FACULDADE DE ENGENHARIA DA UNIVERSIDADE DO PORTO

Evaluating Harmonic Features in Automatic Music Mashup Creation

Noémia Cardoso Ferreira

Mestrado em Engenharia Eletrotécnica e de Computadores

Supervisor: Gilberto Bernardes

October 14, 2022

Resumo

Neste estudo pretende propor-se um novo sistema que melhore tanto a eficiência como a eficácia na determinação de compatibilidade entre faixas musicais dentro de um dataset. Os modelos existentes apresentam algumas limitações em termos de eficácia, não só porque os métodos são principalmente focados no ritmo, mas também porque os métodos que levam em consideração a harmonia não assumem o seu carácter como variável. Para além disso, ao eliminar a necessidade de processar uma música inteira, o custo computacional da determinação da Compatibilidade Harmónica diminui e o processo é significativamente mais rápido, o que significa que a tecnologia está estaria mais próxima de uma aplicação em tempo e contexto reais. Ao abordar os problemas anteriores, visamos, em última análise, a aplicação de métodos computacionais para compatibilidade harmónica em cenários de aplicação da vida real. Para completar o estudo, pretendemos avaliar se é possível determinar a compatibilidade harmónica a partir de uma janela temporal menor e determinar ainda sua localização ideal numa faixa.

É essencial determinar uma *ground truth* para os dados utilizados em testes, o que revela a importância do teste de escuta. Ademais, este teste também permite determinar qual métrica (distância euclidiana, distância de cossenos ou entropia) se alinha melhor a classificação de quão agradável o resultado de uma mashup é.

Em relação aos resultados do trabalho, que serão detalhados neste documento, a distância euclidiana mostrou-se novamente como a melhor métrica para avaliação. Para aém disso, os melhores resultados foram obtidos para a análise de toda a música quando comparados a outras duas janelas de tempo em locais diferentes.

Keywords: *harmonic mixing, tonal interval space, mashup, compatibilidade, mir.*

Abstract

In this study we aim to propose a new system which improves the efficiency and efficacy in retrieving compatible musical audio tracks from a predefined dataset. Existing systems present some limitations in terms of efficacy, not only because the methods are mainly focused on rhythm but also because the methods that take into consideration the harmony don't account for its variation. Moreover, by eliminating the need of processing an entire song, the computational cost of the Harmonic Compatibility determination decreases and the process is significantly faster, which means that this technology is a step closer to a real-time application. In tackling the former problems, we ultimately target the application of computational methods for harmonic compatibility in real-life application scenarios. In order to complete the study, we aim to evaluate whether it is possible to determine HC from a shorter time window and its optimal location.

It is essential determining a ground truth for the data, which means that the perceptual listening test was very important. Moreover, this test also allows us to determine which metric (Euclidean distance, cosine distance and entropy) aligns better with human enjoyment.

Regarding the results of the work, which will be fully detailed in this document, it is fair to say that the Euclidean distance proved to be the best metric for evaluation. Also, the best results were obtained for the analysis of the whole song when compared to other two time windows in different locations.

Keywords: *harmonic mixing, tonal interval space, mashup, compatibility, mir.*

Vi

Agradecimentos

Gostaria de começar por agradecer ao Professor Gilberto Bernardes que, para além de me ter acompanhado ao longo de todo o desenvolvimento mais técnico da minha dissertação, foi quem me introduziu ao mundo da música dentro da engenharia, tanto através de aulas como palestras ou *workshops*.

Também me acompanharam de perto o Daniel e o Macedo, cuja motivação foi constante uma vez que também desenvolveram o próprio trabalho nesta área, que para os três é de interesse até pessoal.

Gostaria de agradecer também aos meus amigos, porque as histórias e memórias da faculdade não teriam tanta cor. Porque sei que nunca vou estar sozinha. À Eduarda, ao Rafa, à Inês e à Filipa.

A minha mãe também merece especial menção, uma vez que não só é uma referência a nível de trabalho e empenho, mas também porque foi quem me ensinou a crescer e quem me moldou.

Por último, mas sempre importante, ao Bruno. Por todos os pequenos almoços, paciência e dedicação infinitos, pelos abraços sem hora marcada, pela felicidade. E ainda ao meu companheiro de muitos dias e muitas noites, que me ensinou o que é o amor incondicional. Ao Platão.

Noémia

"Music gives a soul to the universe, wings to the mind,
flight to the imagination and life to everything."

Plato

Contents

List of Figures

List of Tables

Abbreviations

CD	Cosine Distance
DFT	Discrete Fourier Transform
ED	Euclidean Distance
EDM	Electronic Dance Music
HC	Harmonic Compatibility
HPCP	Harmonic Pitch Class Profile
MIR	Music Information Retrieval
PCC	Pearson Correlation Coefficient
TIS	Tonal Interval Space
TIV	Tonal Interval Vector
WAV	Waveform Audio File Format

Chapter 1

Introduction

1.1 Context

Musical mashup creation is the process that combines musical audio tracks to create a new work and it is widely associated with DJ practice and electronic music. The search for compatible tracks to be mashup has been pursued computationally to aid musicians browsing and navigating personal or public musical audio data sets. The search for musical compatible audio considers many sound characteristics. This process is essential to the DJ practice, and it has been an active field of research within the Sound and Music Computing and creative Music Information Retrieval communities for some years, particularly in what has to do with harmonic compatibility measures.

In order to better understand harmonic compatibility, its many components and its computational methods, we shall detail the Tonal Interval Space, a space where state-of-the-art harmonic compatibility metrics have been proposed. Using this space, we can not only describe the harmonic content of musical audio with some degree of perceptual awareness, but also infer the harmonic compatibility of musical audio tracks as distances in the space, which is one of the objectives of the dissertation.

1.2 Motivation

In existing state-of-the-art methods many problems have been identified, in particular, the definition of the minimal temporal analysis windows and harmonic features resolution with good accuracy and reduced computational cost. Furthermore, it is crucial for the method used to manage to detect tuning deviations.

Even though there is a wide collection of audio files available, free of composer rights, ready to be used for DJ practice, the lack of annotation makes it very difficult to combine with other tracks, making it very difficult for music production. To solve that problem, it is important to develop a method that gives a value of compatibility whilst not analyzing a file, which would have a high computational demand.

Another central point of this dissertation has to do with the database used for testing. It is important to have a wide range of samples and challenging examples so that the work will prove to have a real-world application.

1.3 Objectives

To achieve novel methods for computational mashup creation with greater efficiency and efficacy in relation the to the state-of-the-art, the following four objectives will be pursued:

1. To define harmonic (i.e., audio signal representations) and distance metrics for capturing the harmonic compatibility;

2. To determine the optimal temporal location of musical audio track (e.g., beginning, middle and end time locations) to assess musical audio compatibility, aiming to increase the efficiency of the methods at scale.

3. To assess the minimal temporal resolution that assures high accuracy in capturing mashup compatibility

4. To validate the proposed model(s) by a perceptual evaluation.

1.4 Methodology

This dissertation aims to develop a method that computes the harmonic compatibility of two or more musical audio files.

To conduct objective 1, to define harmonic descriptors and distance metrics to further compute harmonic similarity of different musical audio, the tonal interval space will be adopted to infer entropy measures which can compute the complexity of the signal at the harmonic and rhythmic domains. Different metrics from the literature will be assessed (e.g, Euclidean and cosine distances).

To conduct objectives 2 and 3, to determine the optimal location of a musical of musical audio track and its minimal temporal resolution, we must first build a small data set of full musical audio tracks with challenging examples for compatibility computation. Then, we will study different segmentation strategies (blind and structurally-informed) to study the minimal temporal resolution that ensures a good model of harmonic and rhythmic compatibility between two or more musical audio tracks. In greater detail, we aim to understand whether it is possible to compute the HC value on very short segments of time of a sample, for example, *5 seconds*, as a representation of the full track compatibility. Having this analysis taken down to a minimum, the computational cost will drastically decrease, which would imply that this approach could be used almost in real time mixing for DJ practice. Furthermore, we will assess which location of a musical sample is better prone to indicate the track compatibility. Given the analysis in different parts of the musical audio

– the beginning, the middle or the end –, it will be possible to determine where in the temporal dimension the compatibility metric best approximates perceptual judgments.

