been designed for the case of low frequency room acoustics; Finally, Bespoke Features describe features that were designed for the scope of this work.

As shown in Figure 4.2, the features are extracted on the fully convolved single hit kick drum + room auralisation. This was done to ensure that the mapping between perceptual responses and features matched, as the listeners were not rating on the impulse response in isolation. The full pattern auralisation was not made use of in the feature extraction process, partially due to many features not being suited to full music excerpts, thus reducing the potential room acoustic features (i.e. features designed for impulse responses) that may be utilised. Furthermore, since many of the features discussed throughout this section are metrics describing the impulse response of a room, the similarities in a single kick drum room auralisation share the impulsive nature of an impulse response and are therefore considered applicable to make use of these features.

Furthermore, in an effort to capture more information in the low frequency region, octave band filtering was implemented using the specification provided in BS 61260 (2014). The purpose behind this decision, was to capture information across different frequency bands, particularly acoustic features, which are calculated on the impulse response such as reverb time, clarity and centre time. Octave bands were chosen over a finer resolution (such as 1/3rd octave bands) to reduce the number of features, where a finer resolution would multiply the number of features. The increase in features may be problematic in the modelling process, due to many highly correlating features and a reduced signal-noise ratio in participant responses and features.

With this tool leveraged, it was then applied to many other temporal features that could be broken down into separate octave band frequency regions. However, while some features could be split into octave bands, others (mainly spectral features) could not and some features (mainly features from literature) could only be assessed using the low frequency region. Therefore, features described throughout this Thesis are split into 3 different categories, Broadband - covering the whole frequency range, Band Limited - describing octave band limited features and Low frequency - which describes features calculated across the region of 20Hz to 250Hz. Finally, octave bands were only computed up to the 250Hz region, as inclusion of higher octave bands would introduce information about the dry signal only, which may potentially arise as a confounding factor in the modelling if a particular kick drum was rated with a significant difference.

Frequency region	Encoded label
Broad Band	FeatureName.BB
Band Limited	FeatureName.32, FeatureName.63, FeatureName.126, FeatureName.251
Low Frequency	FeatureName.LF

Table 4.1: Table denoting the encoded labeling used throughout the methodology and modelling process; Where broadband denotes a feature that exists across the whole frequency range, band limited describes a feature that has been octave band filtered and low frequency describes features that are measured between 20-250Hz

4.3.1 MIR Features

The MIR feature library used throughout this section is the Essentia Library (Bogdanov et al., 2013), where the version used in this methodology is the Essentia 2.1 beta4 release (Music Technology Group & Universiat Pompeu Fabra Barcelona, 2019). The Essentia library includes a large feature set comprised of envelope, spectral, timbrel, tonal, rhythm, statistical and dynamic (loudness) based features (Universiat Pompeu Fabra Barcelona & Music Technology Group, 2019). However, for this use case not all features are useful, applicable or practical. For example, most rhythm based features are designed to extract tempo information from full music excerpts and were therefore not used as part of this methodology.

To therefore decide which features were included as part of the methodology, a review of each feature was made to only extract features that were applicable to short single shot samples, single figure metrics and made sense for the case of this research. While this approach is a manual selection route, an example of this approach may be HFC (high frequency content), where it can be computed on an auralised kick drum and fits the criteria for a single figure metric. However, HFC as a feature may become problematic in the analysis and modelling due to the issues of characterising the non-auralised portion of the signal, which deviates from the research problem and would highlight the unrealistic nature of the "anechoic" high frequency. The final feature list is shown in Table 4.2, where detailed descriptions of each feature can be found in the Essentia documentation for algorithms (Universiat Pompeu Fabra Barcelona & Music Technology Group, 2019).

4.3.2 Room Acoustic Features

A practical approach to a machine learning problem is to include features that help to describe the problem, where the feature space can provide reliable mapping to the output variables. Therefore, in an effort to utilise this approach, room acoustic metrics were included from BS 3382 (2009) excluding curvature, which was included from (Davy, 1989). The full feature list is shown in Table 4.3.

There are 2 main issues that must be addressed when using features intended to be calculated from the impulse response. The first is due to a convolved kick drum sample being utilised for feature extraction and the second is due to the typical range of interest for surveying the acoustics of a room residing between 125Hz and 4kHz. Whereas the frequency range of interest for this work is 20Hz to 250Hz.

To address the issue of using a convolved kick drum and impulse response, two main approaches can be taken; One approach would be to have a separate pipeline where the RIR is used to calculate impulse response based features; and the second

Feature name	Feature Label	Category	Frequency Range
Temporal Centroid	TemporalCentroid	Envelope	Band Limited
Strong Decay	StrongDecay	Envelope	Band Limited
Dynamic Complexity	DynComplexity	Dynamics	Band Limited
Loudness	Loudness	Dynamics	Band Limited
Attack Start	attStart	Envelope	Band Limited
Attack Stop	attStop	Envelope	Band Limited
Log Attack Time	LotAttTime	Envelope	Band Limited
Max To Total	MaxToTotal	Envelope	Band Limited
Temporal Centroid To Total	TCToTotal	Envelope	Band Limited
Spectral Centroid	SpectralCentroid	Spectral	Broad Band
Pitch Salience	PitchSalience	Tonal	Broad Band
Flatness (dB)	FlatnessDB	Spectral	Broad Band
Max Magnitude Frequency	MaxMagFreq	Spectral	Broad Band
Strong Peak	StrongPeak	Spectral	Broad Band

Table 4.2: Complete list of MIR features included in this Methodology.

Feature name	Feature Label	Category	Frequency Range
Decay	Decay	Acoustic	Broad Band, Band Limited
Early Decay Time	EDT	Acoustic	Broad Band, Band Limited
Clarity	Clarity	Acoustic	Broad Band, Band Limited
Definition	Definition	Acoustic	Broad Band, Band Limited
Centre Time	CentreTime	Acoustic	Broad Band, Band Limited
Curvature	Curvature	Acoustic	Broad Band, Band Limited

Table 4.3: Complete list of Room Acoustic features included in this Methodology.

approach is to make use of single kick drum hits that are a similar signal to the impulse response. However, the first option becomes problematic due to deviation of results that the input sample may cause. Previous work has shown a strong effect in choice of sample when assessing room acoustics (B. Fazenda et al., 2012) and may lead to a difficult decision of how to compensate for the effect of sample. One solution may be through multiple ratings that share the same coordinate in the feature space or aggregating perceptual ratings across the sample to remove the effect. Furthermore, the issue with using a single hit over the room impulse response, is that it deviates from the feature's design and intended use, which may have limited perceptual relevance. However, calculating an IR based feature on a signal that is more closely linked to a typical impulse response, rather than the auralised pattern presented to the listener, is deemed as a worthwhile trade-off in potential perceptual relevance over an insufficient use-case. Therefore use of single-hit over a continuous kick drum pattern will be used for feature extraction as a compromise between using the acoustic metrics in a novel, untested case, over tampering with scores and potential biases in how the rating or feature space may be manipulated through aggregation of scores.

Features From Literature

To further the available feature pool for modelling the perceptual attributes, features were taken from previous research which were designed to describe the low frequency response as a single figure metric. The figures taken from literature are from (Stephenson, 2012) and are "figure of merits", which describe the frequency deviation from different lines of best fit through the response. The lines of best fit used in this methodology are flat (straight line through the mean magnitude of the response), a smoothed 3rd octave band fit (Vanderkooy) and smoothed 3rd order polynomial fit through the response.

The complete feature list is shown in Table 4.4.

Feature name	Feature Label	Category	Frequency Range
Deviation from flat	FOMflat	Figure of Merit	Low Frequency
Deviation Vanderkooy best fit	FOMvanderkooy	Figure of Merit	Low Frequency
Deviation polynomial best fit	FOMpoly	Figure of Merit	Low Frequency

Table 4.4: Complete list of Figure Of Merit features included in this Methodology.

4.3.3 Bespoke features

Bespoke features are inspired by previous research that introduces a heuristic approach to modelling certain perceptions of low frequency sound. All features described throughout this section are created for the particular use case of low frequency room acoustics. The main focus of these features resides around the Modal Density Function (MDF) described in (Bolla et al., 2019), where a perceptually weighted response is formed using the perceptual modal thresholds. However, while the MDF may be useful for describing the perceptual low frequency response, there does not yet exist a single figure metric that encapsulates information about the perceptually weighted response. Therefore, the MDF features used throughout this

Feature name	Feature Label	Category	Frequency Range
Smoothed Error Late	MDFSmoothedErrLate	MDF	Low Frequency
Number of peaks Early	MDFPeaksEarly	MDF	Low Frequency
Number of peaks Late	MDFPeaksLate	MDF	Low Frequency
Early Cumulative Density	MDFCDensityEarly	MDF	Low Frequency
Late Cumulative Density	MDFCDensityLate	MDF	Low Frequency
Average Exceedance Late	MDFAvgExceedLate	MDF	Low Frequency
Early Decay Ratio	EDR	Acoustic	Band Limited
Exceed Threshold	ExceedThreshold	Acoustic	Band Limited

Table 4.5: Complete list of bespoke created features included in this Methodology.

section were created to describe characteristics of the low frequency response. The full feature list is shown in Table 4.5.

The MDF is split into two sections, early and late; defined by two thresholds, the first 10dB of decay and the remaining decay to -60dB for the early and late respectively.

Smoothed Error is influenced from the Figure of Merits described in Table 4.4, where the same approach used in FOMpoly is used, which describes the deviation from a 3rd order polynomial fit through the response.

Number of peaks is the number of peaks in the relative MDF response.

Cumulative density describes the error between the cumulative sum of the MDF and line of best fit between the minimum and maximum values of the cumulative sum. The underlying aim for this feature is to identify impulse responses that have large peaks that cause sudden large shifts in the cumulative sum of the MDF, where "flatter" MDFs should have a more linear cumulative sum and hence a lower error.

Since the MDF is a function that weights the response relative to the exceedance of the perceptual decay thresholds, Average Exceedance takes the average value of the MDF above the perceptual threshold. The aim for this feature is to outline MDFs that exceed the perceptual threshold and have more audible modes than ones that are under the perceptual threshold.

The early decay ratio is a simple ratio between the early decay time and the reverb time (RT60). The feature was created to show effects of cases where there may be a highly audible mode at the early portion of the response (causing a higher EDT) but may have little effect over the RT60.

Finally, Exceed Threshold describes the ratio between the measured band limited decay time and the perceptual threshold of modal decay. Where 1 denotes the measured decay time equal to the perceptual threshold, < 1 describes decay under the perceptual threshold and > 1 describes exceeding the threshold.

It must be explicitly stated that there may be little to no perceptual validity to these metrics other than the internal investigations performed as a heuristic approach to the research problem. Therefore, it may be difficult to understand the reliability of the metric when investigating the feature importance discussed in Chapter 6. However, since many of the other features included throughout this methodology have never been used in this specific context, a heuristically designed feature that describes certain characteristics in the data may be valid if proved useful in modelling the attributes.

4.4 Machine Learning

All machine learning methodology used throughout this section has used the scikit learn python package (Pedregosa et al., 2011) unless stated otherwise.

The machine learning pipeline used in this methodology is split into 5 major sections; Data preprocessing, feature selection, hyper parameter tuning, model validation and selection. The overview of the machine learning pipeline is shown in Figure 4.6.

4.4.1 Pre-processing of data

Pre-processing of the data is a core part of any machine learning process in order to make the data appropriate enough for the modelling process. The preprocessing steps used in this methodology were the following; Removal of participants due to reasons shown in Section 7.1.3; Shuffling of the data to ensure there were no biases due to a specific ordering in participant ratings; And finally a random split of the data is performed using a 80/20 split for the train and test set respectively.

4.4.2 Models in test

While there are a plethora of machine learning regression algorithms available with a variety of complexity levels, when concerned about the ability to interpret the model the choice becomes more limiting.

A good example of lack of interpretability, comes from a staple model that is used throughout machine learning applications, the neural network. A major advantage of these complex models is the ability to effectively model a wide variety of problems and applications, however due to the black box approach the ability to interpret the input and output relationships is greatly reduced.

Therefore, a simpler model approach is often better in cases where understanding and interpretation of the problem is of key importance, rather than completing a task. Hence, a starting point for most machine learning problems is that of the simple linear regression (or multi-variate regression) model, which describes a direct linear mapping between the input features and output variables. However, an issue with this simplified approach is the lack of complex modelling, where non-linear trends are often poorly modelled.

While simple linear regression is a useful approach, it may be too simplified for the use case of this research problem, as perception of human hearing is often non-linear by nature. Therefore, a more complex approach may be useful, however with the problem of the black box approach in mind, one learning model has a highly useful advantage, the Random Forest. Random Forest differs from other models with use of many decision tree's that allow the observer to directly infer information of the model in question's decisions and mappings. Furthermore, Random Forest by design has feature selection properties, where features that do not split the data significantly are dropped from being used in the decision trees.

Therefore, the models used in this methodology are linear regression and a Random Forest. While it is admitted that there are many more models that may have been used in this research, the two were chosen to represent a simple and complex approach and the results were sufficient enough to limit the model choice and not to increase the complexity.

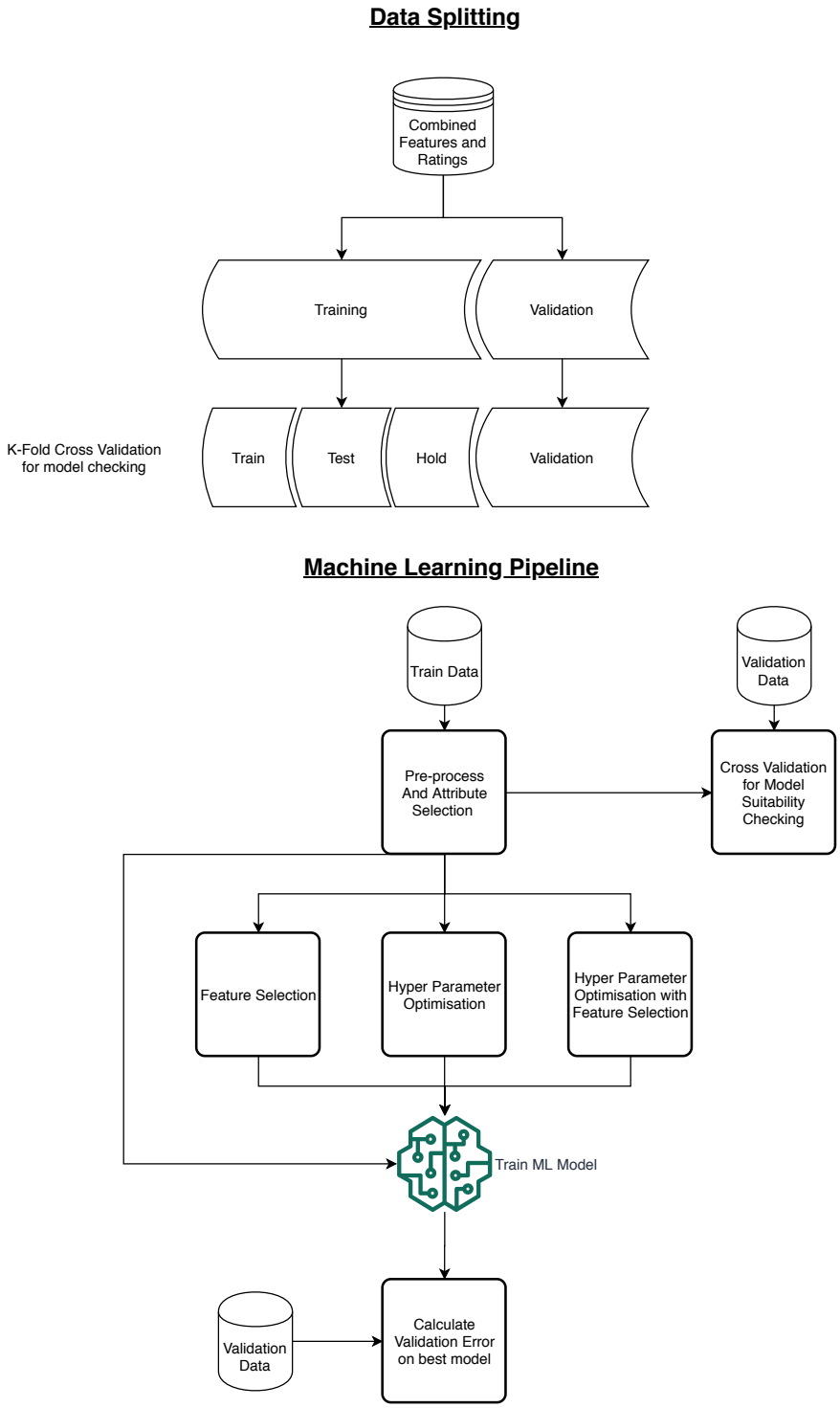


Figure 4.6: High level overview of the machine learning pipeline, where Data Splitting describes the process of forming a 80/20 train validation split of the combined feature and rating data and the further cross validation folds used for model suitability checking. And the machine learning pipeline which describes the modelling approach used to create different optimisations of the perceptual predictors.

Parameter	Values
n estimators	10, 25, 50, 100
max features (%)	5, 10, 30, 60, 90
max depth of estimator	2, 3, 4, 5
(%) minimum samples per split	10, 20
minimum samples at leaf	20, 50, 100

Table 4.6: Grid of hyper parameters used in the parameter optimization process for Random forest.

4.4.3 Feature Selection

The feature selection approach used in this methodology was Recursive Feature Elimination with Cross Validation (RFE-CV). RFE is a wrapper based method, which iteratively removes features across different folds of the data to obtain a best feature reduced model.

While uni-variate threshold based approaches are equally valid for feature selection, such as a k-best selection approach, issues arise however when determining the value of k, particularly when the aim in this methodology is to obtain a small interpretive feature set. Hence, RFE is a useful choice in this example due to the output of the best model with a reduced feature set.

The criterion set for RFECV was to choose a minimum of 5 features, a cross validation of 3 and a step size of 10 features. This was thought to allow for a reasonable feature set without the risk of under fitting with only 1 feature and a suitable step size and cross validation fold number to reduce the time taken for feature selection process without sacrificing too much accuracy.

4.4.4 Hyper Parameter Tuning

The hyper parameter tuning approach used in this methodology is Random Search, which describes the random search across a provided grid. While this approach may not yield the most optimal model, it was deemed appropriate enough with the reduced logistical constraints found in iterating across a large grid to obtain the best model.

The grid used for the random search is shown in Table 4.6, where the chosen number of iterations for the search was 100. This was chosen through a heuristic trial and error approach to obtain better performing models with a course grid to ensure even sampling across each parameter unless changes in that parameter were found to be not as significant.

4.4.5 Model Validation and Selection

A core part of the machine learning workflow is to define that a model is suitable for the intended problem, furthermore a selection criteria is used to obtain the best model. The aim of this methodology, is to make use of audio features to further the understanding of the perceptual attributes that relate to the effects of low frequency room acoustics. Therefore, the selection criteria used will be the model error (Mean Absolute Error), accuracy of prediction on the test set (R^2) and the number of features (N) used in the model. This selection criteria is used to obtain an accurate

model that should not be prone to overfitting with a small amount of features used in the model that may be interpretable.

Finally, to validate that the model approach used in the methodology is able to model the data, cross validation is used to check the models across different folds of the data.

Chapter 5

Experiments

The purpose of this chapter is to outline the methodology and control of variables for the purpose of a subjective listening test. Moreover, a preliminary listening test was performed to provide a greater understanding of the scope of variables used for the primary listening test. It must be noted that the control of variables described throughout this chapter is for the sole use of the primary listening test, where results are shown in Chapter 6. The aforementioned preliminary test methodology and discussion is purely limited to Section 5.4.

5.1 Purpose and outline of test

To reiterate the research objective of this thesis, the aim is to expand the understanding of the perception of low frequency room acoustics through use of perceptual attribute descriptors in a novel specific use case. Where the objective is to be achieved through modelling the perceptual attributes through signal features.

With the features outlined in Section 4.3, to model the perceptual attributes, known ground truths (or labels) are required to map input features onto output ratings. However, with subjective descriptors, there is no objective truth to be applied. Therefore, perceptual ratings are required through the use of subjective testing to investigate the mapping between objective signal features and subjective perceptual ratings.

Hence, the subjective test described throughout this section will consist of auralised kick drums being presented to the listener, where the participants will audition and rate each stimuli with their perceived score of Resonance, Articulation and Bass Energy.

5.2 Choice of Stimuli

5.2.1 Rooms

The variables in question for the choice of room are the following; Room Dimensions, damping/absorption of room boundaries and source-receiver position.

Room Label	X (m)	Y(m)	Z(m)	Volume (m^3)	Schroeder Frequency (Hz)
Rm10	6.6	5.8	2.8	107	191.7
Rm11	10	8	3	240	141.5
Rm12	19	9	3	513	102.6
Rm13	15	16	4	1020	85.9
Rm14	27	13.5	7.5	2734	63.4
Rm15	33	19	10.2	6395	48.2

Table 5.1: Room dimensions and volumes chosen for use in the subjective listening test. Schroeder Frequency estimated using lower bound damping profile (see Table 5.2 to show potential lower limit of Schroeder Frequency.

Room Dimensions

The room dimensions are a critical part of controlling the room volume, which is one of the most crucial aspects of room acoustics, affecting characteristics of the path, modal frequencies, modal density and time between arriving reflections. As previously mentioned, one issue with prior research in the field of room acoustics is the lack of variation in rooms that are used in test, with a strong influence on small listening environments. Therefore, for this methodology, room volumes were chosen ranging from a small listening space to a large performance venue.

To choose the actual environments and dimensions in test, room volumes and dimensions taken from (Adelman-Larsen, 2014), which were used as a resource of many different music venues across the globe, with the exception of *Rm10*, where dimensions were sourced from (University of Salford, 2020).

It must be made clear however that use of these room dimensions is not to empirically model these spaces as accurately as possible. The purpose of using real dimensions is to avoid biases in choosing arbitrary room dimensions and risking a misalignment in chosen environments. Therefore the modelled rooms here are a mere representation of what may be expected of a room of a certain capacity/volume. Table 5.1 shows the chosen room volumes.

Finally, it must be clarified that the room volumes under test represent a wide range of potential enclosed listening spaces that may range from small listening rooms, to medium/large enclosed live music and performance spaces. However, this work and findings may not reflect extremely large enclosed live music venues or open air performance spaces.

Absorption coefficient

While previous work has focused on controlling modal decay time as a factor of the damping of the room, this methodology deviates through use of the acoustic absorption coefficient to define the absorption/damping of the room boundaries. This was done partially due to the addition of a more realistic damping coefficient (use of Equation 3.10 over Equation 3.9), but was also introduced to give a much wider variance in modal decay times across different room-damping profile interactions. It may also be argued that this is a more intuitive approach, as the absorption coefficient is independent of all other room variables, whereas decay time abstracts the interaction between volume and damping of the room boundaries.

However, there is one limitation from using the absorption coefficient for low

Damping Profile	63Hz	125Hz	250Hz
D0	0.13	0.12	0.12
D1	0.28	0.29	0.29
D2	0.45	0.45	0.45
D3	0.75	0.75	0.75
D4	0.97	0.97	0.97

Table 5.2: Acoustic absorption profiles where the absorption coefficient α is defined in 3 octave bands to represent the acoustic absorption of the boundary surfaces.

frequency room acoustics, which is the limitation in measuring the absorption coefficient of materials at low frequencies due to high deviation of damping in the room (Vercammen, 2019). Therefore to account for this, two adjustments were made; absorption coefficient defined per octave band from 63Hz to 250Hz (presumed flat absorption coefficient from 1-63Hz); and a linearly spaced absorption coefficient across bands, which were modified due to the reasons outlined in Section 5.4. The chosen range of absorption coefficients are shown in Table 5.2.

