Master in Sound and Music Computing

Content-preserving reconstruction of Electronic Music Sessions using freely available musical building-blocks

Pablo Novillo Villegas

Supervisors:
Perfecto Herrera

UNIVERSITAT POMPEU FABRA

Abstract

We present a system for creating new versions of a given song ("Seed Song"). The principal novelty in our approach centres on the use of multitrack audio session to extract more accurate and specific information, rather than use the final mix of that song. The system analyses every audio track of the audio session to extract features that will help preserving the melodic\tonal, rhythm and timbre content of the "Seed Song". Then the system reconstructs the song using freely available audios from the Freesound Community[1]. Central to our system is an objective (melodic\tonal preservation, rhythm preservation, timbre preservation) and subjective (originality and interestingness) evaluation via a listening test to examine the listener's acceptance of the songs created by the system.

Keywords: creative MIR, Music remixing, sound similarity, creative reuse, remix culture, open content, interestingness.

[1] http://www.freesound.org

Acknowledgements

I would like to thank Perfecto Herrera for his inspiration, patience and help during this whole year, your advice and direction was invaluable to me. I would like to thank my fellow friends from the master, I learned a lot from all of you. To my family, for their love, support and motivation throughout all this time.

To Laura Andrea,
Benjamín,
Timoteo
and Tomás,
you are my inspiration.
You fill my heart with joy and laughter.

. . : : In Christ alone I stand : : . .

Contents

List Of Figures

List Of Tables

Chapter 1

1. Introduction And Context

In recent years, there has been a significant change in the way people work. Instead of working alone, collaborative work has taken an important role in the music field, developing new ways to create this art, allowing amateurs and professionals to collaborate with each other regardless of their geographical location [1]. Many composers had tried to use this network communication since emergence of the radio and telephone to allow audience participation to become a musical material itself [2], and nowadays with internet it has reached more people [3]. Reusing material to create derivative works (also know as 'remixes') has engaged the community in a new way to create art [4]. Remixing usually involves manual comparison of various perceived acoustic features such as key tempo and loudness. Rhythmic and harmonic compatibility between two segments help the user decide the temporal location of these segments within a musical piece.

Another change in the way people is creating music nowadays, is by sharing their work with the community, and letting others access to the content of their whole sessions not just the final product [5]. It enables people to reuse, remix, work on top of the original ideas, and make suggestions or changes, musicians share finished or in progress multitrack sessions and new artists can continue developing the song. Multitrack sessions refer to a Digital Audio Workstation (DAW) file that contains a separated track for every instrument.

Based on this open sharing emerging community that allow us to reuse its open content sessions to create derivative works, we are going to use Creative Common (CC)[2] licensed multitrack sessions as target and reconstruct them by using new sounds from Freesound[3] [6]. This will generate a reconstructed song that will be defined as a version of a song

[2] http://creativecommons.org/
[3] http://freesound.org

that contains the main idea (melody, rhythm, structure, etc.) of the original song ("Seed Song") and it is reconstructed with different audio clips by another artist. This definition was modified from the one given to remix by Michael Casey [7].

In the following sections, we will elaborate on the motivation of this research and our goals. In chapter 2, we will provide an overview of several areas by which this study has been influenced, with a special focus on Creative MIR. In chapter 3, we will explain the methodology of this project, including the system development and the methodology used in the evaluation stage. In chapter 4, we will cover results and discussions of the system evaluation. Finally in chapter 5 we will draw the conclusions and possible future studies.

1.1. Motivation

Composing and recording music are fields that are conjugated both artistically and technically, but sometimes, classical composers and music producers have frowned upon the use of digital technology to help the creative stage, claiming that the music is created by computers and not by human beings and that this music lacks of inspiration [8]. I have worked in those two areas for more than 10 years, experiencing with different techniques and approaches to generate new music and sonority and also working in different musical genres from folkloric to electronic. I founded that researching in the creative MIR field (described in 2.1.), could allow me to combine both composition and sound creating process with a different perspective, enabling the use of processes, techniques and features that are not humanly easy to obtain.

A lot of research has been made extracting information using the final mix of a song, for that reason I opted to explore a new approach using a multitrack sessions so we can have the advantage of analyze each instrument separately and get more meaningful information. As sound engineer, I found fascinating the idea of reconstructing a song from its audio tracks, because it enables the extraction of more specific and accurate information, and it opens the possibility to generate different versions of a single song.

1.2. GiantSteps Context

GiantSteps[4] is an EU-funded research and development project, coordinated by MTG[5] (Music Technology Group) of the Universitat Pompeu Fabra. It opened the possibility of researching on this topic and provided all the necessary technical support for the development of this project. One of their objectives is "*Create Audio and Music Expert Systems to Guide Users in Composition*", and one of its outcomes is "Melody and rhythm software components that will facilitate the generation and the variation of melodies and rhythmic phrases". This is one of the main elements sought by this project, to explore new ways to generate variations of a given song.

1.3. Open Content Communities

In the context of Creative Common, Open Content Communities are platforms where musicians and producers share in-progress music projects and connect with others to collaborate. Projects can be discovered, previewed, commented on, and more importantly, "pulled" in source format to add to, remix and learn from. One example of these open content communities focused on music are the online communities Splice[6] and Blend[7], here musicians can share finished or in progress multitrack sessions and new artists can continue developing the song. In this project we are going to use Splice as a source of "Seed Songs" (target).

1.4. FreeSound

Freesound is an online collaborative sound database and it is being further developed by MTG. People have shared more than 120,000 recorded sound clips, released under Creative Commons licenses [9] to allow others their reuse. The idea of this research is to take advantage of this open database to gather sound snippets that will assemble the "new song".

---

[4] http://www.giantsteps-project.eu
[5] http://mtg.upf.edu
[6] https://splice.com
[7] https://blend.io/

1.5. Research Goals

- To build a system that analyzes a session of electronic music to generate different versions of that song, preserving its harmonic, rhythmic and timbre content.
- To analyze each audio track from a multitrack session and find a way to remap the information extracted with different audio snippets.

Chapter 2

2. State Of The Art

This project aims to build a system that analyzes each of the audio tracks in a multitrack session in order to create new versions of that song, preserving its harmony, rhythm and timbre. Therefore, this project lies in a field know as "Creative MIR", which aims to research new ways to use the technology for creating music. In this section we will give an introduction to Creative MIR, and we will review the current state of the art and the research areas encompassed by this field. Then, we will discuss about Music Information Retrival (MIR) and the information that can be obtained from audio. Finally, we will address different techniques aimed at analyzing and extracting features from music.

2.1. Creative MIR

A recent emerging area of activity within the music information retrieval (MIR) community is in the field of creative MIR [9]. This new field has the main goal to open new ways for music creation, manipulation and interaction, thanks to the capability of MIR systems to analyze, interpret and extract high level information from music signals.

Notwithstanding, creativity-oriented music technology is not a new research topic; there is increased interest to focus the effort on creative topics and applications in MIR. Some of the most popular research areas nowadays in creative MIR are presented below.

2.1.1. Automatic Mash-Up Systems

Many applications have been developed in order to combine two or more songs together, creating a new entertaining musical results called mashups [10]. The art of creating a mashup consists in finding suitable content and combining different songs in aesthetically pleasing ways. In order to reach that goal there are two fundamental processes needed to

align two songs: rhythmically using time-stretching or harmonically using pitch-shifting, which requires the ability of the composer to perform beat tracking and chord or key estimation.

Therefore, automatic mashup systems are a new creative application which are moving forward the classic MIR tasks; extracting information from audio signals, and using it to build new content. Griffin et al. [11] presented the Beat-Sync-Mash-Coder[8] which is a semi-automated real-time creation of beat-synchronous music mashups. They combine phase vocoder and beat tracker technology to automate the task of synchronizing different audio clips. The application is completely web-based and can be executed in the cross-platform Flash environment. AutoMashUpper is an Automatic Creation of Multi-Song Music Mashups that measures the "mashability" between phrase sections of an input song and songs in a music collection [12], the Wub Machine[9] is a web page and a mobile application, where users can upload a song and it is turned into a Dubstep, Drum & Bass and more styles, ccMixter[10] is another web page to upload a cappellas (pells) or instrumental tracks (samples) and then co-create completed tracks collaboratively under Creative Commons. All this applications are based in two tracks song remixing, it means that they do not use the entire multitrack session, just the final mix.

2.1.2. Recording Tools

Recording music audio is another field of research for MIR, the main idea is to develop new tools to facilitate and automate different tasks, for instance, automatic mixing is the most pursued application in this area having received a sustained research effort for many years. Scott et al. [13] propose a system based on a structured audio framework that can generate a basic mix-down of a set of multi-track audio files using parameters learned through supervised machine learning. Automatic riders are gain controls which adjust the gain on a channel in a continuous and smooth way in order to match a given criterion. Automatic riders have been designed to maintain a fixed loudness level of a target channel by raising or lowering its level, in relation to the rest

[8] http://music.ece.drexel.edu/bsmc
[9] https://the.wubmachine.com/
[10] http://ccmixter.org/

of the mix (e.g. the Vocal Rider from Waves Ltd.). Please see [14] for a wider review in intelligent systems for mixing multichannel audio.

Intelligent equalizers have receive a lot of attention in the recording music field, for instance, Pardo et al. [15] propose a simplified software interface for media production, aligned to the user's conceptual models of equalization. The aim is to quickly and automatically personalize an on-screen slider in the equalizer interface which lets the user manipulate the audio in terms of a descriptive term (e.g. dark, warm). Cartwright in [16] uses a similar approach in order to simplify audio production interfaces focusing in equalization, by allowing the user convey his needs in a descriptive language (e.g. "Make the violin sound 'warmer.'").

2.1.3. Education, Composition and Style

Considerable research has been carried out to make music practice more interactive, and facilitate improvisation during performance, for instance score following software, like IRCAM's Antescofo[11] or Arzt's Complete Classical Music Companion [17]. In contrast, one of the most unexplored topics in MIR is how all the information can be leveraged to help the human learning of concepts in music theory.

In the automated composition field, developed systems engage users in different ways and in varying degrees; consequently, these systems do not have to solve the computational creativity task completely to be useful; nevertheless, a definitive solution to the task of music-computational creativity is still a desirable and ambitious aim. One instance of this task is to create new music that conforms to a given style [18]–[20]. Pearce and Wiggins [18] proposed a methodology for evaluating systems that generate new compositions in a given style. The methodology consists of a listening study in which expert rate how well a composition fits in the style, different creations are presented in random order including authentic pieces from the intended style. A system has succeeded if the difference between ratings for the generated excerpts and authentic pieces are not statistically significant. Collins [19] and Moffat [21] have included additional questions in the methodology (e.g. aesthetic pleasure, or distinguishing the compositions as computer-generated or human-composed) in order to study the impact of listener preferences.

[11] http://imtr.ircam.fr/imtr/Score_Following

2.1.4. Algorithmic Composition

> It takes a good composer to design algorithms that result in music that captures the imagination. - Curtis Roads 1996

Algorithmic composition is the partial or total automation of the process of music composition by using computers. Different computational techniques have been used and developed for algorithmic composition, including techniques related to Artificial Intelligence, such as grammatical representations, probabilistic methods, constraint programming, symbolic rule-based systems, neural networks and evolutionary algorithms. The system proposed in this research project can be seen as part of algorithmic composition since it is an algorithm that analyzes and extracts features (e.g. harmony, timbre, rhythm) from each audio track of a given song and then it creates a new version based on those features, the difference is that our system does not have a process of music composition.

Anderson et al. in [22] present The Generative Electronic Dance Music Algorithmic System (GEDMAS) which composes full Electronic Dance Music (EDM) based on a corpus of transcribed musical data. This corpus data is used to analyze genre-specific characteristics associated with EDM styles. GEDMAS uses probabilistic and 1st order Markov chain models to generate song form structures, chord progressions, melodies and rhythms. Brooks and Ross in [23] discuss two computer-generated compositions which use Musical Weighted Synchronous Calculus of Communication Systems (MWSCCS) to model biological behaviors. Nuanáin and Sullivan [24] explain the design and implementation of a Real-time Algorithmic Composition system that employs a tabletop musical interface for input control. Fernandez and Vico in [25] present a exhaustive review and classification of a wide range of methodological approaches used to implement algorithmic composition. Please see [8], and [26] for further information about the evolution and development of this field.

2.1.5. Music Mosaicing

"Musical Mosaicing is a process whereby a piece of music, called a target and represented by a score or an audio file, is approximated through the amalgamation of many small snippets of audio drawn from sources distinct from the source of the target" [27]. The difference

between the aforementioned process and the one applied in this research is that music mosaicing analyzes the final mix of a song (mono or stereo audio file) to extract the needed information, on the other hand, our system uses each audio track from a session to extract that information. The outcome in both cases consists of a set of audio clips that are joined to represent the original audio.

In electroacoustic music, this technique has been carried out by manually locating, categorizing, arranging, and splicing analog tape or digital samples—a style that Roads [35, pp. 182–187] refers as *micromontage.* Adaptive Concatenative Sound Synthesis (ACSS) is an automated technique for generating and transforming digital sound, then, it synthesizes variations of sounds from short segments of others in the manner of collage based on a measure of similarity. Sturm in [29] proposes the application of ACSS to Micromontage Composition to provide an intuitive way to automate and control this procedure, freeing time for experimenting and composing with this flexible sound-synthesis technique.

Neupert and Gossmann [30] presented an instrument that allows to create music from a collection of audio-visual media fragments. They render a three-dimensional scatter plot from the feature-analysis of an audiovisual recording of an instrumental performance, that information becomes a Theremin-inspired instrument that enables exploration, intuitive navigation and embodied performance of the media fragments on a granular level. Another approach using high level descriptors such as genre, mood, instrumentation, key scale, singer's gender was accomplished by O'Connell [27], to create audio mosaicing, with concatenative sound synthesis (CSS).

2.2. Music Information Retrival

Music Information Retrival (MIR), also known as Music Content Description, is an interdisciplinary area, which through the application and development of computational tools and research, aims to expand the understanding and usefulness of music information. Although this field starts to emerge in the 1960's, it begins to establish itself as a research community in 2000 with the International Conference on Music Information Retrieval [31].