To conduct objective 4, an online survey will be conducted in which users will be asked to either choose from a scale the likability of a mashup or their preference between two different mashups.

1.5 Document Structure

This document aims to present the literature review for the the masters dissertation, as well its developments and its results. The first chapter is the introduction, embracing the context, motivation and methodology. Chapter 2 represents the literature review per se, presenting techniques to calculate harmonic compatibility and methods to evaluate the final project. Chapter 3 consists of the work done, detailing some specifics of the algorithm and the evaluation metrics. Chapter 4 outlines the evaluation procedure and its results. Chapter 5 sums up the conclusions of the work and provides some insight on possible development of future work and current limitation of the purposed model.

Chapter 2

Computer-aided Mashup creation: State-of-the-art

The result of harmonic mixing tends to be very dependent on the harmonic similarity and compatibility of two different audio samples. In order to be able to compute a degree of similarity, an effective way to compare samples is through their representation using distance metrics. So, we relate harmonic compatibility to different forms of distance.

According to the objectives defined for the dissertation, it is important to separate harmonic, timbral and rhythmic compatibility so that all the different approaches to combining melody can be taken into account.

2.1 Tonality and Harmony

Tonality within Western music relates to the concept of an equal tempered tuning, which divides the musical space in 12 notes, or pitch classes (C, C#, D, D#, E, F, F#, G, G#, A, A#, B), and the spacing between each of those notes corresponds to half steps. Its base, the tonal system, includes tonal structures that arrange sounds according to their temporal and spacial structures, such as keys, chords and melody [1]. Dividing them in two large groups, there is harmony, which relates to vertical pitch structures, and melody, which relates to horizontal pitch structures [2]. In other words, harmony relates to the synchronous structures and melody relates to the sequential ones.

2.2 Computational tonal features

This section presents several concepts and methods relevant for the work that describe the tonal dimension of an audio signal.

2.2.1 Chroma Vector

Chroma vectors are widely used to describe the tonal content of musical audio as they represent the sum of the energy of each pitch class[1]. These vectors can then be used to compare different melodies using different computing methods, all of them calculating pitch distances.

In the figure presented bellow, there is an example of a C major scale in four different representations. There is its musical form in *(a)*, and its audio recording in *(c)*. Both *(b)* and *(d)* represent chromagrams, *(b)* having been originated from symbolic representation and *(d)* from an audio file. In fact, the first chromagram shows an ideal representation, being very clean, whereas the second one, *(d)*, counts with both the major scale, represented in red as it has the most energy, and some harmonics residue, which can be seen in orange and in yellow.

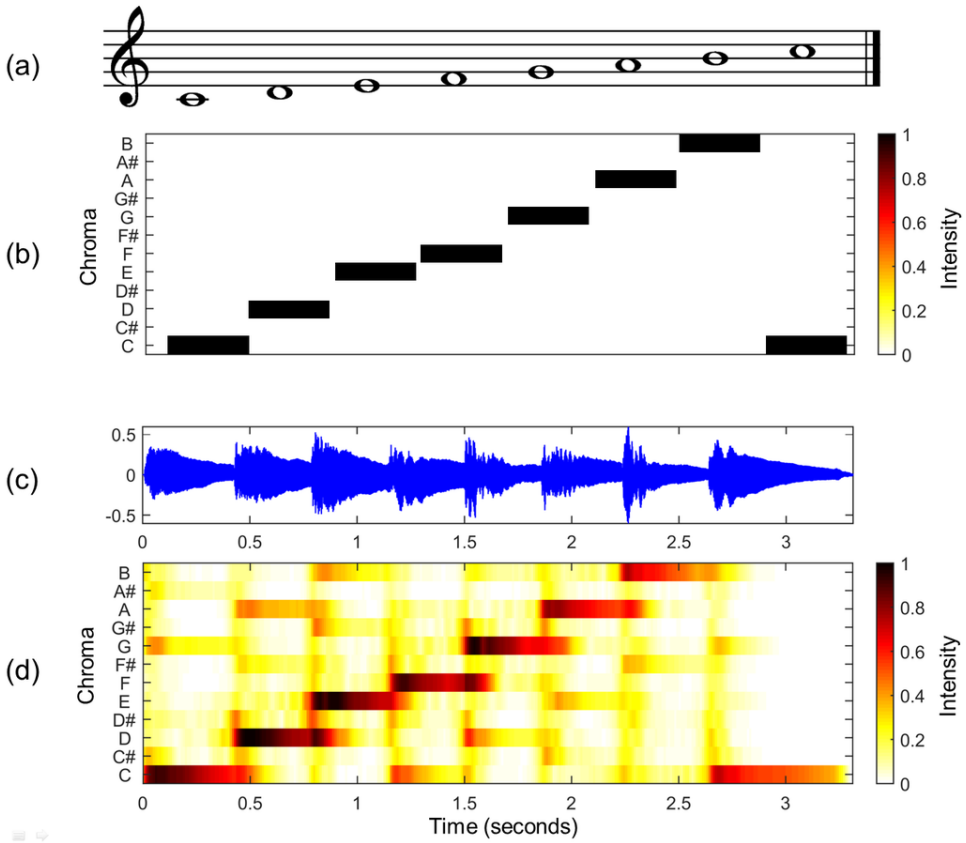

Figure 2.1: (a) Musical score of a C-major scale. (b) Chromagram obtained from the score. (c) Audio recording of the C-major scale played on a piano. (d) Chromagram obtained from the audio recording. Reproduced from Meinard Mueller's original image at "https://en.wikipedia.org/wiki/Chroma feature".

2.2.2 Tonal Interval Space

Chroma vectors have been shown to provide poor indicators in comparing their relatedness. So, as a result, another approach was developed, and that is why the Tonal Interval Space (TIS) accounts for its limitations [1] [3] []. TISs does not only consider pitch distances reflecting human

perception of pitch, but it also gives a tonal pitch consonance indicator.

Different musical properties, such as pitch classes, intervals, chords and keys can be represented in the TIS through Tonal Interval Vectors (TIV), $T(k)$, making use of the Discrete Fourier Transform (DFT), as shown in Equation 2.1. The DFT components are originated from an L_1 normalized chroma vector and are weighted by $w(k) = \{2, 11, 17, 16, 19, 7\}$, as shown in Equation 2.2. There are many different parameters in the equation, being n the dimension, reaching $N = 12$ as it is the total number of pitch classes, the chroma vector pitch class index, k, $w(k) = \{2, 11, 17, 16, 19, 7\}$ and $\bar{c}(n)$ normalized chroma vector.

The result of this operation is a 12-dimensional Tonal Interval Vector (TIV), having each dimension of such vector a direct relation to a music interval. In Figure 2.2 it is possible to observe 6 of the 12 TIV of a TIS computation, in this case a C Major chord [1].

$$T(k) = w(k) \sum_{n=0}^{N-1} \bar{c}(n) e^{-\frac{j2\pi kn}{N}}, k \in \mathbb{Z} \tag{2.1}$$

$$\bar{c}(n) = \frac{c(n)}{\sum_{n=0}^{N-1} c(n)} \tag{2.2}$$

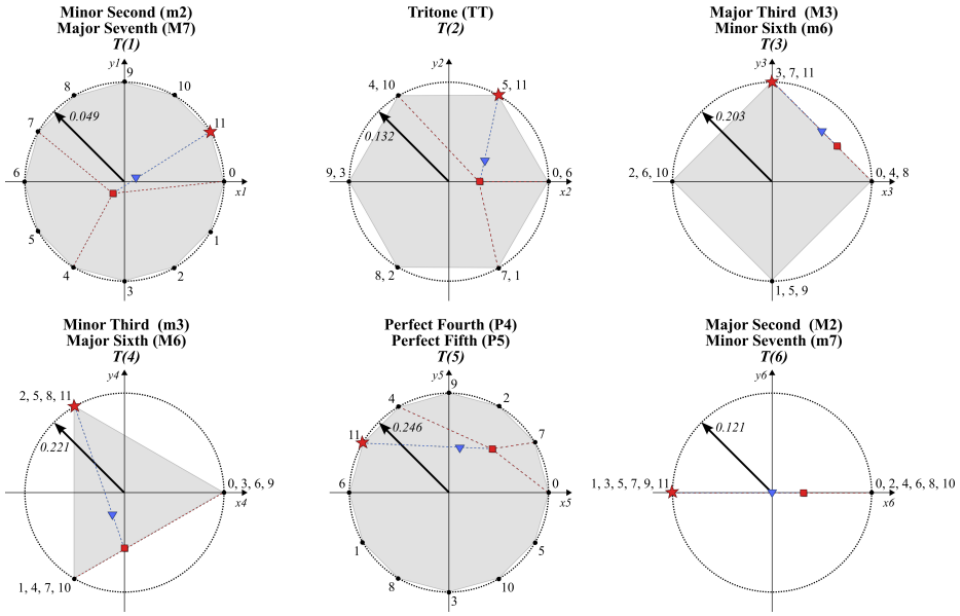

Figure 2.2: Visualisation of 6 of the 12 the TIV as six circles organized according to complementary intervals.