The chosen limits for absorption, while may not be ecologically representative of typical rooms at the higher extremes (i.e. D4), the limits are proposed cases ranging from a poorly treated acoustic space to almost perfect absorption akin to an active absorber solution.

Source Receiver Position

Position of the source-receiver directly influences the modal coupling and therefore, can be assumed to shape the overall room response due to excitement of different modes. An ideal scenario for modal coupling, where all modes are excited is the case of a source in the bottom left corner in the room and the receiver in the opposing upper right corner of the room. However, this example only applies in a theoretical, idealised environment with point sources and receivers and is hence not a useful representation of a typical sound reinforcement setup and is therefore not appropriate for this methodology.

Furthermore, on moving the source and receiver, problematic systematic biases are introduced if the source or receiver are placed on an integer multiple (see Figure 3.2) of the modal wavelength (i.e. half or quarter length into the room) Placing the source or receiver on a pressure null will add a systematic bias in the presentation, where certain order modes are never presented to the listener.

To control for this effect, the source was left placed in the corner, where the receiver was placed in a proposed "Front of House" position, with a random offset and angle to avoid being placed on exact integer multiples of the room dimensions. While this approach does still incur some issues with biased modal coupling, the strongest, low modal orders are not affected by this. Another consideration that would help correct this issue is the use of multiple source receiver positions that may also aid the analysis to investigate the effect of modal coupling. However, due to the issues described in Section 5.4, inclusion of multiple source receiver positions was found to not further the analysis and encroached on the logistical constraints of the amount of stimuli, which was found to be more useful in other variables.

Two examples of the smallest and largest rooms are shown in Figures 5.1 and 5.2 respectively.

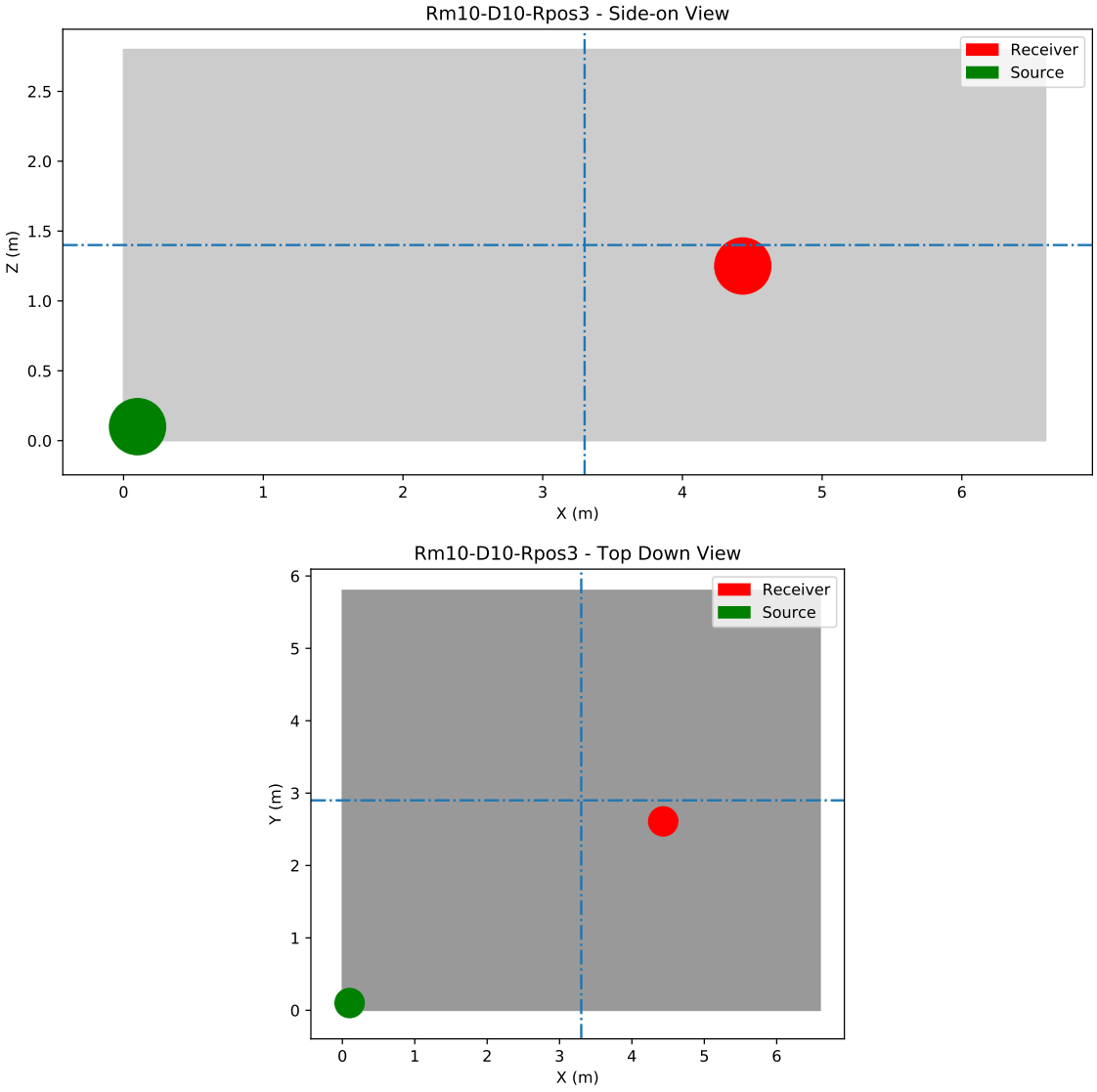


Figure 5.1: Source (shown in green) and Receiver (shown in red) positions of Rm10 showing the side-on elevation view and top down positional view where the dashed line denotes the half way point of the given axis where the source is placed in a corner and receiver is placed at an estimated seated head height of 1.25m at a random offset from the 2/3rds position in the room (simulated FOH position).

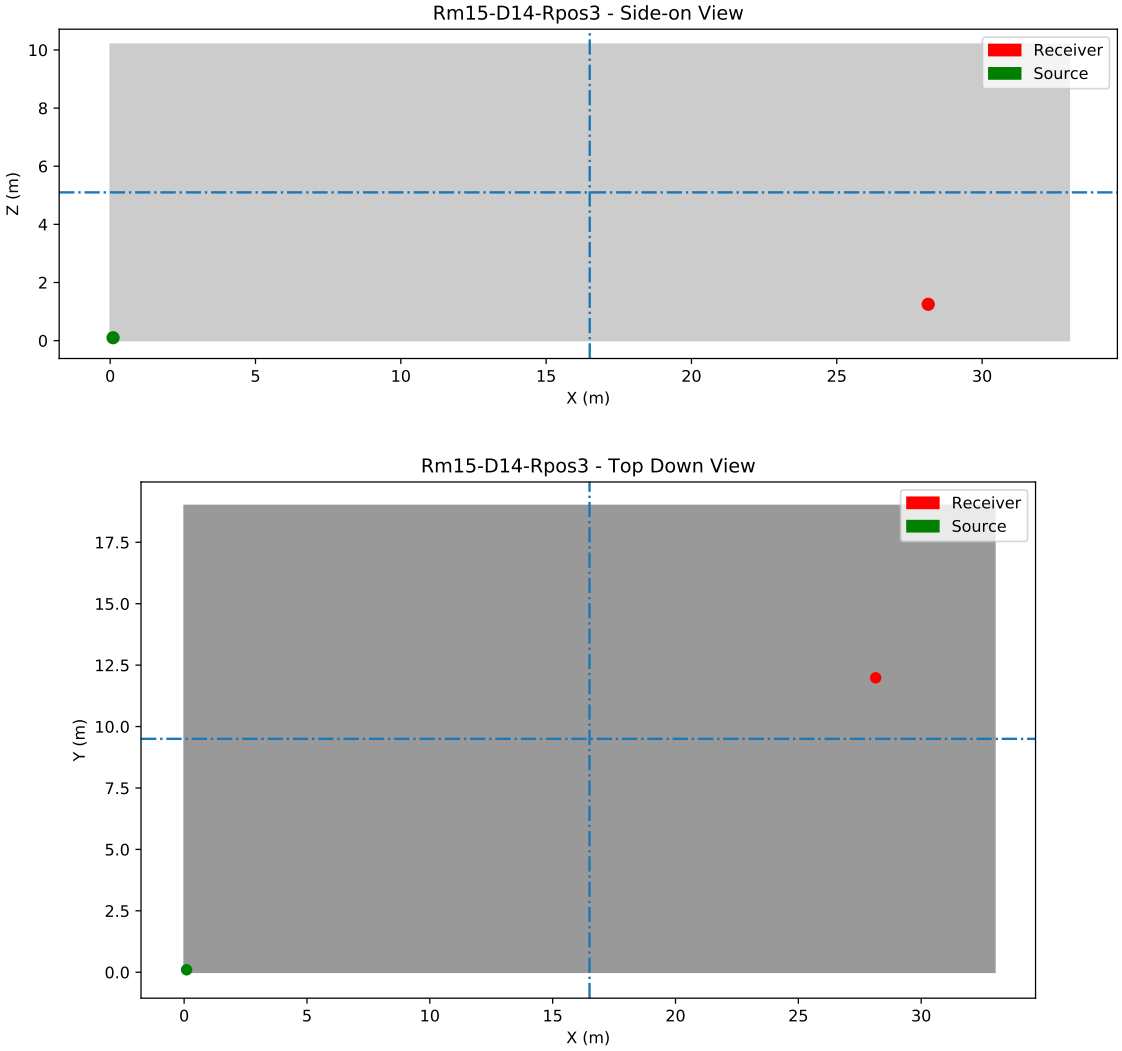


Figure 5.2: Source (shown in green) and Receiver (shown in red) positions of Rm15 showing the side-on elevation view and top down positional view where the dashed line denotes the half way point of the given axis.

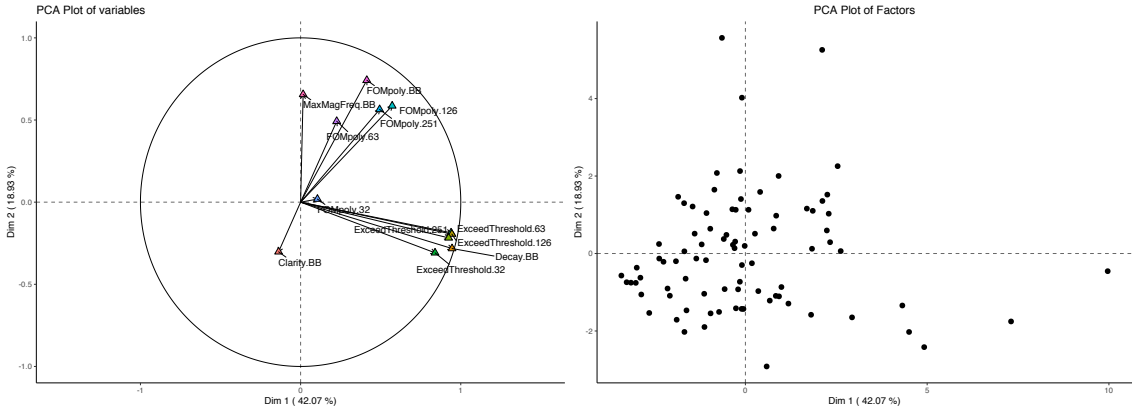


Figure 5.3: PCA of Kick drum space - all kick drums present. Variable plot indicating correlation of features shown on left and factor plot illustrating the mapping of individual kick drums to principal components 1&2.

5.2.2 Kick Drum Samples

Choice of the convolved sample is a complex and important task, as previous research has shown a strong effect due to the choice of music sample (B. Fazenda et al., 2012). Therefore, useful sampling of different kick drums is required to ensure that there is no bias in only investigating a certain type of kick drum.

The method used to select kick drums for the subjective test is similar to (Shier et al., 2017), where audio features and dimensionality reduction methods are used to cluster a sample library into similar characteristics. For this case, all audio features were used and then manually hand picked to obtain a feature space that described the kick drums in components that corresponded to the decay and pitch of the kick drums (dimensions 1&2 respectively shown in Figure 5.3). The corresponding feature space was then auditioned to find samples that covered sufficient variance in pitch/frequency content and decay. Figure 5.4 shows the PCA of the selected kick drum samples, which are taken from "corners" of the two components and auditioned to verify their suitability in representing the extremes of decay and pitch. The aim was to include kick drums that exhibited each combination of low/high pitch with short/long decay, resulting in 4 kick drums.

Another consideration that must be made is the pattern that the kick drums play, as different styles and tempos of patterns may introduce a bias in either looking at slow or fast excitation of the room, potentially biasing certain energy buildup conditions, causing an overemphasis on the effect of the room. Therefore, a pattern was chosen at 120bpm, with a variety of note values and combinations. The sub patterns included were a slow half note rhythm, a standard "four to the floor" quarter note beat, then two mixed patterns of quarter, eighth and sixteenth notes. The order of which was then randomised to avoid any bias due to a gradient of slow-fast or fast-slow. This was done once to form a single 4 bar pattern that was used to auralise the kick drum in each stimuli presented to the listener. The used pattern is shown in Figure 5.5.

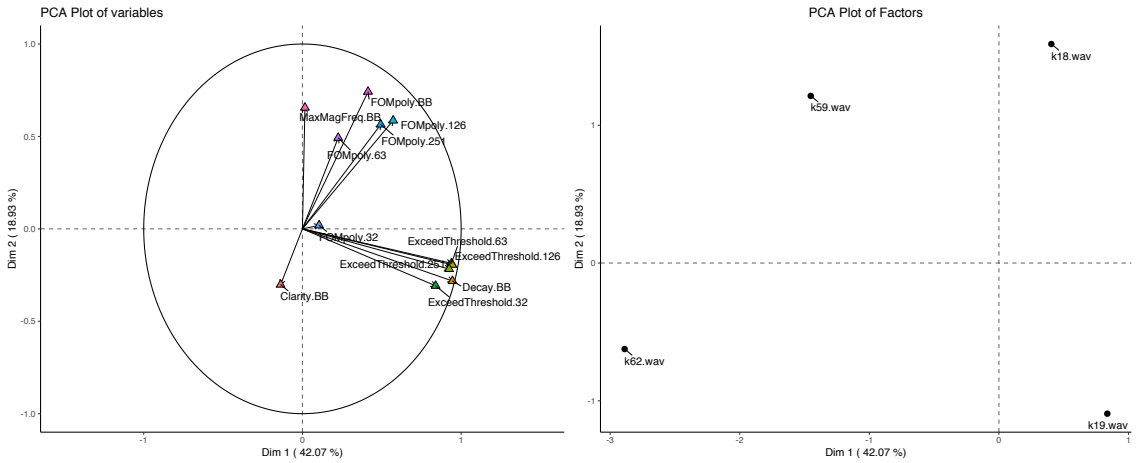


Figure 5.4: PCA of Kick drum space - Kick drums chosen for the subjective listening test, which are chosen from 4 corners of the factor plot (right). Where component 1 represents the decay characteristics of kick drum (denoted by decay based features on left hand variable plot) and component 2 represents features that describe the frequency content of the kick sample (denoted by the spectral based features on left hand variable plot).

Figure 5.5: 4 bar 120bpm kick drum pattern used in the listening test to demonstrate different ranges of note values and rhythmic styles.

5.3 Deployment Design of Test

5.3.1 Test overview

Participants were asked to rate on the three bass quality attributes of Resonance, Articulation and Bass Energy. The attributes were adjusted from their initial descriptors due to some discrepancies between the previous definitions set out in (Wankling et al., 2012). The aim of adjusting the definitions was to ensure that the overall definition did not change in the adapted description, but increased the consistency between attributes. The main changes were to combine the attributes of Strength and Depth to form Bass Energy and to include a definition for the scale extremes of Resonance, which previously only included a definition of high Resonance. The revised attribute definitions are shown in Figure 5.6

To obtain ratings and to allow participants to rate and audition stimuli, a test GUI was created in MATLAB for the listening test, shown in Figures 5.9, 5.7 & 5.8. The test GUI made use of sliders to rate the samples, which were set in the centre by default ranging from a score of 1 - 5 with increments of 0.1 to provide a near continuous scale. Each slider had ticks indicating the integer value of the each attribute rating between 1 and 5. While this may be a bias in centering around these points (Zieliński et al., 2008), it was thought useful to provide a reference to the participant to allow for less variance across perceptually similar samples. Listeners had to audition the stimuli in full at least once and had to move each slider in order to submit their ratings and move onto the next test stimuli.

Furthermore, attribute definitions were also provided in the GUI to maximise the familiarity of the attribute definitions. An example of the shown definitions is Figure 5.8.

5.3.2 Playback Level

The playback level used for the listening test experiment was 85dB SPL (unweighted) to follow the same SPL as used in defining the modal decay thresholds in (B. M. Fazenda et al., 2015).

To ensure playback level was set to 85dB, a similar methodology was used to the previous study. A difficulty arises when considering the playback level of the stimuli, therefore each stimuli was normalised to the same loudness using ITU-R BS.1770-1, where the loudness level was chosen to be -19LUFS which was the highest loudness level achievable without clipping. From this, a 0dBFS 1kHz sine wave was included in the normalisation to become a reference sine tone at -19LUFS.

Furthermore, a 90dB SPL sine tone was recorded through a B&K HATS as the analogue reference sine tone for SPL. This 90dB SPL sine tone was then adjusted to -3dBFS RMS to calibrate the measurement in the digital domain. From this, the -19LUFS sine tone was played through the headphones to the B&K HATS unit through the Sennheiser HD600 in test, where the sine tone was adjusted to read 5dB lower to achieve 85dB SPL playback.

While this measurement setup was sufficient in this methodology, it is admitted that use of a sound level meter over a relative digital change is a less error prone approach to playback calibration with more repeatable results.

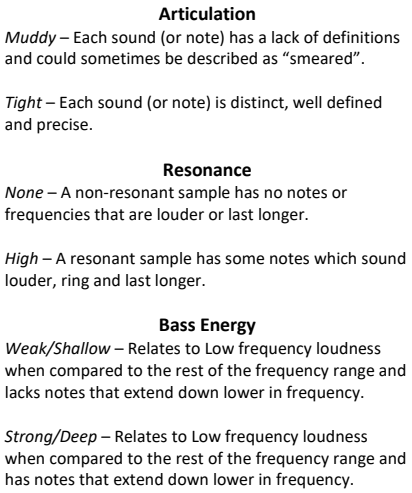


Figure 5.6: Revised attribute definitions as provided to the participants during the subjective listening test - where Bass Energy is formed from Bass Strength and Depth from Table 3.1.

5.3.3 Test Deployment

The equipment used for the test was the Sennheiser HD600 powered by an RME babyface Pro to control playback level. The listening test was conducted in a small listening room at Music Tribe offices with a low background noise level, where 13 participants took part in the final listening test, all of whom were employees at Music Tribe in the Machine Learning team. 11 of the participants could be described as "expert" listeners who had a background education in Music Technology/Acoustics and had prior listening test experience and 2 of which had little to no listening test or critical listening experience.

Regarding the logistics of the test deployment, there were 120 total samples auditioned to the participant (6 rooms, 5 absorption profiles, 4 kick drums), to control fatigue this was split into increments of 40 per session split across 3 separate tests. This split of test was primarily done to reduce fatigue in both rating and critical listening and to avoid timetable conflicts with employees having to "rush" through the test if it was too long, pressuring the subject to complete.

Overall, the 40-sample test took roughly 10-15 minutes depending on the participant, where the participants were given a break after 10 minutes to avoid fatigue, a progress bar was also included to allow the user to observe their progress throughout the test. On arrival of the test, the participants were provided an overview of the experiment procedure and were allowed to take a break or ask questions at any point they felt necessary, where they were free to withdraw consent at any time during the test. The test was approved by Salford University and Music Tribe Ethical procedures, where participants signed a consent form found in Figure A.1 and were

Figure 5.7: Listening Test GUI during the test phase of the subjective listening experiment, denoted by the black background and **TEST** at the top centre of the GUI.

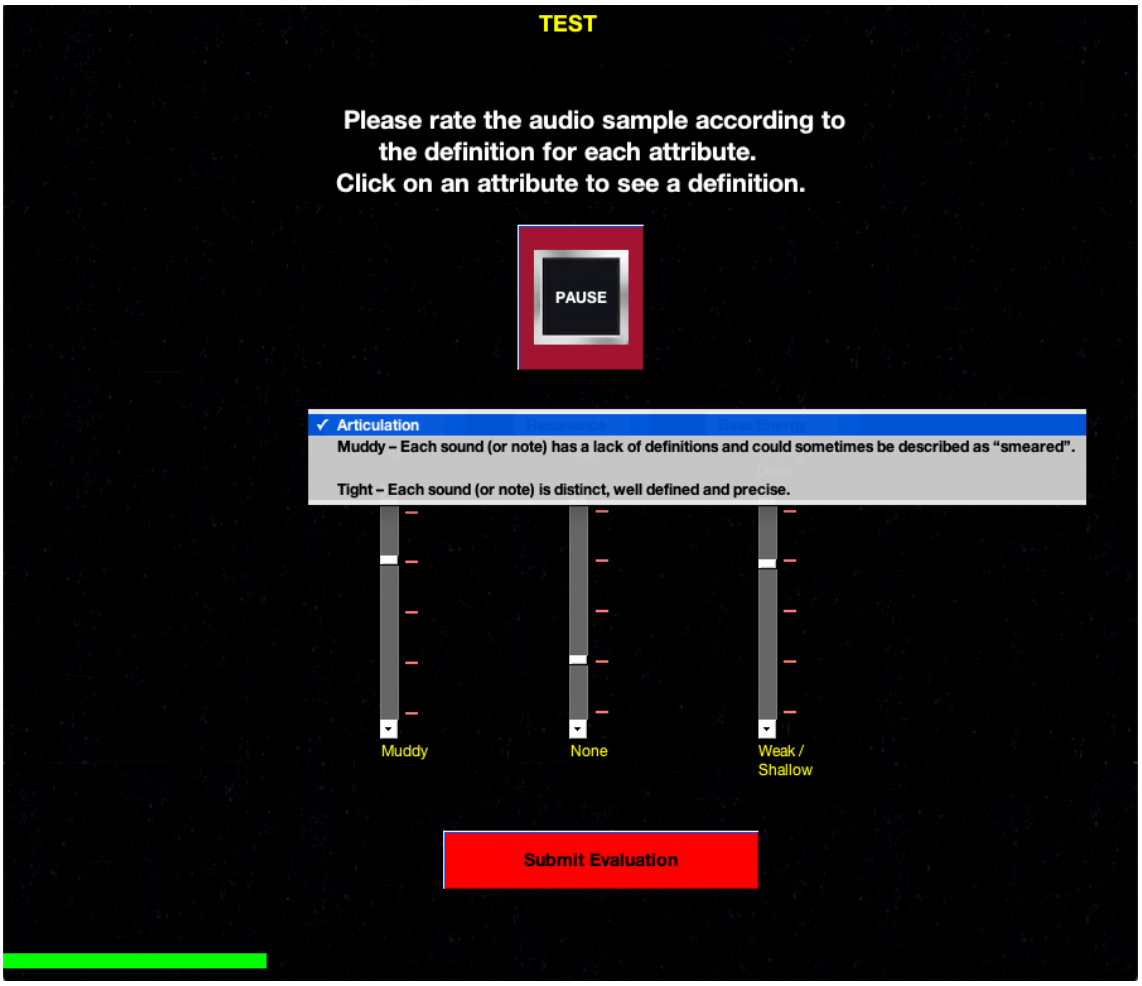


Figure 5.8: An example of an attribute definitions from Figure 5.6 shown as a supplementary method of finding the attribute definitions using the listening test GUI.

Figure 5.9: Listening Test GUI during the training phase of the subjective listening experiment to familiarise the listener with, denoted by the grey background and **TRAINING** at the top centre of the GUI.

provided a copy of the attribute definitions shown in Figure 5.6.

