Escola das Artes da Universidade Católica Portuguesa

Mestrado em Som e Imagem

The Drum Kit and the Studio

A Spectral and Dynamic Analysis of the Relevant Components

Design de Som | 2013/2014

João Manuel Pinheiro de Almeida

Professor Orientador: Professor Doutor Pedro Duarte Pestana

Novembro de 2014

Dedication

I would like to dedicate this work to the musicians and sound engineers that, through their work, helped me choose to follow my dreams of, one day being a part of their world.

Above all, this written work is dedicated to my family and close friends for all their affection and support during these five long and strenuous years of academic endeavours.

Acknowledgements

I would like to thank all those that have supported me during the development of this project. In particularly, the inspiring supervision and willingness to help of my advisor Professor Doctor Pedro Duarte Pestana and the creative, insightful and tireless help from Professor Vitor Joaquim.

My heartfelt gratitude goes to all those who directly or indirectly were involved in the making of this dissertation.

Abstract

The research emerged from the need to understand how engineers perceive and record drum kits in modern popular music. We performed a preliminary, exploratory analysis of behavioural aspects in drum kit samples. We searched for similarities and differences, hoping to achieve further understanding of the sonic relationship the instrument shares with others, as well as its involvement in music making.

Methodologically, this study adopts a pragmatic analysis of audio contents, extraction of values and comparison of results. We used two methods to analyse the data. The first, a generalised approach, was an individual analysis of each sample in the chosen eight classes (composed of common elements in modern drum kits). The second focused on a single sample that resulted from the down-mix of the previous classes' sample pools.

For the analysis, we handpicked several subjective and objective features as well as a series of low-level audio descriptors that hold information regarding the dynamic and frequency contents of the audio samples. We then conducted a series of processes, which included visual analysis of three-dimensional graphics and software-based information computing, to retrieve the analytical data.

Results showed that there are some significant similarities among the classes' audio features. This led to the assumption that the *a priori* experience of engineers could, in fact, be a collective and subconscious notion, instinctively achieved in a recording session.

In fact, with more research concerning this subject, one may even find new a new way to deal with drum kits in a studio context, hastening time-consuming processes and strenuous tasks that are common when doing so.

Keywords: drum kit, idyllic sound, audio descriptors, qualitative analysis.

Resumo

A investigação científica realizada no ramo do áudio e da música tornou-se abastada e prolífica, exibindo estudos com alto teor informativo para melhor compreensão das diferentes áreas de incidência.

Muita da pesquisa desenvolvida foca-se em aspectos pragmáticos: reconhecimento de voz e de padrão, recuperação de informação musical, sistemas de mistura inteligente, entre outros. No entanto, embora estes sejam aspectos formais de elevada importância, tem-se notado uma latente falta de documentação relativa a aspectos mais idílicos e artísticos.

O instrumento musical de estudo que escolhemos foi a bateria. Para além de uma vontade pessoal de entender a plenitude das suas características sónicas intrínsecas para aplicações prácticas com resultados tangíveis, é de notar a ausência de discurso e pesquisa científica que por este caminho se tenha aventurado.

Não obstante, a bateria tem sido objecto de estudo profundo em contextos analíticos, motivo pelo qual foi também relevante originar a nossa abordagem seminal. Por um lado, as questões físicas de construção e manutenção de baterias, bem como aspectos de índole ambiental e de espaço (salas de gravação) são dos aspectos que mais efeitos produzem na diferença timbríca em múltiplos exemplos de gravações de baterias. No entanto, questões tonais (fundamentais para uma pluralidade de instrumentos) na bateria carecem de estudo e documentação num contexto mundial generalizado.

São muitos os engenheiros de som e músicos que alimentam a ideia preconcebida da dificuldade inerente em relacionar este elemento percursivo com os restantes instrumentos numa música. Aliam-se a isto questões subjectivas de gosto e preferência, bem como outros métodos que facilitam a inserção de um instrumento rítmico e semi-harmónico (porque é possível escolher uma afinação para diferentes elementos de uma bateria) numa textura sonora que remete para diferentes conceitos musicais.

Portanto, a questão nuclear que este estudo se foca é: *"será possível atingir um som idílico nos diferentes elementos de uma bateria?"*. Em si só, a ambiguidade desta resposta pode remeter para um conceito dogmático e inflexível, bem como para a ideia de que, até ao momento, nenhuma gravação ou som de bateria alcançou um patamar de extrema qualidade, sonoridade ou ubiquidade que a responda a esta premissa.

Partimos, então, desta interrogação e procedemos a uma análise pragmática de amostras sonoras que fossem o mais assimiláveis possível a um contexto comercial. Reunimos amostras de oito classes pré-definidas: bombos, tarolas, pratos de choque, timbalões graves, médios e agudos, *crashs* e *rides*. As amostras derivaram de bibliotecas que foram reunidas posteriormente à realização de uma pesquisa em busca dos fabricantes mais conceituados, com maior adesão pública e com antecedentes comerciais tangíveis. Daqui recuperamos 481 amostras.

Depois de reunidas, as amostras sofreram um processo de identificação e catalogação, passando também por alguns momentos de processamento de sinal (conversão para ficheiros monofónicos, igualização da duração e normalização do pico de sinal). Em seguida, através do *software* de computação matemática MATLAB, desenvolvemos linhas de código que foram instrumentais para fase da análise de características e descritores de ficheiros áudio. Finalmente, procedemos a uma reunião dos resultados obtidos e a iniciação de suposições que pudessem originar os valores extraídos.

De entre os resultados obtidos, surgiram ideias que, com mais investigação, podem facilitar a compreensão do comportamento sonoro dos diferentes elementos, bem como a criação de métodos de conjugação harmónica entre eles.

É importante referir que, neste estudo, partimos de um conceito qualitativo do som, e como tal, omitimos aspectos físicos que, na sua essência, influenciam substancialmente o som que é emitido. No entanto, este trabalho introdutório pretende retificar de forma preliminar esta falta de conceitos subjectivos com evidências palpáveis. Evidências essas que ainda necessitam de investigação adicional para a sua confirmação.

Palavras-chave: baterias, som idílico, descritores de áudio, análise qualitativa.

Contents

List of Figures

List of Tables

Glossary

ACA – Audio Content Analysis

ACF – Autocorrelation Function

AI – Artificial Intelligence

ANSI – American National Standards Institute

BPM – Beats Per Minute

CD – Compact Disc

COG – Centre of Gravity

DAW – Digital Audio Workstation

dB – Decibel

dBFS – Decibel Full Scale

DFT – Discrete Fourier Transform

DSP – Digital Signal Processing

EQ – Equalization

FFT – Fast Fourier Transform

FOH – Front-of-House

HMM – Hidden Markov Models

Hz – Hertz

IFFT – Inverse Fast Fourier Transform

ISA – Independent Subspace Analysis

KNN – K-Nearest Neighbour

kHz – KiloHertz

MACF – Maximum of Autocorrelation Function

MFCC – Mel-Frequency Cepstral Coefficients

MIDI – Music Instrument Digital Interface

MIR – Music Information Retrieval

ms – Milisecond

OH – Overhead

PSA – Prior Subspace Analysis

PZM – Pressure-zone Microphone

RBF – Radial Basis Function

RMS – Root-Mean-Square

RTA – Real Time Analysis

RTAS – Real Time Audio Suite

s – second

SC – Spectral Centroid

SF – Spectal Flux

SPL – Sound Pressure Level

STFT – Short Term Fourier Transform

SVM – Support Vector Machines

1 Introduction

1.1 Research Subject and Formal Aspects

The current text will focus on one of the most common instruments in modern music: drum kits.

It is important and relevant to mention that this research rose from a personal quest to further understand the subject at hands in an objective context with a prior scientific knowledge to support it. We are convinced that in order to attain the best possible result in a drum kit recording session one must intimately understand the relation among the various classes of drum kits, as well as their relationship with other instruments in a musical context.

Furthermore, we believe that there has been a lack of academic attention concerning this subject. As such, this research will address questions that we hope will contribute to a further understanding of the sonic properties of drum kits and other instruments alike.

This research will contain five chapters. The current chapter will deal with cultural and historical aspects of music. It is important to understand how music has developed through the ages and how creativity and technology has led to its development. Because of its widespread influence and natural differences among cultures, it is important to understand what led popular music to affect worldwide targets. Alongside this, we will be contemplating the fundamental concepts we deem important for this research.

On the second chapter, we will deal with the evolution of percussion instruments and the inventions and necessities that have led to the creation of a drum kit: its roots and how it has developed to become the intricate instrument that we know today. Still in this chapter, we will make a reflection on the academic research that has been done regarding percussion instruments and drum kits. We will close this segment with the general considerations and research questions we will be addressing.

The third chapter will focus on the subject at hands from the sound engineer's point of view. We will explain the research methodology and the analysis we performed to address the questions raised. We will be making a thorough description of the steps that we took in the early stages of this process as well as the problems and solutions we encountered during the sample pool gathering. Furthermore, for the analysis part, we will explain the reasons for the methods of choice and the results that we extracted from their use.

The fourth chapter will concern the results that we mustered on the third chapter, while making comparisons among them, their interpretation, as well as, some tentative assumptions

they elicit. These assumptions will lead to a discussion of what we have achieved and what we believe that should be relevant for further research.

The final and fifth chapter will be the conclusive arguments of this written research, where we will contemplate the whole process.

1.2 Music in the Modern Age

Music has withstood the test of time. It is one of the biggest heritages mankind has continuously passed onto new generations. Its formal origins are still uncertain, and it is unclear how mankind has developed the capacity for music making (McDermott, 2008), but has been widely recognised to be as natural as breathing is.

Nevertheless, and despite all its mysterious roots, the further we understand it, the further we adapt through countless changes and mutations. From the variations, new instruments and genres have emerged, reflecting music's versatility and the equally relevant transformations of society.

Loy (2006, p. 3) stated, "even though our senses are connected directly to the world, our inner experience of phenomena is not identical to the stimuli we receive". From this premise, musical experience diverges from person to person, making it an idiosyncratic expression taste, as well as a direct result of the social context.

According to Cross (2003), being such an important part of our cultural heritage, it should "be possible to understand music by identifying and applying general principles of the type found within formal and scientific theories". In fact, along the ages, history and theory have devised tools, rules and languages that have been the main driving force for composing, writing and reading music.

Similarly, society and has also been the cause for these many changes in the way music is presented to the listener. This social context has permeated on to music, giving it a singular identity that distinctively makes it different. With new instruments created, new ways to insert them in a piece of music were devised; with it, music flourished and composers and musicians rose in ingenuity and creativity, culminating in massive indoor orchestras playing music until the early years of the 20th century.

Still, on this modern age, with social factors such as population and economic growth and being considered a typically elitist activity, music began to breed an industry more dedicated and inclined to please the less erudite, growing part of the population. Heine (2003) mentions the changes in concert life as a direct result of changes of musical taste, while Larkin (2011) discloses that comparisons between both are not few, making popular music suffer "from an inferiority complex based primarily upon colour and class".

An industry revolving around the public nurtured and allowed the possibility of leaving opera houses and theatres and but being able to enjoy music, waging it all in the development of technology that allowed people to reproduce it at home.

At this time, music became popular, and began to branch out, developing multiple identities and, in turn, creating social movements. It saw aspiring young musicians to rise from humble roots to the category of cultural icons, moving millions of people from around the world and developing a whole new boosting economy, spiralling around music sales. No longer was music a strictly erudite form of art, as it had become a part of large segments of the public.

Because of these social reasons, and because of easier access to instruments, new genres derived from the most common forms of popular music, turning the music making industry into something ill-regarded by scholars and academics, as has been reinforced by Middleton (2003). An example given by Frith (1992, p. 174) concerns the rise of the number of rock

groups and bands in the eighties leading to the public embracing the genre to that point that it was “either parasitic or (…) a spontaneous, folk-like activity”.

Nowadays, according to Serrà et al. (2012) modern popular music established a new set of underlying patterns that make it sound very similar, but the regularities that were found may have been “potentially inherited from the classical tradition”. Maintaining such patterns and adding variations may invoke memories from the listeners, which tend to correspond to their expectations (Serrà et al., 2012) as well as conveying emotion and guaranteeing a intimate and personal link from the listener to the music (Juslin & Sloboda, 2011).

Now, on the digital era, anyone can have these tools to create music. A complete orchestral music piece or a pop-rock song can be written, performed and produced in a single laptop computer. On one hand, technology has led to continuous improvements in recording quality and processing power capability, but on the other hand, the easier access to these tools, led to a homogeneous characteristic sound, where everything, even the most creative-based elements such as the melody line, may sound dauntingly similar to many others.

1.3 The Search for Standards in Music

Research concerning characteristics and features of sound and music has been done since the last decades of the 19th century, an this older than a century quest has continued to determine what could apply for similarity and what could not, what is standard and what is an exception.

We could certainly debate, philosophically speaking, qualitative factors such as beauty and pleasure as a general way to determine a common factor in the equation for finding similarity among music. On this subject of aesthetics, Aristotle (2004) once defined the main characteristics of beauty as being "order and symmetry and definiteness, which the mathematical sciences demonstrate in a special degree".

Measuring perception and subjectivity through mathematics and physics can be possible, but still only up to a certain point. According to Oxenham (2012), the German-born scientist Gustav Fechner was credited as the father of psychophysics, and by extension, the field of psychoacoustics, i.e., "the attempt to establish a quantitative relationship between physical variables (e.g. sound intensity and frequency) and the sensations they produce (e.g. loudness and pitch)".

From here, we cross academic field boundaries and make our way into neuroscience and neuropsychology. Concerning the altering of one's emotional state changes as a direct result of a chemical variation on the brain in response to musical stimuli. As Peretz (2011) suggested, it is a part of human nature.

Additional areas of study as sociology and anthropology also debate this question and suggest an even wider gap in the search for an unanimous definition of music, due to geographic, social and cultural differences. Thompson and Blakwill (2011) stress the ethnomusical distinctions that western and eastern music present, are a cause for a "different perspective on the concept of music", and therefore, the sensations they produce on the listener.

Although, being an extremely prolific matter to debate on many areas of thought, and despite existing abundant work attempting to further achieve closure, neither a universal concept was reached, nor can we expect to easily find it; what is the exact standard definition of music or how it should sound like? Yet, it is our intentions to tackle this subject from an objective and practical point of view.

Yet and despite the much-expressed sameness of music in modern settings, we can also assume that the equality trait could lie not on the raw material, but on the way that it is used in musical contexts.

Musical form has been somewhat a cause for this. Musical form could in fact be described as musical architecture, i.e., how songs in its several pieces work together in order to achieve a coherent outcome. Benward & Saker (2007, p. 361) have pointed out the similarity of contemporary songs to the ternary design, commonly designated the verse-chorus-verse. In fact they even state that most songs from the mid-twentieth century and onwards have singled out to use only the chorus.

With worldwide overabundance of raw material to create music, we can expect several common elements in a wide variety of musical subjects. This may lead to an undeniable similar result among songs from different authors. From this, we can still dare to say that music is still far from being totally equal. It depends almost exclusively on the composition

and song writing skills of their authors, and these may in fact be the reason why music in the modern days has been described as “sounding the same”.

This research does not ponder on the possibility of achieving a universal standardisation of music. On the contrary, we plan to approach the raw material used in music (in our case drum kits), analyse their underlying spectral and temporal characteristics and search for common elements and similarities encountered on modern music, so that a fast and reliable reproduction of their attributes can be achieve inside the studio.

With the results to be extracted, we hope to broaden the horizons concerning research on the sonic interaction of drum kits in modern popular music. In the future, we hope that recording engineers world-wide will be able to focus less on the strenuous task of reaching the perfect drum sound for their record and will have more time to deliver the greatest possible result, both creatively and technically.

Still, all this is a subjective approach to a highly objective task, because, as Senior (2008) pointed out, “even the studio greats [sound engineers] disagree about which technique ‘sounds best’, so ultimately it’s up to you to decide which works for you”.

1.4 Fundamental Concepts

Throughout this research we will be dealing with several music, sound and audio related concepts and terminology that should be explained up front. On a first look, the systematic process of studio recording is an individual and ever-changing process that is performed by recording engineers worldwide. As Huber & Runstein (2009, p. 29) summed up the "differences between people and the tools they use allow the process of recording to be approached in many different ways". On the other hand, Katz (2007) identified and outlined five distinct and rather inflexible moments on the music industries' processing chain (Figure 1).

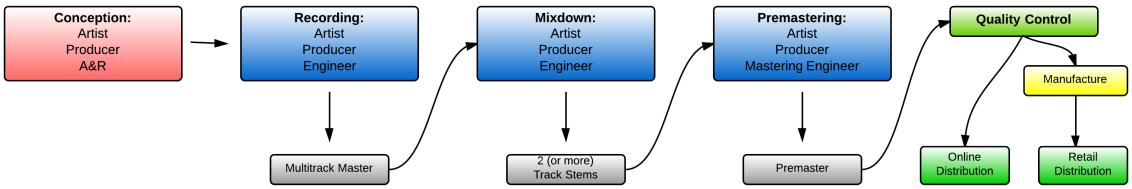

Figure 1.1 The music production processing chain (adapted from Katz, 2007 p. 20)

The purpose of our research is to define a unitary moment that included both the Recording and the Mixdown (post-production moment that comprehends editing and mixing) stages. It is our intention to reach a point where it is possible to identify, on a real-time basis, the fundamental components that incorporate a sound when emitted by its source. As such, we will be giving considerable importance to the spectral response and the dynamic response of individual drum kit elements for substantiating our research.

Analysing the spectral and dynamic components of audio signals has been common practice in the sound business. The gathering of the raw material is fundamental for the recording industry, since it is the basis for the whole project. Still, sound engineers throughout the world use extensively tools such as equalization (EQ) and compression to adapt that raw material to their needs. This is common practice early in the processing chain (editing stage) because it allows them to manipulate certain aspects that could not be controlled during the recording session. This way, by controlling the spectral and dynamic domains they can achieve the intended sound output.

Other concept we would like to stress is the qualities of sound. Schaeffer's (1966) work has dealt, extensively, with the subject of categorizing and representing complex sounds. He presented a three-dimensional category diagram, where Intensity (dB), Frequency (Hz) and Duration (s) intertwine and help confer sound a characterization.

The three proposed categories are the Dynamic (or of Forms), the Melodic (or of Textures) and the Harmonic (or of Timbres). With these various characteristics an easier verbal expression of sound is enabled.

Sonnenschein, on the other hand, proposed several bi-dimensional qualities in sound (Table 1.1) that are perceived and "governed by the physiologic limitations of the hearing apparatus" (2001, p. 65). These categories, despite subjective, allow for a better exchange of ideas among sound engineers and those they wish to convey their ideas to.

Sound Quality	*Extremes*
Rhythm	Irregular — Rhythmic
Intensity	Soft — Loud
Pitch	Low — High
Timbre	Noisy — Tonal
Speed	Slow — Fast
Shape	Impulsive — Reverberant
Organization	Chaotic — Ordered

Table 1.1 Sonnenschein's porposed sound qualities and respective extremes

Despite the subjective nature of these categories, we shall deal with some of them throughout this dissertation. But, those we which to emphasise for the time being are the dynamic envelope and the frequency response.

The time domain of audio signals corresponds to the way sound acts in the course of its duration. For understanding it, one must know the concept of amplitude. It can be related to the pressure of sound (or intensity) and determines how loud or how quiet we perceive sound, the sensorial perception of loudness. Regarding the spectral domain, one must first understand the notion of frequency and its relationship with pitch, defined in 1994 by the American National Standards Institute (ANSI):

> *"Pitch is that auditory attribute of sound according to which sounds can be ordered on a scale from low to high. Pitch depends mainly on the frequency content of the stimulus, but also depends on the sound pressure and the waveform of the stimulus."*

Analysing the spectral domain of a recorded sound allows the user to fully grasp the behaviour of the object (whether it is a voice, an instrument or a soundscape). This can become easier to grasp when the sound is further decomposed in frequency bandwidths. The most common ways to analyse this particular characteristic of sound is through spectrum analysers or spectrograms that are build upon Fast Fourier Transform (FFT) functions. They enable the transposing of a signal's frequency into a visual representation graphic, and in most cases, this happens in real time. The reason for using the functions when comparing to others such as with the Discrete Fourier Transform (DFT) functions are that the latter are less efficient, more time consuming and therefore require much more processing power for lengthy time blocks (Lerch, 2012, p. 197).

Along the years, spectral analysis has been the subject of many researchers, because they enable the opportunity to understand how "the human ear responds to the tonality[1]" (Katz, 2007, p. 104). Furthermore, studies have been developed, analysing the spectral distribution of songs in their entirety.

[1] According to the Oxford Dictionary of Music (Kennedy, Kennedy, & Rutherford-Johnson, 2012), tonality is directly related to the key of a piece of music, i.e., the musical notes (frequency) that relate well with a first note (the tonic).

One of this cases was a study by Pestana et al. (2013) which presented the spectral distribution of the number one records in the American and British top music charts since the 50s.". Their analysis – divided by decade and genre – showed, for example, a significant increase in the low-end frequencies along the years (a rise of over 20 decibels (dB) in magnitude). Their conclusions were expected because "hip-hop's more prominent loudness and extended bass response is evidently related to the fact that post-2000 songs share the same tendency".

They noticed that, although existent, there has not been undertaken a "consistent academic study that tackles the question of how generally similar is the spectral response of critically acclaimed tracks, nor has anyone analysed the surrounding factors upon which it depends."

Unlike Pestana et al. approach, we do not wish to analyse the final product, instead we are aiming for more specfic parts of a song when instruments are recorded individually. Therefore, in our research, Real Time Analysis (RTA) of sound signals takes a prominence. It allows the user to grasp a visual output of the frequency, in an almost instantaneous time interval (through either filter-bank approach or FFT functions), so that the frequency behavioural aspects of the sound during the recording can be understood. Because of this almost instant visual response, necessary adjustments from external factors, such as the environmental and technical traits, can be made; these include choosing the microphones by their frequency response and proximity to source, adapting to the room's acoustic properties and identifying the instrument's timbre and tuning.

In spite of being a major and defining characteristic of sound, frequency is not the only thing engineers might be concerned about. There must be significant understanding of the behaviour of sound when is emitted from its source. This has been called the dynamic envelope and it is a quantifiable "way of diving events and flows into meaningful frames" by considering "the change of energy occurring within a system" (Farnell, 2010, p. 89).

Considering the aspects of dynamic response and behaviour of sound along time, we can divide any sound in four separate moments: attack, decay, sustain and release. In some cases, these different stages may be shorter or longer, depending on the instrument and the way it is played. These are the individual aspects that compile an ADSR envelope, a tool commonly used in synthesizers in order to achieve an efficient emulation of this behaviour that sound presents (Figure 1.2).