2.3 Harmonic Compatibility

Harmonic compatibility is typically defined as the resulting perceptual enjoyment between the harmonic content of two (or more) musical excerpts, when played together (i.e., overlapped)[4]. Recently, some studies have been expanding the concept to equally address the harmonic compatibility between temporal sequences (i.e., horizontal compatibility) in addition to the more tradition overlapping compatibility (i.e., vertical compatibility). In this work, we will mostly address vertical compatibility when referring to harmonic compatibility, unless it is explicitly referred as sequential compatibility. [2]

To compute harmonic compatibility many strategies have been proposed in the literature, and the main three will be detailed in the next subsections. Key affinity is a method that determines HC and it is the most used technology. Both chroma matching and the tonal interval space have to do with spectral similarity, and these are the best options when it comes to technicalities. Finally, the psycho-acoustic dissonance methods give a perceptual data, as they have a user classifying measures of consonance and dissonance.

2.3.1 Key Affinity

In a musical context, the perception of distance takes into consideration intervals between keys, relating major and minor keys, respectively, to what is called a minor relative. Also, the number of alterations needed to go from one key to another is related to their distance in the circle. A visual way to show this relation is the *Circle of Fifths*, presented in figure 2.3, having its major keys represented in red with their respective minor relatives represented in green. In here, it is possible to organize all twelve pitch classes as a sequence of perfect fifths (the interval between a pitch class must be exactly three steps and one-half step). According to this representation, it is very straightforward to get from a major model to a minor one and to get fifth intervals, but other intervals are poorly represented, which means the Harmonic Compatibility representation is very limited [5].

Within the DJ community, this is also referred to as the Camelot Wheel.

In a commercial environment, there are tools like Mixed In Key's Mashup2 [6] which is the best state-of-the-art software in key notation [4] and Traktor [7], a DJ software mostly used for live performing and streaming, which used techniques such as time stretching and pitch shifting.

2.3.2 Chroma Matching

Chroma matching, or chroma vector similarity, inspects the cosine distance between chroma vector representations of different pitch shifted versions of two different audio tracks. Because this is usually computed in a beat-level, it provides a small-scale alignment [1].

Some of the technologies developed so far are loop-based, like Mixmash[8] or Automashupper [9], developed in a more academic setting, both having a generative upbringing, which can lead to scaling barriers.

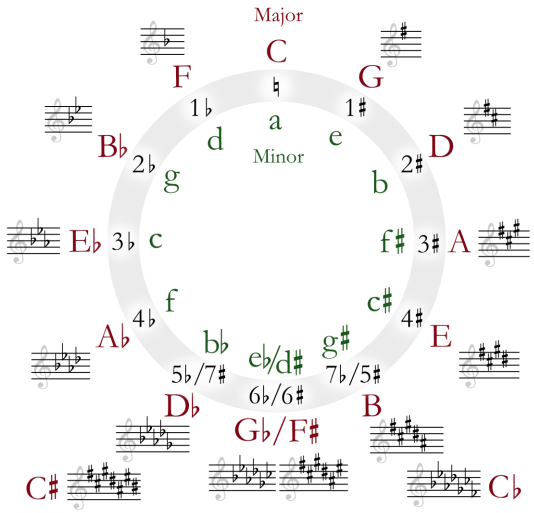

Figure 2.3: Circle of Fifths. Represents relations between keys.

2.3.3 Psycho-acoustic Dissonance Methods

Another approach to Harmonic Compatibility relies on a more subjective method, which is related to consonance and dissonance. As it is widely known in the music community, two consonant sounds tend to be more appealing to the human hearing than two dissonant ones [10].

There have already been developed different methods that represent this perception, them being pure dyads and roughness [10]. In music theory, a dyad consists of a chord of two notes that are played simultaneously, being a pure dyad a chord of two pure notes. In this context, roughness consists of a sound not being perceived as a pure note.

2.4 Rhythmic and Timbre Compatibility

The first models that were developed for mashup creation were mainly focused on the rhythmic part of audio samples [3], and this is still currently used in commercial softwares, such as Traktor [7] or Mixxx [11].

Another approach, developed by Lee et al. [2], consisted on stretching the beats through a phase vocoder so that they would fit the tempo of the input file. Davies et al. [9] used downbeat tracking for track alignment based on onset detection functions.

The most popular approach to tackle rhythmic representation is onset detection, but there are other methods such as the auto-correlation function or even the beat spectrum. While all techniques reveal periodicity, the type of information each one represents can vary from local maximums in onsets to limited copies of a signal with auto-correlation. The Fluctuation Patterns as presented by Pohle et al. [12] depict in a semi-step space a pattern based on onsets. As a result, there is a matrix called onset patterns.

Even though timbre perception is still poorly understood and understudied by the Sound and Music community, it is thought to be every aspect of sound that is neither related to harmony nor rhythm, commonly having associated attributes like rightness, roughness, density, and fullness [13] [14], these being independent from pitch and loudness.

Chapter 3

Computational Harmonic Compatibility

This Chapter will present the metrics adopted to compute Harmonic Compatibility (HC), along with their computational implementation, used to calculate the HC between two songs. Furthermore, in the following sections, two methods will be presented: 1) a basic TIV-basic metric and a 2) newly proposed TIV entropy metric.

All the developed software is available online at the following GitHub repository: https://github.com/matematicadesaramago/HC-model. It includes *Python* scripts, the database used for the listening test and also a brief introduction to the work. The software has multiple dependencies, such as numpy [15] and scipy [16], and three representative libraries for signal processing: *Librosa* [17], *TIV*[1] and *Essentia* [18].

3.1 HC Method Pipeline

In this section, an overview of the method for computing the HC from basic TIV metrics will be presented. Figure 3.1 shows the architecture of the proposed method.

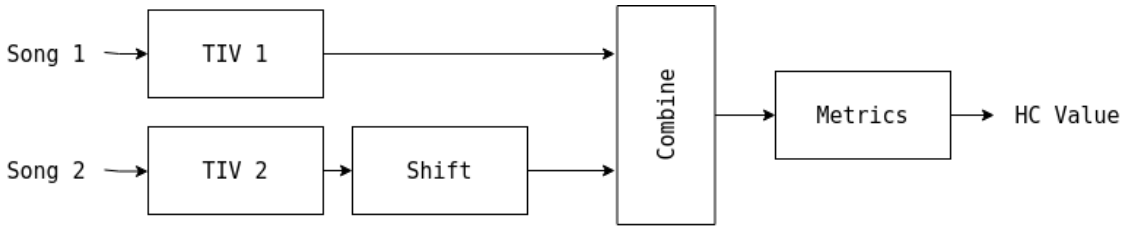

Figure 3.1: Diagram of the HC model

The method assumes two musical audio songs as input, defined by a file path. Using both *Librosa* and *TIV*, the sound file is loaded. Regarding the audio processing restrictions, it is possible to define a sample rate directly using the *sr* variable - when there is no specification, the sample rate assumes its native value. The libraries also allow as an input different audio formats, such as

WAV, FLAC, OGG and MAT, as long as they are supported by SoundFile [19]. As a result, some formats such as mp3 are not supported. In the context of this work, all files had a sample rate of *44.1 KHz* and their format was WAV.