Participants were first met with a training phase to familiarise themselves with the test interface, stimuli and attribute definitions. The training stimuli was comprised of a randomly shuffled, pre-defined list of room/kick combinations that were also including in the test stimuli. This was done reduce bias between listeners, in cases where some may be presented a skewed distribution of stimuli during the training phase when presented with only poorly scored rooms and vice versa. The training stimuli was made up of 10 samples that the listener had to rate. The GUI background was coloured bright grey with *TRAINING* printed across the top of the interface (Figure 5.9). Once complete, the GUI screen changed to black and the text changed to *TEST* at the top of the interface.

Finally, a pop-up dialogue box appeared on completion of the test to inform the listener that they had completed the split of the listening test.

5.4 Preliminary Listening Test

5.4.1 Preface

An initial listening test was conducted as an exploitative effort to better understand the relationship between room acoustic variables and quality attributes. However, on completion of the test, it was found that an artifact of using non-linear phase filtering in the auralisation and improper trimming of the impulse response (prior to the inclusion of the method outlined in Section 4.2.1), caused an exaggeration of the impulse response buildup and length. Therefore, the presented auralisation was a misrepresentation of the intended modelled room. However, useful results were gathered to inform the approaches set out in the primary listening test, where the main findings are set out in Section 6.1. Furthermore, it must be made clear that due to the aforementioned issues, results are not reliable to make conclusive findings and are **not** included in the model and analysis presented in Chapter 6. Therefore, only results that were used to refine the control of variables are shown in this section.

Although the reliability of results shown are distorted by the exaggeration of impulse response, the initial parameters for independent variables resulted in a skewed distribution across scores towards high resonance and low articulation. This distribution effect is best illustrated through observing the Resonance scores across all presented stimuli in Figure 5.10, hence Resonance will be used to illustrate the effects throughout this section for this reasoning. Therefore, the primary objective of the primary listening test was to adapt the independent variables to cover a more even distribution of the Attribute scales.

Henceforth, the following section describes how the preliminary listening test results were used to justify change in independent variables. First the decision on receiver position is discussed, followed by the change of scope of room volumes and how the respective acoustic damping conditions were adapted to optimise the results of the test. Finally the outcome of test refinement is shown in Section 5.4.4.

5.4.2 Effect of Receiver Position

In the preliminary test, 2 receiver positions were used in an attempt to reduce systematic bias in modal coupling. However, it was found to be a complex variable to analyse and interpret in results. Consequently, the trade off to avoiding systematic modal coupling meant that a logistical constraint arose when adding a second receiver position, doubling the stimuli presented to the listener. This accumulated in sacrificing the sampling of other variables that were clearer to analyse and exercised the scale in a more direct manner.

When investigating receiver position, it was found that there was little effect across different volumes and damping profiles, particularly in larger room volumes. Inclusion of multiple receiver positions primarily introduced higher variance into the results as shown in Figure 5.11, particularly in smaller room volumes. On combining receiver positions in the analysis, the two distributions are converged together leading to results to become difficult to interpret when not factoring receiver position. Figure 5.12 shows the effect when excluding receiver position from the analysis, which leads to a greater distribution across ratings in smaller rooms. This effect is observed clearly in room volume 1440, when the average absorption is 0.42.

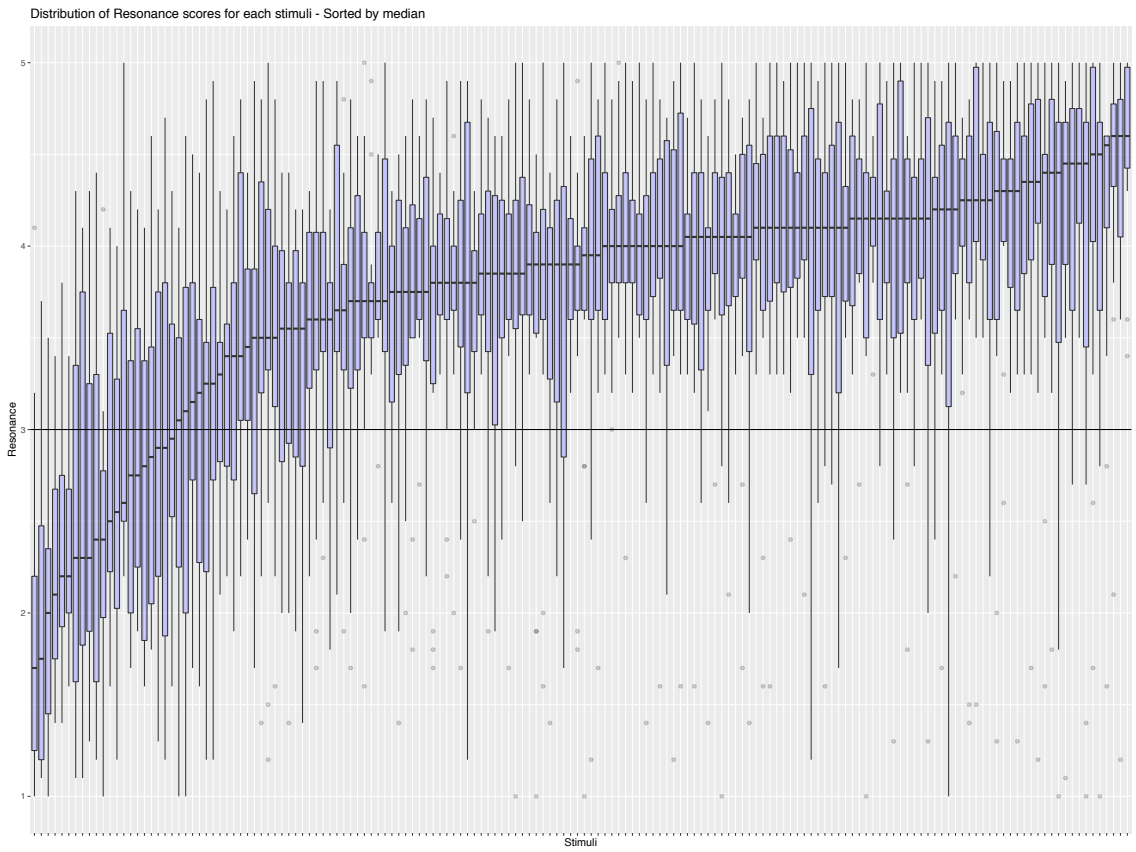


Figure 5.10: Preliminary listening test ratings across all presented stimuli (shown on x-axis) and participants ordered by the median Resonance rating. The black line denotes the mid point of the Resonance scale to highlight the skew of distribution into higher Resonance ratings

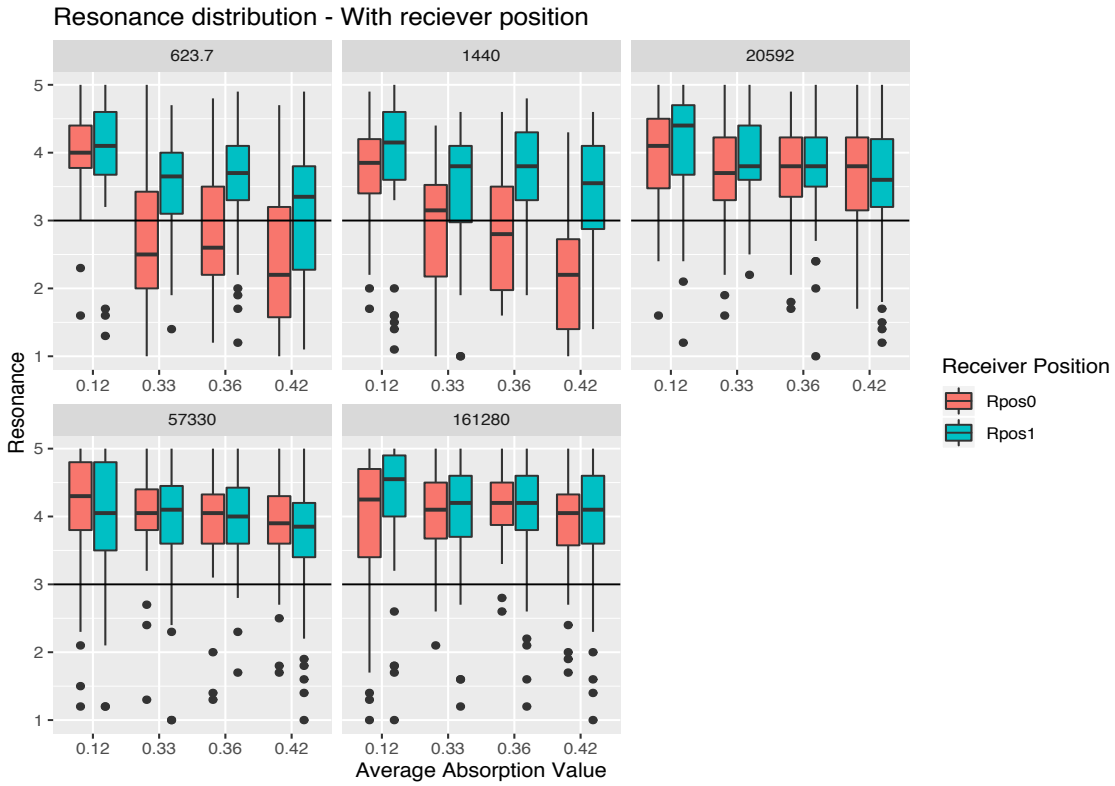


Figure 5.11: Illustrating the effect of including multiple receiver positions - where Resonance scores are split across volume (m^3) denoted in grey bar above each sub-plot and acoustic absorption coefficient of the room (x-axis). The black line denotes the mid point of the Resonance scale to highlight the skew of distribution in scores.

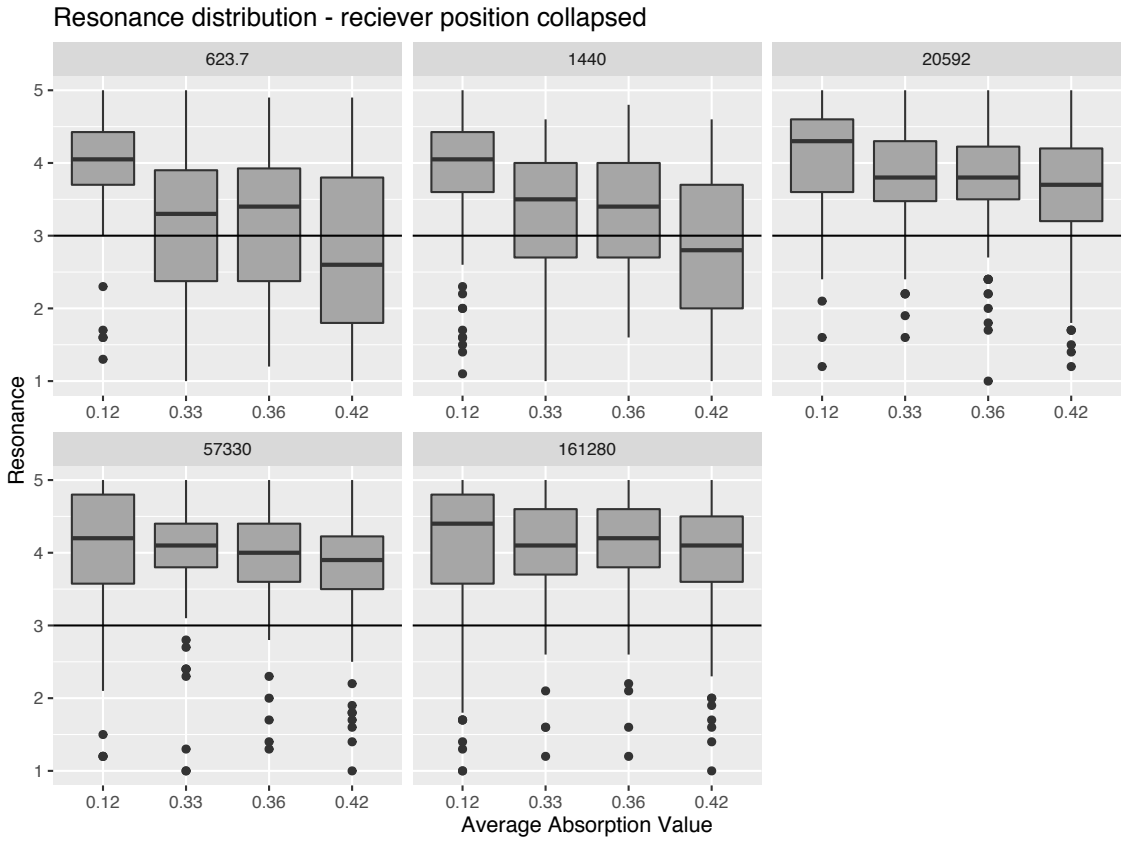


Figure 5.12: Illustrating the effect of variance of ratings due to collapsing multiple receiver positions together over splitting across receiver positions as in Figure 5.11 - Where Resonance scores are split across volume (m^3) denoted in the grey bar above each sub-plot and acoustic absorption coefficient of the room (x-axis). The black line denotes the mid point of the Resonance scale to highlight the skew of distribution in scores.

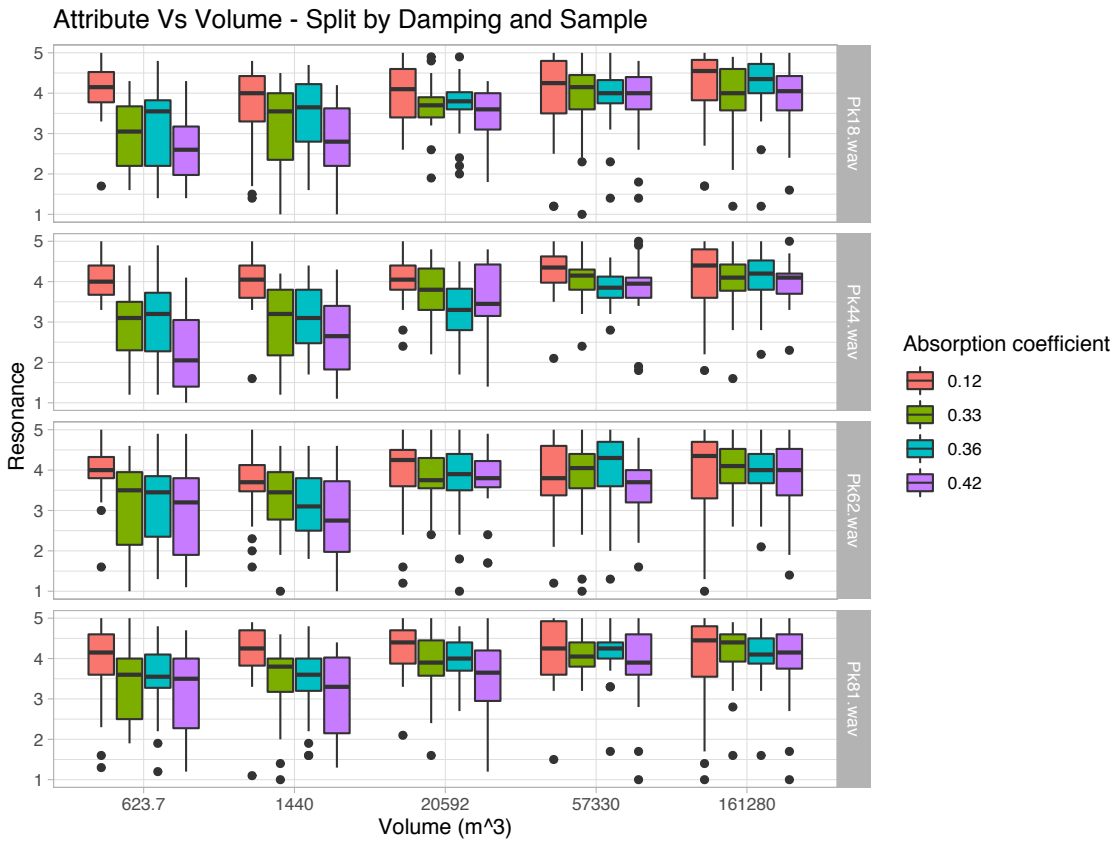


Figure 5.13: Illustrating the skewed distribution of Resonance scores due to the inclusion of large room volumes and limited scope of absorption coefficient in the preliminary listening test - Where Resonance scores are split across Sample (denoted in the grey box to the right of each sub-plot), acoustic absorption coefficient (shown in the coloured legend) and volume (across x-axis).

In this case, the effect of receiver position shows 2 clearly separated distributions, whereas this is smeared when receiver position is not factored.

Furthermore, receiver position is room dependent for modal coupling and cannot be compared directly across two different room volumes. For example, testing Rpos1 in room 1 and Rpos1 in room 2 are not comparable. This is mostly due to the changing room geometry rather than a particular effect of physical volume alone.

Therefore, when considering the trade-off between systematic bias of modal coupling and having a fixed receiver position, it is assumed that inclusion of more room volumes with a fixed position would aid in both introducing another variation in modal coupling and room volume. However, it is understood that this is a factor to consider when interpreting results.

The initial values chosen for room volume were aimed at representing a variety of musical performance spaces ranging from small 300 capacity spaces to large 12,000 capacity arenas using dimensions from literature (Adelman-Larsen, 2014). On analysis of the selected room volumes, a compression effect was found when increasing the room volume. The compression effect occurred in considerably large environments, where an increase in room volume caused no significant change in attribute score. This effect is illustrated in Figure 5.13, where both damping and sample have little effect over the perceived Resonance at the larger room volumes.

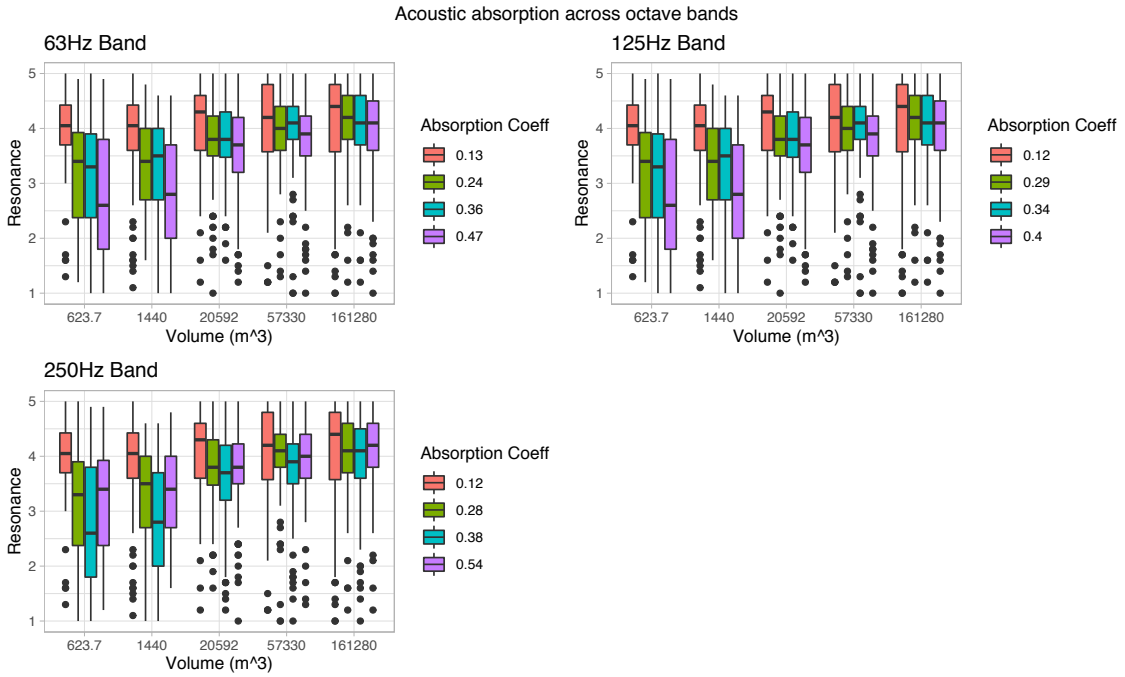


Figure 5.14: Illustrating the effect of varying the acoustic absorption coefficient between octave bands (split in each sub plot) - Where Resonance ratings are split between volume (x-axis) and acoustic absorption coefficient (coloured legend).

Therefore, when refining the test design, a different sampling approach was made to include smaller rooms leading up to the larger volumes (i.e $\sim> 20,000m^3$ in this example), full table of Volumes is shown in Table 5.1.

5.4.3 Scope of Acoustic Absorption

Two major considerations were made when defining the acoustic absorption coefficients in the rooms. The first is that acoustic absorption coefficients are defined per octave band and the second is the respective absorption coefficient chosen.

For the design of the preliminary test, representative values for real rooms were calculated using the reverberation times given in (Adelman-Larsen, 2014). From these times, damping profiles were created where different acoustic absorption values were used across bands. This effect is illustrated in Figure 5.14, which shows how the absorption values vary across volume using resonance to highlight the changes.

There are two significant shortfalls of the approach used here. First, wide discrepancies across the bands makes interpretation of the desired profile more complex, where a room may have problematic resonances in only one band. Secondly, the range of values that were used was not sufficient to exercise the full use of the scale, where in no case did the absorption cause the median value of resonance to fall below 2.5.

Therefore, the outcome of this analysis was to keep the in-band absorption coefficients similar to allow more transparent analysis. Finally, the analysis suggested a wider range of absorption coefficients should be utilised to properly exercise the full use of the scales.

5.4.4 Outcome of Test Refinement

To conclude this section, it was found that the initial approach to the research objective resulted in a non-uniform distribution across the ratings, where there was a skew towards high Resonance, low Articulation stimuli. Therefore, careful consideration of both the scope and control of all the variables was used to aim towards a more balanced test design, where the objective was to test the full extent of the attribute scales.

Figure 5.15 showcases the outcome of the test refinements described throughout this section, which highlights both a more even distribution of scores, as well as less variance and fewer outliers across each stimuli.

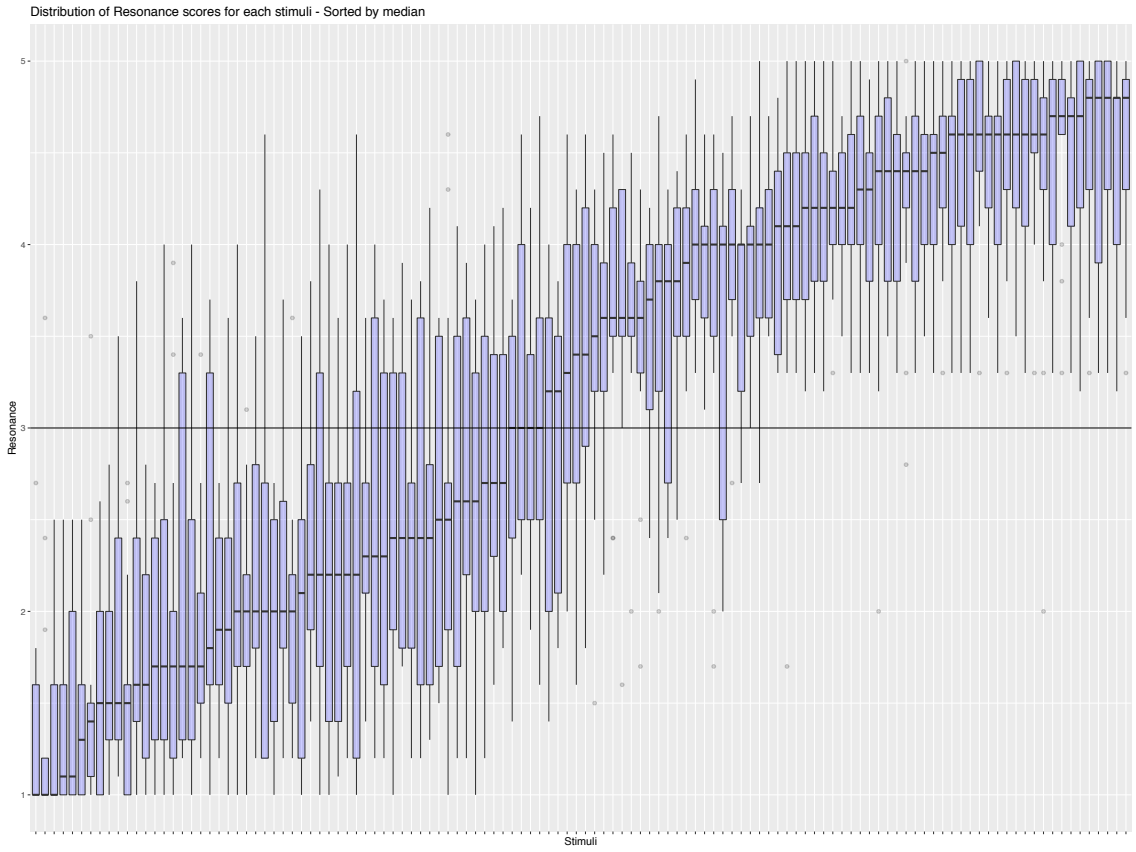


Figure 5.15: Distribution of all Resonance scores in the main subjective listening test, illustrating the improvements made from using the findings from the preliminary test by tweaking the scope of variables - Where the x axis represents each labeled stimuli ordered by median Resonance score across all participants. The black line indicates the mid Resonance score to emphasise the more even distribution of scores compared to Figure 5.10.