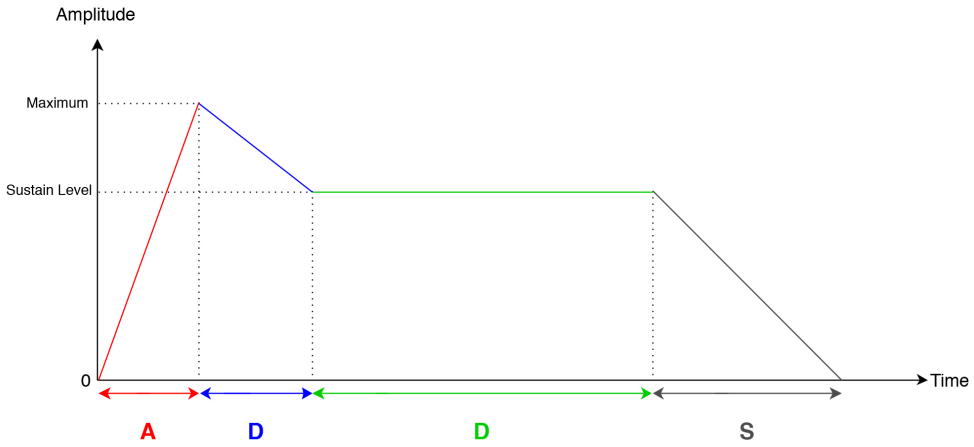

Figure 1.2 Diagram of an ADSR envelope

Sound takes a period of time from silence to peak, or from zero to maximum energy. This is called the *attack* and can be further divided into two stages: transient and rise. The first, shorter and usually louder, is the "excitation stage". The second, carries the energy still present after the sound is produced. As the "energy continues to be supplied after the transient stage" (Farnell, 2010, p. 90) we call this the *decay*, which leads to a stable flow of energy or *sustain*. Finally, when the system stops receiving energy, there is still sound until it vanishes completely. This is the *decay* period of the envelope.

Further specific concepts shall be addressed throughout this dissertation.

2 State of the Art

2.1 Cultural and Artistic Context

Since its inception by the hands of Thomas Edison, recording and reproducing sound has been a practice that has moved thousands of inventors, engineers, technicians and musicians. After years of perfecting high quality sound recording equipment and high fidelity reproduction systems, the multinational corporation Phillips introduced in the early eighties a decisive and impactful turning point in the industry: the Compact Disc or CD (Morton, 2006). With it came the dawn of a digital era in sound.

The digital domain has captivated most of the general population with its significant changes in the technological field. The world of sound and its professionals were also fascinated by this mesmerizing effect. All around the world, sound mixing consoles were replaced by computers and digital controllers, effects racks were swapped by digital software and even the industries' professionals were overlooked in favour of people with little know-how of the trade. This was due to easier access to equipment that could reproduce what was once exclusive to high-end studios (Bartlett & Bartlett, 2009, p. 5).

The impact of digital technology has enabled endless possibilities in sound production as well as more reliability in the equipment and at a much lower price. Watkinson (2001, p. 2-3) defines that an ideal recorder (analogue or digital) provides a "transparent" recording and reproduces "the original applied waveform without error". He further remarks that they "both fall short of the ideal", but the digital domain does so by a "shorter distance (…) and at a lower cost". In the later decades of the 20th century sound was predominantly recorded in digital multi-track environments and being distributed in digital format[2] (Fine, 2008, p. 11).

These and other technological advancements compelled companies to develop in order to meet the requirements of the consumer. These led to the development of the digital audio workstation (DAW), whose "overwhelming advantages" became the "standard in contemporary audio production" (Savage, 2011, pp. 3–6). With the evident weight of software-based workstations, we must understand and transpose the physical phenomenon that is sound as digital representation of data in a computer screen.

[2] By this, we imply the concept of converting sound in digital data, not the sales medium. This came to a peak in 2001 with Apple's presentation of the iPod. CD was no longer a viable, practical option. As it declined, Apple assured top position in the control of the music market through MP3 sales in its virtual store: iTunes (Peng & Sanderson, 2013).

A basic intuitive point in the process of perceiving sound and music is that each instrument has its own peculiarities. In both old and modern music, no instrument has had such an impact and worldwide use as have had the percussion instruments; indeed, since long time ago up to this day and age, the drum kit has held this undisputed top position. For this reason, and also because percussion is the basis of rhythm, as it creates a "perceptually isochronous pulse to which one can synchronize with periodic movements" (Patel, 2010, p. 97), we have decided to focus our attention on its impact and performance in commercial recordings.

With such an important role in music, it is worth to deepen our knowledge on the roots that set ground to the development of these instruments, and relate their archaic build with modern innovations. Also, we think it is a vital point for our work to muster and comprehend which are their essential characteristics and how research has addressed the study of its particular behaviour when compared with other instruments.

Historically speaking, musical instruments such as bone flutes and rasps can be traced back to more than 100,000 years ago. For this reason, we "could assume that" pre-historic tribes might have "crafted percussive instruments" since rhythm retains itself as a humane characteristic; some might add that it is genetic (Peretz, 2011). Dean (2012, p. 5) argued as well the veracity of these concerns, but no factual proof has demonstrated that drums appeared that early, because "the wood and animal hide often used for drums would have perished thousands of years ago".

The author further states that early crafting of drums (or membranophones, a stretched membrane which, when hit, resonates with the surface that is in contact with) could be attributed to the ancient Mesopotamian and the Egyptian civilizations, as early as 3200 BC. This was due to their tradition in taking religious and symbolic items to their tombs, which prevented the decomposition of materials. However, archaeological studies debate such assumption. Excavations in China uncovered alligator and pottery drums dated to the Neolithic period. In fact, according to Liu (2004, p. 122) the "earliest examples" could be traced to a "large burial site in Dawenkou" (4100 BC – 2600 BC) in China.

In the course of centuries, percussive instruments suffered various mutations, both in terms of design, construction and materials as well as in terms of playing styles. Everywhere around the world, tribes, societies and civilizations created resonating devices that could be considered part of the percussion class. Arguably, necessity drove men to produce the drum from the different resonating materials and its development made it way to make them a fundamental element in countless forms of musical expression.

Many times, the need came to incorporate more than one percussionist for guaranteeing musical cohesion. One such case that exemplified this multi-drummer approach were brass bands.

By late 19th century most American towns had one. Much like classic orchestras, they included snares, cymbals and bass drums and many others. Still, no single musician could play them all at once. "When these groups moved inside, the standard instrumentation was cut down somewhat for practical reasons. Because of this, the need for two or more drummers decreased and resourceful inventions began to flourish" (Fidyk, 2011). One of such inventions

was the bass drum pedal (or kick drum[3]), an invention by German-born William F. Ludwig, and his bother Theobald, in the early years of the 20th century.

After being granted a patent for their pioneering device, the first pedal (Figure 2.1) was presented to the public in 1909. Such invention was a milestone on the path to sit down the drummer, allowing him mobility and capacity to play with all four limbs and outline the modern drum kit setup (Dean, 2012, p. 199).

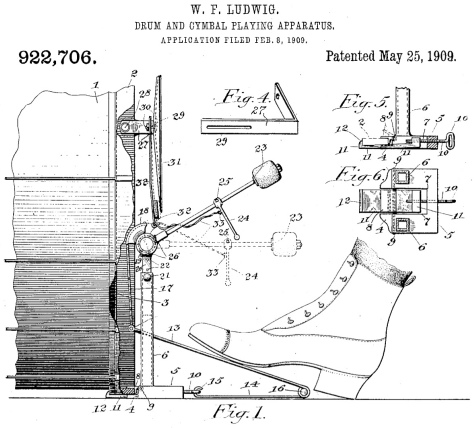

Figure 2.1 Detail of *Ludwig's* 1909 patent (from http://commons.wikimedia.org/)

Nowadays, drum kits may appear in many forms; one of the most common in most music genres is the five piece[4] kit (Figure 2.2), but still, they can range from as few a three piece to multiple drums per kit, especially in progressive rock/metal.

Figure 2.2 A five piece drum kit (created using *DW*'s *Kitbuilder*)

Other complementary instruments could be included in the drum kits of different genres, such as cowbells, chimes and gongs. In some cases, electronic drums are added to add a wider range of possibilities.

[3] The origins of the term "kick drum" could be traced back to this time as sometimes instead of a pedal, a cord was attached to the drummer's foot who had to perform a kicking motion in order to activate an upside beater or mallet (Dean, 2012).

[4] An *n* piece drum kit relates to number of instruments (membranophones) that the kit displays. For example, a one kick, one snare and one floor tom kit with multiple cymbals is still considered a three-piece kit.

Since drum kits are such a vital element in modern music, much research has been dedicated to their sound behaviour, acoustic response and pattern recognition and transcription. They are important aspects for areas such as Music Information Retrieval (MIR), Instrument Acoustics and Automatic Mixing.

2.2 Scientific Context and Analytic Research

Drum kits have been the subjects of extensive study in the academic field. Among researchers, most analysis addressed matters as the sound they produce, their interaction with other instruments in music and their pattern recognition inside songs. We will expose some recent research progress and strategies that have drum kits and percussion instruments as their main focus.

Still, we must first make mention to what we will be debating on the following sections. Lerch (2012, p. 1) stated "*Audio Content Analysis* (ACA) is the extraction of information from audio signals such as music recordings stored on digital media". Over the years, the creation and improvement of methods and tools that extract and analyse such information, have enabled researchers to understand sound more comprehensibly. Each of them has specific objectives that range from broader to a focused approach on the subject at hand.

The present study focuses mainly on the instantaneous features of sound, i.e., characteristics that do not necessarily define a sound as musical or complement the perception of the listener, but significantly allow for an understanding of its behavioural patterns. That is the basis of sound as a physical and perceptual auditory experience. Some of those features include the frequency response domain, time domain, acoustical properties of rooms and materials and psychoacoustic auditory perception. For this research, we will be focusing on time domain and the frequency domain.

Furthermore, we aim to find an idyllic way for an easier process of recording drum kits inside the studio. With this clear objective in mind, we made an analytical research on subjects that contribute greatly to understanding the instrument's behaviour in the studio and its intrinsic characteristics. Still, an investigation on their effect and qualities can complement and aid in introducing a standard and semi-automated approach on recording in the studio. Since we are dealing with a rather seminal piece of work, many of the research gathered for substantiating it does not directly respond to our research questions. Yet, they hold great importance when deciding the subjects to research as well as the concerns we intend to approach on our methodical analysis.

Rossing (2001) has much emphasized "drums have played an important role in nearly all musical cultures". Upon his research on instrument acoustics, he presented a systematic approach to divide percussive instruments into two fundamental categories: membranophones and idiophones. This first includes mostly the resonating drums (which have membranes, thus the name), while the second category comprises mostly cymbals and other metal instruments. He further introduced two distinct pitch-related category classes: definite or indefinite.

Following this line of thought, he discussed the reasons why some membranophones (which vibrate in many modes, i.e., with many harmonics and partials) such as the orchestral timpani or the tabla, can "convey a clear sense of pitch". In the example, it is now known that the timpani achieves a tonal pitch by shifting the "inharmonic partials of a membrane in vacuum into a harmonic relationship", due to the "mass loading by the air in which the membrane vibrates".

On the other hand, the "indefinite pitch" drums, such as the snare drum, present an alternate way of producing sound due to its coupling resonating properties and the contact between the lower membrane and the snares (strands of wire). The coupling resonance effect occurs due to

the enclosed air within the snare and the outer shell, creating pairs of modes of vibration. In fact, Rossing (2000, p. 26) stated that indefinite pitch drums (which include most membranophones used in drum kits), do not or "convey a much weak sense of pitch".

On the second fundamental class, the idiophones, Rossing (2001), found that a single stroke could produce 100 modes of vibration on a crash, while studying their effect on cymbals. From this test, three prominent features could be retrieved: "the strike sound that results from rapid wave propagation during the first millisecond (ms), the buildup of strong peaks around 700–1,000Hz in the sound spectrum during the next 10 or 20 ms, and the strong aftersound in the range of 3–5kHz that dominates the sound a second or so after striking and gives the cymbal its 'shimmer'."

He also demonstrated the nonlinear propagating effects displayed by cymbals when struck by a beater, specifically stating that the conversion of energy from the excited low-frequency modes results in the high-frequency vibrations for which cymbals are characteristic. In addition, they exhibit a "chaotic behaviour" which is led by the initial harmonic generation and followed by subharmonic generation. Moreover, he pointed out that this nonlinear behaviour of cymbals is much more difficult to replicate synthesis-wise, a reason why percussionists are asked to use real cymbals when using electronic drum machines.

Upon further research, Rossing et al. (2004) presented a basis for physical modelling of cymbals employing mathematical analysis. Studies showed that there were "between 3 and 7 active degrees of freedom" when "using nonlinear signal processing methods", the same number of equations for physical modelling. He gives one possible procedure by calculating the Lyapunov[5] exponents "from experimental time series, so that the complete spectrum of exponents can be obtained".

Following Rossing's (2001) studies on the acoustics of percussion instruments, Toulson et al. (2009) presented a paper regarding the perceptual importance of standardising drum kit tuning both in a live and studio context. By employing a methodology of interviews to professionals and performing digital signal analysis, they explained how, frequency-wise, one could create bandwidth boundaries to separate independent elements within the drum kit. This lack for a definite and standard tuning can present itself as a fundamentally creative characteristic of the instrument and a menacing and frightening task for the post-production engineer.

They stated, "Musicians and producers will spend a number of hours achieving a preferred drum sound prior to a performance". This is however a "rather subjective matter" since no such preferred standard tuning has been defined for the drum kit unlike most other conventional instruments (such as strings, keys, winds, etc.).

From their interviews, Toulson et al. realized that a drum setup to achieve the desired drum sound prior to a recording "can account for 15-25% of the entire project". This is not a minor point, as the music industry economics is a factor to take into account, since no record label is "willing to spend large funds on recording projects". Nevertheless, the researchers further stressed that a good drum setup can enhance the swiftness and simplicity of a recording session, and still allow possibility for creativity according to genre.

[5] In mathematics the Lyapunov exponent or Lyapunov characteristic exponent of a dynamical system is a quantity that characterizes the rate of separation of infinitesimally close trajectories.

They also gave an example of different tuning in terms of genre, as drums, in Jazz music, are "generally tuned higher and with longer decay than drums for Rock music". Pitch correction in post processing of drums can also lead to problems, since "it is only possible to enhance frequencies that are evident in the original audio signal". This then leads to the tedious and very common process of replacing the original by pre-recorded samples.

On their research they also took into account the testimony of drummers divided into three groups: advanced, novice and hobbyist. The opinions of members from each group were generally similar. In the first group, professional percussionists regarded tuning as "an essential part of their craft" and that "they can always tune a drum kit to a desired sound by ear alone, but they might not be able to achieve exactly the same sound every time". On the other hand, novice drummers found tuning "very challenging" preferring to "concentrate on their playing technique"; still, they were aware of the importance and benefits of a precise drum kit tuning. The third group only had a general idea of the significance drum-tuning can have in a drum kit, and admitted their inability to perform such task.

After discussion with performers, Toulson et al. presented a quantitative way to tune drums. On their research they investigated fundamental aspects to have in mind while presenting a standardised way to define pitch in drums. The first one is the drum's behaviour, where they outline the centre and edge impulse responses (both dynamics-wise and frequency-wise) of a selected tom drum.

On the centre, they found the fundamental frequency to be 147 Hz (D3 on the musical scale) while on the edge the most common had a frequency of 220 Hz (A3 on the musical scale). The first one relates to the displacement of air inside the drum shell, and is dependent of its size and tension between batter and membrane. The second frequency, on the other hand, is solely dependent of the "dimensions and tension of the batter head alone" since it is "more localised". They then finally proposed tuning to a specified fundamental frequency. This means "performers and record producers can tune an entire drum kit to a musical scale or reference" facilitating the drum kit setup process.

The dynamic envelope of drum sounds are also deeply covered and debated. While a drummer prefers a higher decay time in the drums, that is able to sustain the tone, the producers favour faster decays since they worry about the cohesion of the drums with the rest of the instruments – many producers also try to correlate the decay of the drums with the beats per minute (BPM) of the song. A precise tuning "allows the decay of the response to be in key with the music, so longer decay times can possibly enhance a recording." Other tuning factors and their effects on the drum performance were also considered, such as drum dimensions, materials, head types and cymbals.

These conclusions, albeit qualitative are pertinent in the context of a rather subjective topic as this. Moreover they allow further expansion into inherent and external variables such as music genre, drummer playability, drum kit manufacturing, among others. Adopting a multiple definite drum kit tuning for variable purposes could expand this tactic.

A form of standardised tuning would be a benefit for several areas of study. Over the years, the field of MIR has tackled questions and defined strategies in drum transcription and pattern recognition. Some of those techniques are discussed here for their added interest and for setting boundaries to this research.

Concerning drum recognition, Sillanpää (2000) presented the theory that "members of different drum classes differ from each other by frequency content". The classes proposed

were divided as follows: kick drums, snares, tom-toms, hi-hats and cymbals. By having different samples analysed with different time-frequency characteristics he was able to determine a model for a shape matrix of each sound that could enable pattern recognition.

His study met 87% success rate in accuracy for detecting isolated drum sounds, but as the number of simultaneous sounds increased, the accuracy rate decreased (49% for two and 8% for three simultaneous sounds). This clearly shows that frequency-based drum recognition possible if we are dealing with one single instrument. Yet, when mingling two or more different sources results become unclear.

Ensuing this, Herrera et al. (2002) proposed a more advanced method to categorise the instruments. There were two super-categories: membranes and plates (drums and cymbals respectively). Then they sub-divided in Sillanpää's classes and renamed them "basic-level"; they introduced sub-categories in the toms (high, medium and low), the hi-hats (open or close) and in the cymbals (crashes and rides). For sample control purposes, they disregarded "deviations from a 'standard sound' such as brushed hits or rim-shots"(Herrera et al., 2002, p. 71).

For their analysis, Herrera et al. considered factors of attack, decay, relative energy and Mel-Frequency Cepstral Coefficients[6] (MFCC) to allow for more accurate pattern recognition. Their methodological approach met with extremely high rates of success in recognition when mixing the super-categories (99%). Basic-level (kick and snare) mixing as well as basic-level and sub-categories mixing also met an exceptionally high level of success (97% and 90% respectively). From there, they paved the way further to automatically recognise different drum kit elements even when mixed together.

On the other hand, Yoshii et al. (2004) stated that "automatic description of contents of music is an important subject to realize more convenient music information retrieval" and that the "characteristics or typical drums patterns are different among genres (e.g., rock-style, jazz-style or techno-style)" (Yoshii et al., 2004, p. 184). They further pronounced the importance of drums in contemporary music by considering the role of the drummer and the mood of the song as a differentiating factor in the sound of drums. It is his playing ability that creates the groove/swing of the music that adds to its emotion and feeling.

This presents an obvious setback, concerning the findings of Sillanpää and Herrera et al., since their theories, do not consider deviations from the standard drum sounds, nor do they include the human factor.

Yoshii et al.'s method of pattern recognition also differed from the previous as it was verified in actual musical context. Instead of mixing samples, they used a free research-oriented music database[7] to gather ten music excerpts available online.

Their approach resorted to two different methodological approaches: "base method" and "adapt method". The first one is a "single seed template" that corresponds to each piece of the drum kit, while the second had a custom made algorithm of "template-adaption-and-

---

[6] "The MFCCs have been widely used in the field of speech signal processing since their introduction in 1980 and have been found to be useful in music signal processing applications as well. In the context of audio signal classification, it has been shown that a small subset of the resulting MFCCs (...) contains the principal information — in most cases the number of used MFCCs varies in the range from 4 to 20." (Lerch, 2012)

[7] RWC Music Database is a database with popular, classical and jazz music "available to researchers for common use and research purposes" (Goto et al., 2002).

matching" that identified the onset of each drum using a corresponding template. The onset times of the recognition tests had to be posteriorly hand-adjusted.

The results of the "adapt method" met with a significant increase of pattern recognition when compared to the "single seed" base method both in the kick drum (the F-measure rose from 0.67 to 0.90) and in the snare drum with a smaller but still significant rise of success (0.74 to 0.88).

Although Music Information Retrieval deals extensively with the way to identify musical properties of songs (tempo, meter, key), these approaches of instrument identification and recognition can also be applied in other areas. One of such cases is the possibility to transcribe a piece of music into musical notation.

On this subject, Fitzgerald et. al (2003) introduced a method of source separation by calling it Prior Subspace Analysis (PSA). Fundamentally, they reason that the consistency of the recognition increases when previous knowledge of the sources to be separated is input on a machine-based approach.

This proved to be a far more consistent method when analysing multichannel test samples. In their paper Fitzgerald et. al, presented a comparison between PSA and Independent Subspace Analysis[8] (ISA). The latter proved to be less reliable as the number the sources to be separated increased. The fact that ISA deals with invariant basis amplitude and frequency functions made the separation less accurate when having two sounds at the same time, since these characteristics vary from source to source.

Despite relying on the same notions, such as "overall mixture spectrogram" resulting in "the sum of a number of independent spectrogram", that "can be represented as the outer product of a frequency basis function and an amplitude basis function", PSA differs from ISA by creating Prior Subspaces obtained by "analysing large numbers of each of the sound source of interest".

In this case, the test subjects were snare drums, kick drums and hi-hats. The results showed a similar outcome on both the snares and kick drums when using the PSA and the ISA methods (90.5% and 100% respectively). Yet, the hi-hat recognition and transcription was 5.1% higher in the PSA method. Fitzgerald et al. (2003) further explained that the amount of incorrect identifications on the snares in both cases was due to amplitude modulation (which could be possibly solved by changing the algorithm) rather than incorrect identification.

On the other hand, Gillet & Richard (2004), present a different strategy for drum transcription. They promptly stated that most studies approach isolated sounds, with a prior analysis and a posterior recognition and transcription. Their research presents several methods for drum transcription as they correspond to real-world possibilities "encountered in modern audio recording (real and natural drum kits, audio effects, simultaneous instruments,…)".

[8] ISA is based on redundancy reduction techniques, representing "sound sources as low dimensional independent subspaces in the time-frequency plane". This allows for a time-frequency representation of a single channel mixture. The overall spectrogram results from the superposition of a number of unknown statistically independent spectrograms, which can be "represented as an outer product of an invariant frequency basis function and a corresponding invariant amplitude basis function" (Fitzgerald et al., 2003).

They based their method on Hidden Markov Models[9] (HMM) and Support Vector Machines[10] (SVM). Since their aim is to transcribe drum loops, they present a system architecture based in three major parts: segmentation and tempo extraction, features extraction module and classification module.

Firstly, dividing a drum loop into singular and individual events (either single stroke or multiple stroke) and knowing that "drum loop signals consist in localized events with abrupt onsets" their segmentation algorithm (consisting of "associating a filter bank with an onset detector in each band and with a robust pitch detection algorithm") obtained "very satisfying results".

Secondly, the feature extraction was based on statistical cluster classification algorithms – K-Nearest Neighbour[11] (KNN). Gillet & Richard then evaluated the mean of 13 MFCC 20ms frames followed by the 4 spectral shape parameters (spectral centroid, spectral width, spectral asymmetry and spectral flatness), defined by the first four order moments (mean, variance, skewness and kurtosis respectively). Finally, the signal was divided in 6 band-wise Frequency content parameters "chosen according to a meticulous observation of the frequency content of each drum instrument".