The first step in the method is to represent each song as a TIV from their chroma vector and then compute their combined TIV, i.e., the TIV of the resulting mashup, using *TIV* library. This computation uses the Harmonic Pitch Class Profiles (HPCP) [20] by *Essentia* [18], and it is possible to define many parameters such as the window type or size. The resulting mashup of the two song's TIVs is done by combining their TIV weighted by their energy. To illustrate this process the visualization of the individual songs TIVs and their combinations is shown in Figure 3.2.

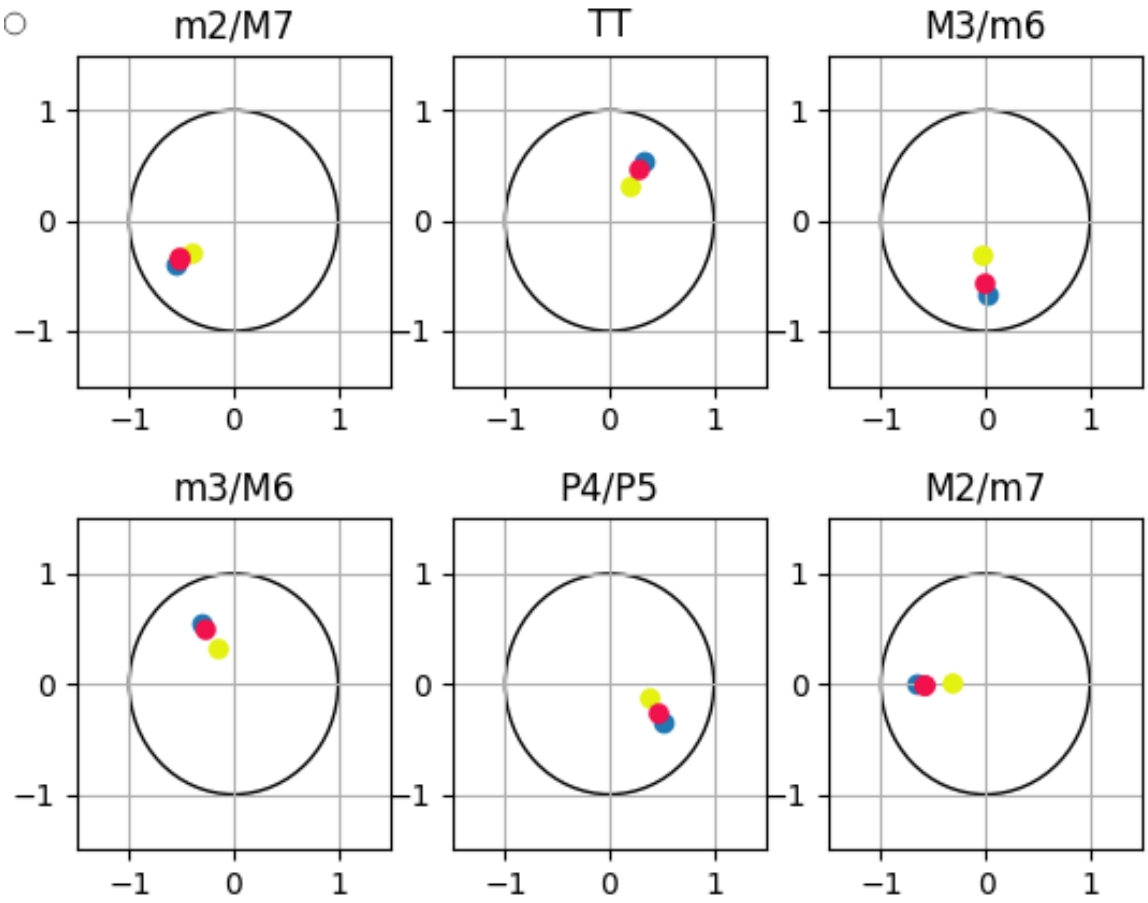

Figure 3.2: Example of combination of two songs: song 1 and song 2. Representation of song 1 TIS in blue, song2 TIS in yellow and its combination in red.

The previous TIV combination is done iteratively in twelve steps to account for the possible twelve pitch shifts of an input song. In other words, twelve combinations of the resulting mashups are created, each of them with one of the songs in a different key, by transposing the musical audio signal by all twelve possible pitch shifts. This process is transposing a given song to assess the best HC in all possible keys within the original mode (i.e., major or minor). As transposing

musical audio signals typically introduce some residual artifacts in the signal, it was assumed that this shifting would happen in the interval [-6, 5], to minimize distortions in the signal other than the pitch shift. In Table 3.1 is an example of a starting note, C, and its outcome after different pitch shift values.

Original Note	Shift	Shifted Note
C	-6	F#
C	-5	G
C	-4	G#
C	-3	A
C	-2	A#
C	-1	B
C	0	C
C	1	C#
C	2	D
C	3	D#
C	4	E
C	5	F

Table 3.1: Pitch Shift Example centered in C

3.2 Harmonic Compatibility Metrics

Based on the TIVs of the songs and their resulting mashup, in this section we detail the metrics adopted to compute the HC value. We propose three different metrics. The first two are the common Euclidean and cosine distances between the songs TIVs and the third is the Shannon entropy of the mashup TIV.

3.2.1 Euclidean distance

The Euclidean distance, one of the most common distance metrics, between two TIVs, $T_1(k)$ and $T_2(k)$, is mathematically defined in Equation 3.1. In the context of this work, Euclidean distance is used to determine the relatedness of two songs given their TIVs, namely capturing the distance between their interval and pitch class content [1].

$$d_{Euclidean}\langle T_1(k), T_2(k)\rangle = \sqrt{\sum_{k=1}^{M}(T_1(k) - T_2(k))^2} \tag{3.1}$$

The rational behind this metric is that smaller Euclidean distances correspond to better mashups, meaning the higher the value of the Euclidean distance the higher the HC also is.

3.2.2 Cosine distance

The cosine distance, also referred to as angular distance, between two given TIVs, $T_1(k)$ and $T_2(k)$, is mathematically defined in Equation 3.2. This metric is used as an indicator of how well pitches relate to each other, almost classifying proximity between tonalities [1], as it computes the number of common pitch classes between the two song TIVs.

$$d_{cosine}\langle T_1(k), T_2(k)\rangle = 1 - \frac{T_1(k)\cdot T_2(k)}{||T_1(k)||_2||T_2(k)||_2} \tag{3.2}$$

The rational behind this metric is that smaller cosine distances correspond to better mashups, which typically are within the same key and enforce the same pitch classes.

3.2.3 Entropy

A different metric was introduced to calculate HC, the Shannon Entropy, as it captures the most commonly used harmonic objects [21] [22]. By using the TIS, which derives from Fourier coefficients as explained before 2, there is a higher-level music meaning than there was with pitches or beats. According to Amiot, there is a relation between the Fourier magnitude values and the degree of musical complexity. The Fourier entropy $H\langle T_c(k)\rangle$ is calculated using previously calculated combined songs, as shown in Equation 3.3. $P_{T_c(k)}$ is the set of normalized Fourier coefficient magnitudes of $T_c(k)$.

If the entropy is maximum, it means that the interval distribution is flat, whereas a null value of entropy means every coefficient is different. In a music context, having a low entropy translates in having a very complex sample and a high entropy means there are not many different elements in the same sample, having its maximum value corresponding to one pitch class, meaning one note.

$$H\langle T_c(k)\rangle = -\sum_{k=1}^{M-1} -p_{T_c(k)} log p T_c(k) \tag{3.3}$$

Chapter 4

Evaluation

This chapter details the evaluation of the multiple metrics presented in our method to calculate the HC value of two given musical audio songs. To this end, a perceptual test was conducted to assess the perceptual judgments of human listening when exposed to two-song mashups with different criteria. The collected data is then adopted as the perceptual estimate to inspect the following threefold criteria in the computational method:

1. Determine which is the best metric to calculate HC compatibility;
2. Define the minimal temporal window that captures, with high accuracy, the compatibility between two songs;
3. Determine the best location in a song, in a temporal scale, that captures HC given a temporal resolution.

With that in mind, comparing the perceptual test with the results obtained from the algorithm with different inputs will shed some light on whether it is possible or not to compute HC with a reduced computational cost. Having a shorter time period being analysed means that determining the HC value will either take less resources or less time, and it is crucial to infer if the methods' results align with human perception.