Chapter 6

Results

6.1 Listening Test Results

6.1.1 ANOVA analysis

To begin the analysis, first an ANOVA was conducted to understand the statistical significance between the independent variables and the attributes in test. To verify the data is suitable for an ANOVA, first the distribution results are shown in Figure 6.1 and the results of a Levene test (suitability through homogeneity of variances) are shown in Table 6.1. All Attribute results have p-values < 0.05 and can therefore be assumed suitable for ANOVA analysis.

The ANOVA results for Resonance, Articulation and Bass Energy are shown in tables 6.2, 6.3, 6.4 respectively. It must be noted that the ANOVA effect sizes are interpreted throughout this section using the rule of thumb values shown in (University of Cambridge, 2019). Where referring to small, medium and large effect sizes corresponds to eta squared values of 0.01, 0.06 and 0.14 respectively.

The resonance ANOVA shows that there is a statistical significance between Resonance and absorption, room volume and the kick drum sample. Effect size as described by Eta squared, shows that the absorption profile used in the room model provides the the highest significance showing a large effect size. Furthermore, the effect size of volume is encroaching on a medium effect size as the second most significant variable. Finally the kick drum sample, although significant, only has a small effect size on the perceived resonance.

Second order interactions are found between the acoustic absorption and volume and the acoustic absorption and sample. While significant, both second order interactions have a lower effect size.

The Articulation ANOVA again shows significance for absorption, volume and sample, where a significant second order interaction between absorption and Vol-

Attribute	F value	Pr(>F)
Resonance	1.468	0.00118
Articulation	1.707	8.016e-06
Bass Energy	1.300	0.02025

Table 6.1: Results from Levene test for suitability for ANOVA - $p < 0.05$ highlighted in red.

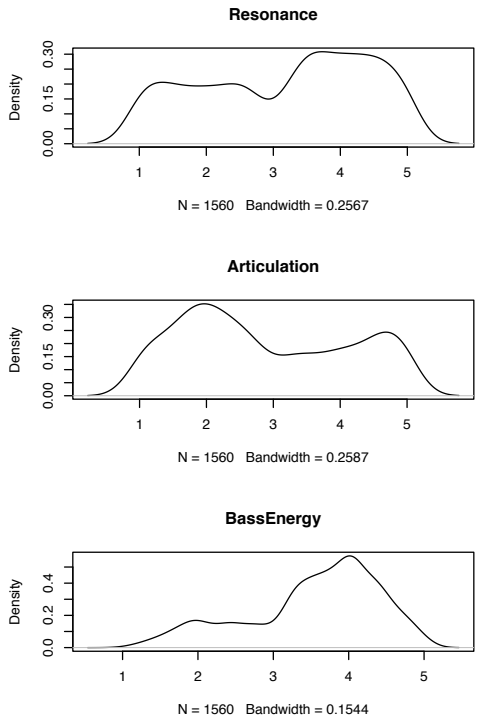


Figure 6.1: Density plots of each perceptual attribute rating obtained in the subjective listening test, highlighting the modality and distribution of ratings across the attribute scale (x-axis).

	Df	Sum Sq	Mean Sq	Eta Sq	F value	Pr(>F)
Absorption coeff	4	1355.05	338.76	0.564	652.10	0.0000
Volume	5	138.33	27.67	0.058	53.26	0.0000
Sample	3	29.56	9.85	0.012	18.97	0.0000
Absorption coeff:Volume	20	68.02	3.40	0.028	6.55	0.0000
Absorption coeff:Sample	12	26.50	2.21	0.011	4.25	0.0000
Volume:Sample	15	10.37	0.69	0.004	1.33	0.1754
Absorption coeff:Volume:Sample	60	26.31	0.44	0.011	0.84	0.7963
Residuals	1440	748.07	0.52			

Table 6.2: Analysis of Variance table for Resonance scores - $p < 0.05$ highlighted in red.

	Df	Sum Sq	Mean Sq	Eta Sq	F value	Pr(>F)
Absorption coeff	4	972.81	243.20	0.399	297.73	0.0000
Volume	5	89.82	17.96	0.037	21.99	0.0000
Sample	3	101.64	33.88	0.042	41.48	0.0000
Absorption coeff:Volume	20	57.31	2.87	0.024	3.51	0.0000
Absorption coeff:Sample	12	10.87	0.91	0.004	1.11	0.3479
Volume:Samples	15	6.68	0.45	0.003	0.55	0.9156
Absorption coeff:Volume:Sample	60	23.12	0.39	0.009	0.47	0.9998
Residuals	1440	1176.28	0.82			

Table 6.3: Analysis of Variance table for Articulation scores - $p < 0.05$ highlighted in red.

ume. Similarly to Resonance, the absorption profile used has the largest effect size for perceived Articulation, where both Volume and sample and the second order interactions had smaller effect sizes. Interestingly, although both small effect sizes, Sample has a greater effect size on Articulation than Resonance.

	Df	Sum Sq	Mean Sq	Eta Sq	F value	Pr(>F)
Absorption coeff	4	68.03	17.01	0.056	24.48	0.0000
Volume	5	49.12	9.82	0.041	14.14	0.0000
Sample	3	32.85	10.95	0.027	15.76	0.0000
Absorption coeff:Volume	20	18.46	0.92	0.015	1.33	0.1503
Absorption coeff:Sample	12	9.24	0.77	0.008	1.11	0.3488
Volume:Samples	15	4.81	0.32	0.004	0.46	0.9595
Absorption coeff:Volume:Sample	60	25.62	0.43	0.021	0.61	0.9911
Residuals	1440	1000.52	0.69			

Table 6.4: Analysis of Variance table for Bass Energy scores - $p < 0.05$ highlighted in red.

Finally, Bass Energy shows significant main effects of absorption, volume and sample, however there are no second order interactions found for Bass Energy. An interesting outcome is found here, although significant relationships are shown, the effect size observed is small across all significant results, with the exception of absorption, which is close to a medium effect size.

6.1.2 First order interactions between attributes and variables

The first order interactions are shown in Figure 6.2 to provide a visual aid to the ANOVA results. The Resonance results show clearly the large effect size due to the absorption coefficient, where a strong relationship is found in decreasing perceived resonance on increasing acoustic absorption. A medium effect size exists between Resonance and volume, where an increasing room volume corresponds to an increase in perceived resonance, however there is high variance among room volumes, where the minimum and maximum Resonance scores are observed across all room volumes. Furthermore, the small effect size is shown in the sample presented to the listener, where there is little variation on the median Resonance.

The Articulation scores show a similar but inverse relationship to Resonance scores, where the high effect size of absorption coefficient correlates to an increasing

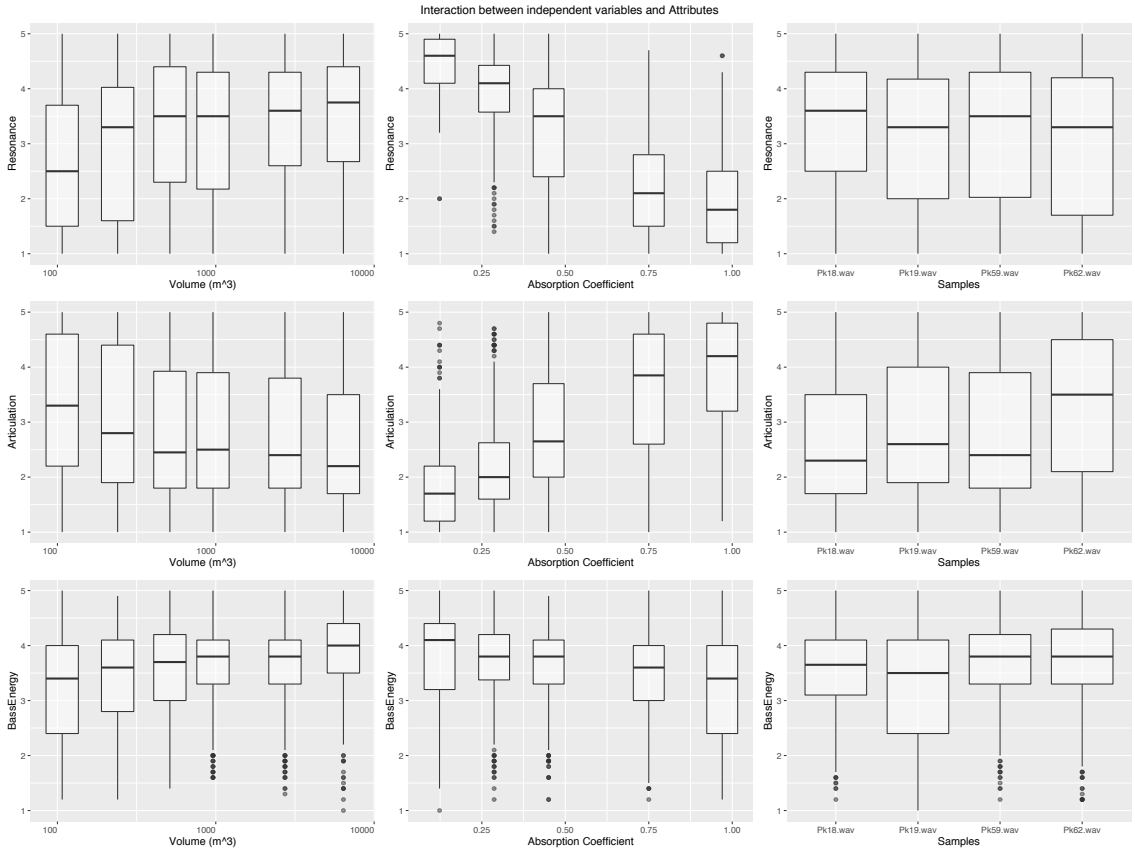


Figure 6.2: First order interactions between attributes (y-axis) and independent variables (x-axis split across sub plots).

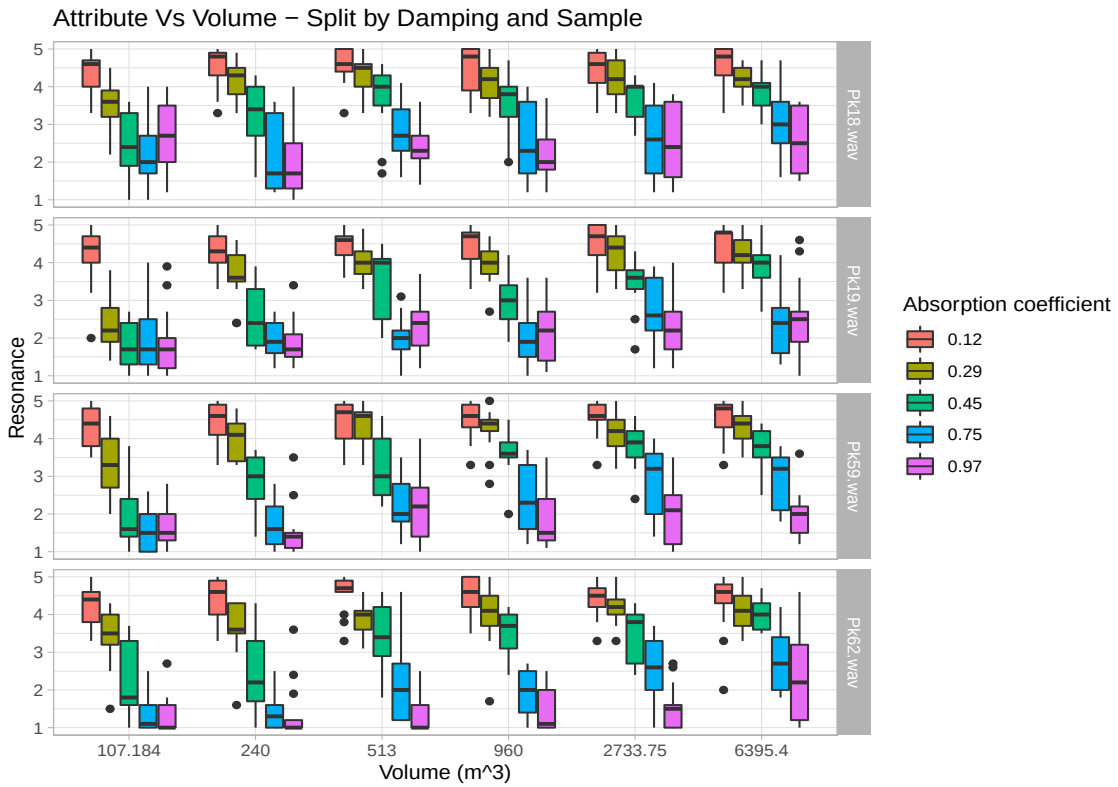


Figure 6.3: All second order interactions between variables and Resonance - Where sample is split across sub plot (denoted by the corresponding grey boxes), volume of room is shown in the x-axis and absorption coefficient is split by colour (shown in plot legend).

Articulation rating on increase in absorption. However, as stated earlier, there was a higher effect size on sample for Articulation (although still small). This larger effect size can mostly be attributed to a single sample (Pk62.wav), which corresponded to a much higher articulation score than the other presented samples.

Finally, turning the attention to Bass Energy, the small effect size becomes clear on observing the results, where perceived Bass Energy does not appear to be significantly changed, where the median score is always towards the higher end of the scale (> 3). However, there are significance from the variables in test, where it can be seen that there is some effect of higher bass energy in larger room volumes and with lower absorption coefficients. It must be noted that it is difficult to therefore draw conclusions with Bass Energy, due to the small effect size found.

6.1.3 Further interactions between variables

Section 6.1.1 found significance for second order interactions between variables. Figures 6.3 and 6.4 show the second order interactions between for Resonance and Articulation respectively. There were no significant second order interactions found for Bass Energy and will therefore not be presented in the results.

The ANOVA results in table 6.2 showed significance between Absorption and Volume and Absorption and Sample. The Resonance scores in figure 6.3 show again the large effect of the acoustic absorption on perceived resonance, where a

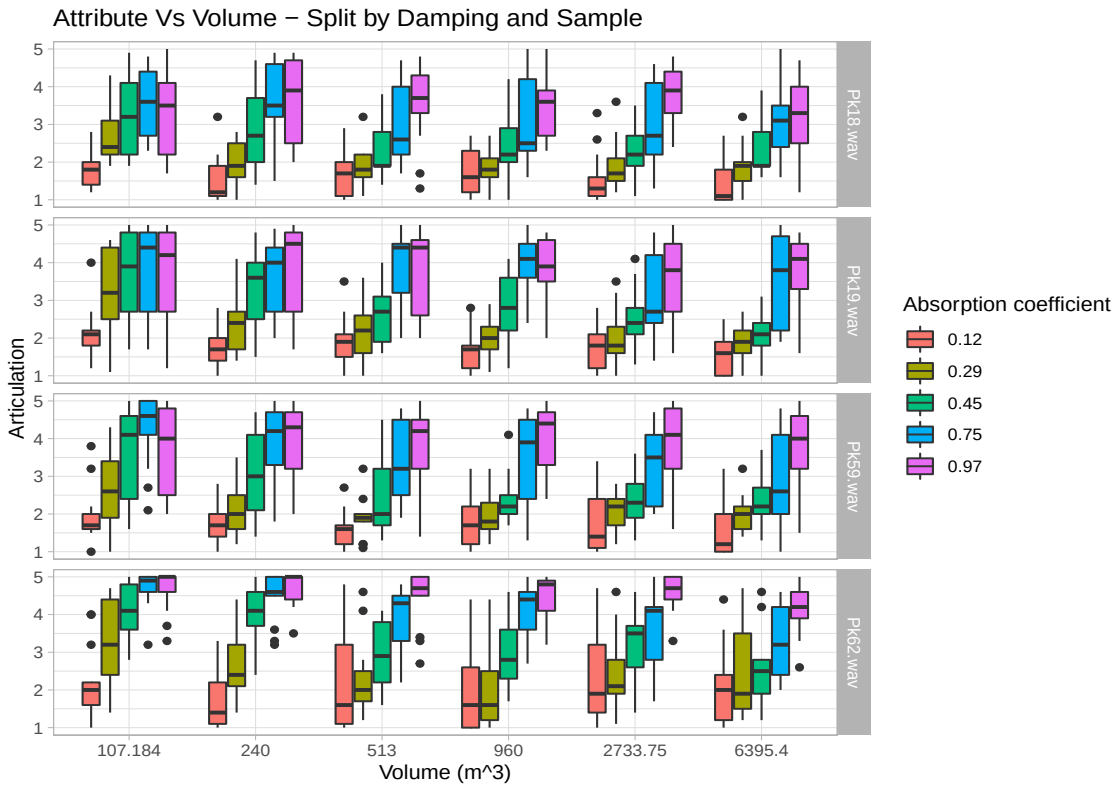


Figure 6.4: All second order interactions between variables and Articulation - Where sample is split across sub plot (denoted by the corresponding grey boxes), volume of room is shown in the x-axis and absorption coefficient is split by colour (shown in plot legend).

similar trend is found for increasing absorption corresponds to a decreasing resonance score. However, the significant second interaction due to sample and room volume appears to change the shape of the resonance and absorption interaction. This is best observed across the smallest room, where to achieve a lower resonance score, less absorption is required, whereas the largest presented room shows that even the highest absorption may not correspond to the lowest resonance score. This effect is also observed across sample, which shows a subtle effect of changing the aforementioned interaction on absorption.

Again, similar but inverse results are found on observing the second order interactions for articulation. However, only the acoustic absorption and volume was found as a significant result with a small effect size. The significant effect of volume and absorption have a similar outcome to the Resonance second order interactions and are observed in Figure 6.4, where the volume changes the distribution of the interaction between Articulation and absorption.

6.2 Results of Perceptual Modelling

This section outlines the results of modelling the perceptual attributes of Resonance, Articulation and Bass Energy. The models explored throughout this section are of a two-fold approach, where the first is between use of a linear regression model, aimed at a simple approach of modelling the attribute; and a more complex approach using

ensemble learning with a Random Forest model. The second part of the investigation incorporates optimisation of the modelling process through investigating both feature selection and hyper-parameter tuning. It should be noted that due to the simplicity of the model, linear regression contains no hyper-parameter tuning, only feature selection.

Abbreviations are used throughout this section for the modelling, which describes the model used, any hyper parameter tuning and feature selection applied to the model. Each model abbreviation is encoded in the following order *Model - Hyper Parameter Tuning - Feature Selection*. Table 6.2 shows the abbreviations and corresponding model used with any optimisations.

Abbreviation	Model	Hyper Parameter tuning	Feature Selection
LinReg-AllFe	Linear Regression	N	N
LinReg-RedFe	Linear Regression	N	Y
RandFor-Default-AllFe	Random Forest	N	N
RandFor-Default-RedFe	Random Forest	N	Y
RandFor-RandSearch-AllFe	Random Forest	Y	N
RandFor-RandSearch-RedFe	Random Forest	Y	N

Table 6.5: Abbreviations used to describe the model in test with the corresponding hyper parameter optimisation and feature selection process applied.

Furthermore, it should be clarified that the feature selection algorithm used throughout this section is that of RFECV (Recursive Feature Elimination - Cross Validation). The hyper parameter tuning method used throughout this section is that of random search cross validation.

All mentions of cross validation have been performed with a k-fold of 3 on the data.

6.2.1 Suitability of models

The first step of investigating how the perceptual attributes can be modelled, is to outline how each regression model performs on a subset of the data. This approach of model suitability is shown in Figure 6.5 by investigating the cross validation error across different train test splits of the data.

The results show that both Articulation and Resonance can be modelled with similar error across different regression models and little deviation between implementations.

While this does not imply on the effectiveness of the chosen models, it does imply that all models have roughly equal suitability in modelling Articulation and Resonance.

Finally, it is clear to state that none of the chosen models are sufficient to model Bass Energy with the given feature set.

6.2.2 Model Evaluation

To evaluate the regression models used, the train and test errors are shown in Figure 6.6.

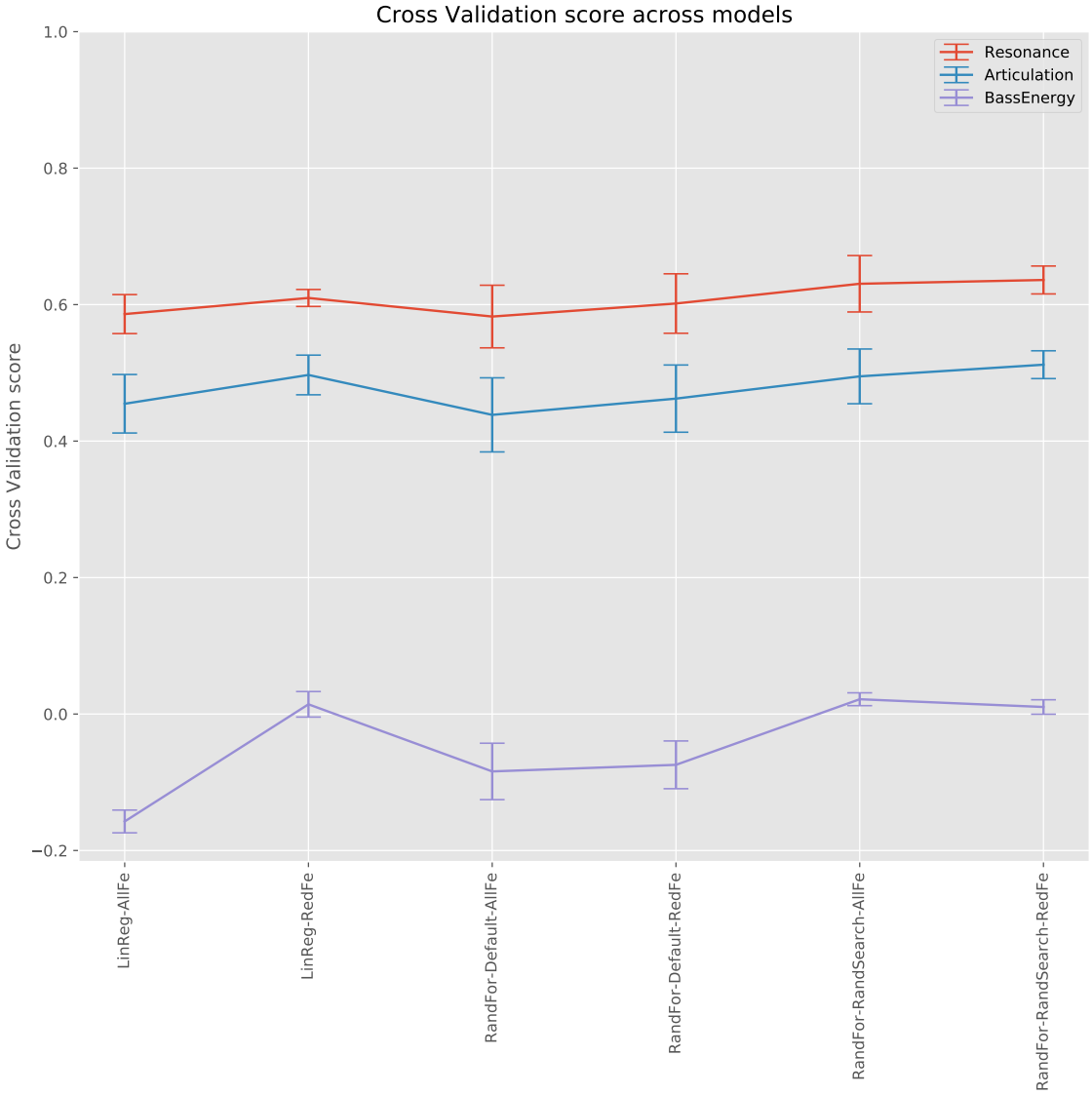


Figure 6.5: Checking the model suitability through cross validation $k = 3$ for Resonance (red), Articulation (blue), Bass energy (purple) across each model (shown in x-axis) - Where each point denotes the mean cross validation score across folds and the upper and lower bound lines denotes the standard deviation across all folds.