The research field of Music Information Retrieval has mainly focused its attention in extracting and inferring relevant information (also called descriptors or features) from the music in order to classify it in different categories (e.g. genre, mood, place of origin, artist), search for specific songs (e.g. Shazam [32], query by humming, query by tapping, cover identification) and recommend music to users (e.g. Last.fm[12], Pandora[13]). This features are foremost extracted automatically from the audio signal (music content) but nowadays they are combined with data related to the music context that cannot be extracted directly from the audio but can be found in external sources for instance web pages [33].

Signal processing along with machine learning techniques are the most widely used tools for extracting and inferring features from the music content, which are classified by Leman et al. [34] in three groups: i) low level, ii) Mid level and iii) high level features, and the structure of the music audio-mining taxonomy is presented in Figure 2.1.

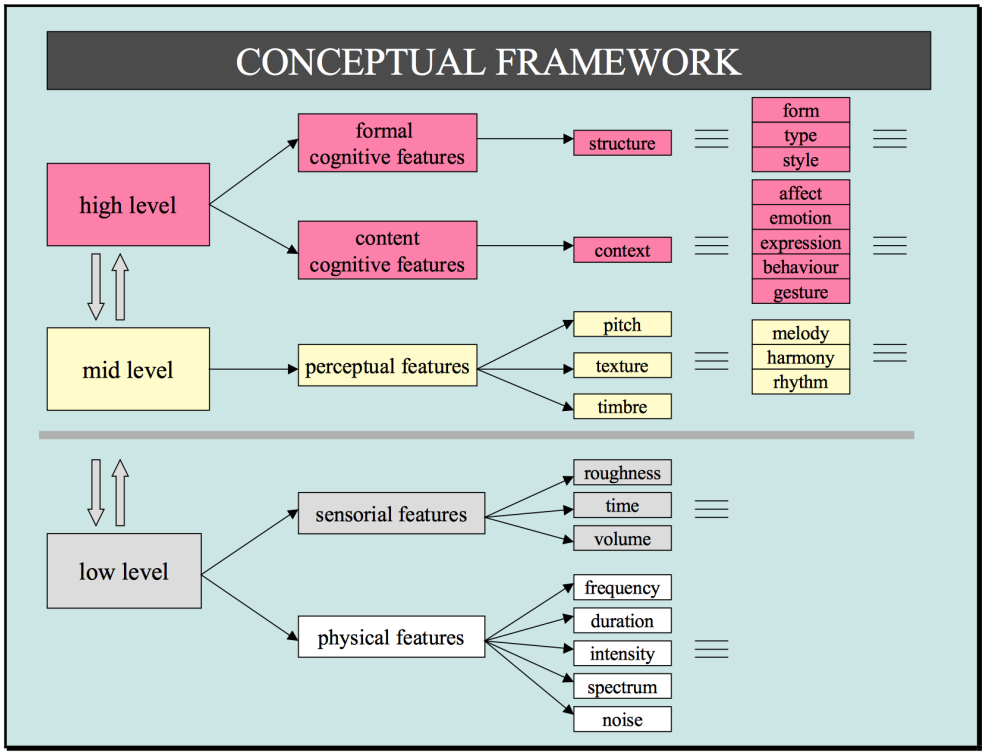

Figure 2.1: Horizontal structure of a typical MIR analysis system (taken from [34])

[12] http://www.last.fm
[13] http://www.pandora.com

2.2.1. Low Level Features

Low level features describe content that is close to the acoustical or sensorial properties of the signal for instance frequency, duration, as well as loudness and roughness. This features are extracted directly from the audio using signal processing techniques specially the fast Fourier transform (FFT) which converts a time domain representation of the audio signal to frequency domain and vice versa.

Some features extracted in this stage (e.g. spectral centroid, spectral flatness, skewness) may be difficult for humans to relate them to something meaningful without some mathematical knowledge of the process applied in the signal. Peeters's [35] report makes a description of these low level features and how they are calculated.

2.2.2. Mid Level Features

Mid level features imply time-space transformations and constrained context dependencies within a time-scale of the musical present (the 'now') (< 3 seconds). This level extends the meaning of the low level descriptors by applying a perceptual models in order to obtain specifications of the musical signal content in spatial terms (e.g. timbre, pitch, chords) and temporal terms (e.g. beat, meter, rhythmic pattern) [34].

Perceptual models are less abstract in concept and relate more closely to semantic musical features (melody, harmony, rhythm), generally they rely in the use of additional preprocessing of the spectrum (specially filters), which simulates the attenuation of the human middle ear in order to weight every frequency bands in a way directly related to human perception (e.g. Bark bands or the Mel scale).

2.2.3. High Level Features

High level concepts deal with structure and interpretation and may be related to cognitive as well as emotional and affective issues that are not only directly associated with the signal properties, but also with properties of subjective feelings and interpretations. These concepts involve learning and categorization beyond the representation of the 'now' (> 3 seconds). This level is highly determined by the cultural

context and processing related to long term memory processes (e.g. mood, genre or instrumentation) [34]. High level features combine different techniques from signal processing, machine learning, music theory, statistical and physical modeling, etc. in order to extract content objects from the audio.

2.3. Automatic Music Analysis

The system proposed in this research project will extract information from audio files, and then, this information will be analyzed automatically to create audio segments that will be replaced with different audio clips. This analysis is carried out to obtain parameters and values that will help us to preserve the melodic\tonal, rhythm and timbre content. For melodic\tonal preservation we will study melody extraction and chroma features; for rhythm, onset detection and for timbre we will need to deal with timbre similarity.

2.3.1. Melody Extraction

In recent years the task of automatically extracting a representation of the melodic line from a recording of a piece of music has been addressed extensively. Mauch and Dixon [36] carried out more recent work in this area, modifying the YIN algorithm for fundamental frequency (f0) estimation, proposing the Probabilistic YIN (PYIN) algorithm, which works in monophonic audio signals. Salamon and Gómez [37], suggested a melody extraction approach for polyphonic audio music signals based on pitch contours and time continuous sequences of pitch candidates grouped using auditory streaming cues". This approach was implemented in MELODIA – melody Extraction[14].

2.3.2. Chroma feature extraction

Chroma features, as illustrated in Figure 2.2, are the representation for music audio in which the entire spectrum is projected onto 12 bins representing the 12 pitch classes of an equal-tempered chromatic scale semitones (chroma). Since, in music, notes with exactly the double frequency value (one octave apart) are perceived as similar, for that reason, knowing the distribution of chroma even without the absolute

[14] http://mtg.upf.edu/technologies/melodia

frequency (i.e. the original octave) can generate useful musical information, for instance chords or key estimation.

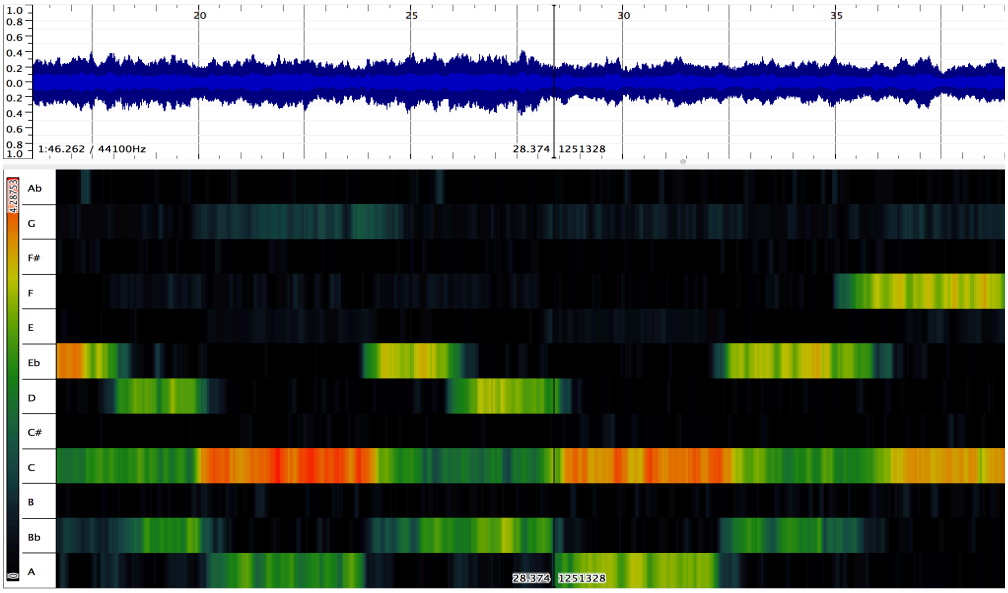

Figure 2.2: Chromagram (time-chroma representation) computed for a given music signal using the approach proposed by Mauch.

In the approach presented by Gómez [38], following the steps illustrated in Figure 2.3, the signal is analyzed to obtain its frequency domain representation. Then, the main frequency components (e.g., spectral peaks) are mapped to pitch class values according to an estimated tuning frequency. The chroma vector is computed with a given interval resolution (number of bins per octave) and it is finally post-processed to obtain the final chroma representation [33].

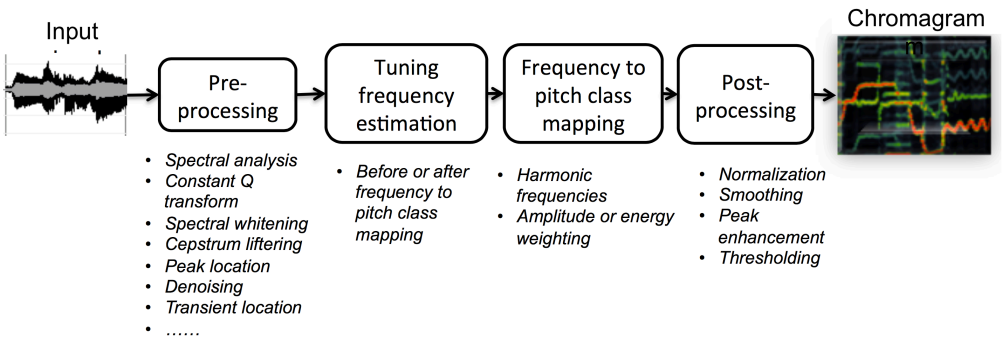

Figure 2.3: Block diagram for chroma feature extraction including the most common procedures.

Mauch [39] also proposed an automatic chord transcription from audio using an existing technique to solve non-negative least squares problems (NNLS). To improve the chord recognition, Mauch propose the

use of NNLS-based approximate note transcription, preceding to the chroma mapping. The chroma representations of the chord are obtained by calculating a log-frequency spectrogram, pre-process it and perform approximate transcription using the NNLS algorithm. This transcription is then wrapped to chromagrams and beatsynchronised to summarize frame-wise features that occur between two beats. McVicar et al. [40] can be referred to for a deeper review of automatic chord estimation from audio.

2.3.3. Onset detection and segmentation

"Humans tend to organize perceived information into hierarchies and structures, a principle that also applies to music. Even musically untrained listeners unconsciously analyze and segment music with regard to various musical aspects, for example, identifying recurrent themes or detecting temporal boundaries between contrasting musical parts" [41]. Many different studies have been developed for segmenting and structuring music audio. Three main approaches have reached good results based on three different concepts repetition, novelty, and homogeneity. In this research project is primordial to take into account different musical dimensions for instance melody, harmony, rhythm, and timbre to perform the onset detection and segmentation processes. The two main areas of research related to novelty detection in MIR are: *onset detection* and *audio segmentation*.

An audio onset is related to the beginning of a musical note and its percept is caused by noticeable change in intensity, pitch or timbre. The significance of the onset detection task lies in the fact that it is essential for many higher-level MIR algorithms and systems, such as tempo and meter identification, automatic audio transcription, audio editing and synchronization. Bello et al. [42] provide a widespread review of the challenges and approaches for onset detection. The main difficulties highlighted by the authors are the presence of slow transients, ambiguous events in the presence of vibrato, tremolo or glissandi and polyphonies.

The most common way to detect onsets is to look for variations in the signal, for instance a sudden burst of energy, a change in the short-time spectrum of the signal or in the statistical properties, etc. Tan et al. [43] present a system that combines the energy and pitch variations in order to detect onsets of different categories of instruments. As

illustrated in Figure 2.4, the first step is to obtain its time-frequency representation. Next, general note characteristics are estimated in the onset categorization block, and they are classified into three main onset categories. Simultaneously, energy-based and pitch-based processing are performed to find possible onsets, which are then integrated in the onsets integration stage.

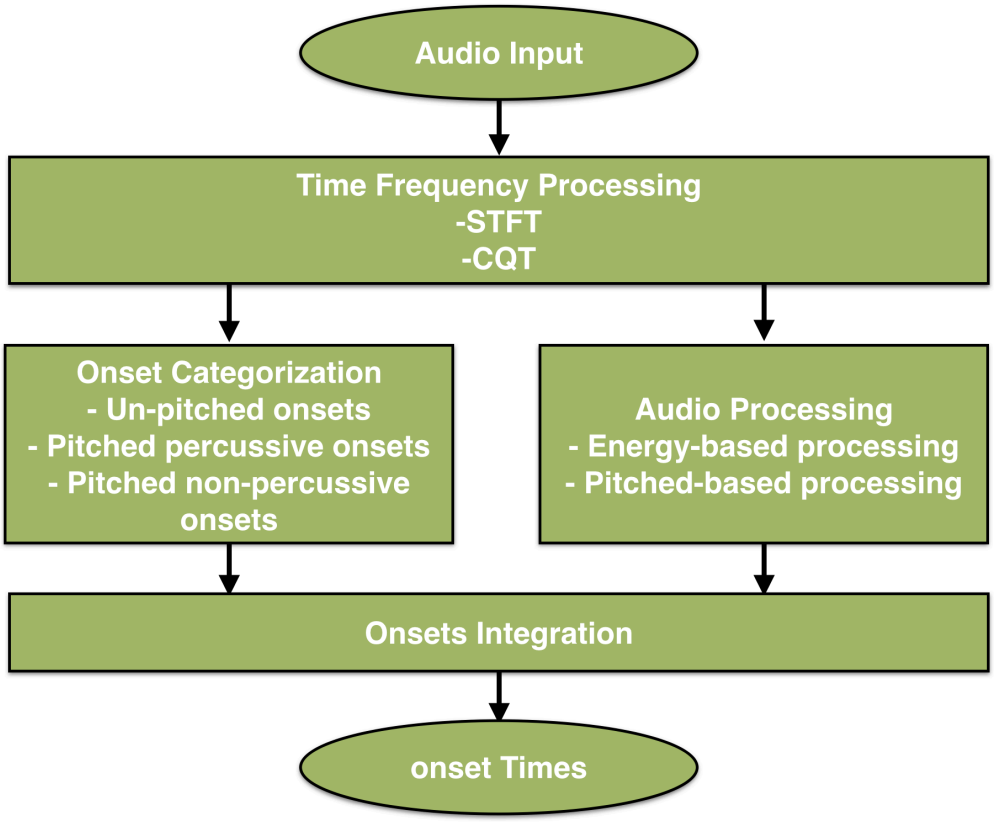

Figure 2.4: Onset detection system presented by Tan et al. [43]

Foote in [44] describes different methods for automatically locating points of significant change in audio or music, this can achieved by analyzing local self-similarity. These methods can be modified to find boundaries between individual notes or natural segment boundaries such as verse/chorus or speech/music transitions, this can be accomplished despite the absence of cues such as silence.