4.1 Perceptual Test

A standard approach to evaluated the "enjoyment" of a mashup is through user studies, namely a listening test, i.e., having a person evaluating the quality of the mashup according to subjective parameters [23, 9]. In this context, in order to perceptually assess the music mashups, a listening test was conducted online and its objective was to determine the enjoyment of a mashup.

4.1.1 Mashup stimuli

One of the important points to tackle in the design of the listening test is the definition of the stimuli. It is very important to have as few as possible confounding variables so, because of that,

it was only used one genre, songs with similar *tempos* and also similar instruments in order to eliminate eventual different timbral influences. This method allows the test to only evaluate the HC degree between two different songs and eliminates other possible variables.

Moreover, the stimuli should mimic as closely as possible the ones that could be used in real world contexts, making the result of this work useful. Because this work is mainly focused on Electronic Dance Music (EDM), the results will be obtained using this genre of music.

Last, one of the challenges regarding the stimuli definition lies in choosing the part of the track to be analysed so that the system can compute its compatibility score with another track. In other words, in order to reduce as much as possible the computational cost, it is crucial to reduce the analysis of the total duration of a track to a shorter part of it - it is yet to be determined which part, if the beginning, the middle or the end of the track.

As a starting point for the stimuli, there was a collection of 8 different songs, all focused on EDM with similar *tempos* and without many variations regarding instruments. This collection of audio samples was handpicked from different royalty free sources, namely free stock music [24]. With the 8 different songs, 4 combinations of two songs were generated and then each pair is used to create three different mashups variations - maximum (Max) HC, minimum (Min) HC and random (R) HC. The result is 12 combinations, as represented in Table 4.1. Regarding the stimuli duration, they all last between 1 and 3 minutes.

It is necessary to provide some details for the generation of the different mashups variations. As detailed previously in 3, there are twelve possible combinations for the two input songs and twelve different mashups are created as a result and their compatibility assessed using the get maximum compatibility function from the *TIV* library [1], developed by Bernardes et al. This function uses the Euclidean distance to determine the compatibility value as, at the moment, it is the literature's reference metric. Given the best compatibility value, the maximum HC combination is defined. The same process is completed for the minimum compatibility, being the lowest value of maximum compatibility, which generates the minimum HC. The pitch shift values of the maximum and minimum HC are collected and, in order to determine the medium HC mashups, a random value within the interval that the previously pitch shift values account for is chosen. That is how the medium HC mashups are created. This process was developed so as not to force a specific pitch shift, thus avoiding a tendency in the results.

Song 1	Song 2	HC	Reference
A	B	Max	1
A	B	R	2
A	B	Min	3
C	D	Max	4
C	D	R	5
C	D	Min	6
E	F	Max	7
E	F	R	8
E	F	Min	9
G	H	Max	10
G	H	R	11
G	H	Min	12

Table 4.1: Example of samples used to the listening test

4.1.2 Procedure

Prior to the beginning of the listening test, conducted online using Lime Survey, the participants are explained how the test should be conducted and what is expected of them, i.e. listening to the whole music mashup before answering or listening to every mashup. Participants are asked to adjust the loudness of the playback using an audio excerpt, and they are told how long the test will take, that there are twelve different mashups and that no personal data will be collected.

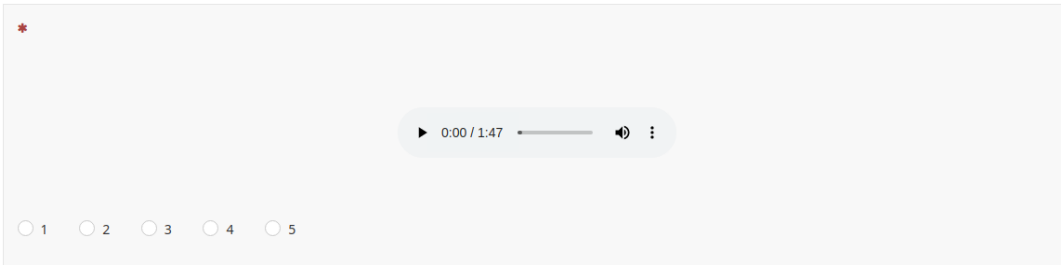

Figure 4.1: Example of a question from the listening test

After the above introduction, participants were asked to evaluate, on a five-point Likert scale, from 1 to 5, their *enjoyment* while listening to 12 different mashups. Each mashup is evaluated independently. After listening to the entire mashup, participants must assess their enjoyment. In Figure 4.1 there is an example of the playback interface of the online listening test for a mashup track and the respective evaluation scale. After this listening part, subjects were asked some socio-demographic information. While this questions were not mandatory, it was requested information concerning their age, gender, and music expertise. The test was made available in Portuguese and English.

4.1.3 Listening Experiment - Details

When preparing the listening test there were many details taken into consideration and this section serves the purpose of detailing them.

4.1.3.1 Random order of display

There are several reasons why music mashups stimuli should be presented in a random way. Starting with the tiring nature of the test due to its repetitiveness, the order in which music mashups stimuli are presented should not be the same for every participant. As the test goes by, the attention span of the participants starts to decrease, which could have an implication on the results as studied by Diemo Shwarz [25]. Moreover, it should be considered that, at first, the participants do not know what to expect. As they answer the questions, participants get to listen to more music mashups stimuli, and the ones that they have previously listened to could influence the classification they would otherwise give. So, in order to avoid order effects and balance out the results, the music mashups are presented in a random order.

4.1.3.2 Rhythm and structure

One of the main concerns while selecting the songs for the mashups was its rhythmic alignment. Although it was fairly easy to obtain songs with the same tempo, defined as beats per minute (BPM), it was not as easy to find songs with that perfect alignment in terms of phrasing, sections, or structure.

In order to minimize the impact of rhythm as a confounding variable for the results along with all remaining timbre and stylistic attributes, the stimuli were developed so that each group of mashups with different HC were originated from the same two original songs. Therefore, three stimuli will have the same elements concerning all structural aspect other then the harmony.

4.1.3.3 Energy levels

To balance the energy levels of the stimuli, which largely correlates with the perceived intensity or loudness of the songs, we first normalized each song individually using *Audacity* [26]. Some manual adjustments were applied to ensure that there was not a predominance of one song over the other.

4.1.3.4 Duration

Finding songs with the exact same duration was not trivial. In order to overcome that problem, the shortest song was taken as a reference and, using an additional software, *Audacity* [26], a temporal point of the longest song was chosen so that the end point of that song would perceptually make sense and it would not be perceived as an abrupt ending.

4.1.3.5 Participants Data

The participants socio-demographic data collected is as follows:

- Age: numerical data inserted by the participant (mandatory);
- Gender: selection between male, female, other;
- Musical training: selection between none, amateur, professional, no answer.

.

By collecting this data it is possible to determine if the collection is varied enough, specially when it comes to age and musical training. This information might help when analysing the answers or help explaining some trends, namely if musical training has some influence over the perception of the mashups enjoyment.

4.2 Results

4.2.1 Listening test

Regarding the listening test, it was completed by 94 participants of various ages and music training background, 41 being amateurs, 6 being professionals and 44 having no experience. Due to the lack of an expressive number of trained musicians, the analysis will not differentiate groups and treat all participants uniformly. The results divided by question are presented in Figures 4.2, 4.3, 4.4 and 4.5. In Figures 4.2, 4.3, 4.4 and 4.5, the plot displayed on the left corresponds to the maximum HC of two songs, the middle plot corresponds to a random value of pitch shift, meaning a random medium value of HC, and the plot on the right corresponds to the minimum HC of the same two songs.

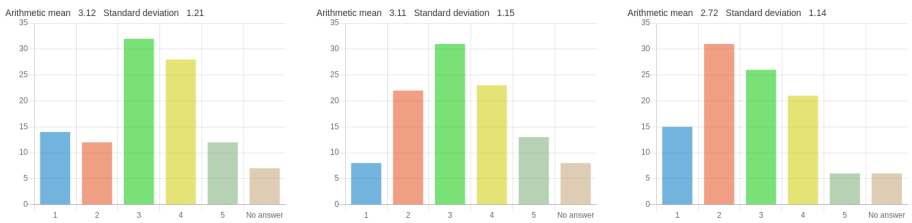

Figure 4.2: Perceptual test results example 1 for three different mashups regarding the threefold levels high, mid and low HC for the same base songs, respectively. A classification of 5 means the participant found the mashup very pleasant and 1 not pleasant.