Thirdly, regarding the classification technique, Gillet & Richard proposed that "drum signals exhibit some kind of context dependencies". This is because the sound originated from a stroke "may continue while the following stroke happens" and thus it may have an "impact on the spectral characteristics of the following events". As such, they resort either to HMM (since they present a viable solution that integrates context and time dependencies) or to SVM (whose design works well with "binary problems classification") algorithms.

Because of an added interest in segment labelling among the instruments, two different approaches are possible: a "2n –ary classifier" (in which only one classifier is used with each possible combination) and an "n binary classifiers" (one binary classifier is "trained" for each instrument to decide whether it is played or not in each segment). They also presented a "drum kit dependant approach" specialized in four kinds of drum kits according to genre (electro, light, heavy and hip-hop styles) because of "high variability of the data".

Their evaluation method included 315 drum loops with 5327 strokes altogether. They also identified manually the source for each stroke by instrument (bass drum, snare drum, hi-hat, etc.) and formed eight categories. For their experiments, Gillet & Richard presented two different taxonomies: a "detailed" one, "defined where each combination is characterized by a label", and a "simplified" one that "gathers some instruments in a reduced number of categories".

Results showed that the SVM approach had a significantly better performance in pattern recognition, which was justified by "the fact that the rather simple acoustic model used with

---

[9] HMM is a statistical Markov Model chain. These "chain techniques are sensitive to their immediately preceding context, so they can create contextually appropriate outcomes. Markov chains use recently chosen states to influence the probability of subsequent choices." (Loy, 2011, pp. 363-364).

[10] "State-of-the-art classifiers transforming the features into a high dimensional space and finding the optimal separating hyperplane between the closest data points; SVMs can nowadays be considered a standard tool in musical genre classification." (Lerch, 2012, p. 155)

[11] "Classifier which evaluates the number of the closest training examples in the feature space." (Lerch, 2012, p. 155)

HMM cannot cope with the high variability of the dataset." For both binary approaches SVM method achieved around 65% for the detailed taxonomy and 83.5% for simplified taxonomy. The "drum kit dependent" approach had a 60.6% average for the detailed taxonomy and an 81% average for the simplified taxonomy.

Gillet & Richard concluded that further research and work on the subject would lead to a better algorithm that could incorporate the advantages of both HMM and SVM approaches into a more robust and reliable way for drum pattern recognition.

Still, Spich et al. (2010) determined that most algorithms up to the time of their research, could be designated as low-level transcription algorithms, because "they do not rely on data post-processing for error correction". In support of their research, they showed developments and improvements on Fitzgerald et al. PSA method claiming a nearly 15% increase of accuracy in transcribing popular polyphonic songs.

After performing the PSA, their method then consists on identifying the tatum (the lowest metrical level of the tempo) and creating a tatum grid "where all the possible onset times for drum events are bound to lie". By doing this they have the guarantee "that the transcription results will always be consistent with the tempo".

They are subsequently forced to perform an error correction, based "on the identification of a reduced set of plausible patterns that best describe the musical excerpt". To be as wide and extensive as possible, the tested songs showed intrinsic variations (genre, date of release and recording techniques). The time signature of the songs was also considered for the research, and all of them showed either 4/4 or 6/8 patterns.

Spich et al., also mentioned that further extension and development in this method is possible as it was not self-contained. In fact, according to the authors, due to being a "high level technique", it does not depend on the choice of PSA and as such many other low-level transcription techniques can be easily applied. As a complement, they mention the relevance of definite pitch, since it would substantially aid in an automatic way to find the most rich and interesting drum sound possible.

On the other side, more practical fields of study would enrich our research with concepts and notions from which we can draw influence. One of such areas concerns automatic mixing, which is, without a doubt, an ever-growing area in terms of research and economy.

Automatic mixing is quite blatantly and "emerging field of multichannel audio signal processing where the inter-channel relationships are exploited in order to manipulate the multichannel content" (Reiss, 2011). As such, in our search for idyllic drum sounds, our study can benefit from the characteristics automatic mixing offers, as being able to stipulate several aspects of a mix such as EQ or compression and achieving them automatically.

The drum kit presents unusual characteristics that have also been subject of research and study, unlike most other pop music instruments, such as guitar or bass guitar or even keyboards, and, because of this, it needs special attention.

Terrell & Sandler (2012) developed a research for an automatic way to create monitor mixes in a live performance context. In their research, the approach they took regarding drum kit drum kits in this context was to separate each individual piece of the kit and considering them singular instruments. They chose the methodology because the usual drum kit recording session encompasses a multi-microphone setup with multiple discrete audio signals, each

containing significant amounts of bleed (the amount of sound from near instruments that enters the instrument's microphone).

On a previous research with Reiss (2009), Terrell determined the Root-Mean-Square[12] (RMS) Sound Pressure Level (SPL) model for each instrument independently at the performer's "point of view" by "combining RMS SPLs of the acoustic signals from the direct signal path and the reinforced signal path in anechoic conditions". Despite several carefully placed boundaries to avoid feedback and maintaining RMS SPL within the fixed limits, they found that on the mix for the drummers location was far from the expected target.

Still, Terrell & Sandler devised a model taking into consideration the acoustic sources of sound; for example, they mention the amplifier of a electric guitar as being part of the acoustic response. Also speaker radiation (cardioid) was considered as a purely acoustic response, as was the case of the drum kit, which emanates in an omnidirectional pattern.

To substantiate their assumption, Terrell & Sandler used their model in a practical case study. They used a venue (with similar acoustical properties to those of a usual concert venue) with individual monitor mixes for each of the four performers and a front-of-house (FOH) mix for the audience. When performance was finished, they analysed the RMS SPL for each of the seven individual instruments in each of the five mixes.

After considerable analysis and error correction, the authors concluded that it was possible to deliver "properties of multiple target mixes to multiple locations on a venue" without defining an exact RMS mix individually. Moreover, they verified the preference drummers' show for listening to mainly rhythmic instruments (such as bass guitars) on their monitor mixes, somewhat neglecting the melodic elements of the group.

On a similar subject, Scott & Kim (2013) stated that the advantages that digital recording and editing tools convey have "led to a desire for increased automation and efficiency". They also mention the level of expertise and comfort in using such tools may cause "many new-comers from obtaining reasonable results even with a significant amount of effort." This has led to an increased research in the field of automating the process of "analysing audio and improving the perceived quality."

Their study focuses specifically on the drum kit and how the application of basic "guidelines" in the spatial positioning in the stereo image and spectral domains can make the sound "more balanced". Much like Reiss's and Terrell & Sandler's methods, source separation is vital for this type of endeavour. On their model approach they implemented time domain and spectral features to have a broader mechanism of adjustment with basis "on the temporal and inter-track relations between the features" as well as psychoacoustic models of loudness and frequency for an anatomic and perceptual approximation to reality.

The size of the corpus for their tests had a significant added value to the research. They also included a highly relevant number of genres (acoustic, dance, jazz, rock, etc.…) and a sizeable pool of practical cases used (135 multi-track songs acquired in the internet for educational and investigation purposes). Professional and student opinions were extracted from interviews and were the human factor that complemented the proposed approach on automatic drum mixing.

[12] "The RMS is one of the most common intensity features and is sometimes directly referred to as the sound intensity." (Lerch, 2012, p. 74)

The tests' features measurements included three processing areas: level balance, stereo panning and EQ. As the first implied caution when balancing the mix, they opted for a parallel fader approach in which each individual track was evaluated against the rest of the mix. In addition to this, the decision to include overhead (OH) or room microphones called for a cautious method. They used the close-microphone recordings as primary sources for the mix.

On their model, spatial positioning of sound sources was a less difficult task to account for. They used current and standardised professional techniques used in stereo recording. These included a centre snare and kick and spread the toms and cymbals along the stereo image. After first determining the source position on the stereo image of the OH microphones, they then, according to their needs, moved the primary sources to their position.

EQ applied was deemed "minimal" and was the fruit of research "obtained from the interviews of engineers". They also mention the engineers' issues "about making generalizations without hearing the source material". Still, Scott & Kim developed a generalized filter scheme, whose purpose was to "boost frequency ranges that often need boosting and cut frequency ranges that often need attenuating". From their point of view the ideal approach would be an adaptive tool for comparing bandwidth "energy ratios" and "making adjustments accordingly".

For the fulfilment of this machine-based approach, a necessary identification of each individual drum source was inevitable, for the algorithm to precisely mix the drums. They used a SVM classifier with several features (time and spectral domain related and MFCC) to be identified via a radial basis function[13] (RBF) for four classes: kick drum, snare drum, tom-tom and overhead. The major problems encountered concerned the bleed from other instruments on the multi-track material, specially coming from the tom-tom tracks' primary source, due to the "relative infrequency" that it is used.

For their model of evaluation, a listening test was devised. The participants were asked to select their preference between a monaural sum and a mix generated by the model. The results showed that in six of the ten examples the automatic mix was favoured over the sum. Scott & Kim's approach on automatic drum mixing employs significant techniques which "based on prior knowledge of instrumentation" propose an introductory method for drum kit multi-track automatic mixing.

Notwithstanding, these research areas that have led to newer and more reliable procedures and tools for recording and post-processing sound, there are still many plug-in manufacturers and sound library companies investing extensively in research and development of high quality sample recordings.

Thanks to the long-lasting Music Instrument Digital Interface (MIDI) protocol, many musicians have opted to use those commercial samples in their music. Alongside with Artificial Intelligence (AI) agents, the field of automatic decryption and recreation of the human factor in music has enlarged. Drum programming is ever more present in music of recent years. Its simplicity and substantially lower cost makes it an obvious choice for many musicians.

[13] "The radial basis function method for multivariate approximation is one of the most often applied approaches in modern approximation theory when the task is to approximate scattered data in several dimensions". (Buhmann, 2003, p. ix)

Tidemann & Demiris (2007), explain that despite such advantages, virtual drum programming commonly lacks "the *groove* that a human drummer will provide" since they "will always have small variations in both timing and velocity". Their research implies the creation of an automated learning AI agent that extracts rhythm of a human drummer and becomes able to replicate it virtually. By adding a humanizing factor to a programmed drum and using highly developed sample libraries, music can substantially become better in both quality and emotional involvement with the listener.

Despite claiming an automatic musical extraction of the groove, the authors chose, for initial trial purposes, to directly input MIDI patterns on the robotic model. They claimed this way to be a "tempo-less representation", and could, therefore, "be played at a different tempo than it was demonstrated".

By operating in a bi-scalar system (small-scale for main beat and large-scale for deviant forms) the machine model was easily able to grasp small simple patterns (e.g. verse) as well as other filling patterns (such as chorus and fill-ins).

Also, they devised a test in which the authors solicited the skills of three different drummers by playing the previous drum patterns with a metronome (verse, chorus and bridge with variations). Each drummer's groove was subsequently integrated in the software. Finally, to compare their results, 18 participants were asked to listen and identify the drummer of the loop their were listening; they first listened to a 15-second drum sample of the musician and then the generated samples.

The results met positive success in identification (94.4% for drummer A, 88.9% for drummer B and 83.3% for drummer C) and showed a pertinent way to demonstrate that human listeners can differentiate different drummers by their playing style.

2.3 General Considerations

The present research began with the idea of devising a faster, reliable and automated strategy to record drum kits in the studio. Towards this goal, we defined an exploratory approach that implied a research and analysis of the practical work of worldwide notorious recording engineers. Their knowledge and experience has steered the creation of important samples and construction kits that have been used for music production all around the world.

Over the years, the easier access to these tools and technologies allowed the growth of independent musicians and producers to create and innovate the world of music. Yet, at the same time, this access could also be the cause for the downfall of music. The limited know-how many emerging musicians and producers demonstrate has created a growing market for automated music post-production tools, which have been the subject of much research and development.

Furthermore, the need for cataloguing music has never been as evident as now. Tools such as ID3 tags and music fingerprinting as led to the development of applications (*Shazam*, *Soundhound*, *LastFM*, *Echonest*, etc.) that instantly recognise and find the information of music on the information grid. Additionally they even advise similar or possible matching artists to the liking of the user.

Despite, being a rather seminal work we have tried to organise how drum kits have been included in the research field on these past recent years. In fact, targeting rather specific aspects of sound is a challenging effort, and, for that we have tried to analyse as many approaches as researchers have seen relevant to investigate and develop over the years.

Yet, although this research deals with pre-established and accepted notions on the field of Audio Content Analysis, in the end the results we are trying to achieve tend to be considered highly subjective. Therefore, we make prior notice that this research and its results should not be considered dogmatic in a recording session. In fact, the human factor is essential for any creative endeavour. Humane critical awareness is essential for an artistically credible result, despite the developments of highly realistic psychoacoustic models that "exploit of human perception" (Brandenburg et al., 2013).

Still, although the questions raised are many, we will be focusing our attention in the underlying characteristics engineers subconsciously achieve by objectively listening to drum kits prior to a recording session.

3 The Drum Kit in the Studio

3.1 The Engineer's Point of View – The Art and the Science

Renowned engineer Alan Parsons (2010) stated: "There is one thing that you can be absolutely certain of hearing on just every record between when Rock'n'Roll started in the fifties and the present day. And that is drums".

Drum kits have benefitted from countless changes and mutations change since they were first introduced in modern music, and, as a result, new ways to play and record this instrument have been developed. The importance and significance of the drum kits is of such high relevance that they are considered by many to be the "foundation of modern music, because it provides the 'heartbeat' of the basic rhythm track" (Huber & Runstein, 2009, p. 155).

Being one of the very few purely acoustic instruments used in popular music, it is also one of the "most difficult, most problematic, an the most misunderstood" (Major, 2014). This leads to a painstakingly process of recording.

Owsinski (2009, p. 111) stated, "Engineers seem to obsess over" the drum recordings in the studio, because "if the drum sound in the room doesn't cut it, there's not much the engineer can do to help". One of such examplesx in the recording studio was the re-location of drummers from small isolated rooms, often with "dead acoustics" (that resulted in dull and uninteresting outputs), to large, ample and reverberant rooms, where the drums would merge with the room's acoustics and create a "larger-than-life" sound (Huber & Runstein, 2009, p. 156).

Apart from the instrument peculiar characteristics and the environmental requirements for its appropriate recording, the human factor, i.e. the drummer, is a key factor. His performance holds a significant weight in the final outcome of the recording, as is denoted by Mark Linett[14]: "it's one of those instruments where the technique of the player really matters" (as cited in Owsinski, 2009, p. 143).

In fact, the drummer is of such importance that sometimes, if the band agrees to it, a session drummer may be hired to perform in the record. A famous example of such extreme conditions was the case of The Beatles' drummer Ringo Starr, whose contribution for the

[14] With over half a century of experience in the music recording business, he became deeply associated with The Beach Boys, by producing most of their albums. He has also work with many other acts, including Jimi Hendrix, Los Lobos and Jane's Addiction.

recording sessions sometimes would be "to play the tambourine instead" of the drums (Lewisohn, 1988, p. 6).

The musical richness of drum kits led Wyn Davis[15] to state that they "are sort of like an orchestra" (as cited in Owsinski, 2009, p. 142), thus expressing how the sound of the whole instrument is the result of the different pieces interaction. For some time, the collection of such diversity however was thwarted by technology. For example, in the sixties, multi-microphone drum kit recordings would be impossible due to the reduced amount of tracks available in a recording machine. In fact, while recalling the original session notes of the *The Beatles*' at Abbey Roads Studios. Lewisohn (1988, p. 54) described how instruments were recorded simultaneously (without overdubbing[16]), and, sometimes, on the same track (e.g. drums and bass guitar).

Nowadays, and thanks to the technological evolution, multi-microphone recording is possible with an almost unlimited number of tracks available to record a single instrument; engineers have the time and possibility to ponder and decide on specific layouts for specific projects. Nevertheless, despite the possibility of applying microphones to virtually every individual element (thus, widening the possibilities in the mixing stage) many engineers tend to prefer a minimalistic approach to recording the drum kit.

With the evolution of recording machines, so have the microphone designs evolved. Today, manufacturers create microphones with characteristics accounted for specific purposes, in which they take into account built-in frequency response, SPL threshold, polar pattern and membrane type. For example, the kick drum microphone variety was designed with this in mind, as is the case of the classic and widely used AKG D112 microphone, whose frequency response is depicted in Figure 3.1.

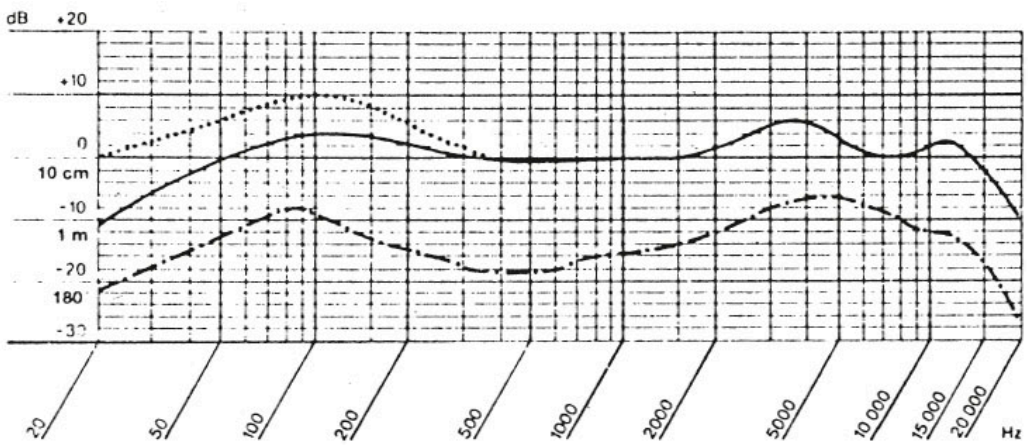

Figure 3.1 *AKG D112* frequency response curve (extracted from service manual)

At a glance, in the response curve, one can see a significant boost in the frequencies around the 100 Hz mark (where the fundamental frequencies of kick drums are commonly found in modern tunings). We can also observe a second boost in the 5 kHz to 6 kHz zone, that helps

[15] California based sound engineer with a prolific discography curriculum in the Hard Rock genre, having worked with such groups Dio, Dokken and Great White.

[16] "The overdubbing process used widely in multitrack recording requires musicians to listen to existing tracks on the tape whilst recording others." (Rumsey & McCormick, 2009, p. 178)

for the *click*/*snapiness* of the kick, and a drop around the 300 Hz to 600 Hz zone, where usually one can find the "dull" sounding zone of the kick (Huber & Runstein, 2009, p. 159). Other visible and significant aspect of the curve is the "*High Pass Filter's*" slope factor that allows for the recording of low-end sub-bass frequencies but also attenuates them, allowing for a less "muddy" recording.

This is the objective data that the analysis of drum frequency charts may provide us. But, that is only one intervenient feature in the process of recording a great drum kit piece of sound. In fact, as small details may take long time to improve, some engineers may spend hours (or even days), choosing the adequate drums and cymbals for the recording session they have at hands. After recording, even longer time may be necessary to master the quality of the sound collected, to the level that our sensitivity may understand it as near-perfect.

The science of sound recording depends closely of the relationship between an engineer and its knowledge of a certain instrument. In other words, it is their experiments in the studio. Knowledge of recording, therefore, comes *a posteriori*, with the experience of countless hours of trial-and-error efforts in achieving the "larger-than-life" drum kit, as Huber & Runstein have mentined.

Although the topic may be considered rather subjective, the reality is that on account of common external factors (time, budget, room response, etc.), many engineers follow simple guidelines that have been laid down to them, previously by their peers. In most drum recording sessions, these guidelines are sure to work well.

For example, one commonly used technique in contemporary recordings is the top-bottom double microphone technique on the snare drum (Figure 3.2), which provides clear recordings of both the skin on top of the drum and the shaking of the snares below it, adding "more of the rattling" (Savage, 2011, p. 98) without the problems of phase cancelation.

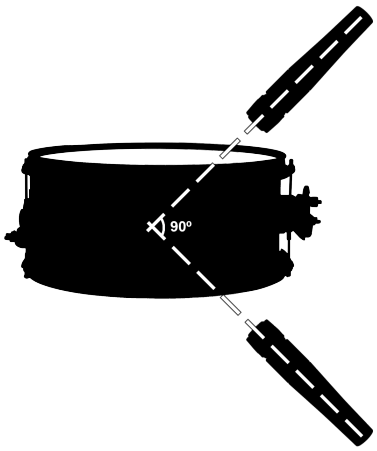

Figure 3.2 Representation of the snare up-bottom double microphone technique

This technique is achieved by using two microphones with identic properties, such as the Shure SM57, "the standard snare microphone for years" (Owsinski, 2009, p. 160), in an opposed 90º angle position. As the lower microphone in the drum kit setup is the only one pointing upwards, it may sum destructively with the other microphones. For avoiding this, polarity inversion of the signal is used (Savage, 2011, p. 98).

In contrast, kick drum microphone placement offers many possibilities for retrieval of different outputs. The big drum shell behaves differently in certain positions; in fact, much

like the snare drum some engineers opt to use microphones in both the front head (the one that is opposite to the drummer) and the inner head, outside of the shell pointing at the beater. This way the recording engineer guarantees both a resonating sound from the shell and a sharp click form the beater.

Boundary microphones, or pressure-zone microphones (PZM), are commonly used for additional kick drum recording. By having an omnidirectional capsule, close proximity could have a disruptive effect on the recording; yet, by being a pressure microphone, such consequences are avoided. For example, they can capture the kick's transients with little interference from reflected waves in the floor (Borwick, 1990, p. 103).

Furthermore, Ross Garfield[17] (as cited in Owsinski, 2009, p. 152) mentions other less conventional methods, such as using blankets, bath towels or pillows inside the kick shell, which allow for the muffling of highly resonating sounds; this adds further control for the engineer on how he wishes the output of the kick drum to be.

Regarding cymbal recording, engineers also use microphones placed in specific points that help for a clear capture: they are called *Overheads* (OH). For the OHs, engineers opt to use two or more microphones, which in spite of their physical separation, may allow for a better capturing of both the cymbals and the instrument as a whole. In stereo recordings they opt for either a coincident, near-coincident or non-coincident[18] pair setup and, depending on their choice, they achieve a wider or narrower stereo image on the mix (Eargle, 2005, p. 270).

As part of the empirical and also subjective character of sound recording, each engineer develops his own methods and techniques for a specific instrument and for specific recordings. For further comprehension on the subject at hands, we should perceive engineers' thoughts and acts inside the studio; we must be aware of their methods and techniques, their objectives and concerns to best capture the sound of a drum kit within the recording room. The most effective application of these methods depends extensively on the creativity of the engineer and his search for a certain standard that he holds as the correct for that specific situation.

It is our intention to analyse and compare these different technical traits each drum record presents and find common elements in their resulting sound. As Major (2014) defined, in drum recordings "*there is no right or wrong*. It's about what the artist feels in the context of the song".