In Electronic Music, Rocha et al. [45] developed an application that finds similar segments of music, based on three models, one for structural segmentation, one for timbre similarity, and one for rhythm similarity of electronic dance music. Scarfe et al. [46] present an algorithm to identify segment boundaries as faithful as possible to what a human expert would find. Segmentation is performed on a self-similarity

matrix which is derived from normalized cosines of various cost matrices which have themselves been derived from a time-series of Fourier based spectral features.

For multitrack audio, Hargreaves, Klapuri, and Sandler demonstrated that "by applying a particular segmentation algorithm to multitrack data, rather than the usual case of fully mixed audio, it is possible to achieve a significant and quantifiable increase in accuracy when locating segment boundaries". [47]

2.3.4. Timbre similarity

A valuable goal in the field of MIR is to devise an automatic measure of the similarity between two musical recordings based only on an analysis of their audio content. This measure can help to improve the development of systems for classification, retrieval, browsing, and recommendation. However and although timbre and music similarity is completely subjective concept, it must be pursued in support of applications to provide automatic organization of large music collections [48].

The first ideas about "content-based search and classification" were presented by [49], proposing a body of acoustical features in order to classify sound effects and short audio samples. A classification of musical genre into not well human-defined categories, was explored by Tzanetakis and Cook in [50]. Herre et al. [51] describes a system for assessing subjective sound similarity between pairs of musical items by using a number of signal features. Most of the research developed through the years relies on the usage of low-level features primarily Mel-Frequency Cepstral Coefficients (MFCCs), which are extracted from the musical signal within a pattern recognition framework.

In [52], Pachet and Aucouturier present Figure 2.5 to illustrate a classical pattern recognition architecture used for timbre similarity tasks. The signal is first cut into frames, and for each frame, the spectral envelope is estimated by computing a set of MFCCs. The number of MFCCs is an important parameter, and each author comes up with a different number. Then the distribution of the MFCCs over all frames is modeled using a Gaussian Mixture Model (GMM). We can now use these Gaussian models to match the timbre of different songs, which gives a similarity measure based on the audio content of the music.

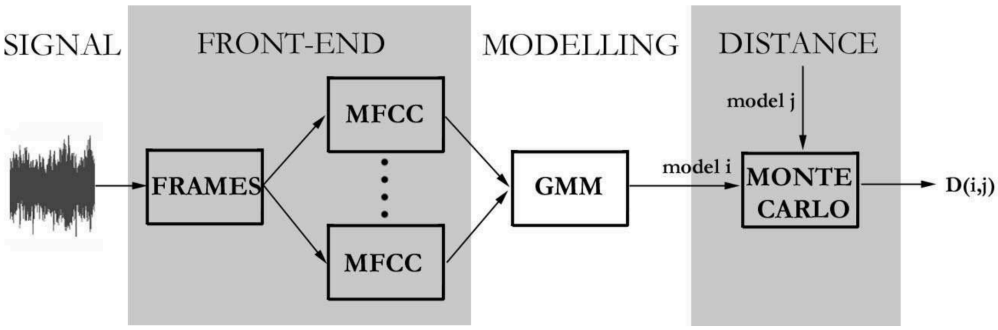

Figure 2.5: Classical pattern recognition architecture for timbre similarity. [52]

Chapter 3

3. Methodology

The goal of this project is to reconstruct a multitrack session[15], called "Seed Song", preserving its melodic\tonal, rhythm and timbre contents using features extracted from each audio track of the target. In this project, reconstruction refers to build a new version of the original song in a creative way.

The genre of Electronic Dance Music (EDM) was chosen for three particular reasons: (1) steady tempo, mostly in the range of 120-150 BPM; (2) a repeating drum pattern is almost always present; (3) its made with computers and electronic instruments that tends to quantize every note in a grid [53]. Nowadays, EDM is mostly created using MIDI (Musical Instrument Digital Interface) events instead of audio. These MIDI events are messages that specify notation, pitch and velocity and many other musical characteristics (e.g. volume, vibrato, audio panning, etc.), these messages are usually sent to a synthesizer, which generates the audio.

Most of the multitrack sessions obtained from Splice have two types of tracks: MIDI and audio. In this project we are going to convert MIDI tracks into audio files using as far as possible the same synthesizers and parameters set by the author of the song to maintain the timbre and the musical idea with which they were created. These audio files will be the input of our system, and from them, we will extract the features to create a new version that will use different audio snippets from the open database called Freesound[16].

As we claim that this is an artistic process and we want it to be creative, all the features to be used for reconstructing the song will be extracted only from the audio files. This will allow us to obtain a different representation of the seed song and also to prevent system overfeed and

[15] A multitrack session consists of all the tracks of a song, individually separated by instruments.
[16] http://www.freesound.org

ending with songs closely resembling the original, or to each other. MIDI events will be used only as ground truth to compare the original arrangements of each instrument with those obtained by our system.

The sounds contained in the multitrack session were divided in four groups (see Figure 3.1):

- **Harmonic sounds:** Sounds with definite pitch, and it includes sounds that produce melodies, arpeggios or chords.
- **Rhythmic base:** This group is conformed by the three main elements of Electronic Dance Music [53]: kick drum, snare and cymbals.
- **Rhythmic fills:** Unpitched percussion sounds (e.g. tambourine, cabasa, bongos, congas, timbales, etc.)
- **Sound Effects:** Inharmonic sounds relatively difficult to identify as to pitch. This category includes environmental sounds (e.g. rain, thunder, wind, etc.), sound effects (e.g. explosions, footsteps, aliens, etc.), noises, etc.

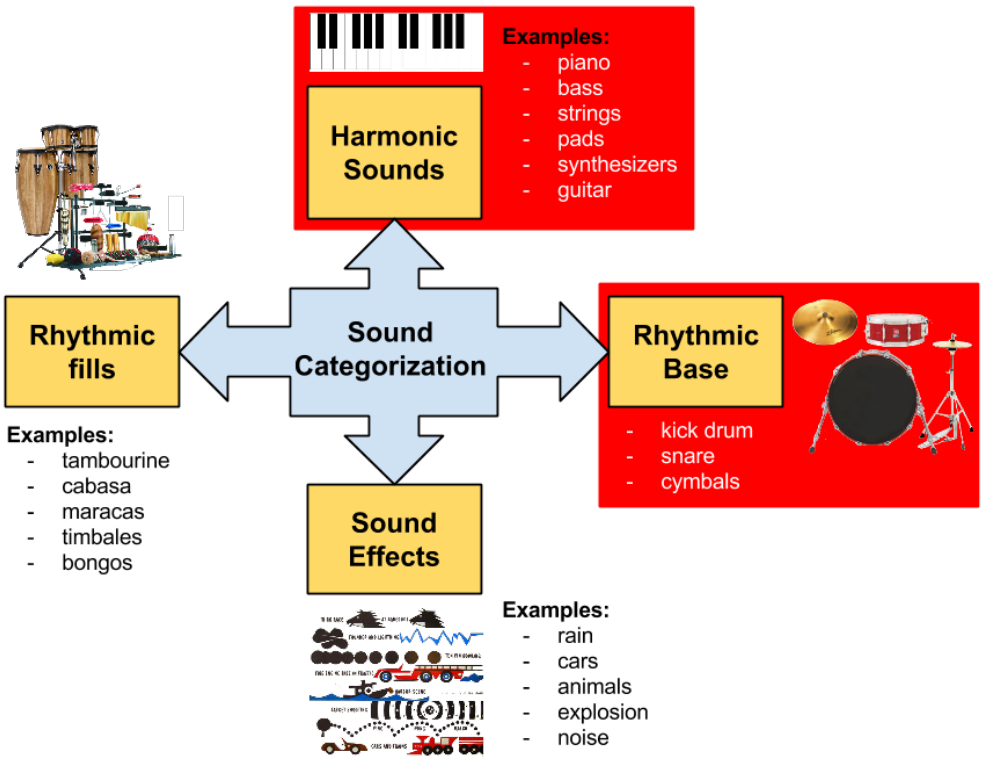

Figure 3.1: Sound Categorization used in this project. The two groups highlighted in red (Harmonic Sounds and Rhytmic Base) will be used in this research.

Since the goal of this project is to reconstruct the melodic\tonal (harmony), rhythm and timbre of the target song, we are going to focus in the first two categories aforementioned, leaving the last two for future research.

The developed system is comprised of three stages: 1) feature extraction and analysis, 2) sound retrieval and 3) song reconstruction, as shown in Figure 3.2. Every stage will be extensible analyzed in the following sections.

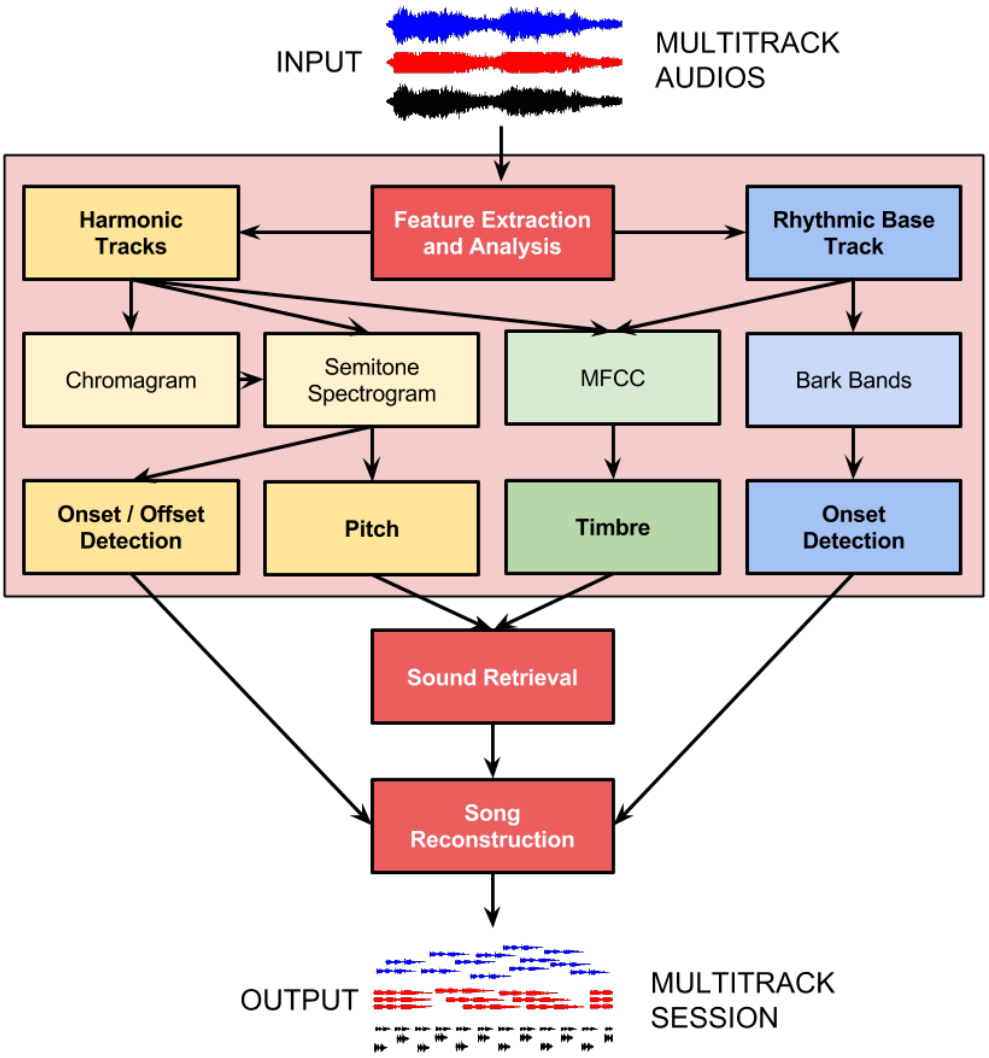

Figure 3.2: System block diagram

3.1. Feature Extraction and Analysis

In Electronic Dance Music is very common the use of MIDI tracks that send information to real or virtual synths to create the sounds that compose the song. It is important to clarify that the system proposed in this project will only use audio files to extract the information needed to reconstruct the "Seed Song". Most of the multitrack sessions used in the development of this system were entirely composed by MIDI tracks, this tracks were converted to audio tracks maintaining all the original setting given by the author of the song. As shown in Figure 3.2, the inputs of our system are multitrack audio files and the output is a Multitrack audio session.