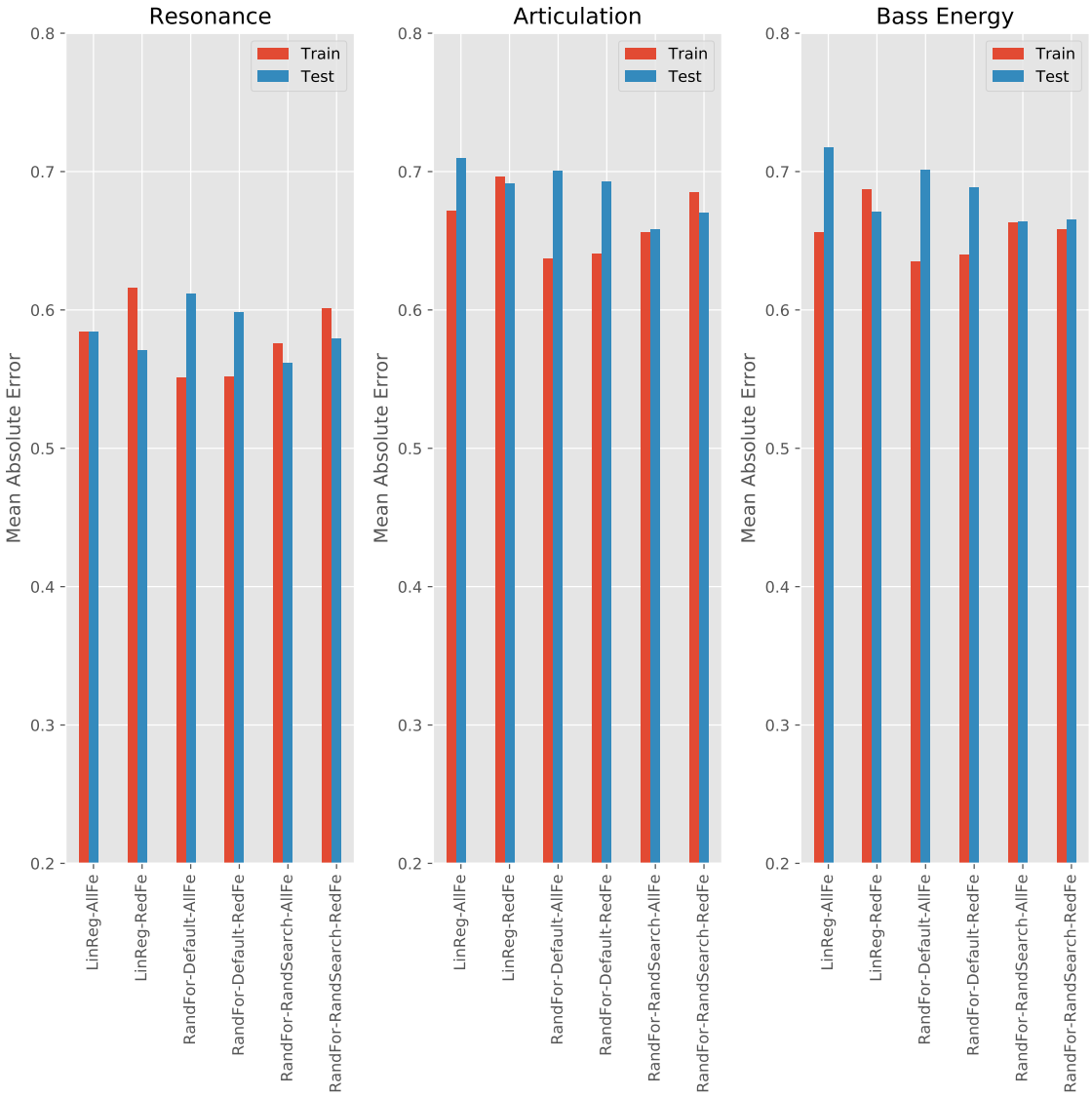


Figure 6.6: Mean absolute train (red) and test (blue) error shown across different models (x-axis) split across each attribute (sub-plot).

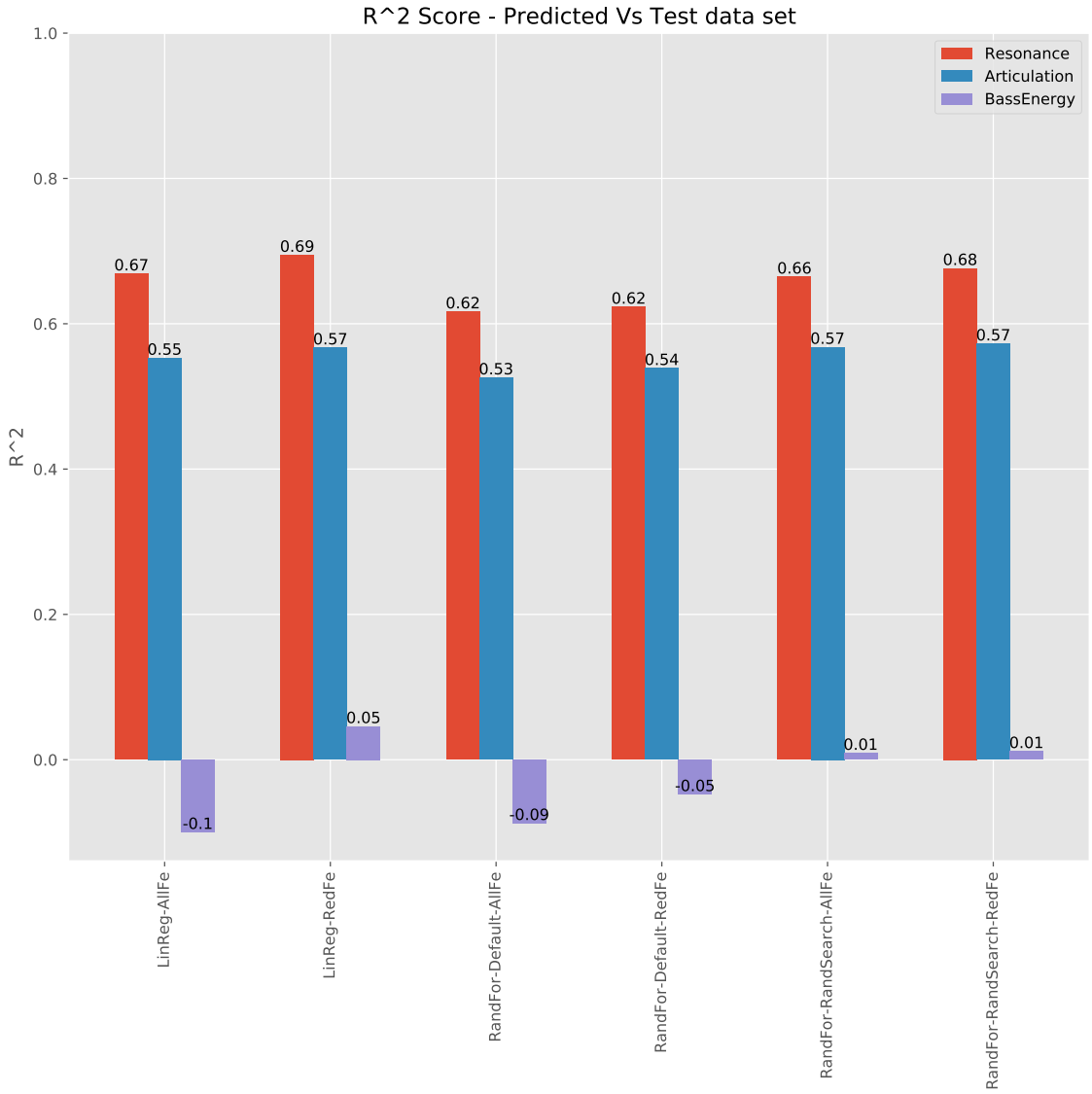


Figure 6.7: Explained variance (R^2) in validation data set shown for Resonance (red), Articulation (blue) and Bass Energy (purple) shown across each model (x-axis).

Further to this, the R^2 score calculated between the predicted and actual scores from the test data set on the corresponding attributes. The R^2 scores are shown in Figure 6.7.

Again, a similar trend is shown, where good values for fit are observed for Resonance and Articulation, whereas Bass Energy cannot be modelled effectively. Furthermore, it is again shown that there is little considerable difference between the different models chosen, although it may be argued that there is a slight decrease in accuracy in the default random forest models that do not include hyper parameter optimisation.

6.2.3 Feature Evaluation

One of the key goals of modelling the attributes, is to find which features are important in describing the perception of attributes. A key part of this, is to reduce

the number of features where only the important key features remain. Therefore, part of this modelling stage is to observe the number of features that were included in each model. Figure 6.8 shows the number of features used in each model, note that the models without feature selection will always report the total numbers of features available in the training data.

From the results shown, the feature selection techniques are successful in reducing the number of features to a very small, interpretative amount of features without incurring a significant loss in accuracy in the models. An interesting comparison now must be made between linear regression and random forest, where random forest seems to account for similar variance in the case of Resonance and Articulation. However it is seen that for the case of random forest without hyper parameter tuning, the feature selection process can use a mere 5 and 8 features, whereas linear regression requires 28 features for both resonance and articulation respectively.

Contrasting this with the error plots shown throughout 6.5, 6.6 and 6.7, it may be clear to evaluate the best model based on the minimum number of features, rather than the absolute error or explained variance in the model, due to the lack of significant deviation of accuracy between each model.

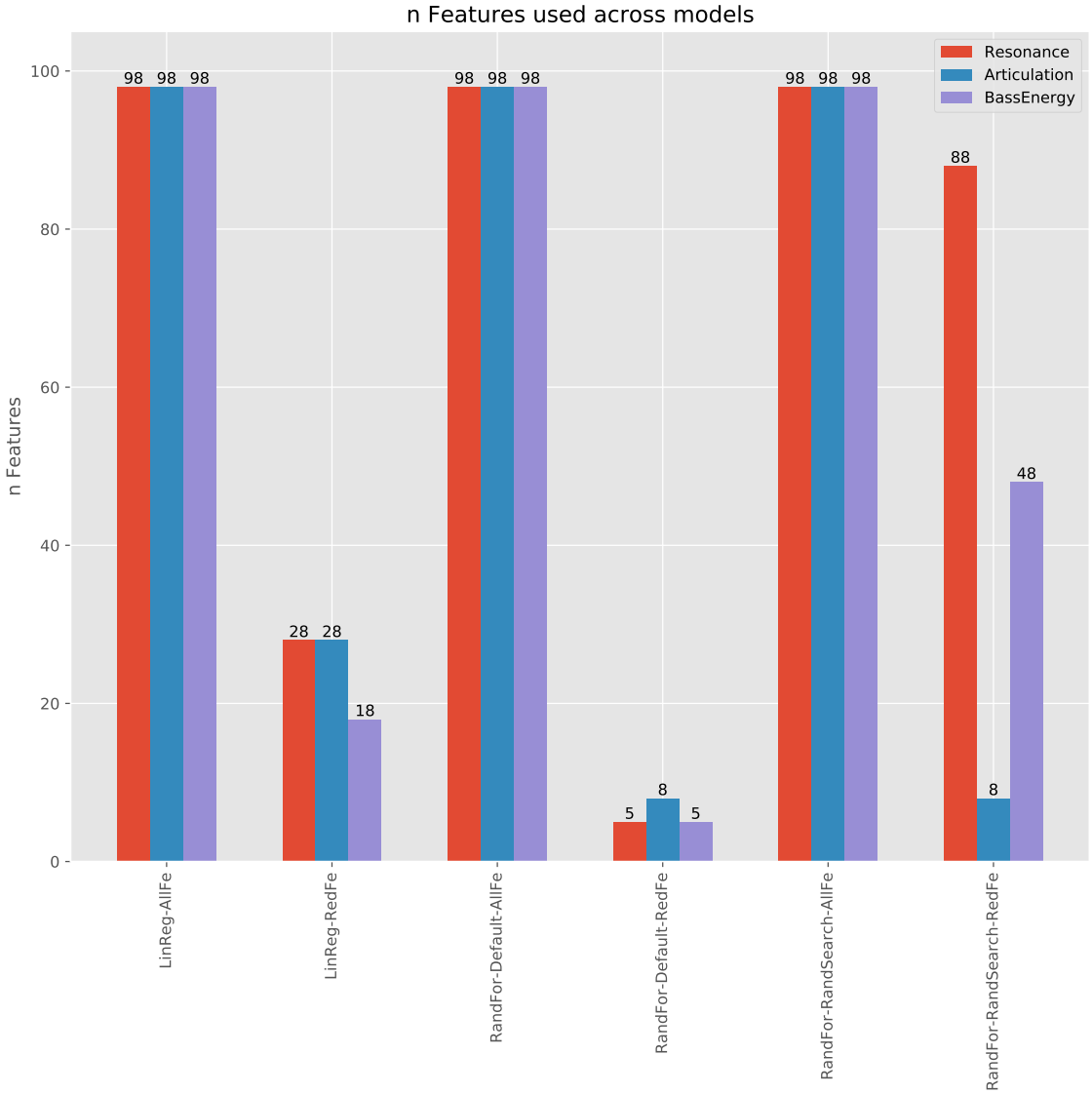


Figure 6.8: Number of features (n) used in each model, showing both cases for feature selection being used and no feature selection applied - Where the y-axis denotes the number of features used across each model and the corresponding modelled attribute (x-axis and colour respectively), with the labelled number of features shown on top of each bar.

6.3 Investigating the perceptual models

6.3.1 Choosing the best model

Section 6.2, showed that across different implementations of both linear regression and random forest, neither provided any benefit in error or explained variance. However, one clear advantage was random forest was more robust using fewer features. The metric to choose the ideal model for investigation will therefore be the random forest model where the best parameters are automatically found (no hyper parameter optimisation applied). This is due to the low number of features for both Articulation and Resonance (see Figure 6.8) and reasonable R^2 score (see Figure 6.7.

Therefore, throughout this section, the models used for Resonance and Articulation will be the random forest with feature selection applied and no hyper parameter tuning (referred to as "RandFor-Default-RedFe" in the figures throughout section 6.2). Furthermore, it should be noted that due to the lack in accuracy in Bass Energy as a metric and the failure to model adequately, this investigation does not include Bass Energy.

6.3.2 Important features in describing bass quality attributes

The reduced feature set used in the perceptual model for Resonance and Articulation are found in table 6.6.

Resonance	**Articulation**
EDT.126	Decay.63
EDT.251	Decay.126
ExceedThreshold.63	EDT.126
ExceedThreshold.251	EDT.251
TemporalCentroid.126	ExceedThreshold.63
-	Exceedthreshold.126
-	ExceedThreshold.251
-	TemporalCentroid.251

Table 6.6: Remaining features that are used in the best model chosen in the RFECV selection process with no hyper parameter tuning applied for Resonance and Articulation respectively.

An interesting observation from both metrics is that they are only temporal based features that are included in the model.

A more detailed way of investigating the features is by using the feature importance of the random forest approach. The feature importance's are shown in Figures 6.9 and 6.10 respectively.

An astounding find from both metrics is that the feature importance relies heavily on Early decay time in the 250Hz band, where most other features have a very low permutation importance score.

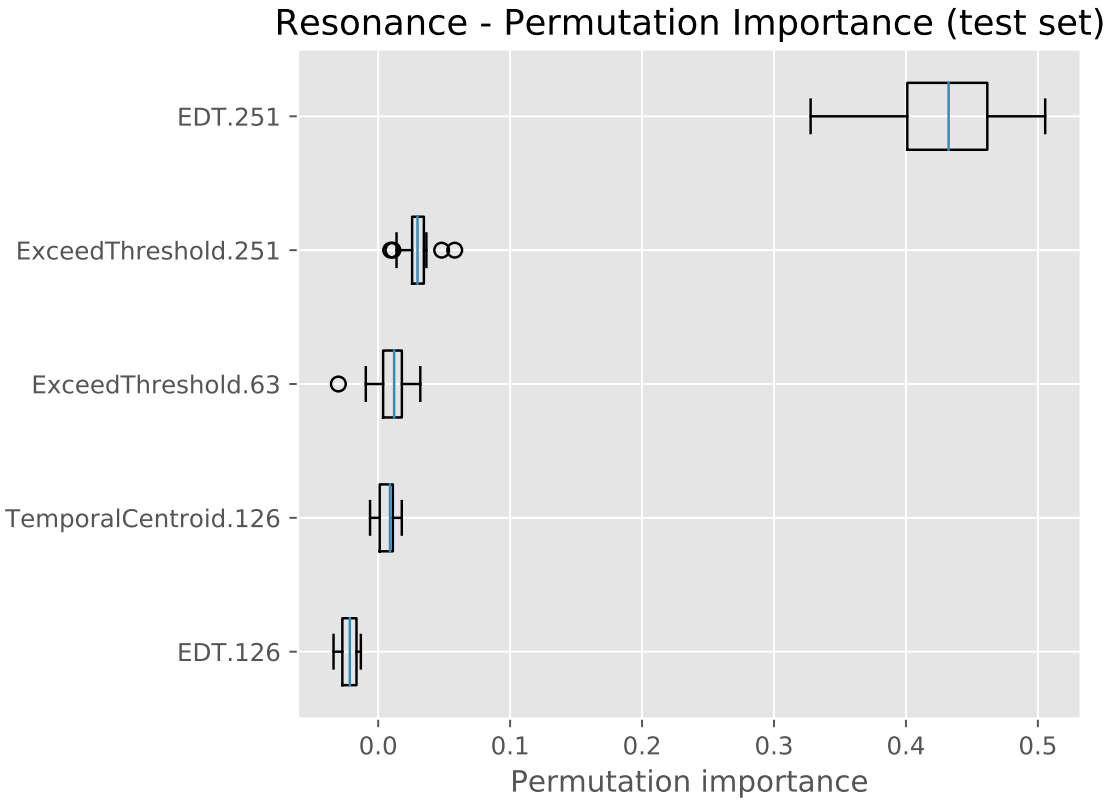


Figure 6.9: Random forest permutation importance for the reduced feature Resonance model shown for the validation/test data set.

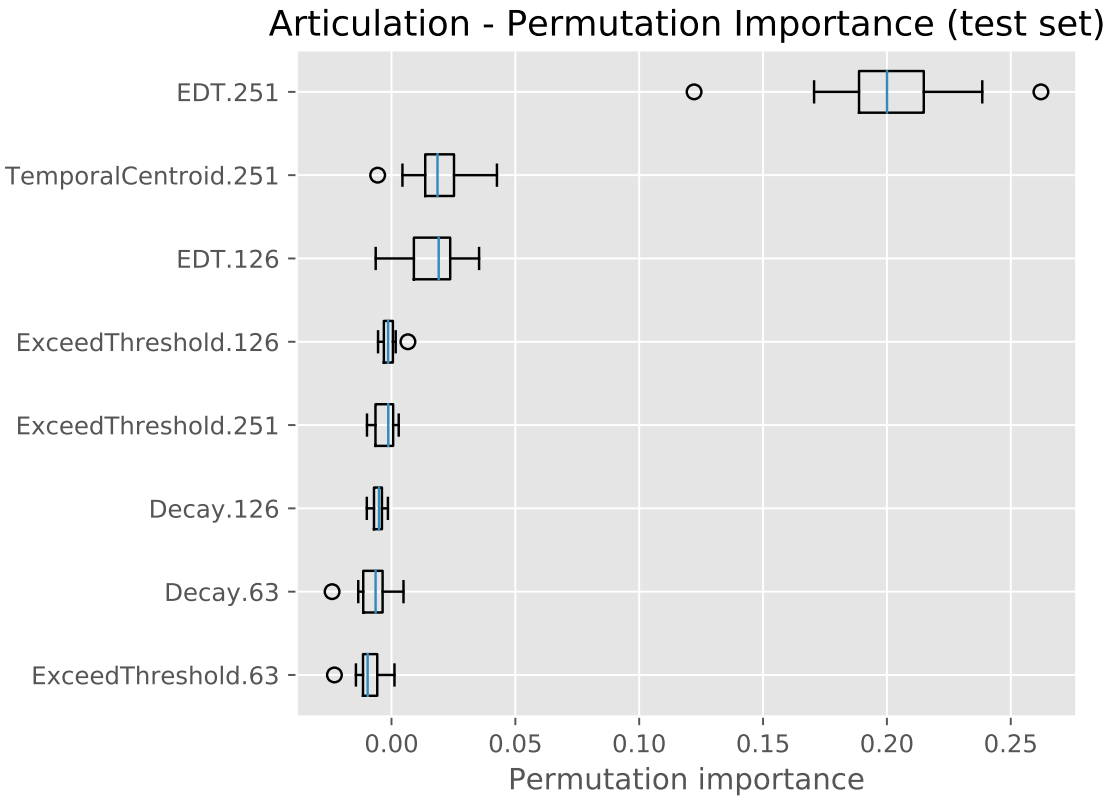


Figure 6.10: Random forest permutation importance for the reduced feature Articulation model shown for the validation/test data set.

6.3.3 Testing the generalisation of the perceptual models

For this section, the results of a generalisation test are presented to understand how the model will generalise onto new, unseen environments. Presented is a test of synthesised room volumes using the auralisation and room generation methodology described in Section 4, where room dimensions were chosen arbitrarily then adjusted to move the room volume to a general range (e.g. 500 or 1500 m^3).

Using the models discussed in Section 6.3.1, Figures 6.11 and 6.12 show the model being tested on unknown environments, where rooms were chosen to display volumes outside the range of test. Each room was then given an acoustic absorption coefficient (equal across each frequency band) in the range $0 < \alpha < 1$ to provide an insight into how both room volumes and acoustic absorption are interlinked through greater sampling.