[17] Popularly known as "The Drum Doctor", owner of the homonymous drum kit shot. According to Owsinski, "his knowledge of what it takes to make drums sound great under the microphones may be unlike anyone else's on the planet". He has tuned and prepared drums for many multi-platinum artists, such as Michael Jackson, Metallica, Alanis Morissette, etc.

[18] The coincident, near-coincident and non-coincident terms relate to the proximity of the microphone's membranes from one another during the recording. An example of microphone setups for each of these cases would be the Mid-Side (MS), ORTF and AB respectively.

3.2 Database: Choice and Development

The development of a sound sample database was a crucial point to cover during the early stages of our research. It was our intention to be broad and wide, in order to cover as much ground for variants as possible. Also, the samples had to have significant impact in modern music making so that higher relevance could be granted to the information that we were about to extract.

Therefore, the first idea that came to mind was to gather up resources for the research directly from the important names of the industry. This assumption headed the attempt to approach and contact world-renown recording engineers with highly acclaimed commercial and technical statuses that have work with influential recording artists. Therefore, for this research to be relevant in both commercial and academic fields, high-level professional contribution was required.

We then devised two similar, but still distinct, paths from where we could gather samples for our sample pool. The first, the long route, was to contact prestigious recording engineers and request recording samples from their most famous works; the second, more direct, was to select the most common and well-reviewed, commercial sound and banks and sound libraries, and extract the samples directly.

Both options would bring advantages to the research, albeit different. On one hand, the first approach would enrich our library, endowing it with different recording processes for different music genres, different choice of equipment, different room response, etc. On the other hand, commercial samples from sound libraries would have the advantage of providing specifically recorded instruments, in optimal conditions, allowing for a fast implementation on a song.

A third option was initially considered. There have been many cases of isolated drum recordings on highly commercial albums. We had contemplated the possibility of extracting such cases for our research but it was soon discarded since this was already a post-mastered material, which included heavy amounts of post processing. Furthermore, the lack of isolation among instruments brought bleed, which caused severe problems in the methodology we were developing.

We then proceeded to face the first two options for the first developments of our sample database.

3.2.1 Sample Gathering

To tackle the first approach, we started to build a solid, strong contact database and established two primary conditions for its development. Firstly, the recording engineers had to be directly involved in the recording of exceptionally high valued commercial artists or groups (sales-wise).

Secondly, they had to have been nominees or recipients for technical and excellence awards (Grammy, TEC, Juno, etc…). From these two conditions, we were tackling both the commercial and the technical sides of the music recording industry.

In its first sketches, our database included the name of 45 recording engineers, well-versed and experienced inside the studio. They were, at the time, alive, active and corresponded to one or both of our conditions. We initiated the first contact with the subjects of the database via electronic mail.

From these initial 45 names, and after extensive research on the Internet, we found no way to contact 2 of them, as they had no information concerning their contacts. Furthermore, 3 others were unreachable through the e-mail addresses existing in their websites/personal pages (we receive a notification that e-mail delivery had failed).

From the ramaining 40 names, 27 had an e-mail address available on the Internet (either personal or from their studio/company), while the remaining 13 were reachable by contacting their managers.

Still, a reply came from just 11 of those 40 names we had initially contacted. But still, not all replies were positive; in fact, only two of them were able to contribute with their samples (both of European origin).

Of the last 9 replies, all from North America, 7 informed us that they could not to contribute to our research, since they had been legally advised not to do it. They reasoned that since they were not the owners of the rights of the music, they could suffer judicial repercussions from record labels (owners of the rights). Finally, the remaining two replies were sent from managers informing that the recording engineers in question would not be available to respond to our request for personal reasons.

Considering this setback, we decided to renew our name database, primarily focusing our attention on the Europeans. Since we had had positive reply from two European recording engineers, we believed our chances to increase our still extremely small sample would be more favourable on that part of the world.

We targeted an additional 20 names from European countries, and, although this second round of requests was swiftly sent, the number of favourable responses was no different. The 4 names that did respond to our e-mail expressed their apologies for not abiding to our request, since most of their work was either legally protected, or was no longer available to them. With such limited database, we were forced to abandon this first approach to expand our sample database. We then concentrated our efforts on the second option and began our research for sample manufacturers.

The main reason for our decision of including pre-recorded commercial sound samples in this research is emphasized by the giant growth that music production has shown during the course of the last years. In fact, it is very common nowadays to employ sample libraries and construction tool kits such as step sequencers in music production.

On the specific case of drum samples, these are commonly used in the process know as re-drumming. It consists in overlapping (or totally replacing) samples and mixing them with the original drum track.

Also, like many live sound engineers, the use of triggers during the recording session can save much time. These are transducers connected to the membranes of drums, which when hit, send information to an electronic drum machine that plays a sample. This allows for pre-recorded samples to match with very small latency (almost imperceptible) when the instruments that are triggered are struck.

Figure 3.3 A drum trigger manufactured by *ddrum* (http://www.ddrum.com/)

With the obvious use of sampling in modern music and because the previous method had produced irrelevant data, we decided to limit our sample database only to commercially released samples. In fact, this process would bring some advantages over the previous, while in turn we would have to put some questions on hold.

There are two major facts that support our decision to include sample-based sounds. The first, deals with the fact that companies hire noteworthy names in the music industry for the recording of samples (some of the names we had initially contacted had signature series on some library manufacturers). Their usage in the music industry is also relevant, as they provide a viable and significantly cheaper option than booking a drum recording session.

The second reason concerns their recording process – one instrument at a time. This detail simplifies our analysis process, since the samples do not include bleed or resonating artefacts from other instruments (this leads to common problems in the editing process, regarding phase consistency of multi-microphone drum recordings). These factors could compromise a clear analysis and precise reading of the qualities of a sound sample.

Furthermore, a third minor, but still significant aspect that brings relevance to our analysis is that these samples are designed to be ubiquitous, and although they are some times genre-specific, they blend with most styles and music genres.

After some research on the Internet, combing the most widely used and best-reviewed drum sample libraries/plug-ins, we determined a list (Table 3.1) of five manufacturers, their products and the initial number of samples that were extracted (a total of 481).

Toontrack **Superior** **Drummer 2.0**	***FXpansion*** **BFD**	***Native Instruments*** **Studio Drummer & Abbey Roads Drums**	***Steven Slate Drums*** **Platinum**	***XLN*** **Addictive Drums**
132	83	107	120	39

Table 3.1 Manufacturers and respective products and number of samples retrieved

Although five seems a small number for a generalist supposition, the chosen library manufacturers included variants within their drum classes, such as drum size and manufacturer.

Most types of sample libraries are commonly implemented in a DAW session through the use of a specific plug-in, whereas some allow for third-party software to do the implementing. In such cases, we used Kontakt 5 (2012) to proceed the extraction.

The DAW we chose for the duration of the process was Pro Tools 10 (2011) which proved to be useful later, to maintain plot similarities. The software contains a function called *Tab to Transient*, which aided (with relative precision) in the identification of the initial transient[19] of the sound signal. From this we were able to guarantee similar plotting along the timeline by aligning the onset start points for each of our sound files.

However, in some cases the software considered the transient point posterior to a visible starting point in the sample. We then proceeded to change them manually, and so, because of this, a 10-sample error margin was considered.

With the samples extracted, we performed a new triage to determine the samples' value to the research. We decided to constrain some of the variants by specifically attacking the most common drum kit elements and the most common ways to play them.

3.2.2 Procedural Approach to the Drum Kit Classes

The triage mentioned previously, was a necessary process, to ensure precise inter-sample consistency among classes. The core of this research follows the assumption that there might be mutual attributes and characteristics of sound in different recordings. Still, different ways of playing the same instruments produce very different results.

Despite existing several playing styles and mannerisms, as is the snare drum's rim-shot, we felt that these deviances would produce significantly different timbres from a same instrument. Other aspects that were considered were the diversity of beaters available (sticks, rutes, brushes, mallets, etc.) and their construction materials (wood, carbon fibre, plastic, etc.). These factors may cause abrupt and distinct changes to the sound produced by the struck instrument because apparently irrelevant things, such as a slight change of intensity of playing, can cause the slightest of differences (small microtonal changes) to alter the whole output.

With these issues in mind, we decided to define a set of boundaries that assured a sensible and consistent methodology for analysis of our samples. First, we opted to focus the analysis on the most common drum kit elements, namely kick drum, snare drum, toms-toms, hi-hats, crash cymbals and ride cymbals. Although other drum kit elements, such as cowbells and chimes, are often used in the most popular forms of music, we decided to exclude them, because of their lack of universal usage and impact of the superscript classes. Also, the choice of playing included common hits with drumsticks.

[19]Farnell (2010, p. 90) describes transients as short and considerably louder parts of sound corresponding to an "excitation stage". Pro Tools' "Tab to Transient" tool allows for a fast identification of this feature along the timeline of an audio file.

We further separated the samples class-wise, similarly to the partition proposed initially by Sillanpää (2000) but went with Herrera et al.'s (2002) class separation. Yet, unlike them, we did not establish difference (unless far too notorious) between opened and closed hi-hats mostly because of the relevant frequency that both ways of playing are used in modern music. Also, for the ride cymbal class, we decided to go for its notorious *ping* sound, instead of a full *splashy* output, similar to the crash cymbals. Table 3.2 shows the final 468 samples, in number and genre (Appendix A 1.1.).

Kick	**Snare**	**Hi-hats**	**Toms**			**Crash**	**Ride**
			Low	**Mid**	**High**		
75	94	37	62	47	49	77	27

Table 3.2 Sample pool quantification and discrimination

3.3 Sample Analysis

The academic and research field has, along the years, defined sets of attributes and sound features for automated machine-learned extraction and analysis. In fact, as Lerch (2012, p. 5) suggested, these features are necessarily required to be "meaningful in a perceptual or musical way", or even "interpretable by humans". Still, they provide much information regarding the behaviour of sound.

The author mentions the distinction applied within ACA to best describe the results that feature extraction delivers: low-level and high-level. The first concerns the features that are imperceptible to human cognition, while the second deals with humanly established characteristics of sound and music (such as tempo, for example).

We have devised a series of process that tackle objective and subjective aspects and characteristics of sound. We intend to extract features that we have considered important for the development of a generally standardised and objective concept that assures consistency to the sound that drum kits display in modern recordings. This way we would be able to find the common ground within sample groups that could later help us achieve idyllic sample uniformity.

For this processes of organization and processing of data, we opted to use the mathematical computing software MATLAB (2012). The choice lies mainly on the nature of this research. It deals with much data and systematic processes which the software's tools aid significantly. Also, due to its wide use in this field of research, much information is available online on script-development and debugging.

Before scripting, we decided to create a uniformed presentation for our samples. Firstly, we used Pro Tools 10 to set the samples' duration identical (Appendix A 1.2.) within each of the eight classes (overall time length). This would later help defining similar time plotting for the results that were to be extracted.

Secondly, to maintain the plot consistencies, we defined a sample rate and bit depth standard for our samples. We chose to convert all samples to the orthodox music standards of 44100 Hz of sampling rate with 16-bit resolution. Furthermore, on some cases, we were dealing with stereo (cymbals classes, for example) and we decided to down-mix them to monophonic outputs, using a MATLAB script with a simple arithmetic mean equation (3.1), where C is the number of channels and i is the sample number (Appendix B1).

$$x(i) = \frac{1}{C}\sum_{c=1}^{c-1} x_C(i) \tag{3.1}$$

3.3.1 The Spectrogram Extraction Process

The first of the processes we have devised for our analysis drew its influence from the work and technique developed by Flanagan & Golden (1966). It is entitled Phase Vocoder. From a computational point of view, this technique would allow the development of a machine-based automated method for both the extraction of the sonic attributes of our samples in a time-

frequency domain, and would allow the creation of a visual plotting of that domain through spectrograms.

Unlike earlier examples of Vocoders (a portmanteau of *voice* and *encoder*), which had the single application of processing voice signals, the Phase Vocoder technique grants the possibility to speech as well as other sounds. This action is performed while preserving short-time amplitude and phase of the original source signal with accuracy. As the authors describe, this process allowed for a more "convenient means for compression and expansion of the time dimension".

Two decades later, Dolson (1986) expanded the concept and explained the potential of the Phase Vocoder technique when applied to the musical domain for "modification of natural sounds". At that time, access to this type of technology was "limited to experts in digital signal processing" because of the amount of processing power needed and the overwhelming size and price of the machines needed to perform it.

According to him, the Phase Vocoding process departs from the assumption that any given signal can be represented as "a model whose parameters are varying with time". The users can preform a "number of useful modifications" that can be synthesised, producing an original sound signal with "high-fidelity time-scale modification, or pitch transposition of a wide range of sounds".

More recently, Arfib et al. (2011) have detailed this process (Figure 3.3), as a way of representing digital signals in a bi-dimensional form (through a Cartesian chart). This leads to the understanding and processing sound in a much more comprehensive and "intuitive" way. Another significant aspect that this process grants is that it enables the user to "modify" sound "in some way and reconstruct a new signal" from the numerical variables extracted in the development of its graphical representation.

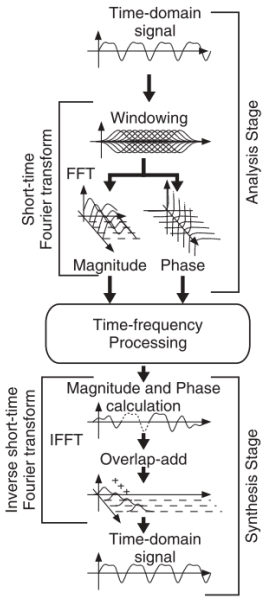

Figure 3.4 Phase Vocoder processing chain (extracted from Zölzer, 2011, p. 220)

The process is achieved through the multiplication of the input signal (in this research will be one of the samples) by a "sliding window of finite length" yielding "successive windowed signal segments". From here we have several time-frames that will continuously overlap for the length of the sample, while averaging their attributes. They are then converted to the

spectral domain through FFTs, which computes the signal depending on the window length (or resolution).

Next, the conversion from the spectral domains to the time domain implies similar processes, although, this time, Inverse Fast Fourier Transform (IFFT) functions are used. The spectrum is windowed once again and these segments are then overlapped and summed. This results in the creation of all-new sound signal. The main purposes for using Phase Vocoders usually concern performing equalization or adding effects (delay, flanger, chorus, etc.) to audio signals. On this research we will be experimenting with these concepts of averaging and morphing in the spectral domain in the attempt to synthesize an original class sound sample. Individual extraction of each sample in each class will allow for the calculation of the class average.

Our approach began by first creating individual spectrograms for each sample. Also, by doing so, we would be extracting the necessary attributes; in this case we decided to extract magnitude and phase values along the time domain. Firstly, we had to develop a way to trace the spectrogram with precise and accurate depiction of the original signal. In addition, it had to have a stable build and be able to adapt itself to minor deviances (e.g. the presence of noise, must result in equally quantifiable changes, and not a large variance of the signal). This way we would be able to achieve "a simple and concise representation of important sound properties which largely simplify the control of synthesis models" (Schwarz & Rodet, 1999).

With this in mind, we developed several MATLAB automation scripts for creating spectrogram representations (Appendix A 2.1.) of each individual sample to be analysed (Figure 3.3), processed and, finally to extract a resynthesized sound for each instrument class.

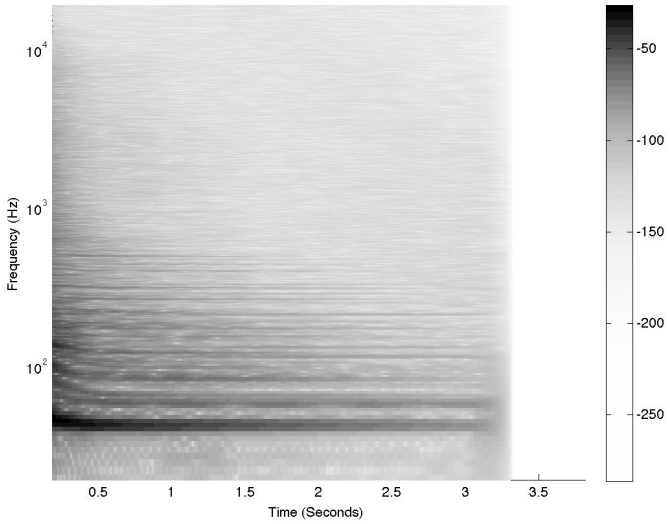

Figure 3.5 Phase Vocoder spectrogram of a kick drum

We defined a time-frequency resolution of 2^{14} samples and a Hanning Window (or raised cosine window) with equal number of points. A hop (an overlapping segment) would later be performed every 32 samples to achieve further control on the spectrogram representation, as well as to avoid first and last zero-weighted window samples of the Hanning function[20].

[20] Roads (2004, p. 255) stated that a spectrum analyser "measures no just the input signal but the product of the input signal and the window envelope". With this, sidebands and clutter are introduced in the resulting product. A way to minimize this would be "a smooth bell-shaped window" such as the Hanning Window.

Equal sample size was also a factor to be considered in the equation, as it was essential for the analysis as well as for the output resolution of the spectrogram of the samples. Despite being possible to add zeros (or silence) through code in MATLAB, we would be consuming an overwhelming amount of processing power from the computer (there was a need for repeated iterations to decide the longest sized sample). Instead, we decided to a DAW to attain same sample length, adding the silence manually. For that, we used Pro Tools 10 once again to make every sample the same length by adjusting all to the longest sized within each class group.

Through Short-term Fourier Transforms (STFT) we were able to determine the phase and magnitude of each sample for each of the several time-frequency windows previously created. During this, we defined and kept two variables named X and $Xmag$. The first indicated a complex value of the spectrum, while the second variable represented the magnitude of the source input signal.

The first variable made it possible for us to synthesize the resulting frequency spectrum, while the second enabled the possibility to create a spectrogram representation of each class sample. This further established a relative point so that the arithmetic mean of the sum of each sample's attributes could be calculated. With this, we would be able to determine the potentially idyllic frequency behaviour of a specific instrument, as well as take assumptions regarding its dynamic envelope. Finally, and more importantly, we would be able to synthesize and be able to listen to such sound.

From the resulting data that was extracted from this first automation script (Appendix B2), we mustered the spectrogram representations (*.jpg* files) and matrices of information (*.mat* files) that contained the X and $Xmag$ variables of each of the 468 sound samples.

One inconvenient we could not resolve during the script writing process concerned the renaming of the matrix files. Hence, we wrote a second script (Appendix B3) that performed this need for renaming the X and $Xmag$ variables that had been extracted. Following the renaming of the variables, the script would gather the $Xmag$ variables of an entire class in a single matrix. This would then allow for the third and following script (Appendix B4) to calculate the average magnitude of the class. From this we were able to plot bi-dimensional representation of a sound wave (with a third dimension represented by the colour fluctuation).

After some reflection, considering the time-frequency resolution of the spectrogram representations that had been extracted, we came to the conclusion that a zero-padding process could, in fact, result better (Appendix A 2.2.). As Roads (2004, p. 255) states, the zero-padding process is a method of analysing digital signals with improved resolution, while in turn, calling for an increased amount of computation and processing resources.

This could then lead to more reliable results as well as grant an easier output to understand and more faithfully represent the sound samples. On one hand, we are dealing with smaller time-windows (with frames usually to the power of two and with half of our desired FFT size), while on the other hand, the function will compute the data up to a certain number of desired samples (usually a multiple of the FFT). For example, in a FFT of 1024 samples, should we desire to compute only the first 256 samples, the following 768 samples would be converted to zeros.

With continuous overlapping add windows, which avoid zero-weight computation on the beginning and end of each time frame, computing these sectioned parts allow the graphical representation of the sound signal to be much less blurry, which in fact was later verified

when compared with the results of the Phase Vocoder spectrogram extraction process. Thus, we wrote a script in MATLAB that performed a fixed time-frequency resolution zero-padding for our samples (Appendix B5). We decided to create an FFT frame size ($fSize$) of 2^{16} samples and a window size ($grainSizeSamps$) of one fourth of that, i.e., 2^{14} samples. The continuous overlap-adds was performed eight times for the duration of the window size (every 2046 samples) with a Hanning window. Figure 3.5 shows the zero-padding spectrogram output for the same kick previously presented in Figure 3.4.

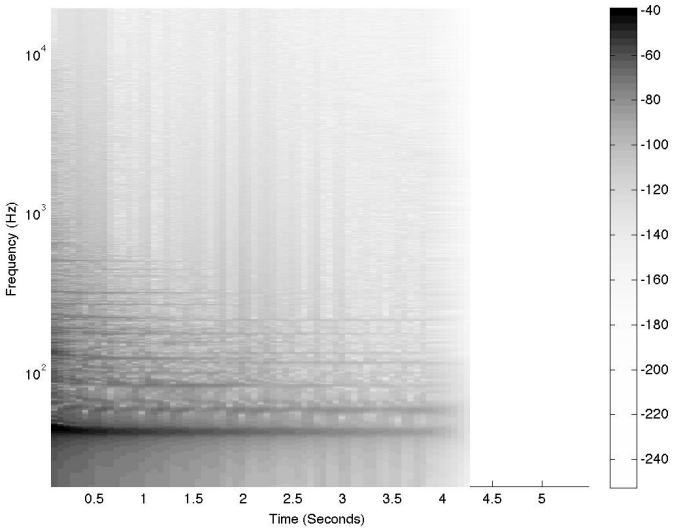

Figure 3.6 Zero-padding spectrogram of a kick drum (same as in Figure 3.4)

By visually comparing both spectrograms, we came to the conclusion that the blurriness of the first process could result in incorrect assumptions, which could lead to an unclear result and understanding of the instruments behaviour. Still, on some cases, the resolution of the Zero-padding spectrogram fell short of the corresponding Phase Vocoder one. Consequently, since we could see pros and cons in both the representations, we decided to use the two ways of plotting the spectrograms as data source for the following research processes that were devised.

3.3.2 The Spectrogram Visual Analysis Process

Given the spectrograms retrieved from the previous processes, we determined to set a number of important characteristics that could be visibly seen and that could somehow help defining standards for the classes.Furthermore, for this research, and unlike Rossing (2000) proposed, we will consider the membranophone classes as definite pitch classes, and the idiophone classes as indefinite pitch. The reason for this lies in the ability of the former classes to be tuned, and to convey a sense of pitch.

This set of visible attributes dealt with low-level features and related to how we perceive the sound of a determined instrument. We searched for a noticeably dominant frequency in the spectrograms (represented by black intensity). We assumed that, if its existence was proven, it meant that a certain instrument could be tuned to a certain pitch and so, we also looked for the fundamental frequency and associated harmonics (two or more harmonics); therefore, with these principles, we went to search for this conveying sense of musicality. Furthermore, we verified that samples that showed relevant energy activity in the sub-bass frequency zone

(between 20 Hz and 60 Hz), which is common in most forms of modern music. Finally we also tried to define a visible higher partials decay zone that appeared above the 2 kHz threshold and that roughly lasted for more than a third of the total sample time.