3.1.1. Melodic\Tonal Content

The NNLS Chroma plugin [54] within Sonic Annotator [55] were used to extract two outputs: 84-bin (7 octave) tuned semitone spectrogram, to extract the notes maintaining its height; and a 12-dimensional chromagram (the distribution of energy across the chromatic pitch classes in a musical octave), that will be used to eliminate inharmonic overtones. These outputs are extracted using the default parameters and their use within this project will be discussed below. Figure 3.3 shows the 84-bin semitone spectrogram and the 12-dimensional chromagram of a piano playing chords conformed by three notes and as we can see in this representation there is also harmonic content of the instrument, not just the fundamental frequencies of notes that are played.

The semitone spectrogram is going to be segmented spectro-temporally in its 84 bins and according to the detected onsets and thereby approximate the beginning and duration of each detected note. The onsets and offsets will be obtained by deriving the time variation of energy in each bin. The chromagram will also be segmented as the previous feature but in this case just in 12 bins keeping just the bins (notes) with higher energy than a threshold that can be configured by the user, this will be used to eliminate inharmonic overtones from the 84-bin semitone spectrogram. This process is illustrated in Figure 3.4.

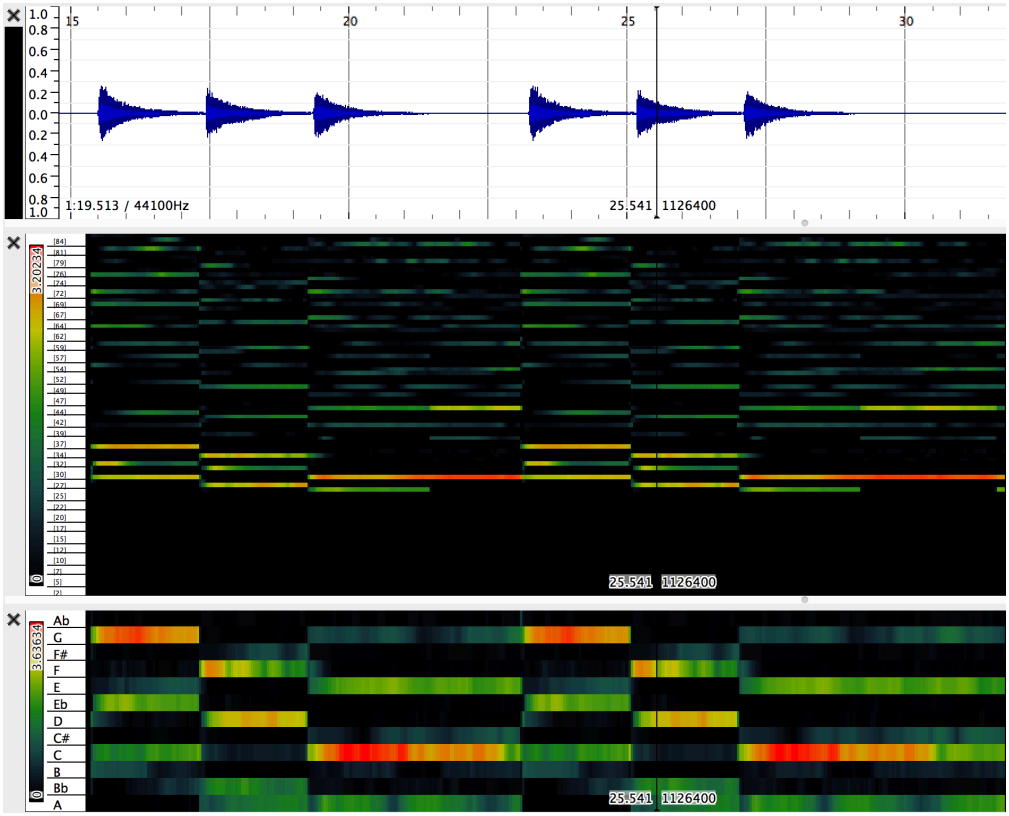

Figure 3.3: Example of the harmonic features extracted from a piano playing chords taked from Sonic Visualizer. Audio Signal (top panel); 84-bin semitone spectrogram (second panel); 12-dimensional chromagram (bottom panel)

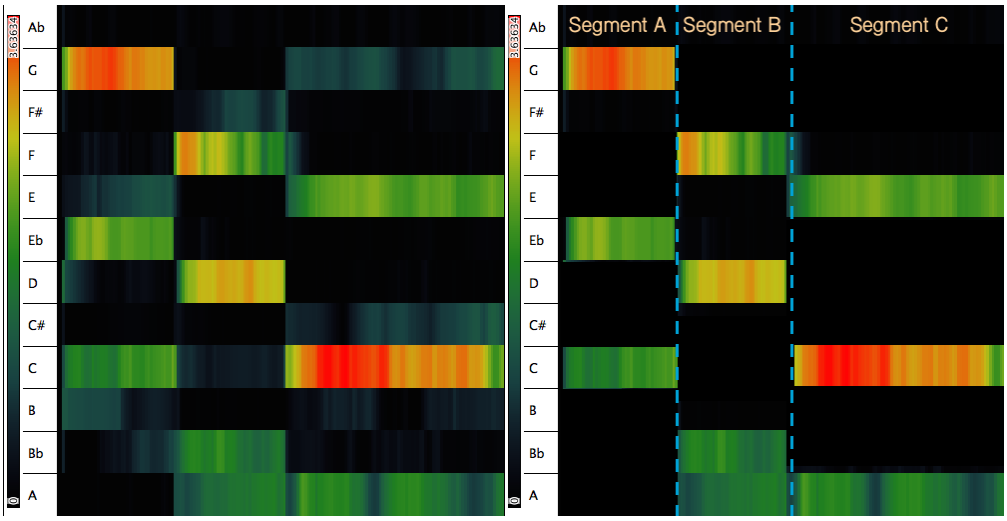

Figure 3.4: Original 12-dimensional chromagram (left); 12-dimensional chromagram after applying the process of segmenting notes (right).

It is important to point out that the number of notes that remain depend exclusively from its energy, this is because we are dealing with audio files and we don not know how many notes are been played in one segment. For instance, in Figure 3.4, only the notes that were kept in segment A (C, Eb and G) will be used in the same segment of the 84-bin semitone spectrogram.

To improve the semitone spectrogram segmentation three parameters that can be set by the user were introduced. These parameters can be adjusted for each song, however, for the evaluation process of this research the same configuration was used in all the versions created.

- ***Length of notes***: This threshold allows to choose the minimum duration of a segment to be considered as a valid note.

- ***Time between notes:*** This threshold allows to remove small silences between notes. If the time between two notes is less than the threshold the two segments will be merged.

- ***Semitone energy variation:*** After finding an onset, this threshold allows the variation of energy of a segment, if the energy is below the threshold, it will be consider as the end of the segment, and when the energy rises above the threshold it will be consider as a new segment.

Figure 3.5 illustrates the process of segmenting the 84-bin semitone spectrogram. The first step (Figure 3.5 (1) - (2)) is to eliminate all the noisy content (e.g. very short notes, very short silences) with the thresholds mentioned above. Then all the notes that were removed from the chromagram in a particular segment will also be removed from the semitone spectrogram (Figure 3.5 (3) - (4)).

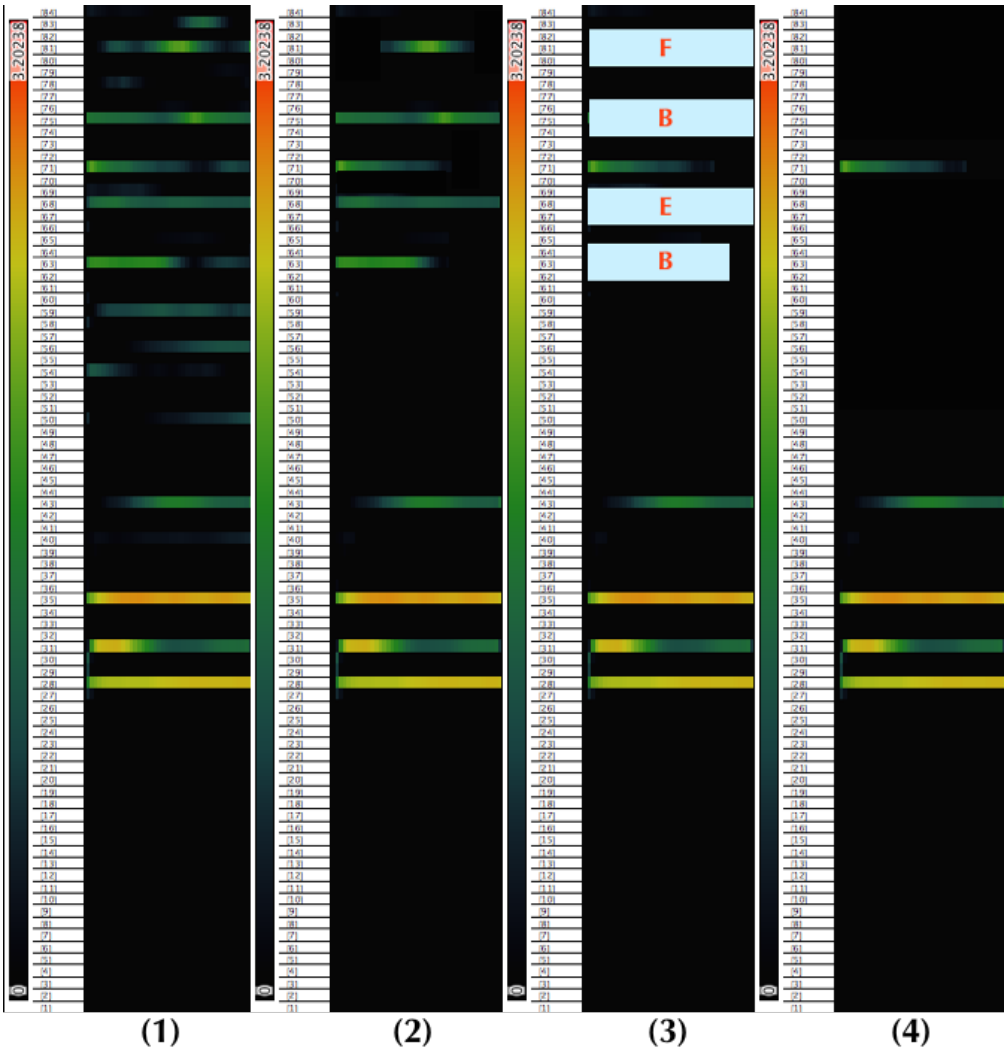

Figure 3.5: Semitone spectrogram segmentation process. (1) Original extraction; (2) semitone spectrogram modified with threshold values; (3) notes that will no be considered as part of the segment due to chromagram restrictions; (4) final outcome of the semitone spectrogram segmentation.

Energy variation is averaged to get a single value for each segment to be used in the reconstruction stage. Figure 3.6 (left) shows the first two bars of the original MIDI file of a piano playing chords, it can be compared with the results obtained by the system showed in Figure 3.6 (right). It can be observed that in this particular case, the harmonic change through time creates octave errors for some notes, due to the fact that sometimes in the attack of a note, the energy in the first harmonic is stronger than the energy in the fundamental frequency.

Figure 3.6: MIDI reconstruction of a piano playing chords. Original MIDI file (left), reconstructed MIDI file (right).

The left hand side of Figure 3.7 presents the first two bars of the original MIDI file of a piano playing a melody comprised of small notes, in the right hand side, the results obtained by the system are showed. In this example, by allowing notes with short duration the system also allows the inclusion of more quantity of false notes that in most of the cases are harmonic overtones (noise, harmonic content). It must be remembered that there is an inharmonic overtones filtering process using the 12-dimensional chromagram, therefore, most of the false notes are within the same harmony, thereby we are introducing variation to the new versions.

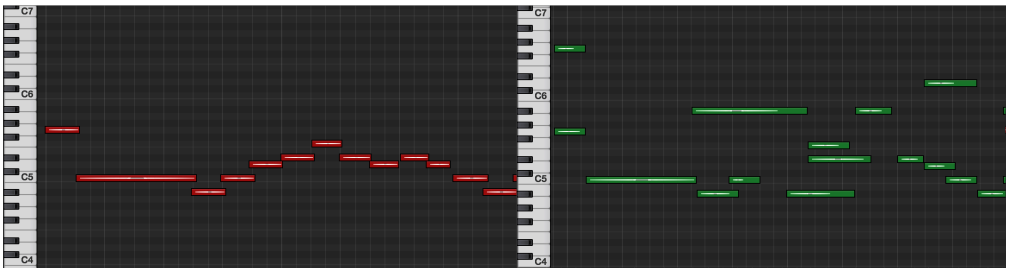

Figure 3.7: MIDI reconstruction of a piano playing a melody comprised of small notes. Original MIDI file (left), reconstructed MIDI file (right).

3.1.2. Rhythmic Content

To reconstruct the rhythm content we are going to use the Bark scale, which is a standardized scale of frequency, where each "Bark" (named after Barkhausen) constitutes one critical bandwidth. The Bark scale ranges from 1 to 24 Barks, corresponding to the first 24 critical bands of hearing [56]. The published Bark band edges are given in Hertz

as [0, 100, 200, 300, 400, 510, 630, 770, 920, 1080, 1270, 1480, 1720, 2000, 2320, 2700, 3150, 3700, 4400, 5300, 6400, 7700, 9500, 12000, 15500]. The published band centers in Hertz are [50, 150, 250, 350, 450, 570, 700, 840, 1000, 1170, 1370, 1600, 1850, 2150, 2500, 2900, 3400, 4000, 4800, 5800, 7000, 8500, 10500, 13500]. These center-frequencies and bandwidths are to be interpreted as samplings of a continuous variation in the frequency response of the ear to a sinusoid or narrowband noise process. That is, critical-band-shaped masking patterns should be seen as forming around specific stimuli in the ear rather than being associated with a specific fixed filter bank in the ear [57].

Bark Bands will be extracted from an audio signal of a drum comprised of kick, snare and cymbals using the MIR.EDU[17] plugin along with Sonic Annotator [55] (see Figure 3.8).