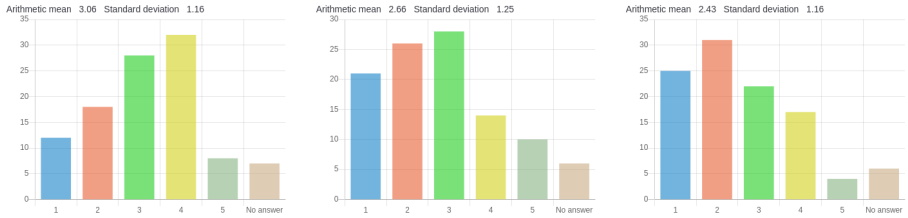

Figure 4.3: Perceptual test results example 2 for three different mashups regarding the threefold levels high, mid and low HC for the same base songs, respectively. A classification of 5 means the participant found the mashup very pleasant and 1 not pleasant.

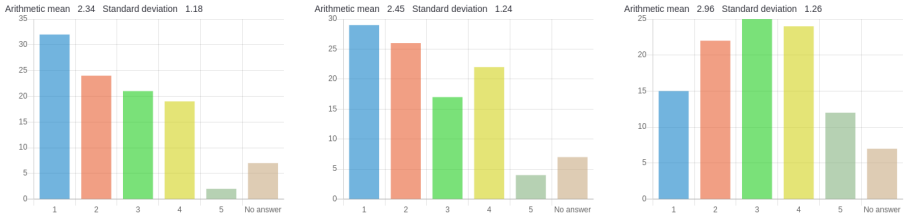

Figure 4.4: Perceptual test results example 3 for three different mashups regarding the threefold levels high, mid and low HC for the same base songs, respectively. A classification of 5 means the participant found the mashup very pleasant and 1 not pleasant.

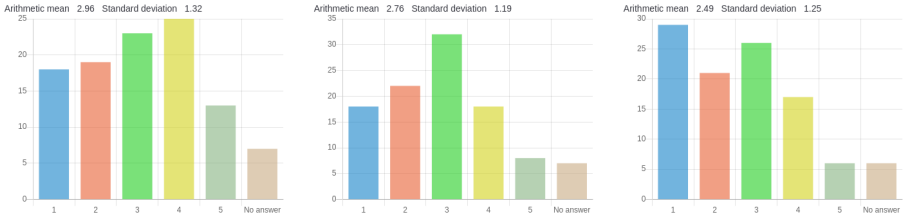

Figure 4.5: Perceptual test results example 4 for three different mashups regarding the threefold levels high, mid and low HC for the same base songs, respectively. A classification of 5 means the participant found the mashup very pleasant and 1 not pleasant.

In Figure 4.6 some descriptive statistics were applied to provide a more intuitive comparison of the data. The same data is now presented in a different way, having the average and the standard deviation for each mashup depicted in the plot. Focusing on the result's average, with an exception on the third group of mashups, the average decreases as HC decreases. Please note that black represents the maximum HC, the darkest gray represents the medium random HC and the lightest gray represents the minimum HC. This, once more, proves that, generally speaking, higher perceptual pleasantness is associated with higher HC. The exception of the third group was then analysed and it was hypothesized that the unexpected outcome of the listening test could be a result of the two base songs for the mashups being more different from each other than in the other groups. The rhythmic component of the mashups having a more inadequate compatibility could also have influenced the results.

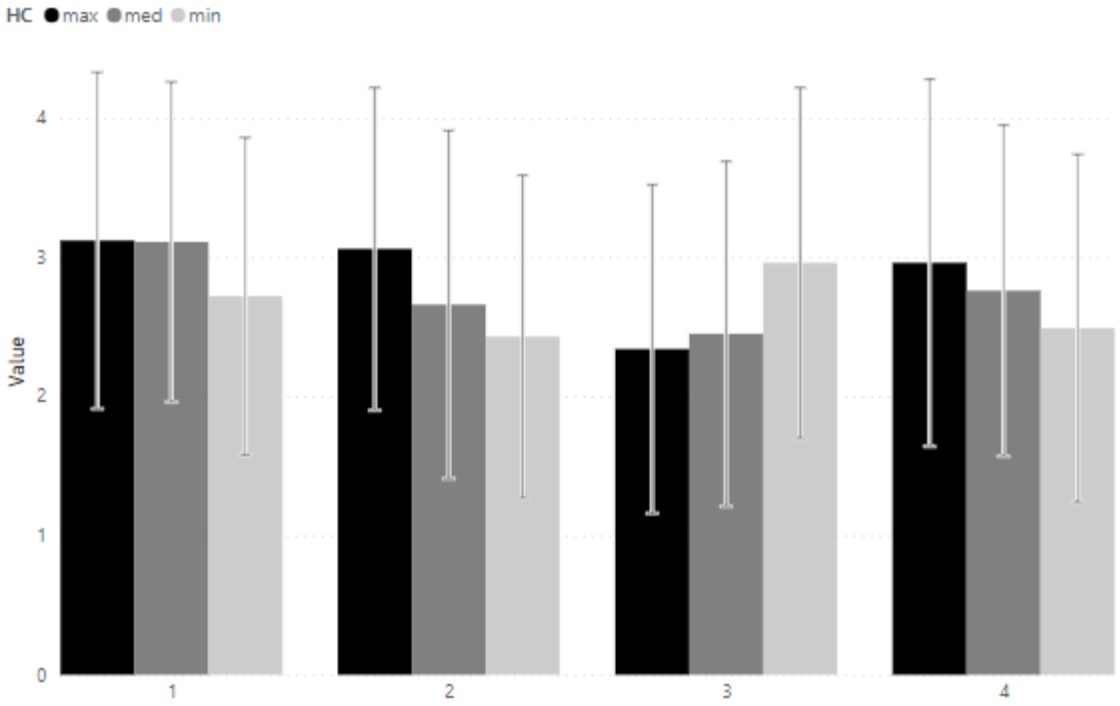

Figure 4.6: Average and Standard deviation for the results of each mashup regarding the participant's enjoyment in a scale from 1 to 5 (represented by the variable value in the vertical axis). The black represents the mashups with a high HC, the darkest shade of gray represents the medium value of HC and the lightest shade of gray represents the lowest value of HC for each group of two songs (represented in 1, 2, 3 and 4 in the horizontal axis).

4.2.2 Assessing the Objective Computational Metrics

To assess the objective computational metrics proposed in the HC method in Section 3, namely the Euclidean distance, cosine distance and entropy, at different time scales and within different structural location of the resulting mashups, we defined structural points and duration for the analysis. Concerning the time scale of analysis, we opted for intervals and time locations, namely 1) the entire duration, 2) 5 seconds and 3) 20 seconds. Assuming the case where only an interval is analysed, the metrics were calculated at the beginning, middle and end of each entire song. In Figures 4.11, 4.8 and 4.9, the tables present the results of the three different metrics adopted, Euclidean distance, cosine distance and entropy, for the whole song and different time windows, respectively. When a time window is applied, its location is specified. In 4.10 there is an example of one iteration of the data collection process, having a clip of the first *20s* of two songs being analysed.

			Perceptual Test		HC metrics		
					Whole		
			Average	Standard Deviation	Euclidean	Cosine	Entropy
1	1	max	3,12	1,21	14,90	1,90	1,64
	2	med	3,11	1,15	15,30	1,98	1,67
	3	min	2,72	1,14	13,36	1,63	1,62
2	4	max	3,06	1,16	18,20	1,84	1,69
	5	med	2,66	1,25	19,61	2,10	1,70
	6	min	2,43	1,16	14,39	1,26	1,63
3	7	max	2,34	1,18	19,88	1,61	1,67
	8	med	2,45	1,24	19,09	1,52	1,66
	9	min	2,96	1,26	19,09	1,52	1,66
4	10	max	2,96	1,32	16,83	1,71	1,60
	11	med	2,76	1,19	15,14	1,49	1,56
	12	min	2,49	1,25	13,88	1,34	1,55

Figure 4.7: Results of the Euclidean distance, cosine distance and entropy when the entire mashup is analysed.