The devotion to these specific visual attributes of the spectrograms meant that we could have a more comprehensive understanding of the frequency behaviour of the classes. Unlike string instruments for example, drum kits do not have a conventional tuning structure. In fact, they are commonly tuned according to the type and tone of the music to be used for. Following this assumption, our search for the classes' dominant frequency and for class-specific pitch (with harmonics) we can expect to achieve a preliminary standardised ratio for drum tuning.

It is well accepted the belief that humans can only hear from a lower threshold of 20 Hz to an upper threshold of 20 kHz. Despite that, many believe that the sub-bass region is not that easily perceived by human ears. Still much study has been done on its effect on the human body (Leventhall, Pelmear, & Benton, 2003). The existence of a sub-bass region, although impossible to reproduce in most home sound systems (it requires specially build speakers, or sub-woofers) holds a driving force in the aspect of physically feeling the low end of a song. For example, a kick drum with a deep bass energy has a different impact on the listener when heard from a subwoofer on a club or in a small speaker system at home.

On the other hand, on the high-end region of the frequency spectrum, the presence of higher partials in an instrument's performance may result in dissonance or an even more chaotic behaviour. If they present high levels of energy, finding a sense of pitch can be very difficult, which then leads to less probable definition of tonality in an instrument. Higher partials are in fact expected in some cases (crash cymbals display a very chaotic motion when struck), but we must search their presence in other commonly tonal instruments (toms, for example).

We selected these features because of their relevance for this study, since they are tangible elements, whose results and assumptions can be reproduced (or taken into account) in a practical approach. One such instance would be determining an average and definite pitch. This could facilitate the process of drum recording and post-processing as has been previously described by Toulson et al. (2009). Also, by comprehending thoroughly the low-end and high-end spectrum of these classes, we can further simplify the engineer's role during the recording and post-processing. Table 3.2 discriminates the percentages of the features existence in our sample pool.

	Kick	Snare	Hi-Hats	Toms			Crash	Ride
				Low	Mid	High		
Dominant Zone (Hz)	60 – 100	120 – 350	N/A	70 – 120	90 – 170	100 – 200	N/A	N/A
Dominant Frequency Identification	61,3	41,5	43,2	100	100	100	51,9	81,5
Pitchy	30,7	42,5	24,3	100	74,5	59,2	38	37
Sub-bass	57,3	12,8	8,1	8,1	0	0	14,3	44,4
High partials decay	100	98,9	45,9	100	100	100	0	0

Table 3.3 Visible characteristics in the MATLAB spectrograms

From the extracted data, we were able to make some assumptions regarding the behaviour of the classes, which then allowed for deeper reflection in the search for reasons and answers that could suggest such results.

Regarding the membranophone classes, kick drums presented an expected significant amount of spectral dominance within the frequency range of 60Hz to 100Hz. The mean average would result in a dominant frequency of around 77,45 Hz. As such, from this we could define a standard kick drum tuning pitch of $D^{\#}_2/E^{b}_2$ (77,78Hz)[21].

This very same reasoning could be applied to other membranophones classes, since they allow precise tuning (by tightening or loosening the lugs distributed along the rim of the drum). Here, the snare drum presents a dominant spectral zone that fluctuates from 120Hz to 350Hz, making an average of 204,93 Hz, which translates roughly into $G^{\#}_3/A^{b}_3$ (207.65Hz).

Among the tom classes, a present sense of pitch was to be expected. As Toulson et al. (2009) explained, they can sometimes "be at odds with that of the bass guitar or other instruments". Still, on our visual analysis, it was evident that the higher the pitch was on the tom, the less present was the sense of pitch (from 100% confirmation of pitch in the low tom class to 75% in the mid and 60% in the high).

Rossing (2001) proposes that it could be explained because of their resonance modes. According to him, drums with larger shells and membranes end up vibrating in harmonic relationship, reducing inharmonic interference.

So, following the previous line of thought, average frequencies for the tom classes could be translated as 91,65Hz for the low (a tuning in $F^{\#}_2/G^{b}_2$ - 92,50Hz), 123,69Hz for the mid (B_2 – 123,47Hz) and 141,42Hz for the high tom class ($C^{\#}_3/D^{b}_3$ – 138,59Hz or D_3 – 146,83Hz). Musically speaking we are in the presence of a root, forth and fifth interval progression in a low to high drum sequence.

It is noteworthy mentioning that the frequencies mentioned and their corresponding musical notes were proposed by comparing the classes' geometric mean. It is simply a mathematical way of trying to define the central frequency tendency by knowing a set of n harmonics and partials (equation 3.2).

$$x\ Hz = \sqrt[n]{\prod_{i=1}^{n} h_i} \tag{3.2}$$

While still concerning to the three tom classes, a slightly accentuated downward curve was observed in the beginning of the samples' spectrograms (Figure 3.5). This could be the result of the coupling effect between the two membranes explained by Rossing et al (2004).

This physical phenomenon occurs due to the resonating modes of the membrane. When struck, it affects the static opposing membrane forcing it to move either in the same or the opposite way at a higher frequency. Thus, the downward curve is the result of the rarefaction of the air inside the drum shell after a "considerable compression" (the hit).

[21] The numbers mentioned correspond to a pitch with a reference A_4 with a frequency of 440 Hz. We chose this because it is the standard and most widely used concert pitch.

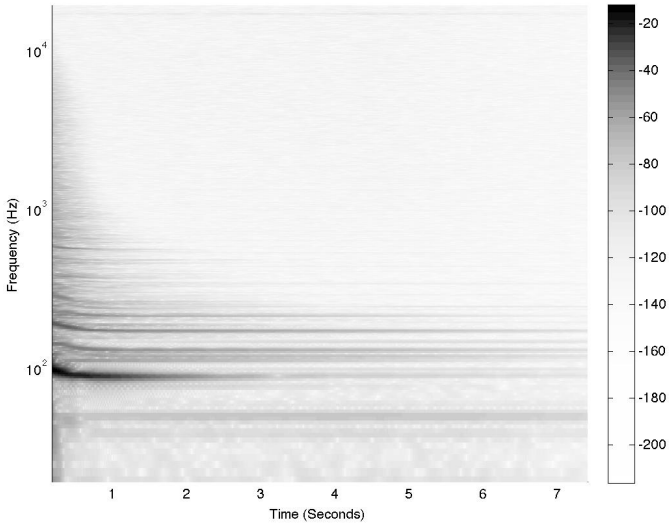

Figure 3.7 Spectrogram representation of a high tom with curve in the attack

On the idiophone classes, sense of pitch was, as expected, significantly lower in comparison to other classes. Nevertheless, the percentages were unexpectedly high for supposedly indefinite pitch classes. Still, and since we were talking about the *ping* sound that ride cymbals produce when struck in a specific way, the values were not making sense. Pinksterboer (1992, p. 26) described a possible explanation for this.

On one hand, crash cymbals "respond quickly with an overabundance of harmonics, and die out quickly". On the other hand, despite having to "be able to produce an identifiable ping", a high quantity of harmonics in the ride cymbal "would conceal definition". Therefore, depending on genre and taste, "the ratio of overtones and pings" may vary significantly, "ranging from a very clear and high cutting heavy ride to a dry, low throaty tick that is surround by washing – but modest – harmonics on a dark jazz ride".

However, the hi-hats class's low percentage in identifying a sense of pitch corroborates the idea that in most common and popular musical genres, hi-hat cymbals are seen as time quantifier, and not "an equal voice of the set" as it is seen in more jazz and fusion genres (Pinksterboer, p. 27). Because of this, hi-hat manufacturers address their concern in producing a cymbal with an identifiable sound, avoiding the masking produced by other instruments.

In terms of searching for a defined pitch, we were certain that the percentage of the membranophones classes would be significantly superior when compared to the idiophone classes. However, according to our findings, this is not so.

Although we can confirm this early assumption in the tom classes, the same could not be said for the kick class. In fact, even the snare drum class spectrograms visually presented a more definite pitched sound, unlike Rossing had described.

One possible explanation lies in the choice for the spectral resolution that may not be enough to represent the harmonics. Yet, taking a look at the table and the other attributes, the presence of a sub-bass zone and the decay of the higher partials on the frequency spectrum could also help to explain this.

One fact remains: the audio samples we have meticulously gathered came directly from their commercial, publicly available source. As such we must take into account the amount of post-

processing that they have been subject to, leaving the manufacturers and engineers no margin for errors in the final product. They have certainly been under heavy equalization and compression, in order to achieve the outcome that is to be expected from a high-budget, multi-resourceful production.

On the subject of equalization, Phil Tan[22] said: "[I use] a simple high-pass filter… on almost everything", "because apart from the kick drum and bass, there's generally not much going on below 120 to 150Hz" (as cited in Senior, 2011, p. 53). This could certainly be a reason for the generally low sub-bass percentages we have presented on Table 3.2.

Like Tan, many engineers opt to do this, because it is common knowledge that the presence of low-end frequencies can cause an "unnecessary muddy" or "rumbling" mix. We could assume that this low-frequency cutting was taken in consideration during the treatment of the recording, prior to the commercial release of the samples.

Furthermore, this common conception could be associated to the fact that companies and manufacturers intend to make their product user-friendly and easy to implement. Drastically equalizing them and removing the unnecessary low-end of the frequency spectrum, simplifies the mixing process further along the way.

In fact, one can even assume that small speaker systems reproduction capabilities were considered during the equalizing of these samples. Since they cannot fully reproduce the sub-bass region, it could have been neglected altogether. Still, this is just an assumption.

Yet, a visible sub-bass region was commonly identified on the ride cymbal class (on 45% of the samples). After some consideration, one possible explanation concerns the way of propagation of the first five or six resonating modes of the cymbal. They move as a wave from the cup of the cymbal (the centre) to the outer bounds, resulting in added low frequency response (Fletcher & Rossing, 1998, p. 650).

Regarding the behaviour of the decay on higher partials, the results extracted matched our expectations. For example, the amount of visible decay of higher partials (or even their existence at all) on the membranophones classes was of 100%. We can safely say that high-end frequency content is generally imperceptible, but may cause problems when summed with other classes' high frequency content.

Likewise, our prior expectations for the cymbal classes met positive results. In the crash and ride cymbal classes, the absence of decay in higher partials falls on the findings of Fletcher & Rossing (1998, p. 650). They are a direct result of the modes of vibration mixing together, leading to challenging identification of pitch, especially on the high-end regions. However, on the remaining class the nearly 50% of visible decay of higher partials in the hi-hat cymbals was somewhat a surprise.

As we have previously mentioned, we chose not to create a distinction between closed and slightly opened hi-hats. This choice of ours can represent the shared percentage of around 50% for decay, when the hi-hats are closed, and 50% of no decay, for the slightly opened hi-hats.

[22] American sound engineer and Three-time Grammy award recipient. He has worked with Number One Billboard artists such as Mariah Carey, Ludacris and Rihanna.

These latter ones usually display a more chaotic behaviour, similar to crash cymbals but, when in an absolute closed position, a certain sense of pitch may be perceived from their sound (Figure 3.5).

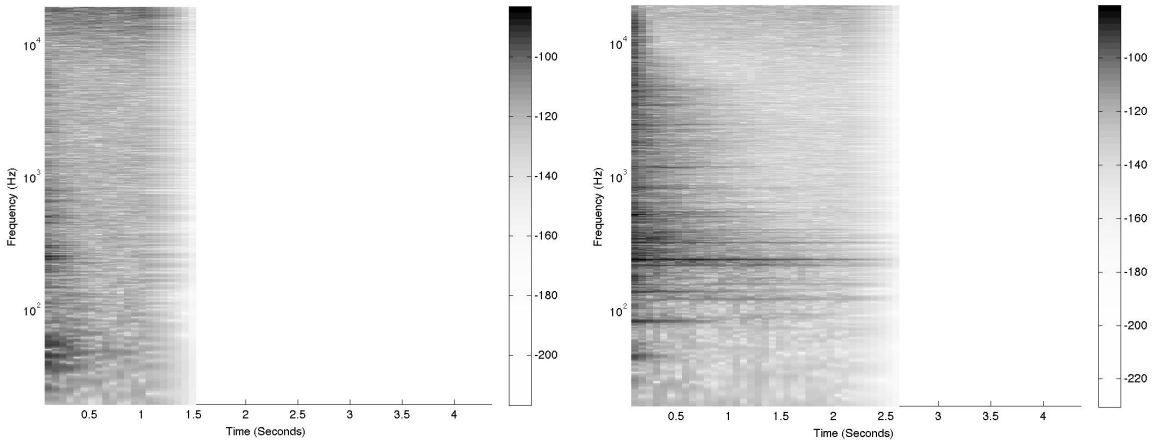

Figure 3.8 Spectrograms of an opened hi-hat (left) and a closed hi-hat (right)

Although this objectively visual analysis presents some interesting data to our research, there are still some factors that need to be accounted, which cannot be obtained from a simple graphical analysis.

3.3.3 The Low-Level Features Wave Analysis Process

Following the previous process, we felt that a visual analysis, although viable, needed a further justification on a machine-based approach. Once again we have used MATLAB for this following analysis.

McAdams et al. (1995) proposed audio features that deal directly with perceptual characteristics of sound, with little concern for the material properties of the source. They deal with the timbre of instruments, and their sound quality. Still, Lerch (2012, p. 31) stated "the number of possible features used in audio content analysis is probably limitless".

He further categorizes the features using different taxonomies, being the most obvious the computational domain of audio content: either the time domain of the frequency domain. Despite these two clear categories, additional sub-categorization can be "difficult to find" since "one feature may fit into more than one category" and because their name changes from author to author (2012, p. 32).

Since Lerch's work goes parallel to this research, and to maintain a simplistic approach to the subject at hand, we will use the author's four "boldly" branded categories: *statistical properties*, *spectral shape*, *technical/signal properties* and *intensity properties.* We felt that although important, the *statistical properties* category would not produce relevant results for this research, since we are not dealing with signal of significant length (such as songs).

Therefore, using the author's code[23], we tackled directly *spectral shape* features of our samples. Essentially, this category's majority of features relate closely to timbre or tone

[23] Available online in http://www.audiocontentanalysis.org/code/ (accessed 22nd July 2014)

colour, i.e. the unique subjective characteristic (alongside pitch and loudness) a voice or instrument that allow a listener to differentiate it from other instruments.

Bregman (1994, p. 92) condemned the lack of coherence of the widely quoted American Standards Association definition of timbre:

> *"That attribute of auditory sensation in terms of which a listener can judge that two sounds similarly presented and having the same loudness and pitch are dissimilar."*

This definition, as the author suggests is "no definition at all", since the lack of a definite pitch would lead to a sound ("such as the scrapping of a shovel in a pile of gravel") not having a timbre. From this, he implied that percussive instruments, such as the one we are studying, could not possess timbre: a conception he openly rejects.

Therefore in his work, Bregman (1994, p. 93) refrains himself from using the term while still keeping the concept, at least, until "the dimensions of timbre are clarified". In spite of this, Lerch's work carries on using the term for audio features extraction. He explains that although taking into account the phenomenon as a direct consequence of "both spectral patterns and temporal patterns" (2012, p. 42), the features he surmises are restricted and directed to "individual monophonic notes" which grows in relevance for our study. The first of these features we decide to extract was the Spectral Flux (SF), and for that we employed Lerch's MATLAB scripts for feature extraction[24]. SF measures the quickness of the shift in the power spectrum of a signal.

Through SFTF the change of magnitude is calculated from two successive frames, restricting its calculations to the frequency bins where energy increases. (Giannoulis, Massberg, & Reiss, 2013).

Since the spectral flux of a sound is measured within a value range between 0 and M, with M being the maximum possible spectral magnitude, the results are dependent of audio normalization.

With this in mind, and in order to achieve a coherent and contained sample pool, performing a peak normalisation for all our samples was required (Appendix A 1.3). We then used MATLAB (Appendix B6), again to perform peak normalization of every sample file, with the defined output value v of -0.3 dBFS (equation 3.2).

$$xNorm_i = \frac{x_i \times 10^{\left(\frac{v}{20}\right)}}{\max_{\forall_i} |x_i|} \tag{3.3}$$

Spectral flux's significance adds importance to our research as it allows observing the spectral changes in similar instruments in a graphic and more intuitive way. This lets transposing the magnitude change to numbers, creating the average flux of any class. We used Lerch's code and employed it (Appendix B7.1.). Figure 3.9 sets the example of the charts extracted (Appendix A 3.1.).

[24] Available online in http://www.audiocontentanalysis.org/code/ (accessed 22nd July 2014)

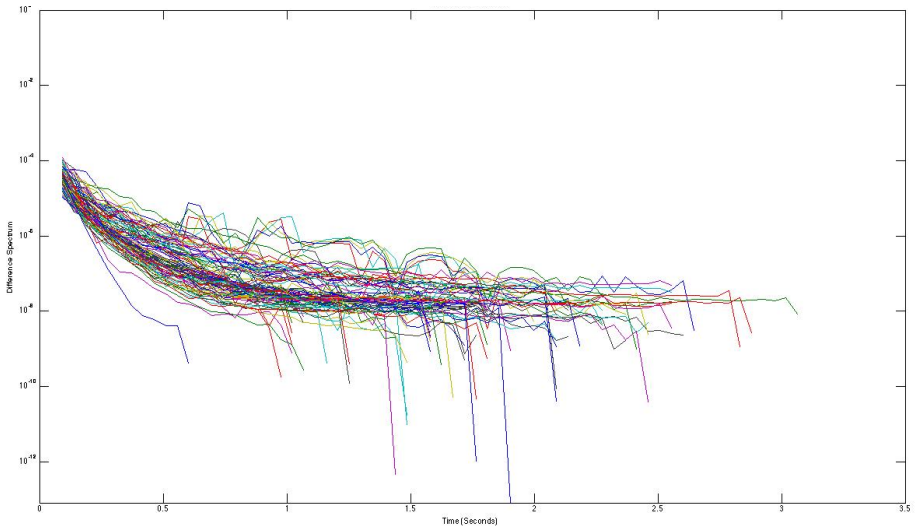

Figure 3.9 Spectral Flux of all the samples in the snare drum class

After this, we felt that understanding how the SF behaved in each class would help, and plotted new graphics with average Spectral Flux in each class (Appendix B7.2.). Figure 3.10 shows the SF in all classes.

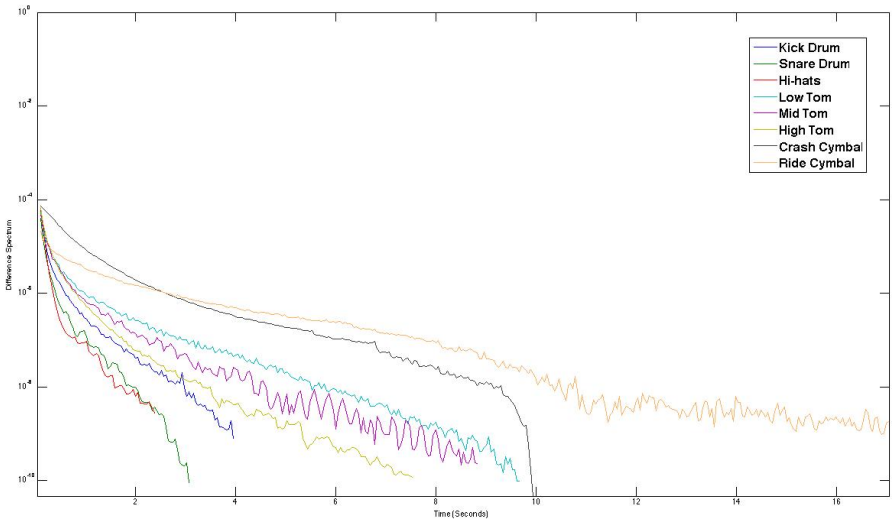

Figure 3.10 Average Spectral Flux of the classes

Results showed that the most accentuated shift in energy happened in the beginning of the samples, the attack phase of the dynamic envelope. This can be related to the behaviour of the instrument when stuck by the beater and the coupling effect explained previously. Following that, during the sustain phase, we could witness a visible and stable decrease in the difference spectrum, which was to be expected.

The feature we decided to tackle next was the Spectral Centroid (SC), which represents the centre of gravity (COG) of the spectral energy. It has been widely used and described as being directly related with an instrument's "brightness" which can be correlated with the amount of high-frequency content in a sound (Schubert, Wolfe, & Tarnopolsky, 2004).

SC extraction results in a bin index, which can then be converted to a parameter range between 0 and 1 or to frequency output values in Hz. The last one helps to understand easily the behaviour of a particular instrument.

Also, the reasons that led us choosing to perform this *spectral shape* feature extraction related to the ability to justify the drum tunings described in the previous process. Although this may not be an accurate descriptor to find a tuning frequency, the distribution of the spectral energy helps us to understand how the instruments work on the frequency domain. Moreover, knowing the COG of this spectral energy distribution, may help us to justify the results retrieved when determining a visible definite pitch and a visible dominant frequency in the Spectrogram Analysis Process.

Lerch (2012, p. 45) also comments on the effect of the input signal and the sound wave's behaviour in the calculation of the spectral centroid. This is most important when the author refers the pauses in the input signal. In fact, it is observable a significant rise in the frequency of the COG in instruments commonly associated with the low end of the spectrum, when they cease to produce sound (Figure 3.9). Once again, we applied the same principles used for the SF extraction in the Spectral Centroid calculation, using Lerch's code (Appendix B8) to extract visual charts of this feature (Appendix A 3.2.)

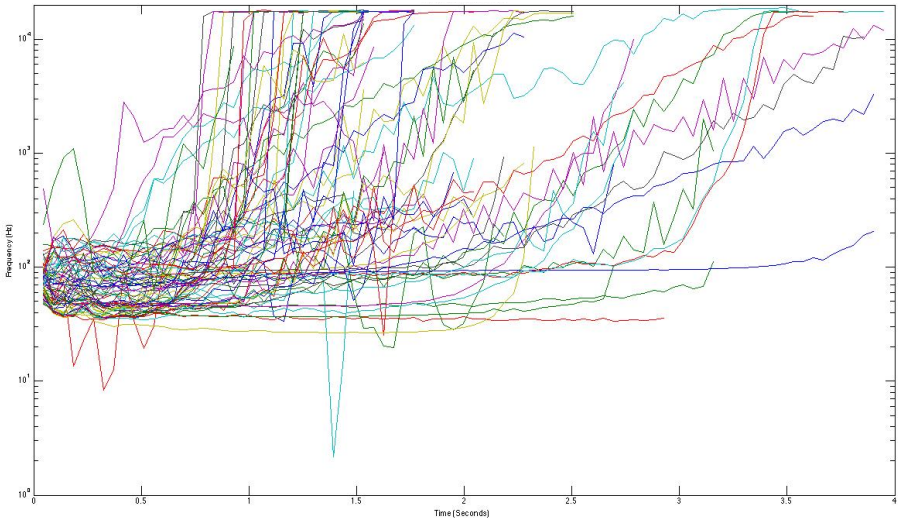

Figure 3.11 Spectral Centroid of the samples in the kick drum class

Yet, after calculating the average and carefully reading the values we then compared them with the graphical representations. We assumed promptly that the results were inconclusive. They were in fact, very far off from the normal values that were to be expected (especially the definite pitch classes). They are shown in table 3.4.