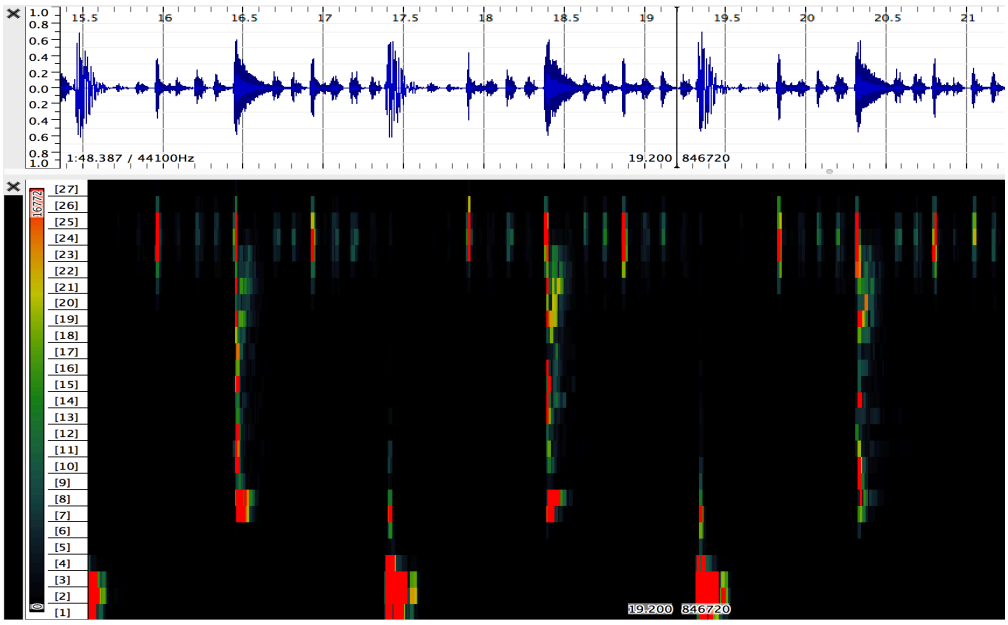

Figure 3.8: Example of the rhythmic features extracted from a drum pattern (kick, snare and cymbals) taken from Sonic Visualizer. Audio Signal (top panel); Bark Bands (bottom panel).

The algorithm used in this plugin has splitted the first two bands [0,100] and [100,200] for better resolution. The first 4 Bark Bands (1-4) will be used for kick onsets, the last 4 (24-27) will be used for cymbals onsets and the remainder (5-23) for snare as shown in Figure 3.9. The output is extracted using the default parameters.

[17] https://github.com/MTG/miredu

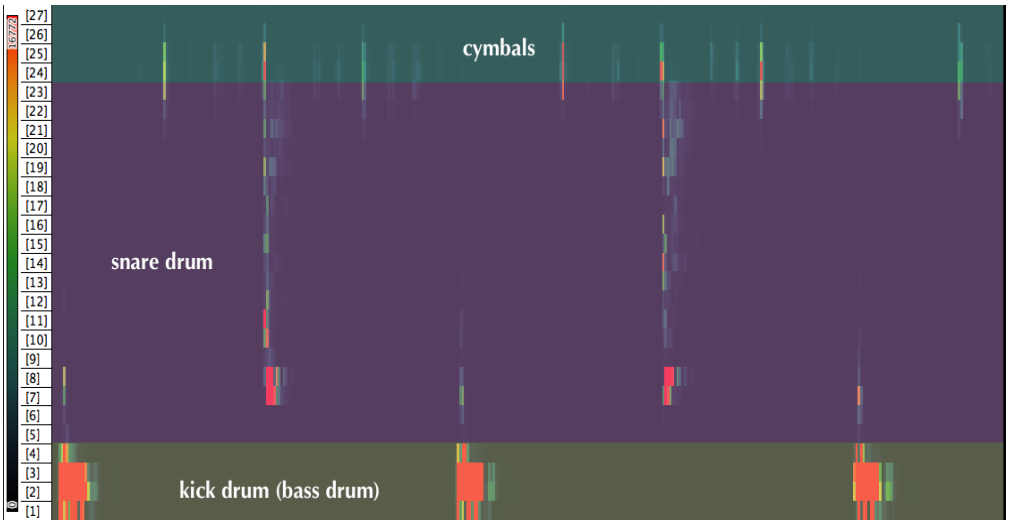

Figure 3.9: Distribution of the Bark bands for drum reconstruction

The energy of each Bark band of these three groups (kick, snare and cymbals) is added to obtain the total energy of each group. Then, these last three values are normalized to calculate the derivative of the energy of each group. A threshold is set for each group, above which, the values of the derivative of the energy would be considered onsets. After some tests we adjust the values of each threshold (0.5 for kick, 0.3 for snare and 0.1 to cymbals) to be used in all the songs that will be evaluated in this research.

Figure 3.10 shows one bar of the reconstructed drum pattern obtained by the system. The dashed lines represent the onsets: blue for kick, green for snare and red for cymbals. It is important to stress that this system will retrieve only one audio file for each of the three elements that comprises the drum base.

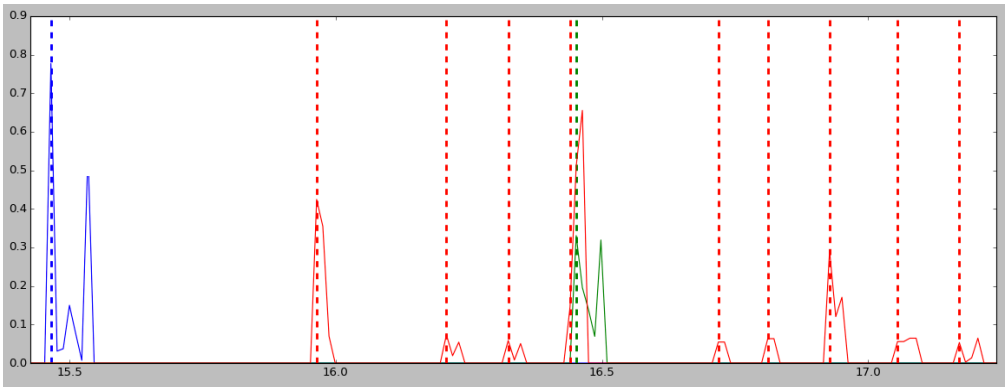

Figure 3.10: Onset placement for kick (blue dashed line), snare (green dashed line) and cymbals (red dashed line).

3.1.3. Timbre Content

As mentioned in 2.3.4, most of the research aimed at timbre similarity relies on the usage of Mel-Frequency Cepstral Coefficients (MFCCs). The MFCCs represent the shape of the spectrum with very few coefficients. The ceptrum, is the Fourier Transform (or Discrete cosine Transform DCT) of the logarithm of the spectrum. The Mel-ceptrum is the ceptrum computed on the Mel-bands instead of the Fourier spectrum. The first coefficient (MFCC[0]) is proportional to the energy [35]. 4 values were calculated from each of the 13 coefficients: average, percentile 75, percentile 25 and standard deviation. Considering that in multitrack audio files it is quite common to find long silences, they were removed before making the aforementioned calculations in order to get more precise values. MIR.EDU[18] plugin along with Sonic Annotator [55] was used to extract this features. The output is extracted using the default parameters.

3.2. Sound Retrieval

This stage was carried out using Freesound APIv2[19], which allows us to browse, search, and retrieve information about the sounds contained in the Freesound[20] database. This API makes possible to find similar sounds to a given target based on the content analysis made on the previous stage and combine it with metadata (e.g. tag, description, duration).

An implemented code available online[21] was modified to perform a combined search with the following parameters:

- **Pitch:** The pitch value taken as target is obtained from the 84-bin semitone spectrogram (each bin represents a musical note).

- **MFCCs:** For each coefficient we use the percentile 25 and percentile 75 as the minimum and maximum values respectively. The first coefficient (MFFC[0]) was not used since it is more related to loudness. After various experiments, it was

[18] https://github.com/MTG/miredu
[19] http://www.freesound.org/docs/api/index.html
[20] http://freesound.org
[21] https://github.com/MTG/freesound-python

necessary to use the coefficients MFCC[1] to MFCC[6]. Using more than those 6 coefficients the system was not able to find similar sounds in some cases.

- **Tags:** The word "*multisample*" refers to single notes recorded from one instrument, in most of the cases these are short events. That word was used to search for harmonic instruments. A second version was created without using any tag, to let the system use a wide range of sounds. For the rhythmic base, the words "kick", "multisample" and "hi-hat" were used for the kick drum, snare, and cymbals respectively.

- **Duration:** This parameter was set from 0 to 8 seconds. Freesound will filter just the sounds that fulfill this condition, so that the time stretching process (discussed in 3.3.) does not require much time. In electronic dance music, most of the notes have short duration, and the same occurs with multisample audios.

Below, it is presented an example of a search query used to find similar sounds with the aforementioned parameters needed for this system:

```
https://www.freesound.org/apiv2/search/combined/?sort=score&ta
rget=lowlevel.pitch.mean%3A55.0&page_size=30&filter=multisampl
e+duration%3A%5B0+TO+8%5D&normalized=1&fields=id%2Cname%2Cprev
iews%2Cusername%2Curl%2Canalysis&descriptors_filter=lowlevel.m
fcc.mean%5B1%5D%3A%5B1.14441e-
05+TO+332.1785%5D+AND+lowlevel.mfcc.mean%5B2%5D%3A%5B-
32.38392+TO+1.6035385%5D+AND+lowlevel.mfcc.mean%5B3%5D%3A%5B-
5.2623295+TO+7.510433%5D+AND+lowlevel.mfcc.mean%5B4%5D%3A%5B-
20.504085+TO+-6.10352e-
05%5D+AND+lowlevel.mfcc.mean%5B5%5D%3A%5B-
14.589985+TO+2.28882e-
05%5D+AND+lowlevel.mfcc.mean%5B6%5D%3A%5B-23.12182+TO+-
2.67029e-05%5D
```

3.3. Song Reconstruction

After retrieving the sounds in the previous stage, it was necessary to perform a time stretching process to modify the length of the downloaded sounds to fit the time of the segment where it belongs. The time stretching was performed using the SMS-tools[22] which are a set of techniques and software implementations for the analysis, transformation and synthesis of musical sounds based on various spectral modeling approaches. The only downside of this tool was the time inverted to do that task, since a song of about 2 minutes has an average of 500 snippets. For that reason we set a threshold, from which, only if the downloaded file was 30% longer or shorter than the segment where it belongs, the time stretching process was performed. This was applied to reduce the number of times that this task should be carried out.

Then every segment obtained in the feature extraction and analysis stage is replaced with the new sound retrieved in the previous stage. To perform that process we use Reaper[23], which is a digital audio workstation that is enabled to load Python scripts via Earscketch[24] API. Figure 3.11 shows the output of the system, a multitrack audio session composed by the different sounds retrieved from Freesound.

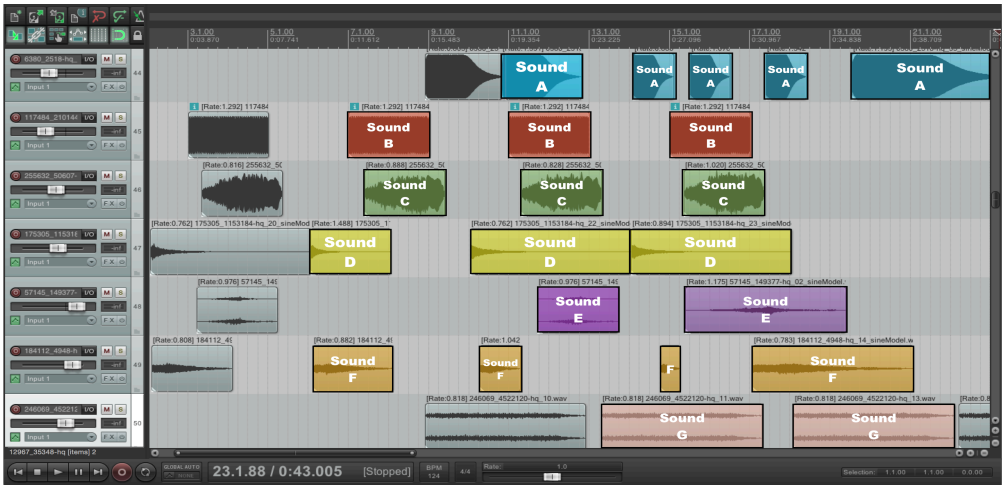

Figure 3.11: Screen Shot of a reconstructed session in Reaper. In one particular track, the same sound is repeated according to the length of the segment. It means that every track corresponds to one particular note of an instrument.

[22] http://mtg.upf.edu/technologies/sms
[23] http://www.reaper.fm
[24] http://earsketch.gatech.edu

3.4. System Evaluation

Since this project is in the field of the creative MIR the evaluation process was subjective following the methodology proposed by Pearce and Wiggins [18]. This evaluation aims to measure the level of preservation (melodic\tonal, rhythm and timbre) that a listener perceives comparing the "seed Song" against the versions created by the system. It also aims to measure the originality and interestingness, two subjective parameters related with the artistic side of this project.

The first part of the test consisted in questions to gather demographic information about the subjects: one question (Q1) about contact information (name and e-mail); two questions about musical background (Q2: Have you ever got some formal musical training (> 1 year)? and Q3: Are you used to listen to songs that use samples or loops?), which are related to the preservation of the melodic\tonal, rhythm and timbre extracted from the "Seed Songs"; and two questions about music creation (Q4: Do you create music by your own? and Q5: Do you use samples or loops in your songs?).

The second part of the test, as shown in Figure 3.12, consisted of a listening study, where subjects first heard the "Seed Song" from which the information was extracted. Five "Seed Songs" were used to create 2 different versions to which the subjects were asked to rate (on a scale from 0/Not At All to 4/Very Much) 5 different parameters:

1. Melodic/Tonal Preservation (Compared to the Seed Song)
2. Rhythm Preservation (Compared to the Seed Song)
3. Timbre Preservation (Compared to the Reference Song)
4. Originality (Compared to the Reference Song)
5. Interestingness (Artistically Speaking)

The two versions presented had one difference: the first version (Multisample version) uses the tag: "multisample" in the Sound Retrieving Stage as explained in 3.2. , and the second version (No tag version) does not use any tag. Every song and its two version were presented in different pages as displayed in Figure 3.12. It is important to mention that the order in which the two versions were presented changed randomly for each song to avoid detection of some possible

trend or systematic difference between the two items presented along the different screens.