			Perceptual Test		HC metrics								
					20 seconds								
					Begining			Middle			End		
			Average	Standard Deviation	Euclidean	Cosine	Entropy	Euclidean	Cosine	Entropy	Euclidean	Cosine	Entropy
1	1	max	3,12	1,21	8,56	1,53	1,19	16,01	1,90	1,70	17,16	1,81	1,53
	2	med	3,11	1,15	9,07	1,69	1,32	14,35	1,63	1,67	15,59	1,59	1,52
	3	min	2,72	1,14	9,07	1,69	1,32	14,84	1,71	1,68	15,29	1,55	1,54
2	4	max	3,06	1,16	22,87	2,10	1,69	16,92	1,79	1,67	18,00	1,90	1,66
	5	med	2,66	1,25	21,61	1,87	1,69	16,92	1,79	1,67	17,96	1,89	1,66
	6	min	2,43	1,16	15,21	0,93	1,68	13,71	1,31	1,64	14,17	1,34	1,63
3	7	max	2,34	1,18	12,97	2,43	1,59	15,38	1,71	1,66	15,10	1,91	1,66
	8	med	2,45	1,24	5,53	0,82	1,29	14,12	1,53	1,59	11,41	1,32	1,58
	9	min	2,96	1,26	5,42	0,80	1,40	11,28	1,17	1,58	6,30	0,67	1,62
4	10	max	2,96	1,32	14,80	1,83	1,70	20,29	1,65	1,63	16,74	1,75	1,63
	11	med	2,76	1,19	14,08	1,71	1,70	21,89	1,83	1,68	18,92	2,11	1,71
	12	min	2,49	1,25	12,57	1,47	1,58	17,41	1,36	1,61	12,42	1,21	1,60

Figure 4.8: Results of the Euclidean distance, cosine distance and entropy when a time window of 20 seconds is analysed at different time locations. The first column is refers to the results of the time window being applied at the beginning, the second column to the middle and the third column to the end of the songs.

			Perceptual Test		HC metrics								
					5 seconds								
					Begining			Middle			End		
			Average	Standard Deviation	Euclidean	Cosine	Entropy	Euclidean	Cosine	Entropy	Euclidean	Cosine	Entropy
1	1	max	3,12	1,21	16,67	1,27	1,63	17,71	1,93	1,66	25,15	1,54	1,67
	2	med	3,11	1,15	17,47	1,40	1,63	17,59	1,91	1,66	28,46	2,07	1,72
	3	min	2,72	1,14	19,86	1,81	1,63	16,44	1,73	1,63	25,08	1,53	1,71
2	4	max	3,06	1,16	22,71	2,17	1,60	17,09	1,73	1,68	29,38	1,94	1,54
	5	med	2,66	1,25	21,90	2,04	1,58	15,68	1,50	1,68	24,77	1,54	1,46
	6	min	2,43	1,16	8,54	0,65	1,48	14,78	1,37	1,67	15,93	0,93	1,49
3	7	max	2,34	1,18	16,59	1,79	1,62	16,04	1,69	1,66	23,77	1,35	1,69
	8	med	2,45	1,24	16,42	1,76	1,63	13,69	1,37	1,52	23,49	1,27	1,63
	9	min	2,96	1,26	16,88	1,85	1,60	10,90	1,03	1,53	23,26	1,21	1,71
4	10	max	2,96	1,32	14,62	1,72	1,70	22,35	1,63	1,51	29,67	1,72	1,64
	11	med	2,76	1,19	16,91	2,13	1,72	21,11	1,52	1,57	27,39	1,55	1,65
	12	min	2,49	1,25	10,01	1,09	1,64	21,11	1,52	1,57	21,39	1,13	1,64

Figure 4.9: Results of the Euclidean distance, cosine distance and entropy when a time window of 5 seconds is analysed at different time locations. The first column is refers to the results of the time window being applied at the beginning, the second column to the middle and the third column to the end of the songs.

```
Compatibility 3.796958477230232e-16, for pitch shift: -6
Compatibility 4.081700393811396e-16, for pitch shift: -5
Compatibility 3.9004921588093043e-16, for pitch shift: -4
Compatibility 4.427212210315757e-16, for pitch shift: -3
Compatibility 4.461402571314549e-16, for pitch shift: -2
Compatibility 4.946864363956244e-16, for pitch shift: -1
Compatibility 6.08719865465593e-18, for pitch shift: 0
Compatibility 5.936039142441207e-16, for pitch shift: 1
Compatibility 3.446161133281269e-16, for pitch shift: 2
Compatibility 3.9905262852275093e-16, for pitch shift: 3
Compatibility 3.6454962369045983e-16, for pitch shift: 4
Compatibility 3.8226174569200536e-16, for pitch shift: 5

Maximum compatibility for pitch_shift of 0

Minimum compatibility for pitch_shift of 1

Medium compatibility for pitch_shift of -2

./orig/song1_20s_b.wav
./orig/song3_20s_b.wav
FOR THE MAXIMUM HC COMBINATION

Cosine distance: 1.8330546934194456
Euclidean distance: 14.802599952317728
Entropy: 1.699475019073465

FOR THE RANDOM HC COMBINATION

Cosine distance: 1.7093079039955468
Euclidean distance: 14.082624802421789
Entropy: 1.6955648393781622

FOR THE MINIMUM HC COMBINATION

Cosine distance: 1.474418453026343
Euclidean distance: 12.57438037506934
Entropy: 1.5757861974164775

mds@mds:~/Tese$
```

Figure 4.10: Example of an algorithm run with inputs of 20 seconds of song 1 and song 3 taken from the beginning of each song.

The Pearson correlation coefficient was adopted to infer the objective HC metric that captures best the perceptual enjoyment of the listening test mashups. The Pearson correlation coefficient is a statistical method that inspects the relationship between two different variables and assesses their linear correlation. It can be computed as shown in Equation 4.1. The coefficient results in a value within the interval [-1, 1]. x and y are the values being correlated, $\bar{x}$ and $\bar{y}$ their respective mean [27]. The PCC presents a perfect linear correlation if its absolute value is one. On the other hand, if the value equals 0 there is no correlation. Table 4.2 present the PCC values guidelines.

$$c_{x,y} = \frac{\sum(x-\bar{x})(y-\bar{y})}{\sqrt{\sum(x-m_x)^2\sum(y-m_y)^2}} \tag{4.1}$$

The plots in Figures 4.11, 4.12 and 4.13 represent the relation between the average results for each mashup and the three different adopted metrics. In the x axis, the average value from the perceptual test is used and then compared with a different variation of either the Euclidean

PCC value	Strength
1	Perfect
[0.5;1[	Strong
[0.3;0.5[	Moderate
]0;0.3[	Weak
0	None
]-0.3; 0[	Weak
]-0.5; -0.3]	Moderate
]1; -0.5]	Strong
-1	Perfect

Table 4.2: PCC classification according to its numerical value.

distance, cosine distance or entropy of different time intervals and locations. In Figure 4.11 there are presented three plots of the relations between the previously detailed metrics and listening test for the whole duration of every mashup. In Figure 4.12 the same analysis is made, but for a time window of *20s*. Each column of three plots represents that analysis at a different time location, being the first column related to the beginning, the second column the middle of the mashups and the third column to the end of the mashups. The same goes for Figure 4.13, where each column represents a different time location, but for a time window of *5s*.

Moving on to some consideration taken from the results in Tables 4.3, it is obvious that in each combination of mashups - in the first column as 1, 2, 3 and 4 - the values for the Euclidean and cosine distances generally decrease as the HC decreases whereas the entropy increases, which was expected.