	Kick	Snare	Hi-hats	Toms			Crash	Ride
				Low	Mid	High		
Average Frequency (Hz)	1084	1331	3243	306	404	501	2104	1271

Table 3.4 Spectral Centroid's COG average frequency for samples' whole duration

For this display of chaotic behaviour, Lerch's answer lies in an abrupt shift in the COG that "requires special consideration". This is due to the SC extraction not considering the lack of input signal from a sound source as silence. Instead, it considers the existence of low-level noise (encountered in any microphone and pre-amplifier, for example) as the reason for this phenomenon.

The result from this can lead to a very difficult reading of the Spectral Centroid results, as we cannot be certain of where the centre of gravity of the spectral energy lies within each class. When plotting a graphic for the average of the class, this abrupt shift is more evidently and easily seen (Figure 3.10).

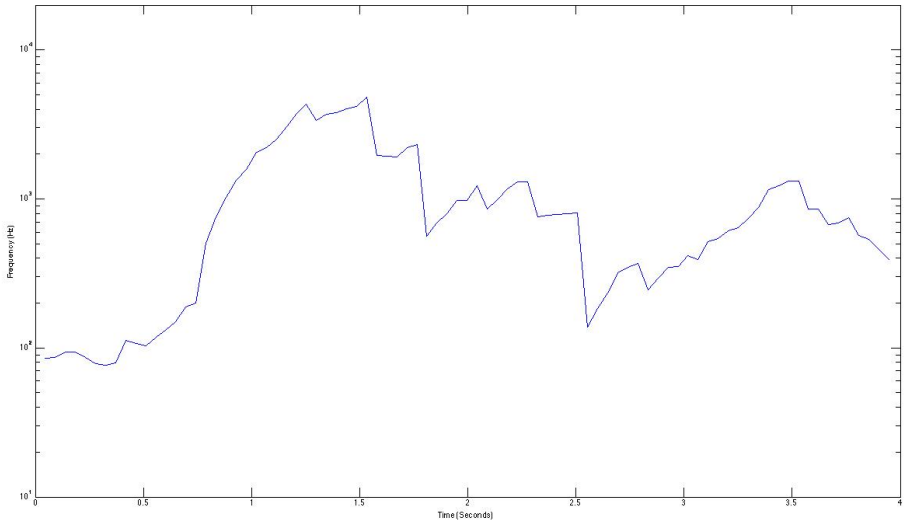

Figure 3.12 Average Spectral Centroid of the samples in the kick drum class

Since the graphical representations do not allow for an accurate extraction of the values of the COG, we felt that by looking at the average values of the MATLAB variable array we could extract more precisely the COG values as well as determining the moment where the shift occurred. We expected that, similarly to the Spectrogram Analysis Process (3.4.2), the definite pitch classes would present the most relevant results.

Still, to take the COG shift factor into consideration, we had to establish a threshold that could ascertain a ubiquitous measurement for all classes of the samples pool. Initially, we were considering COG's frequency as the deciding factor for determining the point in time when the shift occurred. This came to be an inaccurate approach, since determining an upper and lower frequency threshold could lead to two very different frequency bins.

For example, should we calculate the limits of the bin based on octave range, a kick with an initial Centre of Gravity of 100 Hz, would have and lower threshold of 50 Hz and an upper threshold of 200 Hz, thus giving the COG a 150 Hz margin for possible oscillations. The same would not happen on a cymbal with, for example, an initial COG of around 4 kHz, leading to a 6 kHz bandwidth margin (from 2 kHz to 8 kHz). Consequently, we had to consider another way to define a threshold that could be applicable to all the classes in a unanimous way.

Similarly to the proposed approach of Schroeder (1965) for calculating the reverberation time based on the integral of the energy, we opted to measure the spectral energy decay above a defined difference level, thus establishing a limitative threshold valid for every class.

Everest & Pohlmann (2009, p. 153) explain this calculation the reverberation time (RT_{60}) as the time in seconds for sound intensity to decrease 60 dB from its original level. As the authors put it, the 60 dB drop was chosen randomly, but it is a rough estimate of the intensity decrease required for a sound to reach an inaudible level.

Nevertheless, and according to academic circles, due to the increased atmospheric noise, most common measurements determine that drops of 20 dB or 30 dB (RT_{20} and RT_{30} respectively) are relevant and acceptable.

Thereby, we chose to set our threshold to be where sound drops 30 dBFS (DT_{30}) relative to the maximum magnitude of the sample. Once again we used MATLAB to retrieve the individual values of the RMS fluctuation over time (Appendix A 3.4) and then calculated the class average (Appendix B9).

We adopted a window resolution similar to the one proposed by Lerch in his script for Spectral Centroid extraction, as it would allow us to make direct comparison between the time windows defined in both representations. After the script for the RMS calculation was completed, upon the first tested examples, we came to a conclusion that the absence of a sound signal would result in minus infinite values. Therefore, we decided to establish a new overall sound intensity lower threshold to be equal to -96 dBFS.

The reason for our choice lies on the 16 bit quantization of our samples, because it is the most common dynamic range (DR) found on CDs. The quantization formula is given by the expression 3.2, where b is the bit depth (i.e. the number of bits per sample). From here one can create an average of the RMS curve without the $-Inf$ weight.

$$DR = 20\log_{10}(2b) \approx 6 \times b \tag{3.4}$$

After calculating a class average for the RMS curve we were able to extract an approximate time taken for the samples magnitude to drop 30 dB below the maximum magnitude (which occurred in the first time frame). Furthermore, we performed a comparison between the time windows of the RMS drop and the Spectral Centroid results. We then calculated the average of the classes' COG that was contained within that time window (Table 3.5).

	Kick	Snare	Hi-hats	Toms			Crash	Ride
				Low	Mid	High		
Time of Drop (s)	1,11	0,98	0,65	2,83	1,95	1,58	5,34	7,57
Average Frequency (Hz)	573	1533	6133	328	184	160	3109	2241

Table 3.5 Approximate time for RMS and average COG frequency for DT_{30}

From a first glance at the average COG table, the results extracted were far different from the expected. The kick drums COG frequency far exceeded the ordinary boundaries (between 80 and 100 Hz); likewise, the toms' COG was not the expected, since the COG decreased with the increase of the classes' pitch.

As it was equally accepted in academic context, we opted to lower the amount of dBs for the drop to 20 dB of difference (DT_{20}) to see if the COG could be more acceptable for the instruments we were analysing. Table 3.6 shows the new the moment when RMS drops the 20 dB, as well as a new calcultation for the COG average (Table 3.7).

	Kick	Snare	Hi-hats	Toms			Crash	Ride
				Low	Mid	High		
Time of Drop (s)	0,65	0,46	0,37	1,90	1,16	1,07	3,16	6,32
Average Frequency (Hz)	100	1278	6710	93	124	146	3419	2177

Table 3.6 Approximate time for RMS and average COG frequency for DT_{20}

As table 3.7 shows, the results from the DT_{20} were much more conclusive. Although the kick drum's COG was slightly above the expected (around 80 Hz) we were still within the limits of acceptability. Also, the spectral centroid average extracted on the toms displayed a substantially coherent outcome as a gradual rise of the COG's frequency accompanied the rise in the tom's pitch.

Likewise, the crash cymbal displays a COG within the boundaries (3kHz to 5 kHz) that Rossing (2001) assumed to be a part of the strong "after sound" that gives these cymbals their unique "shimmering" sound. This could be a result of the higher partials present on the various resonating modes following the initial strike.

Although he states that initially (up to the first 20 ms of the sound) the energy build-up from the strike should be within the 700 Hz to 100 Hz, we are dealing with time windows of roughly 40 ms, therefore, we could not possibly witness this event on the RMS representations. On the other hand, the remaining classes' COG displayed rather unexpected results.

Firstly, on the snare, the oddly high COG frequency could be a direct result of the rattling provoked by the vibration on the lower skin on the snares. With such a short time window (about 460 ms), the energy distribution of the COG could also result from the sound produced by the metal wires, which, is higher pitched that the skin.

Secondly, the high hats' surprisingly high COG frequency is likely to represent the chaotic behaviour that the multiple partials from the various resonating modes display when they are mixed together. Since the hi-hats' could be considered two small, opposing crash cymbals, the bigger the opening among them is, the more chaotic the sound will be.

Thirdly, on the ride class, the Spectral Centroid's COG exhibited some interesting results. Low-end frequencies are commonly present on cymbals, and since spectral centroid is not such a robust descriptor, this could be a reason. As the time interval from the drop of 20 dB RMS in the ride class took 6,32 seconds (6320 ms), the low-end resonating modes can be weighing significantly on the average.

Although some of the data gathered was more favourable and held more conclusive results, we still felt that a third descriptor could add more substance for this research. Therefore, the final feature we decided to approach was the maximum of Autocorrelation Function (MACF).

Unlike the previous two features that, according to Lerch, belong to the *spectral shape* category, Autocorrelation Function (ACF) fits within the *technical/signal properties* category.

It deals extensively with the concept of *tonalness*. Unlike *tonality*, which is used to describe the harmonic relationship between two frequencies (e.g. musical notes) in a same key, *tonalness*, on the other hand measures a value based on the periodicity of a signal and the inexistence of noisy content (Lerch, 2012, p. 54). Accordingly, the most tonal signal would be a sinusoid wave, whereas the least tonal one would be white noise (Appendix 3.3).

The relevance of calculating the maximum of the Autocorrelation Function lies in a simple estimate of how the classes can be similar in terms of tonalness: the more the value approaches one, the more tonalness the signal has. Notwithstanding, this measure can also prove to be a hindrance in deciding its relevance, since we are dealing with short signals (with lack of periodicity).

Their short duration can lead to an abrupt decrease in the ACF value (similar to the shift experienced in the measurement of the COG in the Spectral Centroid section) as we are dealing with noisy elements present in the signal. Yet, this characteristic also permits to understand better the periodicity among classes.

Still, if we take on the same approach used for the Spectral Centroid's COG calculation on a similarly relevant time-window, this can lead to much more significant results, considering the values of tonalness among samples of the same classes. So, we decided to calculate the maximum of the ACF for the same time windows as used above when the sound energy drops 30 dB (DT_{30}) and 20 dB (DT_{20}) from the maximum magnitude (Table 3.8).

	Kick	**Snare**	**Hi-hats**	**Toms**			**Crash**	**Ride**
				Low	**Mid**	**High**		
MACF DT_{30}	0,51	0,51	0,23	0,77	0,82	0,86	0,33	0,38
MACF DT_{20}	0,51	0,48	0,18	0,79	0,84	0,87	0,30	0,37

Table 3.7 Maximum of ACF of RMS for DT_{30} and DT_{20}

The results were quite promising because, within each class (with some exceptions), the samples showed tonalness. The most encouraging value we got happened in the tom classes with values ranging from over 0,75 to just under 0,90 for both drop times. This can be an indication that manufacturers take a unilateral approach when preparing their setup for recording.

The kick and snare percentages showed a less favourable, albeit still relevant, correlation among the classes' tonalness. They may imply that, among them, the average periodicity of the samples is similar, and, as such, the tone of their pitch could not in fact be related.

On cymbal classes, the results were far less significant than those extracted in the membranophone classes. They were expected though, since the various modes of vibration produced erratic behaviour on the instruments, making similar periodicity much more difficult to attain. Yet, the nearly 0,4 on the ride class can be a forecast showing a more uniform way to manufacture ride cymbals taking into account the *ping* sounds they produce.

3.3.4 The Wave Mix Process

Since we had failed to synthesize a resulting sound for each class in the Spectrogram Extraction Process (3.3.1), we decided to try doing so in a DAW environment by mixing down every sample in each class into a single one.

We decided to use once again Pro Tools 10 to perform this down-mixing process of multiple samples. Since, up to this point, we had been dealing with normalized samples, we had to establish a way to uniformly reduce the volume levels of the samples prior to this process. We used the SPL rule (equation 3.3) where P_{ref} is the pressure reference and P is the input pressure.

$$SPL = 20 \log_{10}\left(\frac{P}{P_{ref}}\right) \tag{3.5}$$

It is know that by doubling the number of sound sources, we can witness a 6dB increment in the amplitude of the resulting sum. Therefore we adapted the equation for multiple sources, translated by the equation 3.4, to determine the SPL difference with any given number n of sources, and consequently, the reduction of amplitude we had to apply to each track.

$$\Delta SPL_n \approx 6 \log_2(n) \tag{3.6}$$

After this was done, we performed a down-mix process that resulted in a single monophonic audio file for each of the eight classes. Then, we applied normalization to the recently retrieved files, in order to maintain inter-process identical modus. At the end, the previously described features were analysed on these eight audio files (Appendix A 3.5., 3.6., 3.7. & 3.8.), so that we could verify whether the results varied significantly when a global approach was compared with a more focalised tactic.

In a pre-analysis listening of the audio files, they had, in fact, produced competent sounding results. Nevertheless, an analysis was necessary to retrieve more information. With a simple and quick visual analysis of the spectrograms retrieved from the mix-down samples, we observed that the range of the dominant frequency zone of the classes deviated very little (or not at all) from the individual visual analysis.

The kick class for example, had its dominant bandwidth similar to the boundaries first described, within the range of 60 Hz and 100 Hz with three discernible high-energy frequency peaks, present around the 60 Hz, 80 Hz and 100 Hz mark.

The snare classes showed a different result from the previous process with a slightly higher pitched dominant zone: a prominently short chaotic zone ranging from around 250 Hz to 450 Hz that lasts for half a second, with additional frequency content going up to 1 kHz in the spectrum (clear peaks around 600 Hz, 850 Hz and 900 Hz lasting up to two seconds are noticeable).

On the other hand, the tom classes' deviation was fairly small: their approximate dominant frequency zones ranged from 70 Hz to 100 Hz in the low tom (with no relevant peaks visible),

from 90 Hz to 200 Hz in the mid tom (with visible peaks around 100Hz and 150 HZ) and from 100 to 220 Hz for the high toms (with no discerning peaks). Again, we performed a geometric mean for the membranophone classes' average frequency; results are exposed in Table 3.8.

	Kick	**Toms**			**Snare**
		Low	**Mid**	**High**	
Mean Frequency (Hz)	78,30	83,67	128,19	148,32	598,37
Nearest Tonal Frequency (Hz)	77,78	82,41	130,81	146,83	587,33
Tuning Frequency	$D^{\#}_2/E^{b}_2$	E_2	C_3	D_3	D_5

Table 3.8 Average frequency and approximate tuning notes for membranophone classes

After considerable thought and reflection, while observing the spectrogram plotting of the hi-hats class, a significant boost of energy in the low-mid frequency zone (300 Hz to 600 Hz) and in the upper high-end of the spectrum (5 kHz and upwards) became obvious. This finding was a direct result from our prior intention of not differentiating hi-hats between opened or closed positions.

In fact, we believe that doing so, in the down-mix process, the hi-hats class sample was able to retain most audio content characteristics from both examples. On the one hand, the lower frequency zone is a direct contribution of the closed position samples. Although feeble, a lone peak near the 400 Hz zone is visible. This can be attributed to that lesser sense of pitch that was discussed previously (3.3.2) as a direct result of the hi-hats position. On the other hand, the higher energetic frequency zone is a result of samples in an opened position. The sample that resulted from the down-mix indicated both a strong identifiable closed hit with and a relevant *hiss* common in opened hits.

Finally, on the cymbal classes, the spectrograms showed results not far from those retrieved in previous processes. Both the crash and the ride classes show a visibly energetic high-end chaotic zone, whose lengths change. On the crash we see a high concentration of energy on the first two seconds, with a fast and steady decay up to the four seconds mark, whereas on the ride, the build up of energy is as steady as its decay (which combined, last up to the eleven second mark).

We noticed a notorious but short abundance of energy above the 10 kHz in the crash class, while on the ride cymbal the spectrum's energy shifted, with two prominent peaks in the sub-bass region (around 30 Hz, 60 Hz) and a third slightly above (70 Hz). These behaviours were similar to what we had seen on previous processes.

Also, in this process, we performed low-level feature extraction on these eight new samples. The first extraction was of the Spectral Flux. This way we could observe the shift in the power spectrum along the time-domain, comparing it with the previous average calculated in Section 3.3.3. Comparing both the Spectral Fluxes, significant similarities appeared (Appendix A 4.2.).

Firstly, the onset values of the SF's difference spectrum were extremely similar for both the class average and the down-mix samples. Still, as we looked further into the time-domain,

artefacts began to show on the latter graphics. This can be explained by the arithmetic mean of all the samples' SF calculated previously, which led to the smoothing of a less altered plotting of the difference spectrum's shift.

In addition, in some cases, the plotting ended (considering the difference spectrum equal to zero) earlier to the time measured on the previous method, explainable by the same reason just mentioned (Figure 3.13). In fact, on the regions that do matter for our research the energy shift is essentially similar to the previous process, which making of it a reasonable effort to create an idyllic sample.

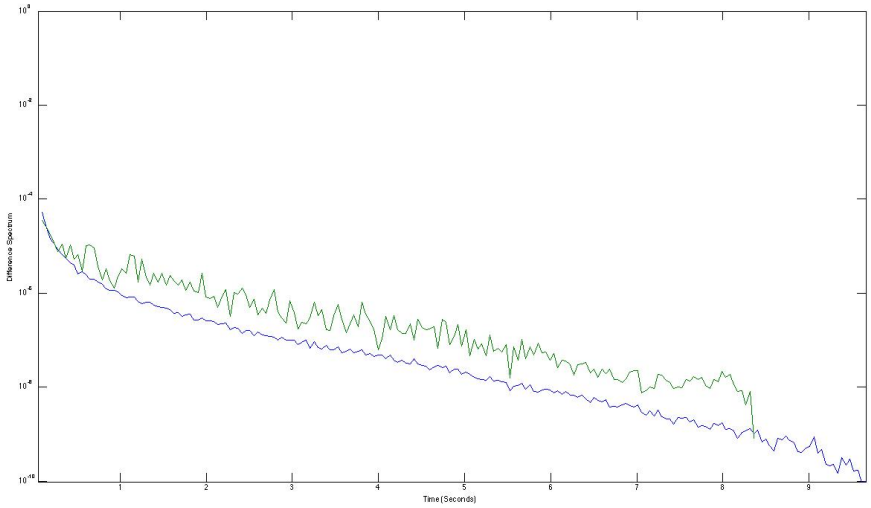

Figure 3.13 Spectral flux mean average (blue) and down-mix (green) of the low tom class

After analysing the Spectral Flux of the samples, we carried on our comparison by extracting the Spectral Centroid of the down-mix class samples. After initial extraction, we were appalled by the results we gathered.

On contrast with the previous extraction, the Centre of Gravity of these samples was far too chaotic when compared with the initial one. A most obvious example of the erratic behaviour of the Centre of Gravity's frequency is in the snare class, on display in Figure 3.14.

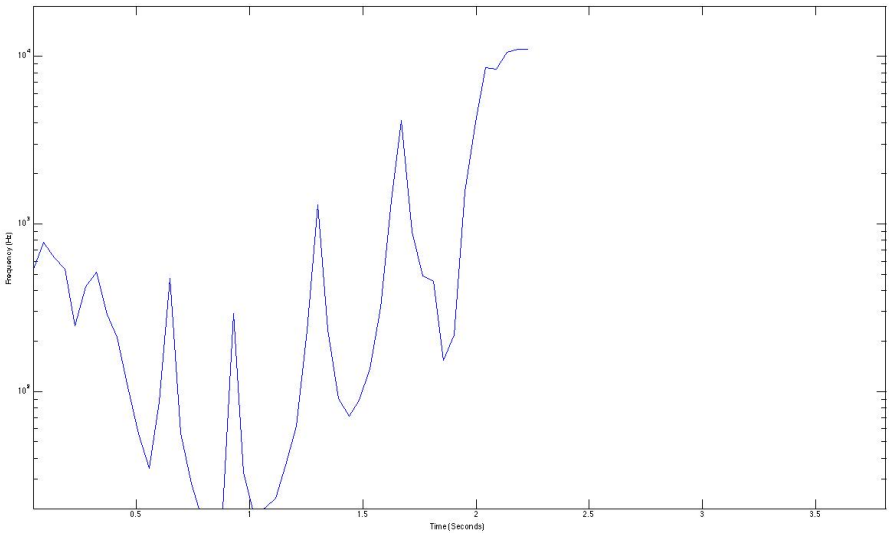

Figure 3.14 Chaotic behaviour of the COG of the SC of the down-mix snare class

Despite these unforeseen results, we used the same method for the calculation of the COG frequency average in the moment when sound drops 30 dBFS and 20 dBFS drop in magnitude. On the 30 dBFS RMS drop time (Table 3.10), similarly to the one exposed in Section 3.3.3, the values extracted led to average frequencies very different noticed during the visual analysis of the spectrograms.

	Kick	Snare	Hi-hats	Toms			Crash	Ride
				Low	Mid	High		
Time of Drop (seconds)	2,14	1,95	1,35	6,73	5,39	2,93	8,03	10,4
Average Frequency (Hz)	80	411	10740	150	160	291	4729	1815

Table 3.9 Approximate time for RMS to drop -30 dBFS and average COG

However, on the 20 dBFS RMS drop time (Table 3.11), the values made much more sense when compared to the previous extraction and postulated assumptions. In fact, those values represented in a much more reasonable way the estimated outcomes of the low-level feature extraction in this section.

	Kick	Snare	Hi-hats	Toms			Crash	Ride
				Low	Mid	High		
Time of Drop (seconds)	0,84	0,93	0,42	3,16	2,60	1,58	3,62	6,78
Average Frequency (Hz)	57	267	7060	75	108	128	3431	1898

Table 3.10 Approximate time for RMS to drop -20 dBFS and average COG

Notwithstanding these slightly different values, we carried on with the high-level features extraction, which led us to the extraction of the maximum of the ACF, and performed a visual analysis to compare the results extracted previously.

	Kick	Snare	Hi-hats	Toms			Crash	Ride
				Low	Mid	High		
MACF DT_{30}	0,47	0,39	0,21	0,69	0,59	0,31	0,18	029
MACF DT_{20}	0,42	0,43	0,14	0,65	0,51	0,56	0,11	0,24

Table 3.11 Maximum of ACF of RMS for DT_{30} and DT_{20}

These values are slightly lower, compared to the ones extracted on Section 3.3.3. Still, they presented some valuable information on how the down-mix process affected the creation of a sample that unified all the attributes of the initial classes' sample pool.

4 Results and Discussion

4.1 Analysis and Cross-Reference of Results

It is important to establish a comparison method that analyses ubiquitously the results that were extracted in the various processes. From here, we can only deduce assumptions of a general behaviour that recording engineers, addressing a specific instrument, usually search during their sessions, in order to bring out the best the instrument can provide.