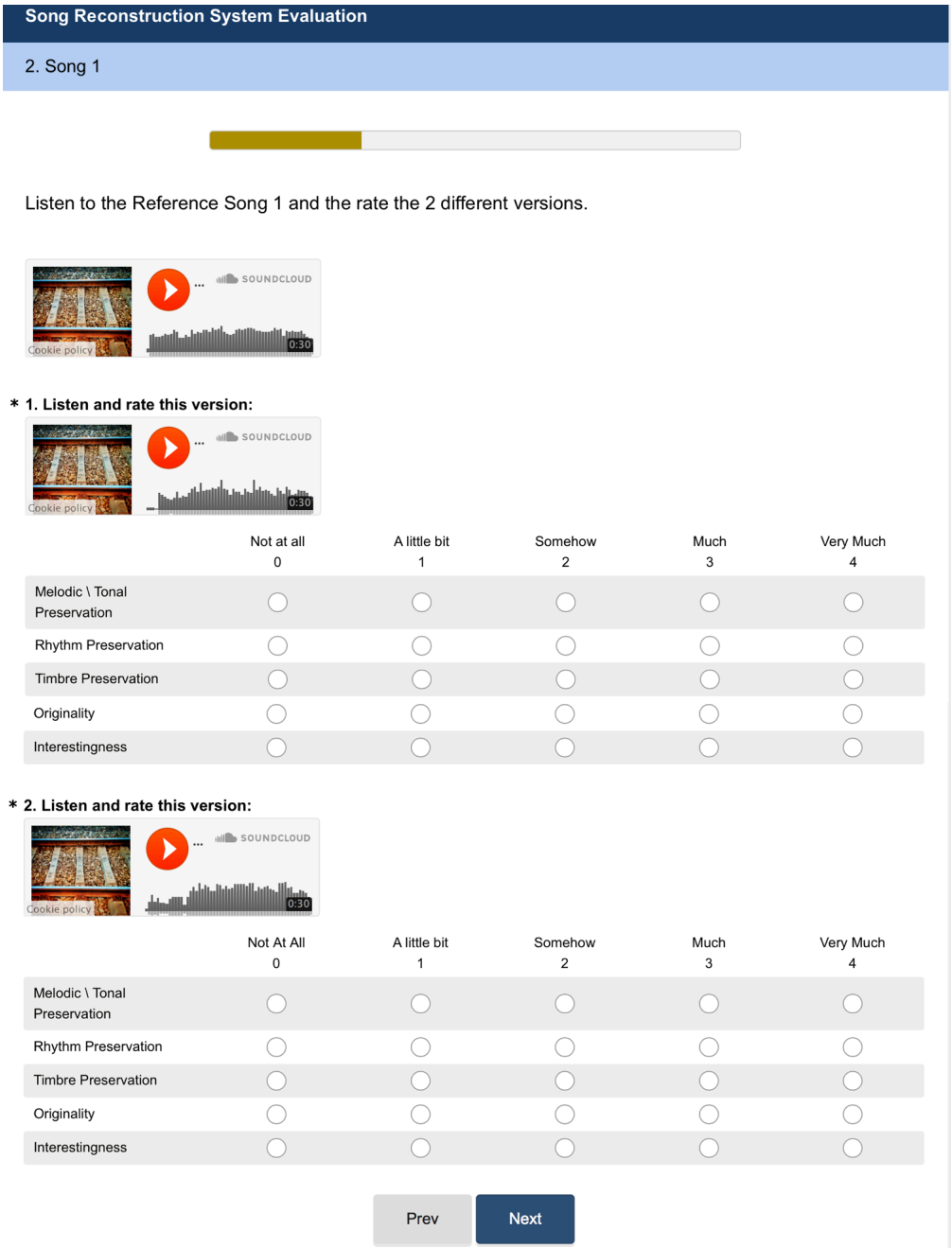

Figure 3.12: Evaluation screen shot. Example of song 1.

To test significant differences between the two versions created by the system we proceeded to make an analysis of variance (ANOVA) for each of the five parameters aforementioned. In the next chapter, we present the analysis of the responses of 25 people.

Chapter 4

4. Results And Discussion

For the evaluation process of this system, an invitation to participate in a survey was sent by e-mail to people related to Sound and Music Computing (MTG) and also to musicians from Ecuador. 25 subjects (11 subjects were from MTG and 14 subjects were musicians from Ecuador) answered the survey which was available online via SurveyMonkey[25]. The text of the demographic questions along with the percentages are shown in Table 4.1.

Table 4.1: Demographic Questions with its respective percentages.

Question	Answer	Percentage
Q1: Contact Information	Name	n/a
	e-mail	n/a
Q2: Have you ever got some formal musical training (> 1 year)?	YES	80%
	NO	20%
Q3: Are you used to listen to songs that use samples or loops?	YES	84%
	NO	16%
Q4: Do you create music by your own?	YES	56%
	NO	44%
Q5: Do you use samples or loops in your songs?	YES	48%
	NO	52%

Questions 1 through 5 (Q1 to Q5) were placed to gather demographic information from the subjects. In Q2: "Have you ever got some formal musical training (> 1 year)?", 80% of the subjects claim to have had formal musical training for at least one year, which suggests that they are familiarized with terms such as Melodic, Tonal, Rhythm and Timbre. Q3: "Are you used to listen to songs that use samples or loops?" was posed to know if the subjects were used to listen to songs that use samples or loops, 84% responded affirmatively, this is also important because audio samples are the foundations of the technique used by the system to create the songs.

[25] https://www.surveymonkey.com/

Since this a system that uses algorithms to create music, it was necessary to gather information to compare if there are differences between the tendencies of the people who compose music and those who do not. Slightly over half (56%) of the individuals compose their own songs, and similarly nearly half (48%) of the subjects use samples or loops in their songs (Q4: "Do you create music by your own?" and Q5: "Do you use samples or loops in your songs?" respectively).

Table 4.2 shows the Means (*M*) and Standard Deviations (*SD*) of the ratings for the 5 parameters evaluated in the survey, the ratings go from 0 (minimum) to 4 (maximum). In nearly all instances, the "Multisample version" has better ratings than the "No tag version". In the case of "Multisample", the parameters related to preservation are in a "positive zone", except for Timbre which mean is 1.7, while in the case of "No Tag", melodic\tonal and timbre preservation fall in the negative zone. The subjective parameters (originality and interestingness) always remain in the positive zone. Standard Deviation values in all cases show that there is no much variability between ratings of each group.

Table 4.2: Means and Standard Deviations on the parameters ratings evaluated of the two versions created by the system ("Multisample version" and "No Tag version")

	Multisample version		No Tag version	
Parameters Evaluated	*M*	*SD*	*M*	*SD*
Melodic \ Tonal Preservation	2.02	1.06	1.35	1.03
Rhythm Preservation	2.47	1.05	2.38	0.91
Timbre Preservation	1.70	1.06	1.14	0.99
Originality	2.65	0.90	2.53	1.00
Interestingness	2.37	1.07	2.39	1.07

Note. The evaluation scale is from 0 (minimum) to 4 (maximum).

Table 4.3 presents a comparison between the global Mean (*M*) and the Mean of the subjects who create music (M_C) of the two versions created by the system. We can observe that although there is not a significant variation, in most of the cases the ratings decrease, except for Melodic\Tonal Preservation, Timbre Preservation and Originality for the "No Tag version", instances in which we observe a small increment in the scores.

Table 4.3: Comparison between the global Mean (M) and the Mean of the subjects who create music (M_C) of the two versions created by the system ("Multisample version" and "No Tag version")

	Multisample version		No Tag version	
Parameters Evaluated	M	M_C	M	M_C
Melodic \ Tonal Preservation	2.02	1.83	1.35	1.40
Rhythm Preservation	2.47	2.29	2.38	2.37
Timbre Preservation	1.70	1.64	1.14	1.26
Originality	2.65	2.59	2.53	2.57
Interestingness	2.37	2.30	2.39	2.36

To test if there are statistically significant differences between the two versions created by the system we proceeded to make an analysis of variance (ANOVA) for each of the five parameters aforementioned. An alpha level of .05 was used for all analyses.

4.1. Melodic \ Tonal Preservation

The analysis of variance for Melodic\Tonal Preservation presented in Table 4.4 shows that there is a significant difference between the means of the two versions, $F(1,248) = 25.27$, $p < 0.05$. The "No Tag version" got a significant lower rate compared with the "Multisample version". This could be due to the tag "Multisample" was chosen to retrieve a higher amount of musical instrument sounds with tonal qualities, and it is very noticeable that many of the sounds retrieve for the "No tag versions" are percussive, human, fx or atonal sounds.

Table 4.4: Analysis of Variance of Melodic\Tonal Preservation between "Multisample version" and "No Tag version"

Analysis of Variance (One-Way) - Melody\Tonal Preservation

Source of Variation	*SS*	*df*	*MS*	*F*	*p-level*
Between Groups	27.556	1	27.56	25.27	9.57E-07
Within Groups	270.48	248	1.09		
Total	298.036	249			

The Mean Square Within also know as *whithin-groups variance estimate* (MS_W = 1.09*)* reflects that there is not too much variability between the ratings obtained within each version. The *between-groups*

variance estimate or Mean Square Between (MS_B = 27.56) also show us that there is a large difference among the group means for the ratings of each version. This is reflected in the Figure 4.1, where almost 40% of the ratings for the "Multisample version" are "2 - Somehow", with a normal distribution, and the ratings for the "No Tag version" has a right-skewed distribution, with more than 40% of the ratings in "1 - A little bit".

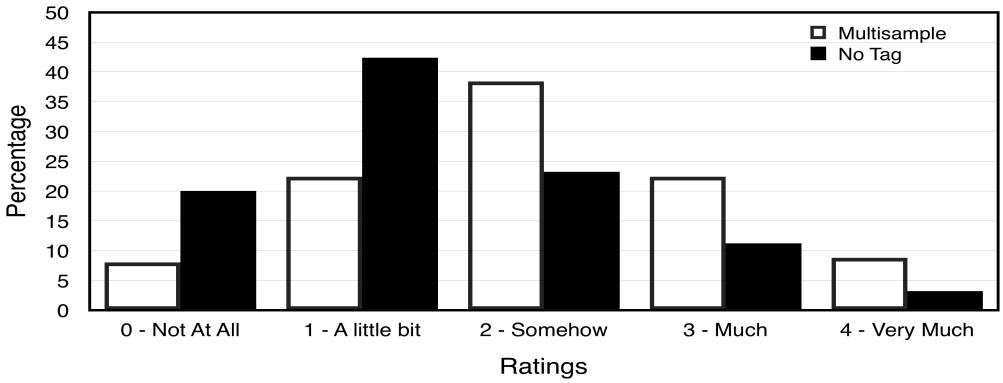

Figure 4.1: Percentage histogram of the rating distributions for "Multisample version" and "No Tag version" on the Melodic\Tonal Preservation parameter

4.2. Rhythm Preservation

In the case of Rhythm Preservation, the analysis of variance contained in Table 4.5 indicates that there is not a significant difference between the means of the two versions, $F(1,248) = 0.59$, $p = 0.44$. For this parameter, the Mean Square Between (MS_B =0.58) and the Mean Square Within (MS_W = 0.97) have values under 1, reflecting low variability of ratings inside each version and between versions. This may be due in part to the rhythm base (kick, snare and cymbals) of both versions was the same.

Table 4.5: Analysis of Variance of Rhythm Preservation between "Multisample version" and "No Tag version"

Analysis of Variance (One-Way) - Rhythm Preservation

Source of Variation	*SS*	*df*	*MS*	*F*	*p-level*
Between Groups	0.576	1	0.58	0.59	0.44
Within Groups	240.48	248	0.97		
Total	241.056	249			

Figure 4.2 shows that the score tendency for Rhythm preservation is towards high values (left-skewed distribution), for instance, in the case of “Multisample version” the ratings “2 - Somehow” and “3 - Much” got more than 30% each one, compare with “No Tag version”, where the same ratings got more than 40% and more than 30% respectively.

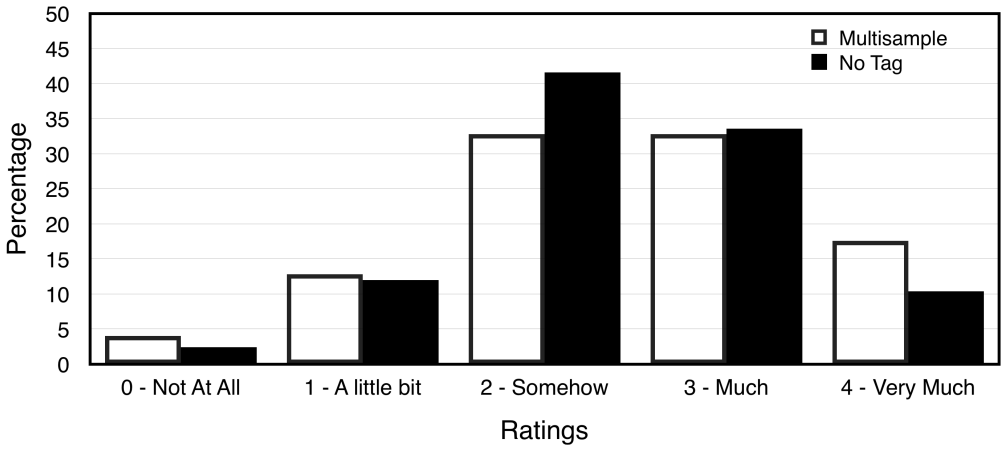

Figure 4.2: Percentage histogram of the rating distributions for "Multisample version" and "No Tag version" on the Rhythm Preservation Parameter

4.3. Timbre Preservation

The ANOVA for Timbre Preservation presented in Table 4.6 shows that there is a significant difference between the means of the two versions, $F(1,248) = 18.72$, $p < 0.05$. The variability within-groups ($MS_W = 1.05$) tells us that there is not too much dispersion between the ratings obtenined within each, moreover, the variability between-groups ($MSB = 18.73$) shows us that there is a large difference among the group means for the ratings of each version.