An interesting effect is found for both metrics, where a subtle floor and ceiling effect is found across rooms. For Resonance, an increase in damping will no longer correspond with a decreasing Resonance score and at the lower end of damping, a decrease in damping will correspond with no further increase in Resonance score. However, an observation into the effect of room Volume can be found here, where the room volume changes the shape of this damping/attribute relationship. This is particularly noticeable in the extreme room cases of $60m^3$ and $13200m^3$, where the room volume changes this relationship somewhat. Observed in Resonance is an effect where much higher absorption is required to achieve a low score in the large room and much less absorption is required to achieve low resonance. The same is

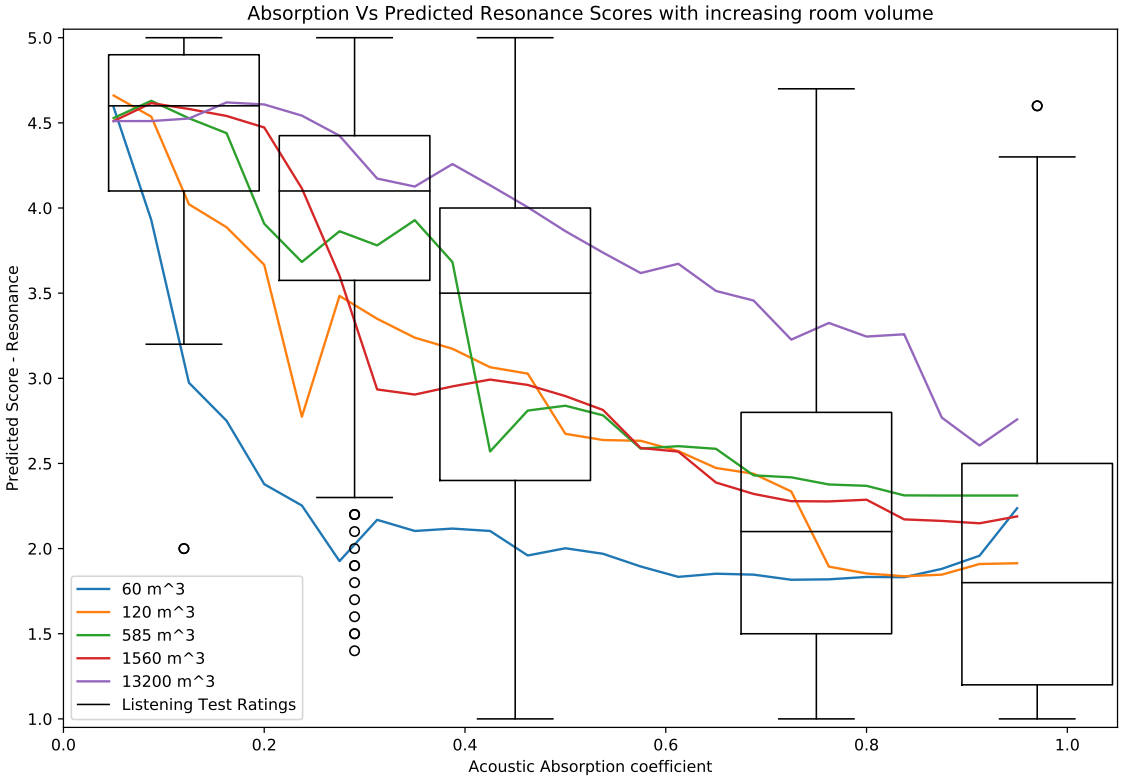


Figure 6.11: Generalisation test of the Resonance predictor on novel rooms and kick drum - Where coloured lines denote the room volume in test, x-axis shows the absorption coefficient used at the room boundaries and box plots show the distribution of resonance scores across different absorption coefficients (where all room volumes are collapsed into this variance) to illustrate the deviation of generalised scores to actual subjective ratings.

found for Articulation, however the converse is true.

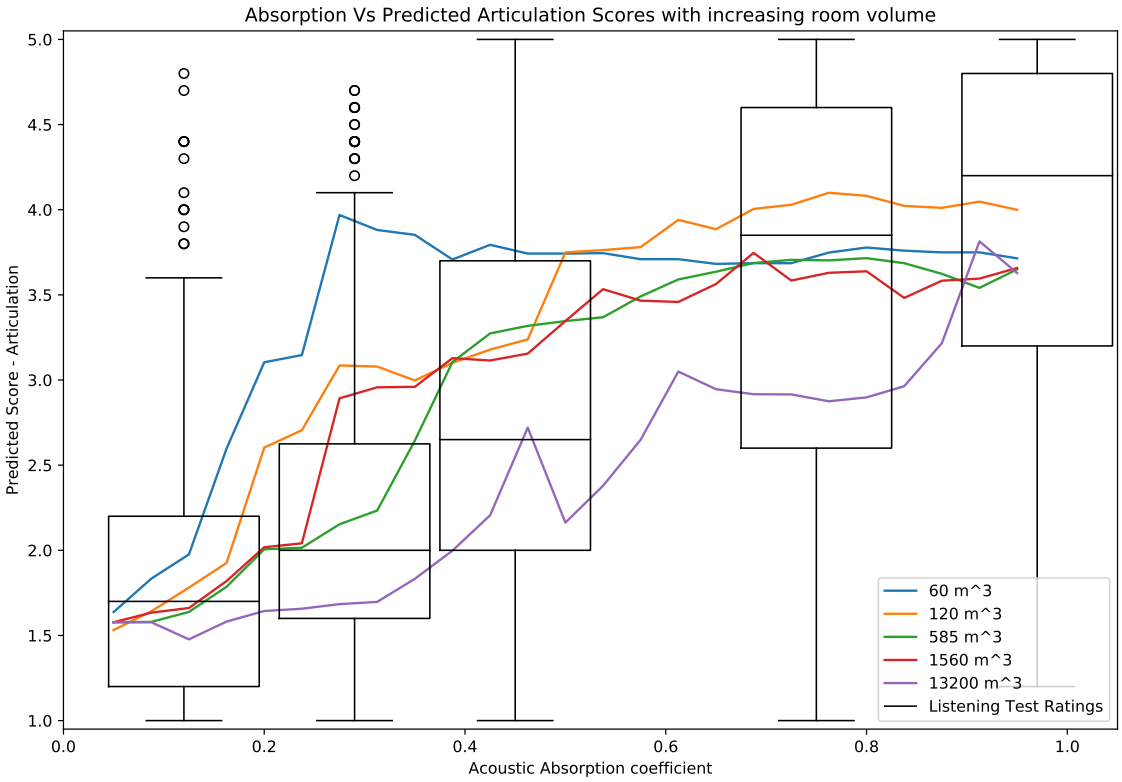


Figure 6.12: Generalisation test of the Articulation predictor on novel rooms and kick drum - Where coloured lines denote the room volume in test, x-axis shows the absorption coefficient used at the room boundaries and box plots show the distribution of Articulation scores across different absorption coefficients (where all room volumes are collapsed into this variance) to illustrate the deviation of generalised scores to actual subjective ratings.

Chapter 7

Discussion

7.1 Discussion of perceptual attributes

7.1.1 Suitability of attributes

To first reiterate the objective of this work, perceptual attributes that were defined in (Wankling et al., 2012) were used to provide a greater understanding of the problem of the perceived quality of reproduced kick drums due to room acoustics, with a specific focus on percussive instruments. This section aims to review the suitability of the attributes when applied to low frequency room acoustics with percussive kick drums.

All attributes were found to be significant when compared against the room acoustic variables (see section 6.1.2), which verifies that the chosen bass quality attributes were useful in describing the effect of room acoustics on the perceived reproduction of sound and thus are suitable for the chosen application.

When considering the modelling of attributes, both Resonance and Articulation were found to be predicted with similar, high levels of accuracy from simplified reduced feature models. Bass Energy however, was not able to be predicted with any accuracy. This result is reflected on closer inspection of the attributes and the interaction with room acoustic variables, where Bass energy has a much weaker interaction than that of Articulation and Resonance (Tables 6.4, 6.3, 6.2 respectively).

One contributing factor to this effect can be seen in the distributions of the attribute scores as given by the test subjects in Figure 6.1. The density plot describes a bi-modal distribution of scores for both Articulation and Resonance around the upper and lower bounds of the scales. Whereas Bass Energy has a much stronger density around the upper limit of the scale, indicating that there was a skewed distribution towards high Bass Energy stimuli presented in the test.

The clustering of high/low Articulation/Resonance can be partially attributed to the cases found in Section 7.1.3, which finds many examples of subjects rating with high Articulation correlated with low Resonance scores. This effect can also be biased by the use of labelling the extremes of the scale displayed on the test interface and providing definitions of the extremes and no definition of the centre of the scale, thus inducing centring bias around the scale limits (Zieliński et al., 2008). However, although this effect may be present, it was implemented to aid listeners in making full use of the scale, which is shown among most participants in Figure 7.4.

Another reason for the lack in accuracy for Bass Energy, may be due to the

very definition of Bass Energy. (Wankling et al., 2012) recommended that Bass Strength and Bass Depth should be combined to create Bass Energy, thus collapsing 2 attributes into a single metric and did not provide a newly formed Bass Energy definition or defined the new extremes of this scale. However, in the presented work, the newly formed attributes were not validated through further subjective testing to confirm the two could be formed as a singular metric, as it was only suggested due to the highly correlating observations in the test. Hence, the result found in this study concluded in many participants reporting that Bass Energy was a confusing metric to rate. Where, a particular case arose where the high frequency would be loud initially due to the attack of the kick drum, however would decay at a much faster rate than the room mode, leading to a dilemma for the participant to question when to define Bass Energy? However, this case is very specific and does not apply to the scenario of full broadband music, where appropriate extension of frequency and loudness in low and high frequency are present.

This highlights another potential area that may contribute to the lack of accuracy in Bass Energy, which is due to the test material and presentation of stimuli. One of the collapsed attributes of the bass energy definition (Bass Strength), refers to the ratio between high and low frequency loudness. However, due to the nature of the listening test, there was little variance in the low/high ratio in energy due to the low frequency kick auralisation and the only variance in high frequency depended on the kick sample. While control over the frequency range was intentional by design to not allow for high frequency content to confound and bias the low frequency, it may have limited the scope of test cases where Bass Energy is an important factor. Hence, it is not applicable to deny the usefulness or validity of Bass Energy as a descriptor in the case of full range broadband music. However the results found in this study suggest that Bass Energy is not a suitable quality attribute for assessing the effect of room acoustics on a percussive instrument in isolation.

7.1.2 Comparison with previous work

The key comparison to be made from this work will be to the origin of the perceptual descriptor attributes in (Wankling et al., 2012), referred to as the previous study throughout this section. However, to broaden the scope and use of suitability of the perceptual attributes, there were many subtle and broad changes to the methodology.

To first reiterate the methodology differences, the previous work was focused on full range music rather than this scope of single percussive instrument excitation. The room volumes under test ranged from $20m^3$ to $1,000m^3$, whereas the room volumes used in this work ranged from $\approx 125m^3$ to $\approx 6,500m^3$. Previously multiple source positions were used throughout each room, whereas this work only made use of a fixed source/receiver position. In the previous study, the modal decay values were under test were set flat across frequency from 0.1 to 0.7 seconds, whereas in this study, modal decay was not specified, instead acoustic absorption coefficients were used ranging from 0.12 to 0.97 to control the decay.

It is therefore important to keep in mind of these differences, which allude to a reduced input stimuli and source/receiver position; an expansion of both the room volumes and decay times; and reduced scope of the excitation of the room through a single percussive instrument.

The main focus of discussion in the previous study was investigating the effect of room volume and decay time on the quality attributes. Therefore, for consistency, the aforementioned discussions outlined throughout this section and the relationships between Articulation and Resonance are discussed in Section 7.1.3.

Effect of room volume

The previous study found that there was a main effect of room volume across all attributes, however there was no general progressive trend found between any of the attributes, thus contesting the previous belief that room volume was tied to perceived quality.

Similarly, this study found that there was a significant interaction between volume and each of the attributes. However, a progressive trend between the median attribute scores and the room volume was observed, this may be in part due to the greater scope of decay values. As observed in Section 6.1.3, a significant second order interaction between room volumes and absorption coefficient was found for both Resonance and Articulation and when investigating the interactions shown in Figures 6.3 and 6.4. This stronger effect due to decay times is shown due to the much higher effect size due to the room damping (Tables 6.2, 6.3).

Effect of decay time

To further the discussion of the effect of decay times, the previous study found that decay times showed a significant interaction to both resonance and articulation, however Bass depth (referred to Bass Energy in this study) was not. Furthermore, a trend of increasing Resonance and a decreasing Articulation was found for an increase in modal decay time.

These findings are congruent with the results of this study, where the largest effect size for both Resonance and Articulation was found to be acoustic absorption (Table 6.3, 6.2). Moreover, the modelling results and in particular the permutation importance, found that the most important feature for modelling the Resonance and Articulation attributes was that of Early decay time in the 250Hz octave band for both (Figures 6.9 & 6.10).

7.1.3 Relationship between Resonance and Articulation

Throughout the results shown thus far, a clear trend has been observed in both Resonance and Articulation. The results in Section 6.1 show an inverse trend between the two attributes, where high resonant cases relate to low articulation and the inverse is also prevalent. To investigate the relationship, first Articulation and Resonance scores are compared against their relevant significant independent variables as described in Tables 6.2 and 6.3.

The first look into how the attributes relate can be seen in Figure 7.1, which illustrates the the relationship between Articulation and resonance split across each room.

Figure 7.1 shows a clear relationship, where each room has a highly negative correlation between Articulation and Resonance scores with statistical significance. Through each room, a similar distribution is found ranging across high and low of both attributes, however there is some variance that is unclear where high Resonance

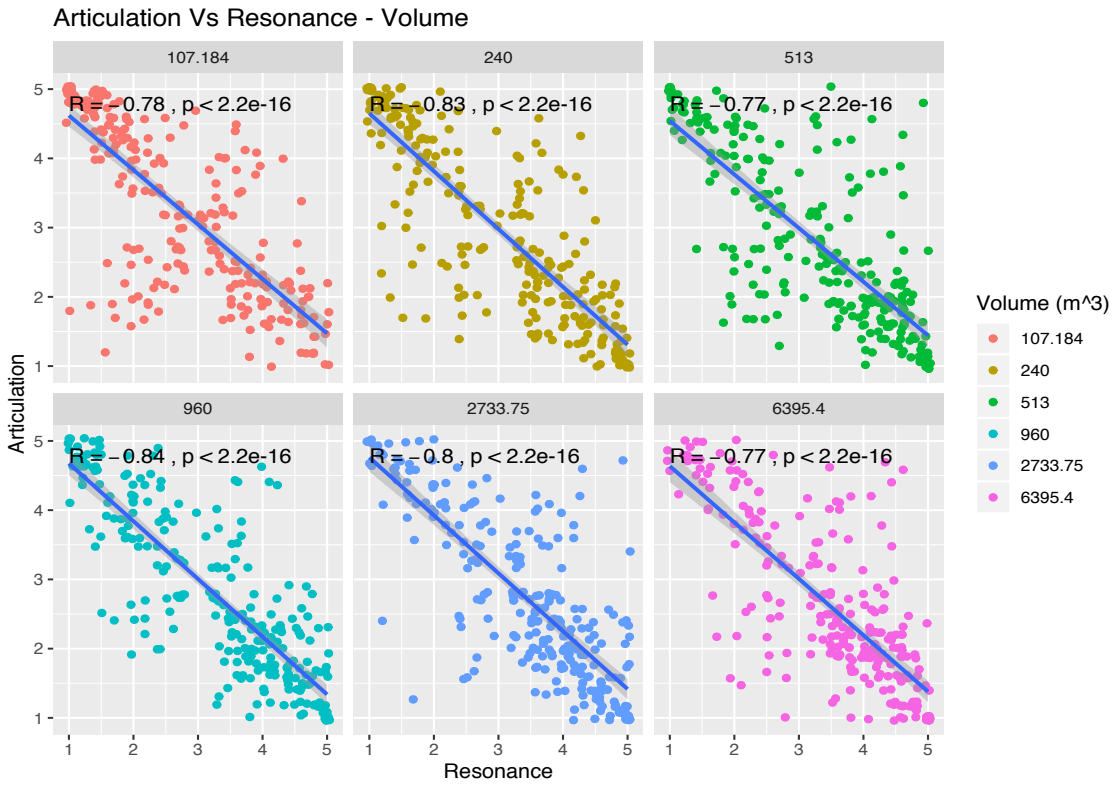


Figure 7.1: Comparing Resonance and Articulation scores split across different Volumes (subplots) where the blue line indicates the line of best fit through the data and error margin denotes the 95% confidence interval, with correlation and significance shown above.

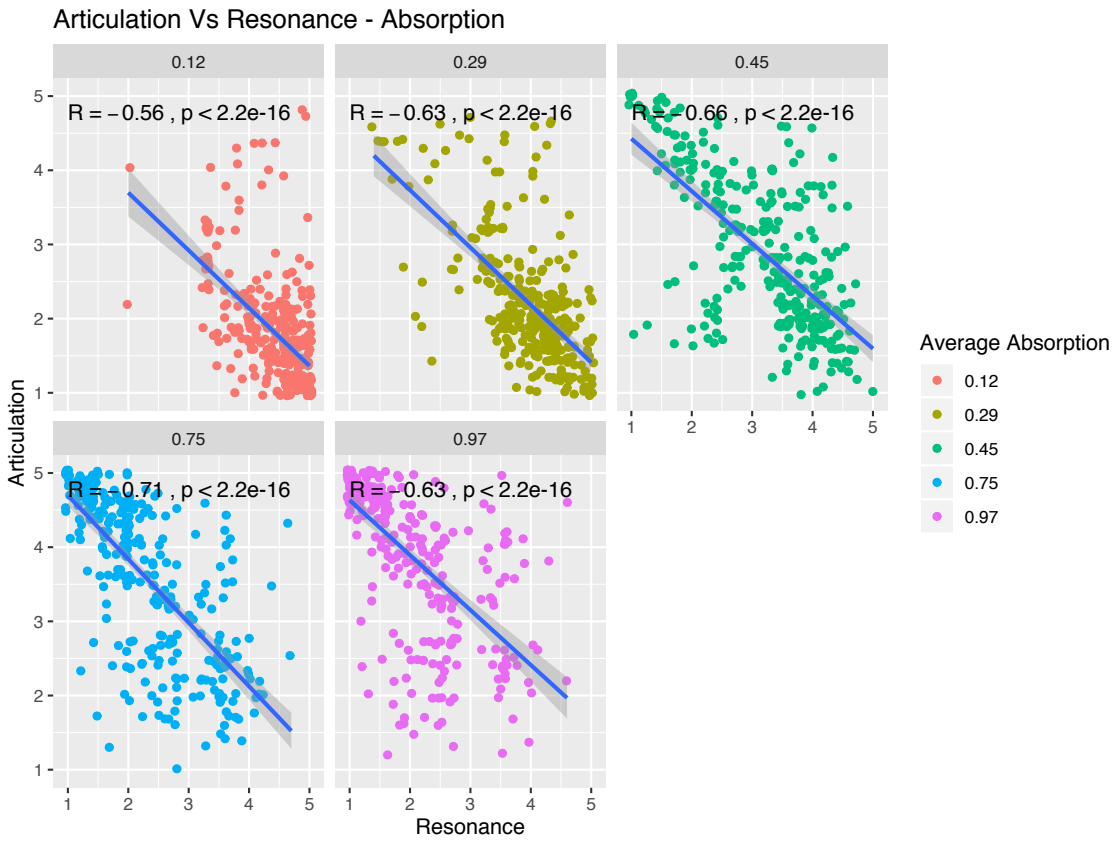


Figure 7.2: Comparing Resonance and Articulation scores split across different acoustic absorption coefficients (subplots) where the blue line indicates the line of best fit through the data and error margin denotes the 95% confidence interval, with correlation and significance shown above.

and high Articulation are both observed (and their inverse counterparts). Control and understanding of these outliers is of great importance due to the modelling process of the perceptual attributes, where ratings are given to the model as ground truths and incorrect labelling through unreliable data will sacrifice model accuracy. Therefore, it is key to differentiate through this section to understand whether the outliers are actual cases of perception that are valid, or are ratings that may not reflect on the actual stimuli that they are rated against.

Again, significant correlations are found now between Resonance and Articulation in Figure 7.2 which show where clustering occurs when splitting the scores by the average acoustic absorption coefficient. The results show that low damping corresponding to High Resonance, low Articulation whereas higher damping corresponds to low Resonance and High Articulation. It should be noted that the lower values for R, indicating a lower correlation between the two attributes is mostly attributed to both the clustering (non-linear trend) and the higher variance around the clusters.

A further split of the ratings for each sample in Figure 7.3 shows a similar trend to volume, where a linear and significant relationship is found between Resonance and Articulation. However, a similar spread of the variance is found of cases for high articulation and high variance.

Therefore, the observed off-trend outliers are not a product of the independent

Figure 7.3: Comparing Resonance and Articulation scores split across different Samples (subplots) where the blue line indicates the line of best fit through the data and error margin denotes the 95% confidence interval, with correlation and significance shown above.

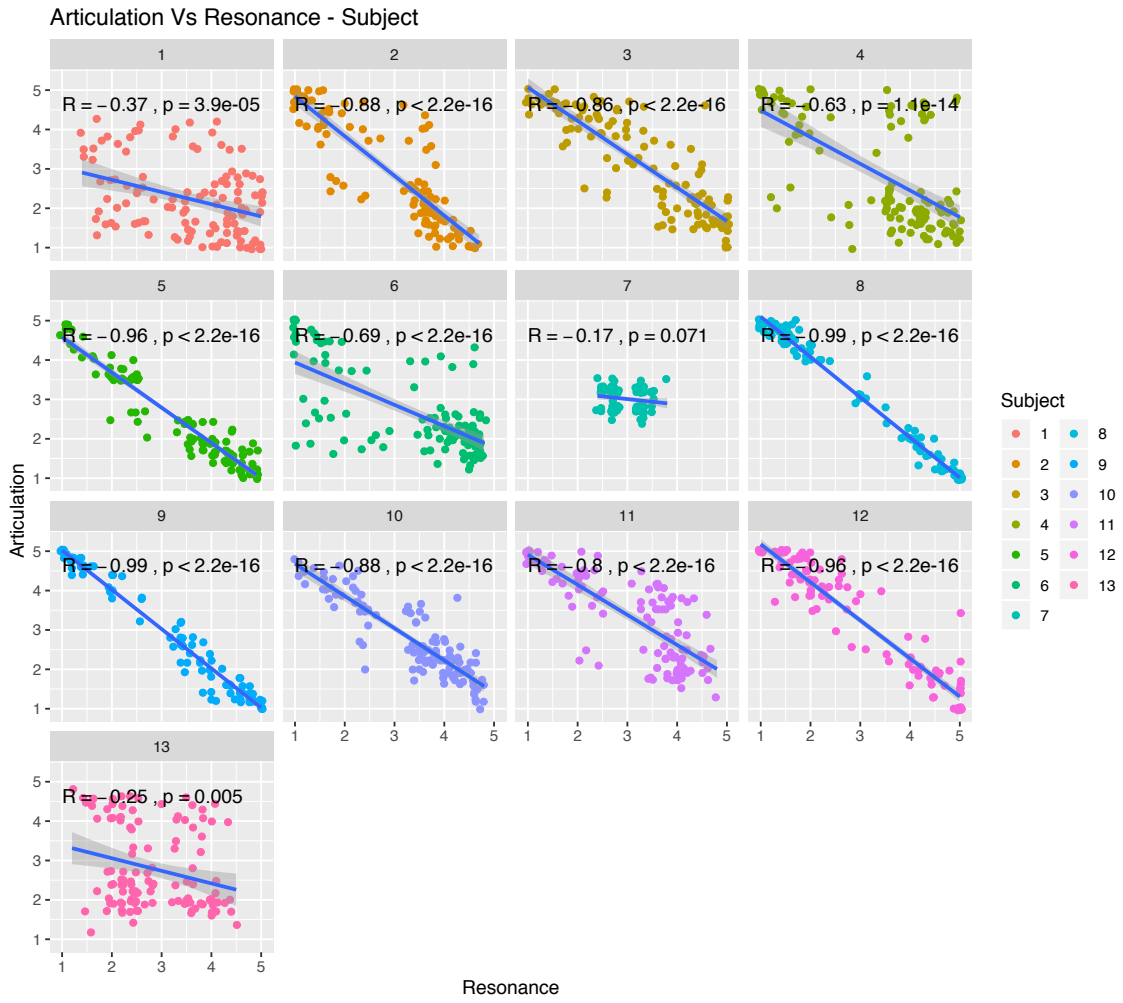


Figure 7.4: Comparing Resonance and Articulation scores split across individual participants (shown in subplots) where the blue line indicates the line of best fit through the data and error margin denotes the 95% confidence interval, with correlation and significance shown above.

variables in test, as each variable in test shows the same significant correlation and outliers between Articulation and Resonance. Finally a look at Figure 7.4 shows the rating relationships exhibited in individual participants.