Furthermore, along this study, we have conveyed some of our expectations towards the outcome of this analysis. In this section we mean to deliver more profound and insightful results likely to reflect the general behaviour of the different drum classes that we decided to tackle. Consequently, the efforts made thus far in order to endow this research with relevant significance, led to a final review concerning all these extractions of audio features and interpret the results that have been gathered. The two very different methods we have been addressing resulted in more data for further comparison.

On the one hand, we performed the extraction on a big sample pool and to arrive to conclusions, we calculated arithmetic averages on the results. From here, we were able to establish a general behaviour within the classes. Since each sample was treated individually, it resulted in multiple probabilities, variables and features. In the end, the arithmetic mean helped forming an average plot of the class. This was the Average Method.

After this, we decided to implement a second method that would rely on a down-mix approach. By mixing all the sound samples in each class, we retrieved a *super-sample*, where the dominant features were elevated and the least significant features were supressed. From this, we made an inductive assumption, from an individual subject, which could be applied to the general universe. We called it the Unified Method.

Although we have previously emphasized the possibility that a standard tuning for these drum kit classes could be achieved, it was not the sole intention, not even the most significant reason to develop in this research. In fact, the musical aspects that we have dealt so far, concern many variables (temperament, reference frequency, musical key), which could in fact undermine a whole project if attempted and dealt with dogmatically.

As we mentioned in the Introduction of the current study, we addressed two domains along this research: the frequency domain and the time domain. From these, we analysed behavioural patterns within classes that, sometimes, fail to display them, due to external variables. As a first one, concerning the frequency domain, we must mention the tuning possibilities of the classes addressed.

As has been mentioned, only five of the eight classes could in fact be included in the category of altering pitch classes. The hi-hats, crash and ride classes must not be included in these due to their lack of possibility in altering their pitch following their manufacturing.

On the other hand, kick drums, snare drums, and tom drums do possess this intrinsic ability to change their overall pitch, allowing for, if necessary, a precise tuning. Upon construction, various lugs are included on the limits of the upper and lower rims of the drum shell (adjacent to the drum membrane); tightening the lugs makes the membrane stretch, which raises the pitch, whereas loosening the lugs contracts the membrane and lowers the resulting pitch.

Toulson et al. (2009) mentioned that this ability to tune the classes to an unspecified fundamental frequency, following a series of frequency relationships, can improve the final sound on a live situation. Masking occurs when various instruments are playing together, influencing the overall performance of each. Furthermore, from this relative comparison, we intend to muster the relationships among the different classes and how they react to one another. By doing this we can, similarly, present ways to avoid inharmonic interference and masking effects.

After visual analysis, we have proposed pitch frequencies that could be applied to the tuneable classes. Furthermore, we also proposed the approximate corresponding musical notes that these frequencies apply to in a twelve-ton equal temperament ratio, with common reference pitch of 440 Hz (A_4). Furthermore, due to its added importance on tempo keeping in music, and its irreplaceable spot in modern music, we considered the kick drum to be the most suitable class to be the root note for relative tuning on a harmonic system. We also decided to establish a sequence based on frequency content (from low to high) for mapping the relative ratios. Table 4.1 describes these relationships and correspondences more plainly.

	Kick	**Toms**			**Snare**
		Low	**Mid**	**High**	
1st Method Frequency	77,45	91,65	123,69	141,42	204,93
1st Method Tuning	$D^{\#}_2/E^{b}_2$	$F^{\#}_2/G^{b}_2$	B_2	$C^{\#}_3/D^{b}_3$ or D_3	$G^{\#}_3/A^{b}_3$
1st Method Semitones	0	3	6	8 or 9	15
2nd Method Frequency	78,30	83,67	128,19	148,32	598,37
2nd Method Tuning	$D^{\#}_2/E^{b}_2$	E_2	C_3	D_3	D_5
2nd Method Semitones	0	1	8	11	35

Table 4.1 Musical properties of the definite pitch classes for both methods.

From a closer look at the table, some assumptions can be made. First, we see that for both methods, the least significant change occurred in the kick drums (less than 1 Hz of difference). Yet, by looking at the remaining classes, we see the frequency difference within them rise, the higher the instrument is tuned. For the tom classes, the frequency fluctuates, with less than 7 Hz of difference, and could cause up to a semitone of difference for the approximate "pitch" of the class. On the snare however, the frequency from method one to method two was raised approximately three times the first value.

Furthermore, it is relevant to see that in the first method, an upwards frequency-wise drum sequence (kick, low tom, mid tom, high tom and snare) would span a little over an octave (plus three semitones) including a tritone[25] exactly in the middle of the roll (mid tom class). On the other hand, a similar drum sequence in the second method would span for nearly three octaves, with the first four classes happening in the first.

Nevertheless, concerning the indefinite pitch classes, control is much more difficult after the manufacturing of the instruments. Still, for our research, we must consider our findings and understand their relationship with the definite pitch classes.

Defining, or trying to define, a root tone frequency in these classes is far more difficult; due to a wide dominant frequency range and the inherent resonating modes that cymbals in general display, our average and unified methods would produce insignificant results.

Notwithstanding this, we must resort to the results obtained in the high-level feature extraction (previous section) and establish the common grounds that can verify our current findings.

Despite a divergence, on some cases, the results gathered verified the tuning assumptions that we made. In fact, on a quick look at the graphics plotting both methods' Spectral Flux, the results evidence only minor disparities between both.

In general, the average and unified SF analysis shows the changes in magnitude are similar for the most part; in contrast, the second analysis presents far more visible artefacts. Still, the variations in the spectral envelope's for the whole duration of the sample are not significant (Appendix A 4.2.); furthermore, the overall magnitude of the unified SF of all classes is slightly higher than in the average plotting, as figure 4.1 illustrates.

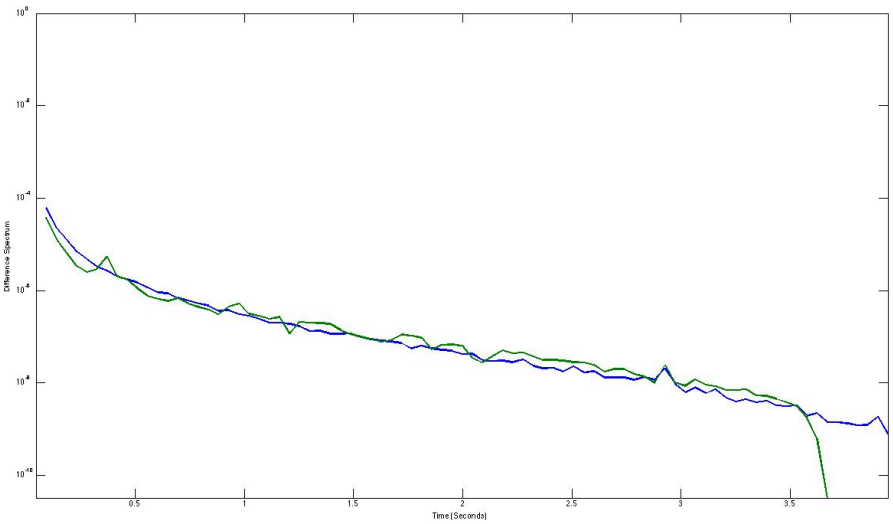

Figure 4.1 Average SF (blue) and Unified SF (green) in kick drum class

On the other hand, in terms of Spectral Centroid extraction, the proportion of differences that we identified was far too relevant not to be mentioned. As we have noted previously, the calculation of the SC's centre of gravity required special conditions in order to achieve a

[25] Name given to any interval composed by three adjacent whole tones. In the medieval and renaissance ages its dissonant sound was considered dangerous and unstable, thus obtaining the term *diabolus in musica*, or *Devil in Music* (Arnold, 1996). Although documented and accepted, they are not commonly used in modern music.

plausible result. We defined the decrease of the samples magnitude up to a certain point (either a 30 or 20 dBFS drop) as a way to exclude noise that caused the COG to rise abruptly and significantly.

Since the time for the magnitude to drop 20 dBFS (DT_{20}) produced much more reliable results we have decided not to include the DT_{30} values for the remaining of the research. The reasons for this decision concerned the amount of time that the magnitude took to drop (e.g. kick class first method: 650 ms for the DT_{20} and 1110 ms for DT_{30}; kick class second method: 850 ms for DT_{20} and 2140 ms for DT_{30}) as well as some outlandish values extracted (e.g. COGs of 573 Hz for the kick class in the first method or snare's COG of 411 Hz on the second method).

Still, we were stunned with the shocking similarity between our visually speculative estimates in the dominant frequency of the tuneable classes and the Spectral Centroid's Centre of Gravity of the same classes.

After performing a comparative analysis between the Spectral Centroid's COG and the estimated average frequencies, accuracy was blatantly visible (Table 4.2).

	Kick (F_0)	**Toms**			**Snare**
		Low	**Mid**	**High**	
1st Method Frequency	77,45	91,65	123,69	141,42	204,93
1st Method SC DT_{20}	100	93	124	146	1287
1st Method SC Tuning	G_2 or $G^{\#}_2$	$F^{\#}_2/G^{b}_2$	B_2	D_3	$D^{\#}_6/E^{b}_6$
2nd Method Frequency	78,30	83,67	128,19	148,32	598,37
2nd Method SC DT_{20}	57	75	108	128	267
2nd Method SC Tuning	$A^{\#}_1/B^{b}_1$	D_2	A_2	C_3	C_4

Table 4.2 Estimated tonal frequency (Hz) and COG (DT_{20}) in membranophones classes

The fact that a possible tone could be so similar to the Spectral Centroid's Centre of Gravity has been discussed by Marozeau et al. (2003) and Schubert & Wolfe (2006). They first used Spectral Centroid as a descriptor for correlating the timbre of equal instruments by their brightness, which the later verifies with a success rate far superior, compared to the correlation factor of the SC's COG and any given fundamental frequency. Still, the findings of Marozeau et al. (2003) and Schubert & Wolfe (2006) pertained to instruments commonly associated with high amounts of high-end frequency content (such as flutes and trumpets).

Furthermore, these values are related to a time-window that could be considered the initial attack phase of the instruments (where energy levels are at their highest). So, from here we can say that the perceived brightness in these instruments (a result of loudness perception as well) is also a direct result of a fundamental frequency, or the tone of the instrument.

An hypothesis could be made regarding this; since percussive instruments (except classes such as the idiophones) do not display such concentrated amount of high-end content, we can assume that on instruments with lower registers (sub-bass, low and mid low range) the

tonality and the Spectral Centroid's centre of gravity could in fact be related. More research could be done regarding the relationship and the ratio of fundamental tone pitch and the frequency of the centre of gravity in percussive instruments (especially membranophones).

On cymbal classes however, the expectedly high centre of gravity (the brightness), was confirmed. On one hand, hi-hat classes held tightly the high-end of the spectrum with average COG frequencies of around 6,7 kHz and 7 kHz, for the average method and the unified method respectively (DT_{20}). Crash cymbals, on the other hand, swayed on the mid-high region of the spectrum showing a COG of 3,4 kHz for both methods. The ride cymbals dangle a little lower on the spectrum with 2,1 kHz for the average method and 1.8 kHz for the unified method.

Still, it is relevant to say, that despite dealing with much higher frequencies than on the membranophone classes, the COG's average maintained itself significantly similar during time (table 4.3).

	Hi-hats	**Crash**	**Ride**
1st Method Time of Drop DT_{30} (s)	0,65	5,34	7,57
1st Method SC's COG DT_{30} (Hz)	6133	3109	2241
1st Method Time of Drop DT_{20} (s)	0,37	3,16	6,32
1st Method SC's COG DT_{20} (Hz)	6710	3419	2177
2nd Method Time of Drop DT_{30} (s)	1,35	8,03	10,4
2nd Method SC's COG DT_{30} (Hz)	10740	4729	1815
2nd Method Time of Drop DT_{20} (s)	0,42	3,62	6,78
2nd Method SC's COG DT_{20} (Hz)	7060	3431	1898

Table 4.3 Average COG and drop time duration in idiophone classes

Still, we must address questions of periodicity on the samples. On this respect, with the maximum of autocorrelation function extraction, we were able to calculate how similar was the sample's behaviour along the time domain.

A problem that could arise from this attempt to define periodicity in our samples would dwell in the time-window resolution. But, since our time frames last 2048 samples on audio files with a sample rate of 44100 Hz, frequencies with cycles lasting longer than 46ms would be problematic (equation 4.1).

$$f = \frac{1}{T} = \frac{1}{0{,}046} \approx 21{,}7\ Hz \qquad (4.1)$$

Therefore, as we are dealing with the human hearing threshold, we could say that this time-frequency resolution would not cause problems for our research.

Furthermore, despite having dealt exclusively with the drop time of 20 dBFS for the MACF extraction, analysing the values obtained on section 3.3 for a drop time of 30 dBFS in magnitude allows us to understand the periodic performance of the samples in their whole duration. It is visible in table 4.3 how the value of the MACF extraction increases with the extension of the considered time-windows (Appendix A 4.3.).

	Kick	**Toms**			**Snare**
		Low	**Mid**	**High**	
1st Method MACF DT_{30}	0,51	0,77	0,82	0,86	0,51
1st Method DT_{30} (s)	1,11	2,83	1,95	1,58	0,98
1st Method MACF DT_{20}	0,51	0,79	0,84	0,87	0,48
1st Method DT_{20} (s)	0,65	1,90	1,16	1,07	0,46
2nd Method MACF DT_{30}	0,47	0,69	0,59	0,31	0,39
2nd Method DT_{30} (s)	2,14	6,73	5,39	2,93	1,95
2nd Method MACF DT_{20}	0,42	0,65	0,51	0,56	0,43
2nd Method DT_{20} (s)	0,84	3,16	2,60	1,58	0,93

Table 4.4 Maximum of ACF for DT_{30} and DT_{20} for membranophones classes

Each method yields different results. With the first method, where a reasonable amount of samples is analysed individually, the average tells us that, in the class universe, the samples are periodic. On the second method however, results prove to be more relevant for our research purposes.

Since a sample down-mix was performed, each individual characteristic of the samples was condensed in a single one. Very different, fundamental frequencies of the class samples would cause severe disturbances in the periodicity of the resulting unified sample. We could expect much lower results if that were the case, but still, and especially on the tom classes, we found significant values.

As stated, the closer to 1 the value of the MACF is, the more periodical the signal in the considered evaluated time is. Therefore, these results can very well show that in the Average method, the tonal properties of the samples within the class can be similar. The same cannot be said for the cymbal classes. In fact, in these classes, the MACF swayed steadily on the lower numbers. On the first method, the maxima of the hi-hats class ranged from 0,19 to 0,23 (DT_{20} and DT_{30}, respectively), showing small periodicity in the class; likewise, on the second method, the class showed dissimilar tonal frequency on the different classes (0,14 for DT_{20} and 0,21 for DT_{30}).

Crash and cymbal classes showed similar fluctuations in their values. For the average method the MACF values in the crash cymbals were 0,31 for DT_{20} and 0,33 for DT_{30}; in the ride cymbals the values were 0,37 and 0,38, respectively. In the unified method, however, the values dropped once again: crash cymbals had 0,11 and 0,18 and ride cymbals had 0,24 and 0,29, for DT_{20} and DT_{30}.

4.2 Discussion

The findings that we have presented previously have exposed some interesting facts that could produce some relevant changes in how drum kits are perceived and how recording engineers should procure the best performance out of the instrument.

On the one hand, in our research, we dealt with and extracted objective and subjective audio features that have shown us behavioural aspects of the widely used commercial drum kit samples. On the other, they showed surprisingly similar results that may be basis for future studies to be developed.

The two methods we have devised tackle similar points but yield very different outcomes when analysed individually. First, more research on tonality and pitch relationships could help define a standard tuning ratio for the aforementioned drum sequence (or similar ones). In this case, no longer would drum kits recur to an instrument specific tuning, i.e., "tuning by ear", to see what would fit best with the song. Secondly, this would potentially follow a semitone-based ratio that would allow for different root note harmonic systems to be employed. This root note (kick drum) should be in tonal context with the song, which allows the easier management of the spectral contents (Pestana, 2013).

Moreover, on the computations performed for spectral centroid extraction, the values retrieved produced the assumption that the likeness of the proposed average frequencies and the spectral centroid on the attack portion (the most energetic part of the samples) in membranophone classes could be related; in fact, it could even be a hypothetically innovative descriptor that analyses audio and returns the fundamental frequency of these classes. Taking this factor into account, finding the definite fundamental pitch of drum classes might become easier. Still, since this is a very seminal suggestion, additional investigation is required to attain definite results, and their implementation in practical purposes (studio, live, etc.).

The signals' periodicity, through the maximum of Autocorrelation Function, has also produced some interesting results over both methods. Through our findings, in the first method, we noticed considerably high levels of periodicity in the samples. This may verify the extent of work taken on achieving a drum tuning that produces precise harmonic content, avoiding partial content. Had it been otherwise, the MACF computation would suffer, resulting in the decrease of the maxima (partial content would produce significant problems on establishing a periodical pattern).

On the other hand, with the second method, by performing a mix down of the samples and merging all their different values of both subjective and objective features, we were able to discern a quantitative value for inter-sample periodicity. The maxima, in this case, conveys an idea that the samples in the same class may, in fact, have very similar tunings, leading to the assumption that engineers may focus more on trimbral qualities of the instruments instead of tonal properties (i.e. a snare could produce a similar tone but with very different outputs if the shell is build from different materials).

The notable exception on the high level features extraction was the evaluation of the spectral envelope (Spectral Flux) that, in both methods, was sensibly the same. This indicates that there is little to no variation of magnitude among the samples, and demonstrates the stability and steadiness of the samples.

The values show that resemblances of drum kits may be analogous in general popular music in terms of sound (we are dealing with a rhythmic instrument), but there are still the physical and human factors to be considered.

It is important to mention that this research has overlooked physical properties of drums over sound quality and sound properties. Nonetheless, for achieving the idyllic sound of drums they must also be considered. A good musician or engineer, who knows the instrument in depth, may work with those physical properties in order to achieve a respectable output. A stable relationship between shell size and depth and tuning frequency can generate significant improvements to the output.

Furthermore, this physical dimension of the instrument can be directly related with tonality; tuning drums to a certain key produces relevant results in the overall mix. If the ratio between size and frequency is to be maximized, the resonating sound of the shell can increase the song's coherence. This could avoid unwanted resonant sounds that cause destructive artefacts, which also take significant time to resolve. Environmental traits of the room can cause similar problems in the resulting recordings.

Further research should tackle these sound, physical and environment properties, in order to reach an ideal ratio between them maximizing the instrument's resulting sound.

5 Conclusion

During the previous chapters we have comprehensively exposed a yearlong research concerning the behavioural aspects of drum kit samples available for in modern music production.

Primarily, we were anticipating the retrieval of the common elements that engineers subjectively and objectively try to uncover while addressing to the instrument, prior to its recording in a professional studio environment. We tackled several audio features in the attempt to best describe a definition of how drum kit elements should sound.

Despite the existing external variables that produce significant changes on the final output of an instrument recording, such as the room conditions, the instrument's materials and manufacture or even the way that the musician plays, we have overlooked them. Instead, we designed a research methodology based fundamentally on the instruments intrinsic qualities or, as has been said, their sound qualities.

Still, the strategies we have endorsed in this research have returned some interesting facts concerning how (maybe instinctively) the instrument is addressed in the studio. Surprisingly, among a wide range of sound samples from different manufacturers, we have seen far too many similarities in terms of tone and frequency content, periodicity and dynamic envelope, to consider it simple coincidence.

In fact, in the future, research in this area could carry added significance to the world of music by endowing professionals and amateurs the possibility of having better results in their drum kit recordings with a slight help from the academic world.

As we have seen, very little research has been undertaken to increase the knowledge concerning drum kit characteristics and their intrinsic inter-class relationships. Being an extremely prolific instrument in modern music, the amount of time dedicated to it by recording engineers worldwide, indicates that a next logical step may be, in fact, the development of tools that would gather these (and future findings) and apply them in real time in studio situations.

With so many developments in the field of post-production and digital signal processing, recording has become a somewhat uncertain and experience-based process, where only those having acute hear will thrive. In fact, we may assume that most recording engineers plan their output so precisely that in the end it becomes far too similar. Nonetheless, we believe that despite the similarities of some qualities, this does not result in less creative results, neither does it make drum kit recording an inflexible process.

Further work could be conducted addressing the following concepts:

- Audio descriptors that address fundamental frequencies of different classes;
- Resonating relationship of drums and environment, considering tuning, drum size and room conditions;
- Drum kit harmonic system ratio, with root note on the kick drum and based on song tonality;
- Management of the frequency spectrum content in a drum kit harmonic system;
- Drum kit automatic systems for multiple applications based on prior knowledge of frequency content and dynamic levels;

In conclusion, this attempt to understand the qualitative aspects that lie beneath the sound hopes to convey to the future researcher the slightly polished canvas that is the sound behaviour of drum kits and the inter-class relationship that they demonstrate. Although we deem this to be a scientific research, where we have extensively dealt with well established, commonly used and academically accepted, sound descriptors and tools, in the end, we are still within subjectivity. Indeed, it is the engineer's ears and perception that make the best possible outcome.

Bibliography

Arfib, D., Keiler, F., Zölzer, U., Verfaille, V., & Bonada, J. (2011). Time-frequency processing. In U. Zölzer (Ed.), *DAFX: Digital Audio Effects* (2nd ed., p. 602). John Wiley & Sons.

Arnold, D. (1996). *The New Oxford Companion to Music* (p. 2017). Oxford University Press.

Bartlett, B., & Bartlett, J. (2009). *Practical Recording Techniques: The step-by-step approach to professional audio recording* (5th ed., p. 633). Oxford: CRC Press.

Benward, B., & Saker, M. (2007). *Music in Theory and Practice, Vol. 1* (8th ed., p. 407). Mcgraw-Hill.

Borwick, J. (1990). *Microphones Technology and Technique*. Oxford: Focal Press.

Brandenburg, K., Faller, C., Herre, J., Johnston, J. D., & Kleijn, W. B. (2013). Perceptual Coding of High-Quality Digital Audio. *Proceedings of the IEEE*, *101*(9), 1905–1919.

Bregman, A. (1994). *Auditory Scene Analysis: The Perceptual Organization of Sound* (p. 773). MIT Press.

Buhmann, M. D. (2003). *Radial Basis Functions: Theory and Implementations* (p. 259). Cambridge University Press.

Clayton, M., Herbert, T., & Middleton, R. (Eds.). (2003). *The cultural study of music: a critical introduction*. *New York and* (p. 368). Psychology Press.

Cross, I. (2003). Music and Biocultural Evolution. In M. Clayton, T. Herbert, & R. Middleton (Eds.), *The cultural study of music: a critical introduction* (p. 368). Psychology Press.

Dean, M. (2012). *The Drum: A History* (p. 463). Plymouth: Scarecrow Press.

Dolson, M. (1986). The phase vocoder: A tutorial. *Computer Music Journal*, 14–27.

Eargle, J. (2005). *Handbook of Recording Engineering* (4th ed.). Los Angeles: Springer.