Table 4.6: Analysis of Variance of Timbre Preservation between "Multisample version" and "No Tag version"

Analysis of Variance (One-Way) - Timbre Preservation

Source of Variation	*SS*	*df*	*MS*	*F*	*p-level*
Between Groups	19.6	1	19.6	18.73	2.18E-05
Within Groups	259.456	248	1.05		
Total	279.056	249			

Furthermore, Figure 4.3 depicts a skewed trend toward low scores for both versions, "Multisample version" (M = 1.7, SD = 1.06) got nearly 40% "1 - A little bit" and more than 30% "2 - Somehow", the "No tag version" (M = 1.14, SD = 0.99) got more than 30 % of "0 - Not At All" and "1 - A little bit". This low score was expected, as explained in section 3.2. , our system uses a range of values (from P25 to P75) from the first 6 MFCCs of each target sounds, because using more than that, the system was not able to found and retrieve sounds in some cases.

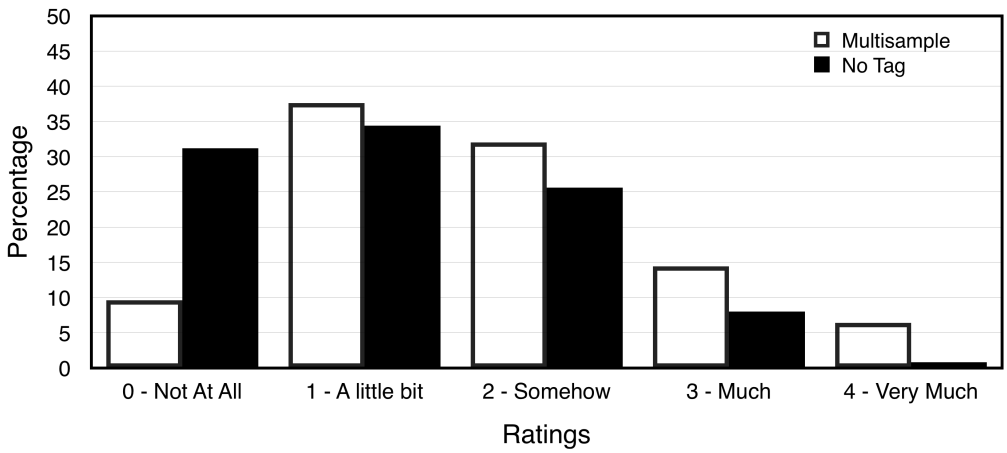

Figure 4.3: Percentage histogram of the rating distributions for "Multisample version" and "No Tag version" on the Timbre Preservation Parameter

4.4. Originality

Since originality is the quality of being novel or unusual, with this parameter we want to measure how far or close the new versions are from the "Seed Song". To evaluate this subjective parameter, the subjects were asked to compare each version with the "Seed Song" and rate how original they thought it was. Therefore, what we expect is that if the scores for Melodic\Tonal, Rhythm and Timbre preservation are low, the scores for Originality should be high and vice versa.

The analysis of variance for Originality provided in Table 4.7, shows that there is not a significant difference between the means of the two versions, $F(1,248) = 1.0$, $p = 0.32$. For this parameter, the Mean Square Between (MS_B =0.9) and the Mean Square Within (MS_W = 0.9) have values close to 1, reflecting low variability of ratings inside each version and between versions.

Table 4.7: Analysis of Variance of Originality between "Multisample version" and "No Tag version"

Analysis of Variance (One-Way) - Originality

Source of Variation	*SS*	*df*	*MS*	*F*	*p-level*
Between Groups	0.9	1	0.9	1.	0.32
Within Groups	223.664	248	0.9		
Total	224.564	249			

The score tendency for Originality illustrated in Figure 4.4 is towards high values (left-skewed distribution), where more than 40% of the ratings are "3 - Much" for both versions. The Mean values presented in Table 4.2 for Originality, "Multisample version" (M = 2.65, SD = 0.9) and "No Tag version" (M = 2.53, SD = 1.0), also show a great acceptance of this parameter, but the latter version was expected to get a higher rating since the sounds used in that version where less tonal and in many cases where not musical sounds (e.g. human speech, sound effects, noises, etc.) which makes them less similar to the "Seed Song" in Melodic\Tonal Preservation and Timbre Preservation, as can be observed in Figure 4.5. One explanation for this issue could be that even though the "No tag version" has lower ratings for the 5 songs they end up being more complex and this could mitigate the musicality perceived by the listener.

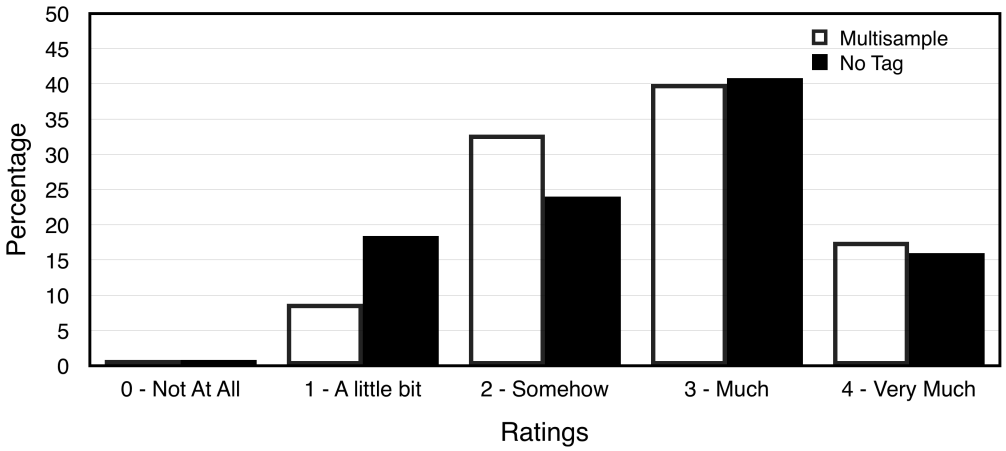

Figure 4.4: Percentage histogram of the rating distributions for "Multisample version" and "No Tag version" on the Originality Parameter

From Figure 4.5 we can also see that there is a direct relation between Rhythm Preservation and Originality, both parameters get similar scores, but there is an opposite trend among Originality and Melodic\Tonal Preservation and Originality and Timbre Preservation,

which is more noticeable for the "No Tag version" except for Song 4. The latter opposite trend was expected as mentioned in the beginning of this section.

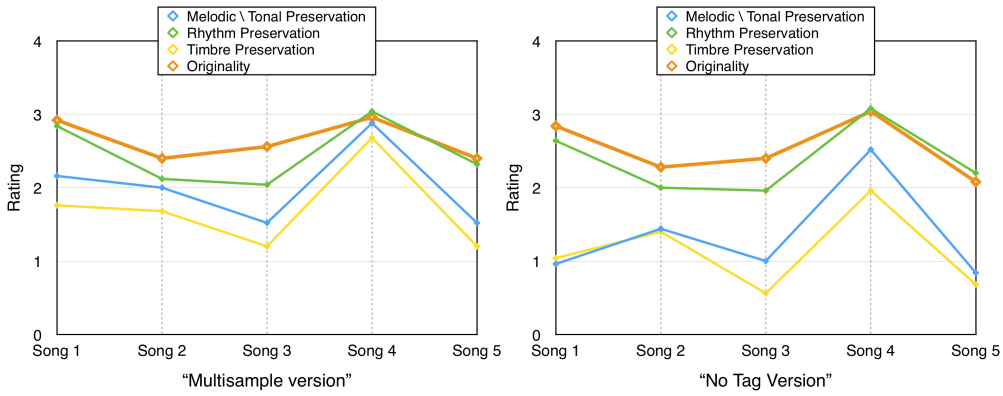

Figure 4.5: Mean ratings for Melodic\Tonal Preservation, Rhythm Preservation, Timbre Preservation and Originality for each of the 5 songs. On the left side the ratings for "Multisample version" and on the right side the ratings for the "No Tag version".

4.5. Interestingness

It is well know that there is considerable resistance to algorithmic composition from all sides, from musicians to the general public [8], for that reason, the idea of evaluate the interestingness is to measure the artistic or musical acceptance level of the different versions created by the system. As mentioned by Schmidhuber in [58]: "What's beautiful is not necessarily interesting. A beautiful thing is interesting only as long as it is new, that is, as long as the algorithmic regularity that makes it simple has not yet been fully assimilated by the adaptive observer who is still learning to compress the data better". A song is interesting it when includes increments and decrements of tension.

With the foregoing expressed, we can proceed to do the analysis of variance for Interestingness. Table 4.8 indicates that there is not a significant difference between the means of the two versions, $F(1,248) = 0.03$, $p = 0.86$. The Mean Square Within ($MS_W = 1.15$) tells us that there is not much variability within the values of each version, and the Mean Square Between ($MS_B = 0.04$) also reveals a very low variability of ratings between versions.

Table 4.8: Analysis of Variance of Originality between "Multisample version" and "No Tag version"

Analysis of Variance (One-Way) - Interestingness

Source of Variation	*SS*	*df*	*MS*	*F*	*p-level*
Between Groups	0.036	1	0.04	0.03	0.86
Within Groups	284.864	248	1.15		
Total	284.9	249			

The aforementioned is also illustrated in Figure 4.6, where we can see that the distributions of ratings are very similar for both versions, and there is not a predominant preference as in the previous parameters. The ratings "1 - A little bit", "2 - Somehow", and "3 - Much" had between 25% and 30% of election each one for the two versions and the Mean values presented in Table 4.2 for Interestingness, "Multisample version" (M = 2.37, SD = 1.07) and "No Tag version" (M = 2.39, SD = 1.07), show a little tendency towards high values but remaining in a neutral zone, which means that the songs created by the system in general are "Somehow" interesting.

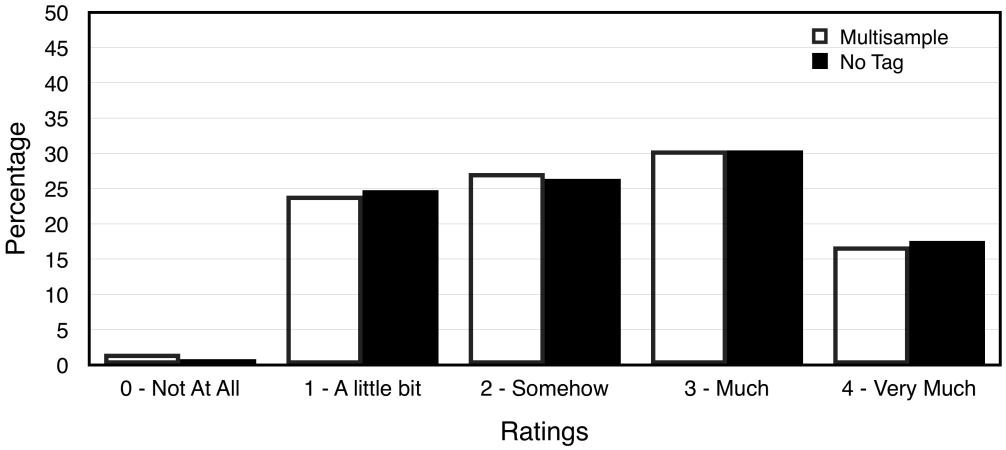

Figure 4.6: Percentage histogram of the rating distributions for "Multisample version" and "No Tag version" on the Interestingness Parameter

From Figure 4.7 we can state that there is a very similar rating tendency between Rhythm Preservation and Interestingness for each song in both versions. Comparing Interestingness with Melodic\Tonal Preservation and Timbre preservation we can observe that in the case of "Multisample version" there is a similar trending in the ratings of the 5 songs and in the case of "No Tag version" an opposite trending can be appreciated, except for Song 4 in which all parameters got the highest scores, this trend is very similar than the one obtained for originality.

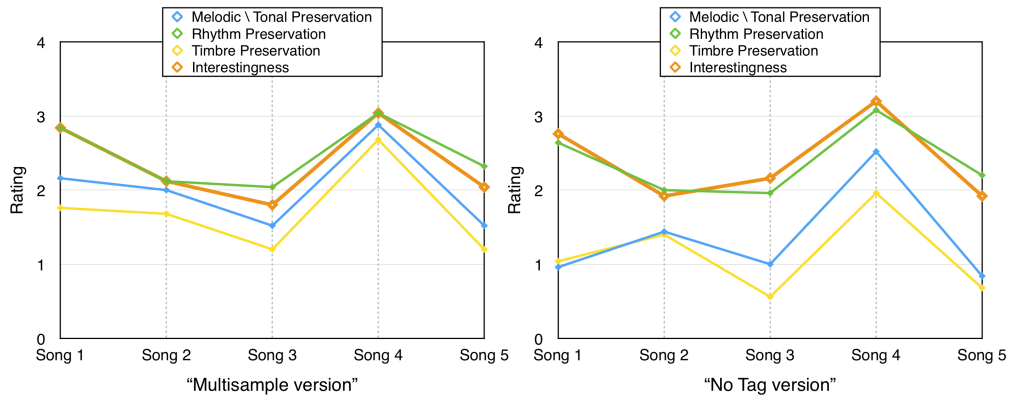

Figure 4.7: Mean ratings for Melodic\Tonal Preservation, Rhythm Preservation, Timbre Preservation and Interestingness for each of the 5 songs. On the left side the ratings for "Multisample version" and on the right side the ratings for the "No Tag version".

These results tell us that although there is not a high perception of preservation (specially Melodic\Tonal and Timbre) compared with the "Seed Song", the two versions created by the system obtained a good level of acceptance that is reflected in the ratings achieved for this parameter, considering that the subjects were asked to rate Interestingness artistically speaking.

Chapter 5

5. Conclusion And Future Work

5.1. Conclusions

We have presented a system for creating new versions of a given song ("Seed Song"). This work has been driven by two main motivations: first, from a creative perspective, we have attempted to demonstrate that the use of MIR techniques can open new musical possibilities and potential advantages for music creation; and second, from a research perspective, to highlight the significant potential of remixing (create new versions of a song) by using multitrack sessions.