This shows an interesting trend where most participants in test (10/13) showed a significant and strong relationship ($|R| > 0.6$), whereas subjects 1, 7 and 13 showed a low trend with no significance (excluding 13 as it is on the confidence limit). Therefore, there is an obvious agreement between most subjects in their rating of high and low resonance. While these graphs do not show if participants agreed on samples down to an individual level, it does show a clear and significant negative trend between their ratings for Resonance and Articulation.

Therefore, the variance can be somewhat reduced by exclusion of removal of participants 1, 7 and 13, of which 7 & 13 were naive listeners. Removal of non-significant subjects was therefore done before modelling to aid in the accuracy in the perceptual models. It must be noted that removing participants from an already small pool of listeners risks the reliability of potential findings, however the large effect sizes seen in the ANOVA results (Tables 6.2, 6.3, 6.4) suggest that the

possibility of observing a type 1 error in this instance is low.

While a strong correlation is shown between Resonance and Articulation, this result differs from the observations in the previous work (Wankling et al., 2012). The previous study found a much weaker correlation between Resonance and Articulation, suggesting the two were independent scales. In the previous study, the weak correlation between Articulation and Resonance was suggested to be attributed by cases where "an individual resonance may be heard in an otherwise tight sample". This may be due to complexities in the presentation in full music samples, where a more complex broadband signal may blur the relationship of perceived Resonance and Articulation, however on testing isolated kick drums, the resonant and articulated characteristics of the samples are much more apparent in a clearly defined envelope.

Furthermore, although a clear trend is shown throughout the previous results, it does not explain whether Articulation and Resonance are explaining the same precept. To analyse this, a look into the feature importance from the perceptual models can show an insight into the relationship of the two attributes. Both Resonance and Articulation (Figures 6.9, 6.10) show that Early Decay Time in the 250Hz octave band is the most important features in modelling the attributes. Therefore, the results from this listening experiment show that there is a clear relationship in both the attributes and their underlying perception.

7.2 Forming a perceptual model

7.2.1 Relationship between audio features and perceptual attributes

It is important to review the features used in modelling the perceptual attributes, this will aid in checking any biases that may arise in due to potential disproportionate feature inclusion or exclusion from different groups that the model was able to pool from. Figure 7.5 shows the breakdown of each feature by the categorical type of feature (i.e. Temporal, Spectral etc..), the frequency range of interest and finally the origin of the feature, where the bespoke category are features that were created for the use of this project.

First, to comment on the range of features, there is an even split of features across the different categories, where the largest amount of features are the ones calculated from the impulse response and the fewest features correspond to the dynamics of the signal.

However, when turning the attention to the frequency range of interest, a skewed distribution is found between band limited (i.e. octave band features) and broad band and low frequency ($< 250Hz$) features. This is due to the duplication of many features when performing octave band filtering. An example of this is found in decay, which is computed across each octave band resulting in several features describing a similar characteristic, however with an emphasis on a specific frequency region. A caveat with this approach is highly correlated features or multi-collinearity, where many of the band limited features can describe similar variance in the data. This must be taken into consideration, particularly when reviewing the selection of features due to inflated variance and the potential to incorrectly identify relevant features (Dormann et al., 2013).

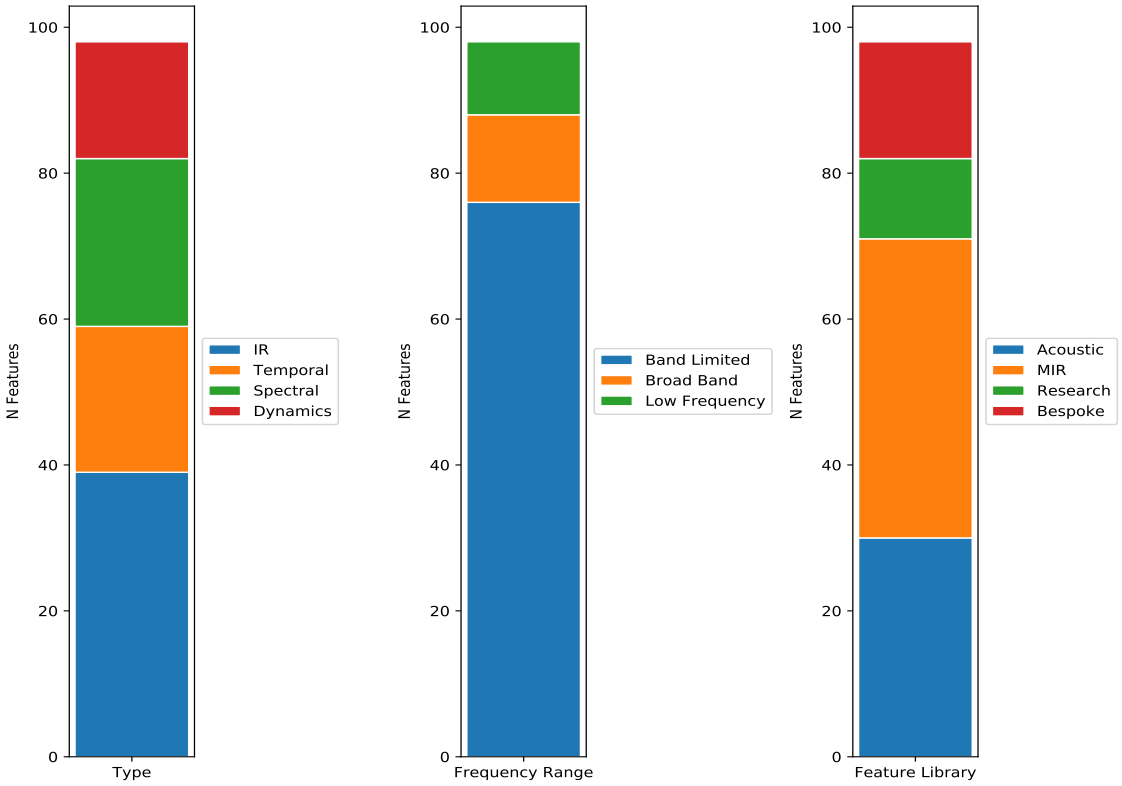


Figure 7.5: Stacked bar plot of the available feature pool in the modelling process - where features are split and broken down by type, frequency range and the source of feature origin.

Finally, the origin of each feature is analysed to represent a potential bias in the amount of features taken from varying sources or fields. The analysis shows that an even distribution of features is available to the model, where the largest feature pool comes from the MIR based features. Meanwhile the smallest feature pool used was from previous research, which were primarily the figures of merits defined in (Stephenson, 2012).

With the understanding of the source and characteristics of features used when modelling the perceptual attributes, the attention must now be turned to the feature importance shown in Figure 6.9. The outcome of the Random Forest modelling shows that Early Decay Time at 250Hz is by far the most important feature alongside Exceed Threshold and Temporal Centroid. When comparing these reduced features to the full feature set, a breakdown of the reduced set can be seen in Figure 7.6. The reduced feature space consists primarily of band limited impulse response characteristics, that are derived from a mixture of acoustic, MIR and bespoke features created in this work.

To compare the findings with previous research, there have been attempts at making use of features to try and achieve ideal room responses such as (T. Cox & D'Antonio, 2001), which explores the use of a cost function to optimise a room response through changing room dimensions to achieve a flatter low frequency response. Furthermore, (Stephenson, 2012) describes audio features (some of which were made available to this model in this study) that describes deviations from an idealised room response from lines of best fit.

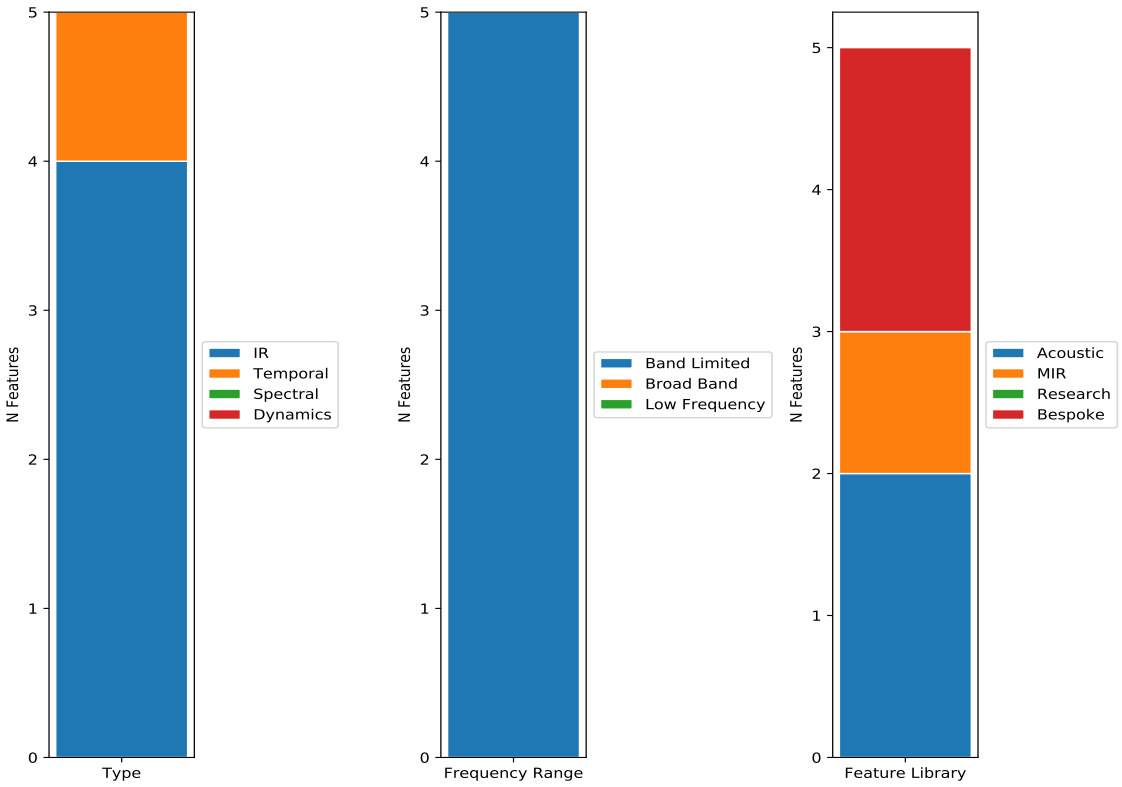


Figure 7.6: Stacked bar plot of the reduced feature set from the reduced feature Resonance predictor (feature list shown in Table 6.6) - where features are split and broken down by type, frequency range and the source of feature origin.

There has been a wide overemphasis on prior research on focusing on the frequency domain and so called "flattening the response", where the temporal nature of the room response is not taken into consideration. However, when using features from research and MIR features that describe spectral characteristics of the room response, no spectral features were included as part of the reduced feature model in this work. The best model found for Articulation and Resonance found that decay and temporal metrics were the most important by far when relating low frequency quality. It could be argued that some spectral information is retained in these metrics through octave band limiting of features, such as 250Hz resulting the most important band for early decay time which corresponds to the most sensitive band in the modal thresholds (B. M. Fazenda et al., 2015). Although this may not be conclusive as discussed earlier, due to potential co-linearity between octave band limited features, it may not be reliable to conclude that one octave band is more important than another.

Again, this result furthers the importance of moving away from the concept of the negative audible effects being a primarily frequency domain based problem and emphasises the importance of modal decay and the audibility of modes when considering the impact on perceived sound quality.

7.3 Predicting the effect of modal density on perceived Resonance

Now that a prediction model has been formed and calibrated (see section 6.3) for the Resonance attribute, it may be used as an "expert listener" in novel cases. The model may be used to enquire the relationship between room acoustic parameters and their influence on the perceived Resonance for an auralised kick drum.

Previous research has often highlighted the importance of modal density as a criteria when designing a room for the purpose of high quality sound reproduction. Where modal density is used (by means of the Schroeder Frequency) to define the transition between "modal" and "diffuse" sound fields.

The key argument for which, assumes that for smaller rooms, the Schroeder frequency falls into the audible range and therefore the modal region will have more of an adverse effect than their large room counterparts, where a diffuse field is assumed due to the Schroeder frequency being below the lowest audible frequency. (Kutruff, 2009) goes as far as to say that in large halls, there is no reason to evaluate eigenfrequencies (modal resonances) due to the Schroeder Frequency residing around the lower audible limit. Many of the previous works assumed that modal density was a useful means of control modal behaviour by moving the transition frequency below the audible range through room dimensions, proposing a room volume criterion defined by the Schroeder frequency (Kutruff, 2009).

However, throughout previous research in the field of low frequency room acoustics, there has been an emphasis on the investigation of low frequencies in small rooms, where small control rooms and listening rooms often fit the "Acoustically Small" archetype (see (Angus, 1997)). Among the previous work, it is often believed that problems in low frequency issues arise from low modal density (T. Cox & D'Antonio, 2001). Where the main contributing factor to modal density is that of room dimensions and room volume, the former being a major factor in much earlier

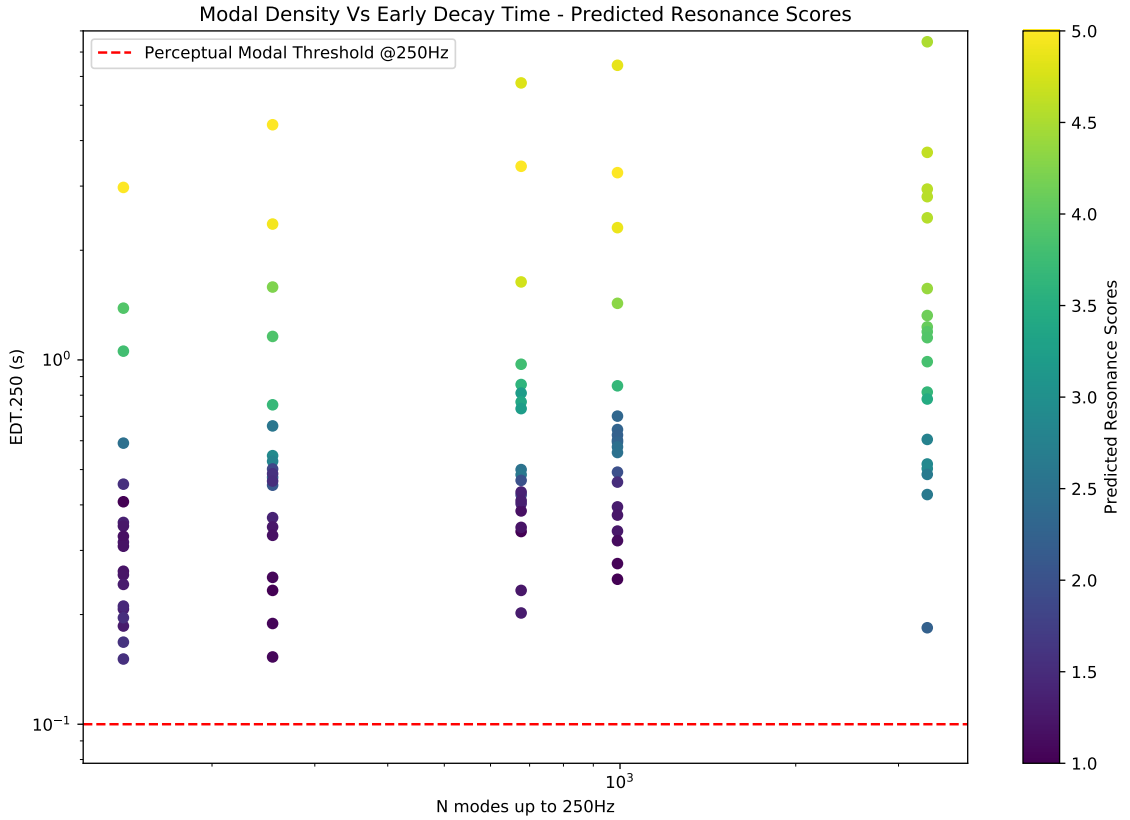


Figure 7.7: Use of the Resonance predictor to compare the effect of increasing modal density (described by the number of modes up to 250Hz x-axis) and Early decay time measured in the 250Hz octave band - Where the predicted Resonance score is coloured and shown in the colour map to the right of the plot.

research, such as the famous Bolt's curve (Bolt, 1946).

Therefore, assuming this relationship between increasing modal density and perceptual quality to be true, then larger room volumes must therefore contribute to higher levels of audio quality. However, the findings from this work suggest otherwise, where larger room volumes correlated with a decrease in quality, where the strongest effect was found in controlling the acoustic absorption of the room and that the quality attributes relied heavily on the decay time of the modes. These findings are congruent with that of (Wankling & Fazenda, 2009), which found that there was no correlation between increasing modal density and found a much stronger influence in decay time regardless of modal density.

Using the perceptual model defined in Section 6.3.1 and verification that the model can generalise well onto novel unseen rooms in Figure 6.11, Resonance scores can be predicted for new unseen kick drum room combinations. The rooms generated for the test in Figure 6.11, are then used to show the effect of Modal Density in Figure 7.7, where early decay time is used as a metric to represent the modal decay due to the high level of feature importance found in Figure 6.9.

Figure 7.7 shows that there is little effect of modal density, where increasing the density exponentionally does not have any effect on the perceived resonance scores, whereas there is a clear relationship between the perceived resonance and Early decay time exhibited. However, in the largest possible room with the highest modal

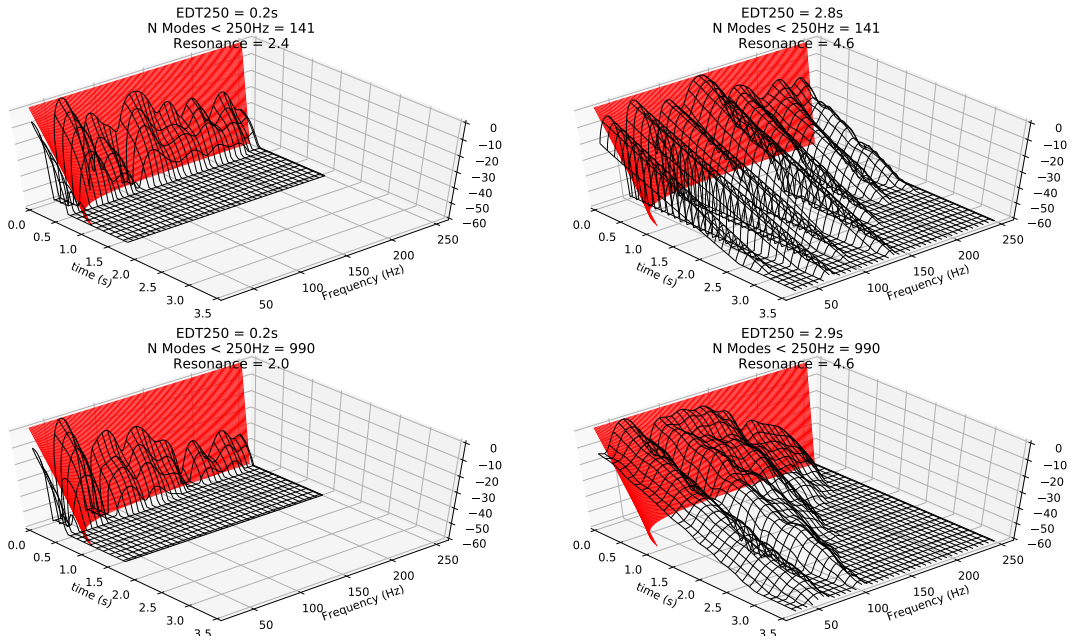


Figure 7.8: Convolved kick drum with room waterfall plot showing 4 rooms which either share the same Early Decay Time or Modal Density, represented as the number of modes under 250Hz. The Perceptual Modal Thresholds have been overlaid as a linear decay marked in red.

density, the only difference found is the perceived resonance is actually higher at similar decay times. It must be taken into consideration to understand that while this may reflect on the acoustic environment, this room was chosen to test a volume outside those used in the listening test. Therefore, the model is generalising on an unseen environment that is outside of the scope of the known ratings and feature relationships and may not be fully reliable to define correlations here.

To investigate the effect of modal density further, an example of two extremes of room volumes are chosen to represent low and high modal density. Furthermore, early decay time is controlled using the acoustic absorption coefficients to match the modal decay times in the low and high modal density cases. To illustrate the effects of modal decay and modal density, a waterfall plot is shown in Figure 7.8 of a convolved kick drum using the $60m^3$ and $1560m^3$ rooms from Figure 6.11. Values for the perceptual modal threshold have been imposed onto the waterfall plots to illustrate any exceedance and therefore audible modal effects.

Shown in 7.8 in the short decay, low modal density case; it can be seen that there are few modes present and are only just exceeding the perceptual modal thresholds. When moving to a higher modal density, but matching modal decay time; the waterfall plot structure is almost identical despite being in a room volume over 20 times larger. Therefore, it seems to show than when sufficiently damped, the room has little effect on the perceived resonance of the auralised kick drum. However, where modal density comes into a factor is at longer decay times, where moving from low damping, low modal density, to low damping high modal density. It can be seen that while there is less definition between individual modes, the long decay time of each mode may be the defining factor as to why higher modal density may not be a useful factor in achieving better perceived quality and may arise as a hindrance in reproduction of low frequency sound in large volumes.

To summarise, it was first discussed that previous research hypothesised that in order to mitigate the effects of Modal sound fields, it was important to increase the modal density sufficiently to obtain a diffuse environment and therefore the problematic effects of a modal sound field are mitigated. This work also refutes the previous belief that perception of "Resonant" characteristics of modes is associated with a lower modal density (B. Fazenda & Wankling, 2008). The findings in this work align with the direction of more recent work that suggest that there is little influence in modal effects due to modal density (Wankling & Fazenda, 2009) and is more critically due to the modal decay characteristics (B. M. Fazenda et al., 2015).

7.4 Summary

When investigating the outcome of the listening tests, it was found that Articulation and Resonance were able to be accurately modelled, whereas Bass Energy was not, suggesting that Bass Energy is not of importance for the band limited case of kick drums. While all attributes showed significance with the room acoustic variables, there was a main effect due to the absorption coefficient used in the rooms indicating a main effect of modal decay time.

On comparing this study to previous work, it was first shown that there were differences in the test methodology. While the results here are not directly comparable, the results highlight both the validity of the test and any differences that may arise due to test methodology. The key differences between this study and the methodology used in the previous study, were the input stimuli of full music rather than the use of kick drums; a smaller range of room volumes and modal decay times; and finally, the previous test included several source receiver positions, as opposed to a single source receiver position used in this work.

Similar results were found, where both the decay time and room volume were significant in determining Resonance and Articulation. Further to this, it was found that while room volume contributed to each attribute, there was no linear progression found, whereas this study found a linear progression but most likely linked to the decay times due to the large effect size and overall influence in the decay times.

Finally, results of the attributes were congruent with the previous test, where resonance and articulation were highly correlated, however Bass Energy (Depth and Strength) were somewhat independent of the other attributes. Therefore, it can be concluded that the use of Articulation and Resonance are useful room acoustic descriptive attributes that generalise well onto different applications of reproduced sound in rooms, however Bass Energy may only be applicable when the entire frequency range is excited in the case of broadband music.

Furthermore, results from this work and the previous study showed a strong inverse correlation between Resonance and Articulation. Significance in Articulation and Resonance ratings were strongly correlated across almost all test participants and modelling the perceptual attributes revealed that both attributes relied strongly on the early decay time in the 250Hz octave band. This suggests that the highly inverse correlation is due to Articulation and Resonance being described by the same underlying percept.

When investigating the created perceptual model, a feature breakdown showed that there was an even distribution of features that were made available from different sources, although there was risk of potential overemphasis on band limited

features that could cause co-linearity issues. However, from this large varied pool of potential features, it was found that the random forest model could remain accurate only using 5 features, all of which were temporal based.