Everest, F. A., & Pohlmann, K. C. (2009). *Master Handbook of Acoustics* (5th ed., p. 528). Mcgraw Hill Professional.

Farnell, A. (2010). *Designing Sound* (p. 664). Cambridge: MIT Press.

Fidyk, S. (2011). History of the Drum Set. Retrieved December 09, 2013, from www.nationaljazzworkshop.org/freematerials/fidyk/Steve_Fidyk_History_Drum_Set.pdf

Fine, T. (2008). The Dawn of Commercial Digital Recording. *ARSC Journal*, *39*(1), 1–17.

Fitzgerald, D., Lawlor, R., & Coyle, E. (2003). Prior subspace analysis for drum transcription. In *Audio Engineering Society Convention 114*.

Flanagan, J. L., & Golden, R. M. (1966). Phase Vocoder. *Bell System Technical Journal*, *45*(9), 1493–1509.

Fletcher, N. H., & Rossing, T. D. (1998). *The physics of musical instruments* (2nd ed.). Springer.

Frith, S. (1992). The cultural study of popular music. *Cultural Studies*, *188*, 180.

Giannoulis, D., Massberg, M., & Reiss, J. (2013). Parameter Automation in a Dynamic Range Compressor. *Journal of the Audio Engineering …*, *61*(10).

Gillet, O., & Richard, G. G. (2004). Automatic Transcription of Drum Loops. In *Acoustics, Speech, and Signal Processing, 2004. Proceedings.(ICASSP'04). IEEE International Conference on* (Vol. 4, pp. iv–269).

Goto, M., Hashiguchi, H., Nishimura, T., & Oka, R. (2002). RWC Music Database: Popular, Classical and Jazz Music Databases. *ISMIR*, *2*, 287–288.

Heine, A. (2003). Music and Mediation. In M. Clayton, T. Herbert, & R. Middleton (Eds.), *The cultural study of music: a critical introduction* (p. 368). Psychology Press.

Herrera, P., Yeterian, A., & Gouyon, F. (2002). Automatic classification of drum sounds: a comparison of feature selection methods and classification techniques. In *Music and Artificial Intelligence* (pp. 69–80). Springer.

Huber, D. M., & Runstein, R. E. (2009). *Modern Recording Techniques* (8th ed., p. 644). Focal Press.

Juslin, P. N., & Sloboda, J. (Eds.). (2011). *Handbook of Music and Emotion: Theory, Research, Applications* (p. 992). Oxford University Press.

Katz, B. (2007). *Mastering Audio: The Art and the Science* (p. 334). Oxford: Focal Press.

Kennedy, M., Kennedy, J., & Rutherford-Johnson, T. (2012). *Oxford Dictionary of Music* (6th ed., p. 976). Oxford University Press.

Kitbuilder. (2010). Drum Workshop Inc. Retrieved from http://www.dwdrums.com/kitbuilder/

Kontakt 5. (2012). Native Instruments. Retrieved from http://www.native-instruments.com/en/products/komplete/samplers/kontakt-5/

Larkin, C. (Ed.). (2011). *The Encyclopedia of Popular Music*. Omnibus Press.

Lerch, A. (2012). *An Introduction to Audio Content Analysis: Applications in Signal Processing and Music Informatics* (p. 248). New Jersey: John Wiley & Sons.

Leventhall, G., Pelmear, P., & Benton, S. (2003). *A review of published research on low frequency noise and its effects*.

Lewisohn, M. (1988). *The Complete Beatles Recording Sessions: The Official Story of the Abbey Road Years 1962-1970* (p. 204). New York: Harmony Books.

Liu, L. (2004). *The Chinese Neolithic: New studies in archeology* (p. 310). Cambridge University Press.

Loy, G. (2006). *Musimathics: The Mathematical Foundations of Music* (Vol. 1, p. 482). Cambridge: MIT Press.

Major, M. (2014). *Recording Drums: The Complete Guide* (p. 400). Course Technology PTR.

Marozeau, J., de Cheveigné, A., McAdams, S., & Winsberg, S. (2003). The dependency of timbre on fundamental frequency. *The Journal of the Acoustical Society of America*, *114*(5), 2946 – 2957.

MATLAB. (2012). Mathworks, Inc. Retrieved from www.mathworks.com/products/matlab/

McAdams, S., Winsberg, S., Donnadieu, S., Soete, G. De, & Krimphoff, J. (1995). Perceptual scaling of synthesized musical timbres : Common dimensions , specificities , and latent subject classes. *Psychol Res*, *58*, 177–192.

McDermott, J. (2008). The Evolution of Music. *Nature*, *453*(7193), 287–8.

Morton, D. (2006). History of the Music Recording Insdustry. Retrieved October 29, 2013, from www.recording-history.org/HTML/musicbiz1.php

Owsinski, B. (2009). *The Recording Engineer's Handbook*. Boston: Cengage Learning.

Oxenham, A. J. (2012). The Perception of Musical Tones. In D. Deutsch (Ed.), *The Psychology of Music* (3rd ed., p. 786). Academic Press.

Parsons, A. (2010). *Art & Science of Sound Recording*. USA: Keyfax NewMedia. Retrieved from www.artandscienceofsound.com/

Patel, A. D. (2010). *Music, language and the brain* (p. 513). Oxford: Oxford University Press.

Peng, Y.-N., & Sanderson, S. W. (2013). Crossing the chasm with beacon products in the portable music player industry. *Technovation*, 1–16.

Peretz, I. (2011). Towards a Neurobiology of Musical Emotions. In P. N. Juslin & J. Sloboda (Eds.), *Handbook of Music and Emotion: Theory, Research, Applications* (p. 992). Oxford University Press.

Pestana, P. D. (2013). *Automatic Mixing Systems Using Adaptative Audio Effects*. Unpublished doctoral dissertation, Universidade Católica Portuguesa, Porto, Portugal.

Pestana, P. D., Zheng, M., Reiss, J. D., & Barbosa, Á. (2013). Spectral Characteristics of Popular Commercial Recordings 1950-2010. In *135th AES Convention* (pp. 1–7).

Pinksterboer, H. (1992). *The Cymbal Book*. (R. Mattingly, Ed.) (p. 212). Hal Leonard Publishing Corporation.

Pro Tools 10. (2011). Avid. Retrieved from www.avid.com/US/products/family/Pro-Tools

Reiss, J. D. (2011). Intelligent systems for mixing multichannel audio. In *Digital Signal Processing (DSP), 2011 17th International Conference on* (pp. 1–6).

Roads, C. (2004). *Microsound* (p. 409). MIT Press.

Rossing, T. D. (2000). *Science of Percussion Instruments* (p. 208). World Scientific.

Rossing, T. D. (2001). Acoustics of percussion instruments: Recent progress. *Acoustical Science and Technology*, *22*(3), 177–188.

Rossing, T. D., Yoo, J., & Morrison, A. (2004). Acoustics of percussion instruments: An update. *Acoustical Science and Technology*, *25*(6), 406–412.

Rumsey, F., & McCormick, T. (2009). *Sound and Recording* (6th ed.). Oxford: Focal Press.

Savage, S. (2011). *The Art of Digital Audio Recording - A Practical Guide for Home and Sudio* (p. 271). Oxford: Oxford University Press.

Schaeffer, P. (1966). *Traité Des Objets Musicaux: Essai Interdisciplines...* (p. 672). Éditions du Seuil.

Schroeder, M. R. (1965). New Method of Measuring Reverberation Time. *The Journal of the Acoustical Society of America*, *37*(3), 409–412.

Schubert, E., & Wolfe, J. (2006). Does Timbral Brightness Scale with Frequency and Spectral Centroid ? *Acta Acustica United with Acustica*, *92*(5), 820–825.

Schubert, E., Wolfe, J., & Tarnopolsky, A. (2004). Spectral centroid and timbre in complex, multiple instrumental textures. *Proceedings of the International Conference on Music Perception and Cognition, North Western University, Illinois.*, 112–116.

Schwarz, D., & Rodet, X. (1999). Spectral envelope estimation and representation for sound analysis-synthesis. *Proceedings of the International Computer Music Conference, Beijing, China*.

Scott, J., & Kim, Y. E. (2013). Instrument Identification Informed Multi-Track Mixing.

Senior, M. (2008). Kick & Snare Recording Techniques. *Sound on Sound Magazine Online*. Retrieved from http://www.soundonsound.com/sos/jun08/articles/kickandsnare.htm

Senior, M. (2011). *Mixing Secrets for the Small Studio* (p. 342). Taylor and Francis.

Serrà, J., Corral, A., Boguñá, M., Haro, M., & Arcos, J. L. (2012). Measuring the evolution of contemporary western popular music. *Scientific Reports*, *2*, 521.

Sillanpää, J. (2000). Drum Stroke Recognition. *Tampere University of Technology. Tampere, Finland*.

Sonnenschein, D. (2001). *Sound Design: The Expressive Power of Music, Voice, and Sound Effects in Cinema* (p. 250). Michael Wiese Productions.

Spich, A., Zanoni, M., Sarti, A., & Tubaro, S. (2010). Drum music transcription using prior subspace analysis and pattern recognition. In *Proc. of the 13th Int. Conference on Digital Audio Effects (DAFx-10)* (pp. 233–237).

Terrell, M., & Reiss, J. D. (2009). Automatic Monitor Mixing for Live Musical Performance. *Journal of the Audio Engineering Society*, *57*(11), 927–936.

Terrell, M., & Sandler, M. (2012). An offline, automatic mixing method for live music, incorporating multiple sources, loudspeakers, and room effects. *Computer Music Journal*, *36*(2), 37–54.

Thompson, W. F., & Blakwill, L.-L. (2011). Cross-Cultural Similarities and Differences. In P. N. Juslin & J. Sloboda (Eds.), *Handbook of Music and Emotion: Theory, Research, Applications* (p. 992).

Tidemann, A., & Demiris, Y. (2007). Imitating the groove: Making drum machines more human. In *Proceedings of the AISB symposium on imitation in animals and artifacts, Newcastle, UK* (pp. 232–240).

Toulson, R., Crigny, C. C., Robinson, P., & Richardson, P. (2009). The perception and importance of drum tuning in live performance and music production. *Journal on the Art of Record Production*, (4), 1–7.

Yoshii, K., Goto, M., & Okuno, H. G. (2004). Automatic Drum Sound Description for Real-World Music Using Template Adaptation and Matching Methods. In *ISMIR* (pp. 184–191).

Appendix A

Contents in Digital Format (DVD)

1. **Samples**
   - **1.1.** Original Samples — **1.1.1.** Kick **1.1.2.** Snare **1.1.3.** Hi-Hats **1.1.4.** Low Tom **1.1.5.** Mid Tom **1.1.6.** High Tom **1.1.7.** Crash **1.1.8.** Ride
   - **1.2.** Same Length Samples — **1.2.1.** Kick **1.2.2.** Snare **1.2.3.** Hi-Hats **1.2.4.** Low Tom **1.2.5.** Mid Tom **1.2.6.** High Tom **1.2.7.** Crash **1.2.8.** Ride
   - **1.3.** Normalized Samples — **1.3.1.** Kick **1.3.2.** Snare **1.3.3.** Hi-Hats **1.3.4.** Low Tom **1.3.5.** Mid Tom **1.3.6.** High Tom **1.3.7.** Crash **1.3.8.** Ride
   - **1.4.** Down-mix Samples — **1.4.1.** Down-Mix **1.4.2.** Normalized Down-mix
2. **Spectrograms**
   - **2.1.** Phase Vocoder Spectrogram — **2.1.1.** Kick **2.1.2.** Snare **2.1.3.** Hi-Hats **2.1.4.** Low Tom **2.1.5.** Mid Tom **2.1.6.** High Tom **2.1.7.** Crash **2.1.8.** Ride
   - **2.2.** Zero-Padding Spectrogram — **2.2.1.** Kick **2.2.2.** Snare **2.2.3.** Hi-Hats **2.2.4.** Low Tom **2.2.5.** Mid Tom **2.2.6.** High Tom **2.2.7.** Crash **2.2.8.** Ride
   - **2.3.** Down-mix Phase Vocoder Spectrogram
   - **2.4.** Down-mix Zero-Padding Spectrogram
3. **Low Level Descriptors**
   - **3.1.** Spectral Flux — **3.1.1.** All in Class **3.1.2.** Average in Class
   - **3.2.** Spectral Centroid — **3.2.1.** All in Class **3.2.2.** Average in Class
   - **3.3.** Maximum of Autocorrelation Function — **3.3.1.** All in Class **3.3.2.** Average in Class
   - **3.4.** Root Mean Square
   - **3.5.** Down-mix Spectral Flux
   - **3.6.** Down-mix Spectral Centroid
   - **3.7.** Down-mix Maximum of Autocorrelation Function
   - **3.8.** Down-mix Root Mean Square
4. **Comparing Charts**
   - **4.1.** Average Spectral Flux in All Classes
   - **4.2.** Sample Pool Average & Down-mix Spectral Flux
   - **4.3.** Sample Pool Average & Down-mix Maximum of Autocorrelation Function
5. **MATLAB Code Files**
   - **5.1.** Alexander Lerch's MATLAB code
   - **5.2.** Original Written Code

Appendix B

B1. MATLAB script for down-mixing.

```
clear all;
path = uigetdir();
files = dir([path,'/*.wav']);

for i = 1: size(files)
    [x, fs] = wavread([path,'/',files(i).name]);
    xMono = nanmean(x,2);
    name = regexp(files(i).name, '\.', 'split');
    newpath = [path,'/', name{1},'_mono.wav'];
    wavwrite(xMono, fs, newpath);
end
```

B2. MATLAB script used for spectrogram generation and *.mat* file saving.

```
% creates mat files with the short term fourier transform

clear all;
path                    = uigetdir();
files                   = dir([path,'/*.wav']);
plot                    = 0;     % zero for not plotting, one for plotting

for i = 1:size(files)
    tic

    % read wav file
    [x, fs]             = wavread([path, '/', files(i).name]);
    if (size(x,2) > 1)
        x               = mean(x,2);          % make it mono
    end

    % STFT constants
    wLen                = 2^14;               % time-frequency resolution
    hop                 = wLen/512;
    win                 = hanning(wLen);
    halfWin             = wLen/2;
    k                   = (1:halfWin+1)';


% Create STFT (through spectrogram)
    [X,f,t]             = spectrogram(  x, win, wLen - hop, wLen, fs);
    Xmag                = 2*abs(X)/halfWin;
    powerMag            = 10*log10(Xmag);
    Xav                 = nanmean(Xmag,2);  % this is the average spectrum

    % plot spectrograms
    if (plot)
        figure()
        % Spectrogram definitions
        a               = colormap('Gray');
        a               = (1-a.^4);           % invert colors / alter range
        colormap(a);

        h               = surf(t,f,20*log10(Xmag),'edgecolor','none');
        view(0,90);
        axis([min(t) max(t) 20 20000]);
        set(gca, 'YScale', 'Log');
        grid off;
        xlabel('Time (Seconds)');
        ylabel('Frequency (Hz)');
        colorbar;
    end

    % save the mat file
    name                = regexp(files(i).name, '\.', 'split');
    newpath             = [path,'/', name{1}];
    save ([newpath, '.mat'], 'X', 'Xmag');

    toc
end
```

B3. MATLAB script used for *.mat* variable name change and variable grouping.

```
% creates a large array with all the STFTs and magnitudes.

clear all;
path                    = uigetdir();
files                   = dir([path,'/*.mat']); % create file name array

for i = 1:size(files)                % loop along the files
    tic

    load([path, '/', files(i).name]); % loads file (X and Xmag variables)
    Xlarge(:,:,i)       = X;
    Xmaglarge(:,:,i)    = Xmag;

    toc
end
```

B4. MATLAB script used for creating the mathematical mean of the variable groups.

```
% uses variablenamechange with no clear afterwards
% creates the mean STFT
% ATTENTION: values for wLen and hop must match those in runSpec

% create the mean file
    tic

meanXlarge  = nanmean(Xlarge, 3);


% STFT constants
wLen        = 2^14;
hop         = wLen/512;
win         = hanning(wLen);
halfWin     = wLen/2;
k           = (1:halfWin+1)';

% build necessary variables
fs          = 44100;
f           = ((1:(wLen/2)+1) * (fs/wLen));
t           = (1: size(X,2))*(hop/fs);
meanMag     = 20*log10(2*abs(meanXlarge)/halfWin);

    toc

% plot the mean
figure()
a           = colormap('Gray');
a           = (1-a.^4);
colormap(a);
h           =surf(t, f, meanMag, 'edgecolor', 'none');
view(0,90);
axis([min(t) max(t) 20 20000]);
set(gca, 'YScale', 'Log');
grid off;
xlabel('Time (Seconds)');
ylabel('Frequency (Hz)');
colorbar;

% save the mean file
save meanXlarge.mat;

% do the IFFT to reconstruct
y = zeros(size(meanXlarge,2)*hop + wLen,1);
for i = 1 : size(meanXlarge,2)-1
    start   = (hop*(i-1))+1;
    stop    = start + wLen/2;
    grain   = ifft(meanXlarge(:,i));
    y(start:stop,1) = y(start:stop,1) + grain;
end
sound(y,fs);
wavwrite(y, fs, 16, '(nameofclass)resynth.wav');
```

B5. MATLAB script used for zero-padding spectrogram generation.

```
clear
path                    = uigetdir();
files                   = dir([path,'/*.wav']);

for i = 1:size(files)
    [x fs]          = wavread([path,'/', files(i).name]);
    x = nanmean(x,2);

tic
    fSize               = 65372;
    grainSizeSamps      = 16368;
    hop                 = grainSizeSamps/8;
    win                 = hanning(fSize);
    halfFrame           = fSize/2;
    myLength            = length(x);

    x                   = [x; zeros(fSize - mod(length(x),hop), 1)];
    nrWin               = length(x)/hop-fSize/hop+1;

grain               = zeros(fSize,1);
    X                   = zeros(fSize,1);
    Xmag                = zeros(fSize,1);
    spectraMag          = zeros(fSize, nrWin);

for k = 1:nrWin
       wStart                   = (k-1)*hop + 1;
       gStop                    = (k-1)*hop + grainSizeSamps;
       wStop                    = (k-1)*hop + fSize;
       grain                    = zeros (fSize,1);
       grain(1:grainSizeSamps)  = x(wStart:gStop);
       grain                    = grain.*win;
       X                        = fft(grain);
       Xmag                     = 2*abs(X)./(fSize/2);
       spectraMag(:,k)          = Xmag;
end

        f     = (((1:fSize)./(fSize)).*fs)';
        t     = ((1:nrWin)./nrWin).*(length(x)./fs);
        clear k wStart wStop grain X Xmag;

            figure
            a = colormap('Gray');
            a = 1-a.^4;
            colormap(a);
            h = surf(t,f,20*log10(spectraMag),'edgecolor','none');
            set(gca, 'YScale', 'Log');
            view(0,90);
            axis([t(1) t(end) 20 20000]);
            title(sprintf('ZP(nameofclass)%02d',i));
            xlabel('Time (Seconds)');
            ylabel('Frequency (Hz)');
            grid off;
            colorbar;
            name                = regexp(files(i).name, '\. i.', 'split');
            newpath             = [path,'/', name{1}];
            saveas (h, [newpath, '.jpg']);

toc
end
```

B6. MATLAB script used for peak normalization.

```
clear all;
path = uigetdir();
files = dir([path,'/*.wav']);

for i = 1: size(files)
    [x, fs] = wavread([path,'/',files(i).name]);

    xNorm = 10^(-0.3/20).*x./max(abs(x)); %normalizes to -0.3 dBFS

    name = regexp(files(i).name, '\.', 'split');
    newpath = [path,'/', name{1},'_norm.wav'];
    wavwrite(xNorm, fs, newpath);
end
```

B7. MATLAB script for Spectral Flux extraction

B7.1. Script for Spectral Flux computation of all samples in each class.

```
clear
path                    = uigetdir();
files                   = dir([path,'/*.wav']);

for i = 1:size(files,1)
    tic


        [X, f_s]             = wavread([path, '/', files(i).name]);

        [fLux(i,:), t] = ComputeFeature ('SpectralFlux', X, f_s);
            % calls for Lerch's Spectral Flux script

         semilogy(t, fLux);

            title('LIB(nameofclass)SpectralFlux');
            xlabel('Time (Seconds)');
            ylabel('Difference Spectrum');


     toc

end

        fLuxAVG = mean(fLux,1); %for plotting the average in the class
```

B7.2. Script for average Spectral Flux computation of each class.

```
% Use following spectral flux extraction

    fLuxAVG = mean(fLux,1);

    semilogy (t, fLuxAVG);

            title('LIB(nameofclass)SpectralFluxAVG');
            xlabel('Time (Seconds)');
            ylabel('Difference Spectrum');
```

B8. MATLAB script for Spectral Flux extraction

B8.1. Script for Spectral Flux computation of all samples in each class.

```
clear
path                    = uigetdir();
files                   = dir([path,'/*.wav']);

for i = 1:size(files,1)
    tic


        [X, f_s]    = wavread([path, '/', files(i).name]);

        [cEnt(i,:), t] = ComputeFeature ('SpectralCentroid', X, f_s);
            % calls for Lerch's Spectral Centroid script

        semilogy (t, cEnt);

            title('LIB(nameofclass)SpectralCentroid');
            xlabel('Time (Seconds)');
            ylabel('Frequency (Hz)');
    toc

end

        cEntAVG = mean(cEnt); %for ploting the average in the class
```

B8.2. Script for average Spectral Flux computation of each class.

```
% Use following spectral centroid extraction

    cEntAVG = mean(cEnt,1);

    semilogy (t, cEntAVG);

            title('LIB(nameofclass)SpectralCentroidAVG');
            xlabel('Time (Seconds)');
            ylabel('Frequency (Hz)');
```

B9. MATLAB script used individual RMS calculation.

```
clear
path                    = uigetdir();
files                   = dir([path,'/*.wav']);

iHopLength              = 2048;
iBlockLength            = 4096;


for i = 1:size(files)
  tic

        [x, fs]         = wavread([path, '/', files(i).name]);

        nrWin           = length(x)/iHopLength-iBlockLength/iHopLength+1;

        t               = ((1:nrWin)./nrWin).*(length(x)./fs);


     for j = 1:nrWin
        i_start         = (j-1)*iHopLength + 1;
        i_stop          = (j-1)*iHopLength + iBlockLength;
        grain           = x(i_start:i_stop);

        i_rms(i, j) = 10*log10(sqrt(sum((grain.^2)/(i_stop - i_start))));

                    if isinf(i_rms(i, j))
                        i_rms(i,j) = -96;
                    end
     end
  toc
end
```

A9.2. Script for average RMS calculation of each class.

```
% Use following rms individual extraction

    i_rmsmean = mean(i_rms,1);


    h = plot (t, i_rmsmean);

            title('LIB(nameofclass)RMSDrop');
            xlabel('Time (Seconds)');
            ylabel('dBFS');
            axis([0 t(end) -90 0]);

            saveas (h, '(nameofclass)drop', 'jpg')
```