The use of the tag "Multisample" helped the system to retrieve more melodic\tonal sounds from Freesound[26], but in the other hand, this limited the amount of the sounds, retrieving the same audio files multiple times within a song and in different songs. Contrary, not using tags increases the variety of sounds but it reduces the melodic\tonal content of them. A similar effect was observed with respect to the number of MFCCs used to preserve timbre, more coefficients used, less variety of sounds, and in some cases no retrievable sound was found.

The results obtained in the survey suggest that the preservation stage should be improved if the goal is to get a closer reconstruction of the "Seed Song", specially the timbre parameter, which obtained the lowest scores.

Since this is a project that is within the creative MIR field, it was important to obtain results that were artistically pleasing to the listeners. The subjective and more artistic parameters (originality and interestingness) got encouraging scores. These results could be interpreted that there is a good level of acceptance musically speaking.

[26] http://www.freesound.org

5.2. Limitations and Future Work

Rhythm fills and sound effects are two categories of sounds that were left aside in this project due to their complexity. These categories might be added into the system to get a more accurate representation of the "Seed Song".

The use of the SMS tools to perform the time stretching process, forced us to introduce the “Duration” parameter (from 0 to 8 seconds) for the audio clips in the sound retrieval stage, because these tools needed a lot of time to perform that process. Other tools, such as Rubber Band Library[27], can be used to improved the time stretching process and therefore to be able to work without this constraint.

Two very important qualities in the sonority of EDM are the “Loudness variation” and “Frequency Variation”; these are widely used resources to create increments and decrements of “tension” within the song. The preservation of these two parameters could be added to the system to thereby obtain a more musical reconstruction.

Extracting mid and high-level music content descriptions from the “Seed Song” and applying musical knowledge, could give the user a deeper interaction and control over the creative process (e.g. allowing the user to change the key of the song). This could be integrated within an interface in which the user has a clearer and more visual control of the parameters that can be handled, giving the user greater choice and decision on the final result.

[27] http://breakfastquay.com/rubberband/

References

[1] A. Tanaka, N. Tokui, and A. Momeni, "Facilitating collective musical creativity," in *Proceedings of the 13th annual ACM international conference on Multimedia - MULTIMEDIA '05*, 2005, p. 191.

[2] S. Gresham-Lancaster, "The Aesthetics and History of the Hub: The Effects of Changing Technology on Network Computer Music," *Leonardo Music J.*, vol. 8, p. 39, 1998.

[3] Á. Barbosa, "Displaced soundscapes: A survey of network systems for music and sonic art creation," *Leonardo Music J.*, vol. 13, no. 2003, pp. 53–59, 2003.

[4] G. Cheliotis and J. Yew, "An analysis of the social structure of remix culture," in *Proceedings of the fourth international conference on Communities and technologies - C&T '09*, 2009, p. 165.

[5] G. Cheliotis, "From open source to open content: Organization, licensing and decision processes in open cultural production," *Decis. Support Syst.*, vol. 47, no. 3, pp. 229–244, Jun. 2009.

[6] V. Akkermans, F. Font, J. Funollet, B. De Jong, G. Roma, S. Togias, and X. Serra, "Freesound 2: An Improved Platform for Sharing Audio Clips," *Int. Soc. Music Inf. Retr. Conf. (ISMIR 2011), Late-breaking Demo Sess.*, 2011.

[7] M. Casey, C. Rhodes, and M. Slaney, "Analysis of Minimum Distances in High-Dimensional Musical Spaces," *IEEE Trans. Audio. Speech. Lang. Processing*, vol. 16, no. 5, pp. 1015–1028, Jul. 2008.

[8] M. Edwards, "Algorithmic composition," *Communications of the ACM*, vol. 54, no. 7, ACM, pp. 58–67, 01-Jul-2011.

[9] X. Serra, M. Magas, E. Benetos, M. Chudy, S. Dixon, A. Flexer, E. Gomez, F. Gouyon, P. Herrera, S. Jorda, O. Paytuvi, G. Peeters, J. Schlüter, H. Vinet, and G. Widmer, *Roadmap for Music Information ReSearch*. MIRES Consortium, 2013.

[10] J. Shiga, "Copy-and-Persist: The Logic of Mash-Up Culture," *Crit. Stud. Media Commun.*, vol. 24, no. 2, pp. 93–114, Jun. 2007.

[11] G. Griffin, Y. E. Kim, and D. Turnbull, "Beat-sync-mash-coder: A web application for real-time creation of beat-synchronous music mashups," in *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings*, 2010, pp. 437–440.

[12] M. E. P. Davies, P. Hamel, K. Yoshii, and M. Goto, "AutoMashUpper: Automatic Creation of Multi-Song Music Mashups," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 22, no. 12, pp. 1726–1737, Dec. 2014.

[13] J. Scott, M. Prockup, E. Schmidt, and Y. Kim, "Automatic Multitrack Mixing Using Linear Dynamical Systems," in *Proceedings of the 8th Sound and Music Computing Conference*, 2011, pp. 177–182.

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

[15] B. Pardo, D. Little, and D. Gergle, "Building a personalized audio equalizer interface with transfer learning and active learning," in *MIRUM*, 2012, p. 13.

[16] M. B. Cartwright and B. Pardo, "Social-EQ: Crowdsourcing an Equalization Descriptor Map," in *ISMIR*, 2013, pp. 395–400.

[17] A. Arzt, G. Widmer, S. Böck, R. Sonnleitner, and H. Frostel, "Towards a complete classical music companion," in *Frontiers in Artificial Intelligence and Applications*, 2012, vol. 242, pp. 67–72.

[18] M. Pearce and G. Wiggins, "Evaluating Cognitive Models of Musical Composition," in *Proceedings of the 4th International Joint Workshop on Computational Creativity*, 2007, pp. 73–80.

[19] T. Collins, "Improved methods for pattern discovery in music, with applications in automated stylistic composition," PhD thesis, Faculty of Mathematics, Computing and Technology, The Open University, 2011.

[20] N. Gonzalez Thomas, P. Pasquier, A. Eigenfeldt, and J. B. Maxwell, "Meta-Melo: A system and methodology for the comparison of melodic generation models," in *Proceedings of the International Symposium on Music Information Retrieval*, 2013.

[21] D. C. Moffat and M. Kelly, "An investigation into people's bias against computational creativity in music composition," in *The Third Joint Workshop on Computational Creativity*, 2006.

[22] C. Anderson, A. Eigenfeldt, and P. Pasquier, "The Generative Electronic Dance Music Algorithmic System (GEDMAS)," in *Proceedings of the Artificial Intelligence and Interactive Digital Entertainment (AIIDE'13) Conference.*, 2013, pp. 5–8.

[23] S. Brooks, B. J. Ross, S. Leonardo, M. Journal, and B. Ross, "Automated Composition from Computer Models of Biological Behavior," *Leonardo*, vol. 6, no. 1996, pp. 27–31, 2014.

[24] C. Ó. Nuanáin and L. O. Sullivan, "Real-time algorithmic composition with a tabletop musical interface," in *Proceedings of the 9th Audio Mostly on A Conference on Interaction With Sound - AM '14*, 2014, pp. 1–7.

[25] J. D. Fernández and F. Vico, "Ai methods in algorithmic composition: A comprehensive survey," *J. Artif. Intell. Res.*, vol. 48, pp. 513–582, Feb. 2013.

[26] M. Supper, "A few remarks on algorithmic composition," *Comput. Music J.*, vol. 25, no. 1, pp. 48–53, 2001.

[27] J. O'Connell, "Musical Mosaicing with High Level Descriptors," Universitat Pompeu Fabra, 2011.

[28] C. Roads, *Microsound*. Cambridge (Mass.) : MIT Press, 2001.

[29] B. L. Sturm, "Adaptive Concatenative Sound Synthesis and Its Application to Micromontage Composition," *Comput. Music J.*, vol. 30, no. 4, pp. 46–66, 2006.

[30] M. Neupert and J. Gossmann, "A remix instrument based on fragment feature-analysis," in *2013 IEEE International Conference on Multimedia and Expo Workshops (ICMEW)*, 2013, pp. 1–5.

[31] P. Herrera, J. Serrà, C. Laurier, E. Guaus, E. Gómez, and X. Serra, "The Discipline formerly known as MIR," *Int. Soc. Music Inf. Retr. Conf. Spec. Sess. Futur. MIR*, 2009.

[32] A. Wang, "The Shazam music recognition service," *Communications of the ACM*, vol. 49, no. 8. ACM, p. 44, 01-Aug-2006.

[33] M. Schedl, E. Gómez, and J. Urbano, "Music Information Retrieval: Recent Developments and Applications," *Found. Trends Inf. Retr.*, vol. 8, no. 2–3, 2014.

[34] D. V. S. M. Leman, L. P. Clarisse, B. De Baets, H. De Meyer, M. Lesaffre, J.P. Martens, Martens J. P, "Tendencies, Perspectives, and Opportunities of Musical Audio-Mining," 2002.

[35] G. Peeters, "A large set of audio features for sound description (similarity and classification) in the CUIDADO project," in *IRCAM*, 2004.

[36] M. Mauch and S. Dixon, "PYIN: A fundamental frequency estimator using probabilistic threshold distributions," in *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings*, 2014, pp. 659–663.

[37] J. Salamon and E. Gomez, "Melody Extraction From Polyphonic Music Signals Using Pitch Contour Characteristics," *IEEE Trans. Audio. Speech. Lang. Processing*, vol. 20, no. 6, pp. 1759–1770, Aug. 2012.

[38] E. Gómez, "Tonal Description of Music Audio Signals," 2006.

[39] M. Mauch, "Automatic chord transcription from audio using computational models of musical context," Queen Mary University of London, 2010.

[40] M. McVicar, R. Santos-Rodríguez, Y. Ni, and T. De Bie, "Automatic chord estimation from audio: A review of the state of the art," *IEEE Trans. Audio, Speech Lang. Process.*, vol. 22, no. 2, pp. 556–575, Feb. 2014.

[41] J. Paulus, M. Müller, and A. Klapuri, "Audio-based music structure analysis," in *11th Int. Soc. Music Inf. Retrieval Conf.*, 2010, pp. 625–636.

[42] J. P. Bello, L. Daudet, S. Abdallah, C. Duxbury, M. Davies, and M. B. Sandler, "A tutorial on onset detection in music signals," *IEEE Trans. Speech Audio Process.*, vol. 13, pp. 1035–1046, 2005.

[43] H. L. Tan, Y. Zhu, L. Chaisorn, and S. Rahardja, "Audio onset detection using energy-based and pitch-based processing," in *Proceedings of 2010 IEEE International Symposium on Circuits and Systems*, 2010, pp. 3689–3692.

[44] J. Foote, "Automatic audio segmentation using a measure of audio novelty," *2000 IEEE Int. Conf. Multimed. Expo. ICME2000. Proceedings. Latest Adv. Fast Chang. World Multimed. (Cat. No.00TH8532)*, vol. 1, pp. 452–455, 2000.

[45] B. Rocha, N. Bogaards, and A. Honingh, "Segmentation and timbre- and rhythm-similarity in Electronic Dance Music," *SMC 2013*, pp. 1–29, Sep. 2013.

[46] T. Scarfe, W. Koolen, and Y. Kalnishkan, "Segmentation of electronic dance music," *Int. J. Eng. Intell. Syst. Electr. Eng. Commun.*, pp. 1–18, 2014.

[47] S. Hargreaves, A. Klapuri, and M. Sandler, "Structural Segmentation of Multitrack Audio," *IEEE Trans. Audio. Speech. Lang. Processing*, vol. 20, no. 10, pp. 2637–2647, Dec. 2012.

[48] A. Berenzweig, B. Logan, D. P. W. Ellis, and B. Whitman, "A Large-Scale Evaluation of Acoustic and Subjective Music-Similarity Measures," *Computer Music Journal*, vol. 28, no. 2. MIT Press, pp. 63–76, 01-Jun-2004.

[49] E. Wold, T. Blum, D. Keislar, and J. Wheaton, "Content-Based Classification, Search, and Retrieval of Audio," *IEEE Multimed.*, vol. 3, no. 3, pp. 27–36, 1996.

[50] G. Tzanetakis and P. Cook, "Musical genre classification of audio signals," *IEEE Trans. Speech Audio Process.*, vol. 10, no. 5, pp. 293–302, Jul. 2002.

[51] J. Herre, E. Allamanche, and C. Ertel, "How similar do songs sound? Towards modeling human perception of musical similarity," *2003 IEEE Work. Appl. Signal Process. to Audio Acoust. (IEEE Cat. No.03TH8684)*, pp. 83–86, 2003.

[52] F. Pachet and J.-J. Aucouturier, "Improving timbre similarity: How high is the sky," *J. Negat. results speech audio Sci.*, vol. 1, no. 1, pp. 1–13, 2004.

[53] M. J. Butler, *Unlocking the Groove: Rhythm, Meter, and Musical Design in Electronic Dance Music*. Indiana University Press, 2006.

[54] M. Mauch and S. Dixon, "Approximate note transcription for the improved identification of difficult chords," in *11th Int. Soc. Music Inf. Retrieval Conf.*, 2010, pp. 135–140.

[55] C. Cannam, M. O. Jewell, C. Rhodes, M. Sandler, and M. D'Inverno, "Linked Data And You: Bringing music research software into the Semantic Web," *J. New Music Res.*, vol. 39, no. 4, pp. 313–325, 2010.

[56] E. Zwicker and H. Fastl, *Psychoacoustics, Facts and Models*. Berlin: Springer Verlag, 1990.

[57] J. O. Smith and J. S. Abel, "Bark and ERB bilinear transforms," *IEEE Trans. Speech Audio Process.*, vol. 7, no. 6, pp. 697–708, 1999.

[58] J. Schmidhuber, "Simple algorithmic theory of subjective beauty, novelty, surprise, interestingness, attention, curiosity, creativity, art, science, music, jokes," *J. SICE*, vol. 48, pp. 21 – 32, 2009.