Finally, the model was used to investigate how perceptual Resonance is effected by the modal density of a space. The results from this investigation found little to support any effect on modal density, where large room volumes with high modal density predicted identical Resonance scores when the early decay time of the environment was fixed. This result further suggests the importance of controlling the modal decay time and moves away from prior research focused on controlling modal density to account for problematic low frequency behaviour.

Chapter 8

Conclusion

8.1 Recapitulation

To begin, the primary aim for this Thesis is to further the understanding of low frequency perception in rooms by using signal features to model perceptual attributes in the novel case of kick drums in large rooms.

The objectives laid out at the start of this Thesis outlined forming a low frequency room model and auralisation pipeline; conducting a listening test for perceptual attribute ratings; creating a feature extraction pipeline for the auralised kick drums; forming a model of perceptual attributes and finally, to use the perceptual predictors to further understand problematic modal behaviour.

The contributions to knowledge from this research are two-fold. The first contribution describes furthering the available knowledge and data for the perceptual attributes in the case of impulsive instruments, including the scope of large room volumes. Furthermore, the second contribution to knowledge was to form a perceptual model which can accurately describe the perceived Resonance and Articulation of an auralised kick drum.

8.2 Overview

The scope of prior research has resulted in an overemphasis of small listening room environments, where large spaces are often discounted as being sufficiently diffuse to avoid modal problems. However, for the case of live sound engineering, these issues are still apparent and problematic, specifically in the example of correction of low frequency instruments in large acoustic spaces.

Furthermore, a group of perceptual descriptive attributes from previous research defined as Resonance, Articulation and Bass Energy, have proved useful in describing the subjective nature of low frequency room acoustics. However, these quality attributes were understood through their subjective definitions, with little knowledge into the generalisation to novel cases and the underlying objective characteristics which the attributes are built on. With the rise of interest in MIR and audio features paired with Machine learning, it became the objective to model the perceptual attributes using signal features to further understand the underlying perception of perceived bass quality.

Therefore, to form a perceptual model for the attributes, a listening test was conducted to obtain attribute ratings on auralised kick drums in a variety of acoustic

spaces. This was achieved through an auralisation pipeline and low frequency feature extraction and machine learning pipeline to create interpretive perceptual predictors for each attribute.

8.3 Furthering the understanding of the perceptual attributes

The first claim to knowledge made at the start of this research, was to further the available knowledge and data of how perceptual descriptors: Articulation, Resonance and Bass Energy are rated. The findings of the subjective listening test showed for the case of auralised kick drums, Articulation and Resonance are the most important attributes in defining bass quality. The outcome of both the listening test and modelling shows that Articulation and Resonance are highly inversely correlated and the underlying perception relies on the same temporal characteristics. This suggests that for the case of percussive kick drums, Resonance and Articulation are the same precept and describe the inverse of the other.

Moreover, it was discovered for the case of kick drums, the attribute of Bass Energy was not important in describing the perceived bass quality, contrary to the findings in full range music. It is suggested that due to the reliance of low frequency extension and ratio between low/high frequency loudness, that Bass Energy is not sufficient for this band limited case.

Therefore, when considering how a room may affect the perceptual attributes of kick drums, Resonance or Articulation are the most important attributes, whereas Bass Energy is negligible.

8.4 A perceptual model for bass quality attributes

The second claim to knowledge described forming a perceptual model of signal features, which can accurately describe the perceived Resonance and Articulation of an auralised kick drum. It was found that the two attributes relied on similar temporal based features with similar accuracy through a reduced feature Random forest model. It was also found that the feature of most importance for both attributes was early decay time in the 250Hz octave band.

Furthermore, a perceptual model based on signal features now allows for interpretation into other avenues of modal behaviour, through use of the model as an "expert listener" to predict the perceived Resonance of novel environments. In this research, the model was used to investigate the effects of modal density on the perceived resonance in large rooms, which prior research would often assume sufficiently diffuse enough to not exhibit modal effects. However, on investigation, the model showed little influence with regards to modal density, where a fixed decay time between a small and large room volume resulted in the same perceived Resonance. Use of the interpretive model has further highlighted the importance of controlling modal decay for mitigation of modal artefacts and aids in aligning the scope of future research by not focusing on controlling modal density to optimise the perceived quality attributes.

8.5 Future Work

8.5.1 Applications of work

The outcome of this work has highlighted key areas to further understand how modal artefacts affect reproduction of kick drums in rooms. The potential applications that may make use of this work include but are not limited to:

- A stronger emphasis on control of modal decay time in optimising bass quality attributes.
- Less importance of Bass Energy related precepts when considering perceived quality of kick drums in rooms.
- A perceptual predictor for perceived resonance for novel kick drum/room auralisations.

8.5.2 Extension of research

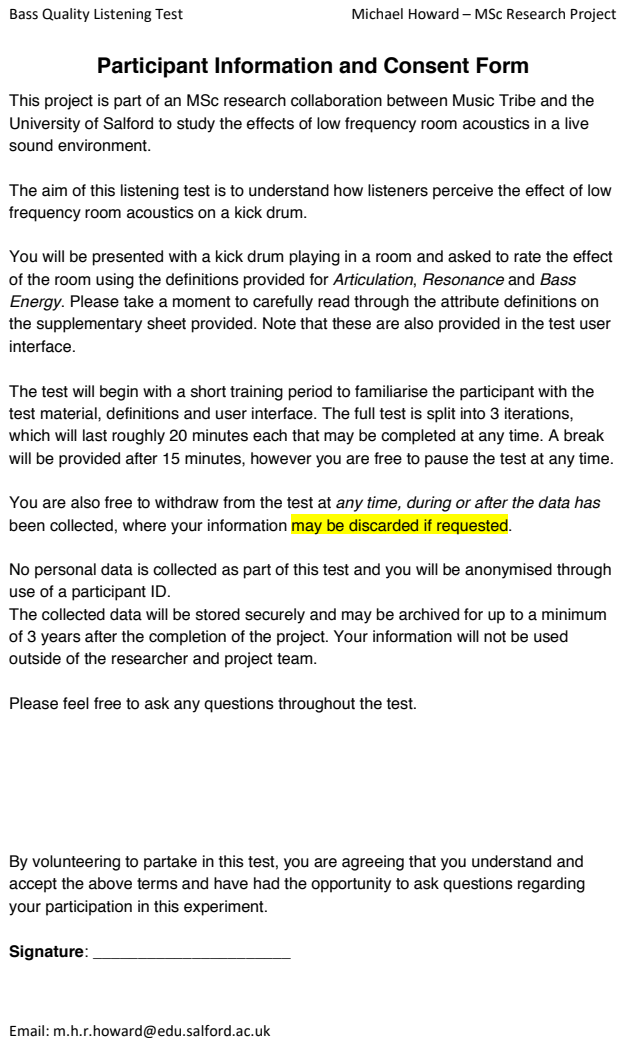
Shown in this work is an example of using the model as an informative tool to question the influence of modal density on the perceived Resonance of an auralised kick drum. A natural extension of this work would be to investigate potential effects of other modal characteristics through use of the model. It should also be noted that while this work is primarily focused on kick drums, extending the research to other low frequency instruments such as bass is a natural progression of understanding the perceptual attributes, as bass excites the low frequency region in a more complex manner.

Another useful extension to this work would be in the form of a "live listener", which may constantly monitor an environment to provide feedback on the perceived Resonance in realtime. Furthermore, the use of a live listener may allow the Resonance predictor to act as an optimisation target for a modal correction tool.

Appendix A

Appendix

A.1 Participant Consent and Information Sheet



Bass Quality Listening Test Michael Howard – MSc Research Project

Participant Information and Consent Form

This project is part of an MSc research collaboration between Music Tribe and the University of Salford to study the effects of low frequency room acoustics in a live sound environment.

The aim of this listening test is to understand how listeners perceive the effect of low frequency room acoustics on a kick drum.

You will be presented with a kick drum playing in a room and asked to rate the effect of the room using the definitions provided for *Articulation*, *Resonance* and *Bass Energy*. Please take a moment to carefully read through the attribute definitions on the supplementary sheet provided. Note that these are also provided in the test user interface.

The test will begin with a short training period to familiarise the participant with the test material, definitions and user interface. The full test is split into 3 iterations, which will last roughly 20 minutes each that may be completed at any time. A break will be provided after 15 minutes, however you are free to pause the test at any time.

You are also free to withdraw from the test at *any time, during or after the data has* been collected, where your information may be discarded if requested.

No personal data is collected as part of this test and you will be anonymised through use of a participant ID.
The collected data will be stored securely and may be archived for up to a minimum of 3 years after the completion of the project. Your information will not be used outside of the researcher and project team.

Please feel free to ask any questions throughout the test.

By volunteering to partake in this test, you are agreeing that you understand and accept the above terms and have had the opportunity to ask questions regarding your participation in this experiment.

Signature: _____

Email: m.h.r.howard@edu.salford.ac.uk

Figure A.1: Participant information and consent form as provided to participants during the subjective listening test - Revisions highlighted

Bibliography

Adelman-Larsen, N. W. (2014). *Rock and Pop Venues* (1st ed.). Springer. https://doi.org/10.1007/978-3-642-45236-9

Akhtar, Z., & Falk, T. H. (2017). Audio-Visual Multimedia Quality Assessment: A Comprehensive Survey. *IEEE Access*, *5*, 21090–21117. https://doi.org/10.1109/ACCESS.2017.2750918

Angus, J. A. S. (1997). The Behaviour of Rooms at Low Frequencies. *102nd AES Convention, Munich*, *4421*.

Bai, M. R., & Chen, M. C. (2007). Intelligent preprocessing and classification of audio signals. *AES: Journal of the Audio Engineering Society*, *55*(5), 372–384.

Bogdanov, D., Wack, N., Gómez, E., Gulati, S., Herrera, P., Mayor, O., Roma, G., Salamon, J., Zapata, J. R., & Serra, X. (2013). ESSENTIA: an Audio Analysis Library for Music Information Retrieval. *International Society for Music Information Retrieval Conference (ISMIR'13)*, 493–498. http://hdl.handle.net/10230/32252

Bolla, C., Palladini, A., & Fazenda, B. M. (2019). Adding the Room To the Mix : Perceptual Aspects of Modal Resonance in Live Audio. *Proceedings of the 5th Workshop on Intelligent Music Production, Birmingham, UK, 6 September 2019*.

Bolt, R. H. (1946). Note on Normal Frequency Statistics for Rectangular Rooms. *Journal of the Acoustical Society of America*, *18*(1), 130–133. https://doi.org/10.1121/1.1916349

Bonello, O. J. (1979). Acoustical evaluation and control of normal room modes. *The Journal of the Acoustical Society of America*, *66*(S1), S52–S52. https://doi.org/10.1121/1.2017814

Bonello, O. J. (1981). New Criterion for the Distribution of Normal Room Modes. *AES: Journal of the Audio Engineering Society*, *29*(9), 597–606.

Breiman, L. (2001). Random forests. *Machine Learning*. https://doi.org/10.1023/A:1010933404324

Cecchi, S., Carini, A., & Spors, S. (2018). Room Response Equalization-A Review. *Applied Sciences*, *8*(16).

Celestinos, A., & Nielsen, S. B. (2008). Controlled acoustic bass system (CABS) a method to achieve uniform sound field distribution at low frequencies in rectangular rooms. *AES: Journal of the Audio Engineering Society*, *56*(11), 915–931.

Celestinos, A., & Nielsen, S. B. (2011). Low frequency sound reproduction in irregular rooms using cabs (control acoustic bass system). *Proceedings of Forum Acusticum*, 293–298.

Cox, T., & D'Antonio, P. (2001). Determining optimum room dimensions for critical listening environments: A new methodology. *Preprints-Audio Engineering Society*, (1), 1–6.

Cox, T. J., D'Antonio, P., & Avis, M. R. (2004). Room sizing and optimization at low frequencies. *AES: Journal of the Audio Engineering Society*, *52*(6), 640–651.

Davey Smith, G., & Ebrahim, S. (2002). Data dredging, bias, or confounding. *British Medical Journal*, *325*(7378), 1437–1438. https://doi.org/10.1136/bmj.325.7378.1437

Davy, J. L. (1989). The variance of the curvature of reverberant decays. *Journal of Sound and Vibration*, *128*(2), 297–305. https://doi.org/10.1016/0022-460X(89)90773-6

Dawes, R. M. (1979). The robust beauty of improper linear models in decision making. *American Psychologist*, *34*(7), 571–582. https://doi.org/10.1037/0003-066X.34.7.571

Dormann, C. F., Elith, J., Bacher, S., Buchmann, C., Carl, G., Carré, G., Marquéz, J. R., Gruber, B., Lafourcade, B., Leitão, P. J., Münkemüller, T., Mcclean, C., Osborne, P. E., Reineking, B., Schröder, B., Skidmore, A. K., Zurell, D., & Lautenbach, S. (2013). Collinearity: A review of methods to deal with it and a simulation study evaluating their performance. *Ecography*, *36*(1), 27–46. https://doi.org/10.1111/j.1600-0587.2012.07348.x

Duncan, P. J., Mohammed, D. Y., & Li, F. F. (2014). Audio information mining - Pragmatic review, outlook and a universal open architecture. *136th Audio Engineering Society Convention 2014*, 36–43.

Eerola, T., Lartillot, O., & Toiviainen, P. (2009). Prediction of multidimensional emotional ratings in music from audio using multivariate regression models. *Proceedings of the 10th International Society for Music Information Retrieval Conference, ISMIR 2009*, 621–626.

Fazenda, B., & Davies, W. (2002). The views of recording studio control room users. *Proceedings of the Institute of Acoustics*, *23*(8), 1–8.

Fazenda, B., & Wankling, M. (2008). Optimal modal spacing and density for critical listening. *Audio Engineering Society - 125th Audio Engineering Society Convention 2008*.

Fazenda, B., Wankling, M., Hargreaves, J., Elmer, L., & Hirst, J. (2012). Subjective preference of modal control methods in listening rooms. *AES: Journal of the Audio Engineering Society*.

Fazenda, B. M., Avis, M. R., & Davies, W. J. (2005). Perception of modal distribution metrics in critical listening spaces - Dependence on room aspect ratios. *AES: Journal of the Audio Engineering Society*, *53*(12), 1128–1141.

Fazenda, B. M., Stephenson, M., & Goldberg, A. (2015). Perceptual thresholds for the effects of room modes as a function of modal decay. *The Journal of the Acoustical Society of America*, *137*(3), 1088–1098. https://doi.org/10.1121/1.4908217

Fenton, S., & Lee, H. (2015). Towards a Perceptual Model Of 'Punch' In Musical Signals. *139th Audio Engineering Society International Convention, AES 2015*, 1–10.

Fenton, S., & Lee, H. (2019). A Perceptual Model of "Punch" Based on Weighted Transient Loudness. *Journal of the Audio Engineering Society*, *67*(6), 429–439. https://doi.org/10.17743/jaes.2019.0017

Groh, A. R. (1974). High-Fidelity Sound System Equalization by Analysis of Standing Waves. *J. Audio Eng. Soc.*, *22*(10), 795–799.

Heddle, J. (2016). Room acoustics for listening. *2nd Australasian Acoustical Societies Conference, ACOUSTICS 2016*, *1*(November), 69–78.

Hill, A. J. (2018). Live sound subwoofer system performance quantification. *144th Audio Engineering Society Convention 2018*, 1–10.

Hill, A. J., Hawksford, M. O., Rosenthal, A. P., & Gand, G. (2011). Kick-drum signal acquisition, isolation and reinforcement optimization in live sound. *130th Audio Engineering Society Convention 2011*, *1*.

Hutter, F., Lücke, J., & Schmidt-Thieme, L. (2015). Beyond Manual Tuning of Hyperparameters. *KI - Kunstliche Intelligenz*, *29*(4), 329–337. https://doi.org/10.1007/s13218-015-0381-0

International Organisation for Standardisation. (2004). MPEG-7 Overview (version 10). *Coding of moving pictures and audio*.

Karjalainen, M., Antsalo, P., Mäkivirta, A., Peltonen, T., & Välimäki, V. (2002). Estimation of modal decay parameters from noisy response measurements. *AES: Journal of the Audio Engineering Society*, *50*(11), 867–878.

Kates, J. M., & Arehart, K. H. (2016). The hearing-aid audio quality index (HAAQI). *IEEE/ACM Transactions on Audio Speech and Language Processing*, *24*(2), 354–365. https://doi.org/10.1109/TASLP.2015.2507858

Kressner, A. A., Anderson, D. V., & Rozell, C. J. (2013). Evaluating the generalization of the hearing aid speech quality index (HASQI). *IEEE Transactions on Audio, Speech and Language Processing*, *21*(2), 407–415. https://doi.org/10.1109/TASL.2012.2217132

Kutruff, H. (2009). *Room Acoustics* (5th). Spon Press.

Liaw, A., & Wiener, M. (2002). Classification and Regression by randomForest. *R News*, *2*(3), 18–22.

Love, B. C. (2002). Comparing supervised and unsupervised category learning. *Psychonomic bulletin \& review*, *9*(4), 829–835.

Lundeby, A., Vigran, T. E., Bietz, H., & Vorlaender, M. (1995). Uncertainties of measurements in room acoustics. *Acustica*, *81*(4), 344–355.

Luo, G. (2016). A review of automatic selection methods for machine learning algorithms and hyper-parameter values. *Network Modeling Analysis in Health Informatics and Bioinformatics*, *5*(1), 1–16. https://doi.org/10.1007/s13721-016-0125-6

Ma, L., Milner, B., & Smith, D. (2006). Acoustic environment classification. *ACM Transactions on Speech and Language Processing*, *3*(2), 1–22. https://doi.org/10.1145/1149290.1149292

Mao, K. Z. (2004). Feature Subset Selection for Support Vector Machines Through Discriminative Function Pruning Analysis. *IEEE TRANSACTIONS ON SYSTEMS, MAN, AND CYBERNETICS—PART B: CYBERNETICS*, *34*(1), 60–67. https://doi.org/10.1109/icmlc.2004.1384586

Mehta, P., Bukov, M., Wang, C. H., Day, A. G., Richardson, C., Fisher, C. K., & Schwab, D. J. (2019). A high-bias, low-variance introduction to Machine

Learning for physicists. *Physics Reports*, *810*, 1–124. https://doi.org/10.1016/j.physrep.2019.03.001

Music Technology Group, & Universiat Pompeu Fabra Barcelona. (2019). Essentia - Documentation. https://essentia.upf.edu/documentation.html

Olive, S. E., Jackson, J., Devantier, A., Hunt, D., & Hess, S. M. (2009). The subjective and objective evaluation of room correction products. *Audio Engineering Society Convention 127*, 1–17.

Olive, S. E., Welti, T., & Khonsaripour, O. (2017). A statistical model that predicts listeners' preference ratings of in-ear headphones: Part 2 – Development and validation of the model. *143rd Audio Engineering Society Convention 2017, AES 2017*, *2*, 593–602.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., & Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, *12*, 2825–2830.

Rindel, J. H. (2015). Modal energy analysis of nearly rectangular rooms at low frequencies. *Acta Acustica united with Acustica*, *101*(6), 1211–1221. https://doi.org/10.3813/AAA.918914

Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, *1*(5), 206–215. https://doi.org/10.1038/s42256-019-0048-x

Rumsey, F. (2011). Live sound things to get right. *AES: Journal of the Audio Engineering Society*, *59*(12), 986–990.

Saeys, Y., Inza, I., & Larrañaga, P. (2007). A review of feature selection techniques in bioinformatics. *Bioinformatics*, *23*(19), 2507–2517. https://doi.org/10.1093/bioinformatics/btm344

Sagi, O., & Rokach, L. (2018). Ensemble learning: A survey. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, *8*(4), 1–18. https://doi.org/10.1002/widm.1249

Savioja, L., & Svensson, U. P. (2015). Overview of geometrical room acoustic modeling techniques. *The Journal of the Acoustical Society of America*, *138*(2), 708–730. https://doi.org/10.1121/1.4926438

Shier, J., McNally, K., & Tzanetakis, G. (2017). Analysis of drum machine kick and snare sounds. *143rd Audio Engineering Society Convention 2017, AES 2017*, *2*, 671–676.

Stephenson, M. (2012). *Assessing the Quality of Low Frequency Audio Reproduction in Critical Listening Spaces* (Doctoral dissertation). University of Salford, Salford, UK.

Stoller, D., Ewert, S., & Dixon, S. (2018). Wave-U-Net: A multi-scale neural network for end-to-end audio source separation. *Proceedings of the 19th International Society for Music Information Retrieval Conference, ISMIR 2018*, 334–340.

Teret, E., Pastore, M. T., & Braasch, J. (2017). The influence of signal type on perceived reverberance. *The Journal of the Acoustical Society of America*, *141*(3), 1675–1682. https://doi.org/10.1121/1.4977748

Thiede, T., Member, A., Treurniet, W. C., Bitto, R., Schmidmer, C., Sporer, T., Beerends, J. G., Colomes, C., Keyhl, M., & Stoll, G. (2000). PEAQ: The ITU

Standard for Objective Measurement of Perceived Audio Quality. *Journal of the Audio Engineering Society*, *48*(1/2), 29.

Toole, F. E. (2013). Sound Reproduction Loudspeakers and Rooms. *Sound reproduction loudspeakers and rooms* (2nd, pp. 197–248). Focal Press.

Universiat Pompeu Fabra Barcelona, & Music Technology Group. (2019). Essentia - Algorithms Reference. https://essentia.upf.edu/algorithms_reference.html

University of Cambridge. (2019). Rules of thumb on magnitudes of effect sizes. http://imaging.mrc-cbu.cam.ac.uk/statswiki/FAQ/effectSize

University of Salford. (2020). Listening Room - University Of Salford Acoustic Testing. http://www.salford.ac.uk/acoustics-testing/labs/listening-room

van Dorp Schuitman, J., de Vries, D., & Lindau, A. (2013). Deriving content-specific measures of room acoustic perception using a binaural, nonlinear auditory model. *The Journal of the Acoustical Society of America*, *133*(3), 1572–1585. https://doi.org/10.1121/1.4789357

Vercammen, M. (2019). On the revision of ISO 354, measurement of the sound absorption in the reverberation room. *Proceedings of the 23rd International Congress on Acoustics*, 3991–3997.

Walker, R. (1992). *Low-Frequency Room Responses: Part 2 – Calculation methods and experimental results* (tech. rep.).

Wankling, M., & Fazenda, B. (2009). Studies in modal density - Its effect at low frequencies. *25th Reproduced Sound Conference 2009, REPRODUCED SOUND 2009: The Audio Explosion - Proceedings of the Institute of Acoustics*, *31*, 62–73.

Wankling, M., Fazenda, B. M., & Davies, W. J. (2012). The assessment of low-frequency room acoustic parameters using descriptive analysis. *AES: Journal of the Audio Engineering Society*, *60*(5), 325–337.

Wegener, S., Haller, M., Burred, J. J., Sikora, T., Essid, S., & Richard, G. (2008). On the robustness of audio features for musical instrument classification. *European Signal Processing Conference*.

Welti, T., & Devantier, A. (2003). In-room Low Frequency Optimization. *115th AES Convention, New York, Convention Paper 5942*, *5942*, 1–15.

Welti, T. S. (2009). Investigation of Bonello Criteria for use in small room acoustics. *127th Audio Engineering Society Convention 2009*, *2*, 1359–1366.

White, P. (2015). *The SOS Guide to Live Sound* (1st). Focal Press.

Wilson, A., & Fazenda, B. M. (2016). Perception of Audio Quality in Productions of. *J. Audio Eng. Soc*, *64*(1), 23–34. https://doi.org/10.17743/jaes.2015.0090

Wilson, R. J. (2006). Can we get the bass right? *Proceedings of the 21st AES UK Conference: Audio at Home*, 1–11.

Zieliński, S., Rumsey, F., & Bech, S. (2008). On some biases encountered in modern audio quality listening tests - A review. *AES: Journal of the Audio Engineering Society*, *64*(6), 427–451.