As far as the results for each music mashup group, given all the results provided by the plots of Figures 4.2, 4.3, 4.4 and 4.5, a brief observation shows that group 2 and group 4 had the closest results to what was expected. However, as shown in Figure 4.6, the spread range of results does not show less agreement in the enjoyment of some stimuli groups more than in others stimuli groups given that the standard deviation is very similar in all groups. Following the above observation, in Figure 4.11, it was also visible that the precision of each linear approximation had its most approximate result when analysing the whole song, as expected. However, even with visible dispersion, the results for the cosine distance and its PCC were significantly superior in the *5s* samples rather than in the 20s samples, as one would suspect. That could be a result of *5s* not being enough time for a phrase of a song to go from the beginning to the end, not to mention its addition to another song which phrases could not be perfectly aligned.

Duration	Location	Metric	PCC
whole	-	Euclidean	-0.12
whole	-	Cosine	**0.59**
whole	-	Entropy	0.14
20s	Beginning	Euclidean	-0.01
20s	Beginning	Cosine	0.10
20s	Beginning	Entropy	-0.26
20s	Middle	Euclidean	0.09
20s	Middle	Cosine	0.29
20s	Middle	Entropy	**0.32**
20s	End	Euclidean	0.20
20s	End	Cosine	0.10
20s	End	Entropy	**-0.32**
5s	Beginning	Euclidean	**0.44**
5s	Beginning	Cosine	0.0.23
5s	Beginning	Entropy	0.28
5s	Middle	Euclidean	0.15
5s	Middle	Cosine	**0.37**
5s	Middle	Entropy	0.07
5s	End	Euclidean	**0.68**
5s	End	Cosine	**0.73**
5s	End	Entropy	0.25

Table 4.3: PCC betweeen the average results of the listening test and the different metrics of the whole mashups, *20s* and *5s* exerts.

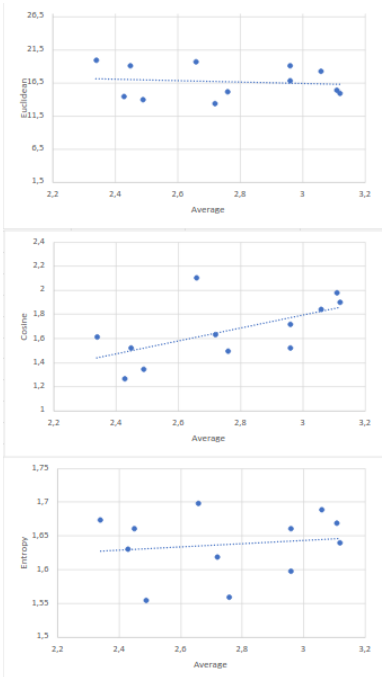

Figure 4.11: Relation between the listening test and metrics results for the whole duration (Euclidean distance, cosine distance and entropy, respectively), resulting in a linear regression.

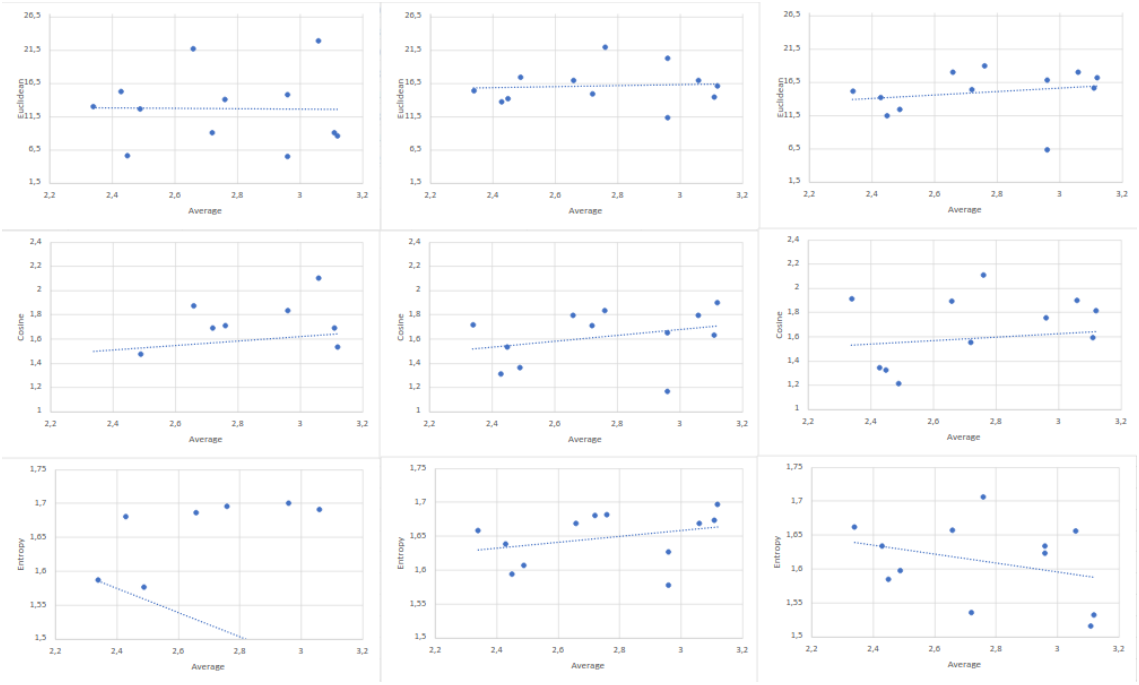

Figure 4.12: Relation between the listening test and metrics results for ta time window of *20s* (Euclidean distance, cosine distance and entropy, respectively), resulting in a linear approximation. The first column shows the results for the time window applied at the beginning, the second column shows the results applied for the same time window applied at the middle and the third column showed the results of the time window when applied at the end of the songs.

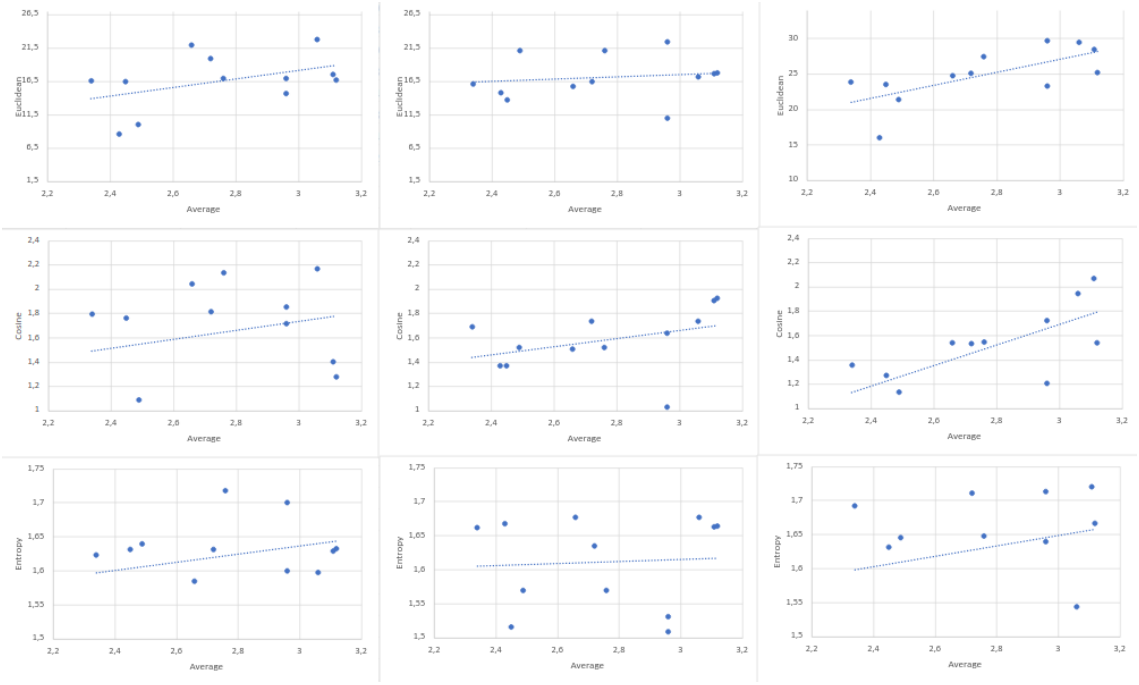

Figure 4.13: Relation between the listening test and metrics results for ta time window of *5s* (Euclidean distance, cosine distance and entropy, respectively), resulting in a linear approximation. The first column shows the results for the time window applied at the beginning, the second column shows the results applied for the same time window applied at the middle and the third column showed the results of the time window when applied at the end of the songs.

A summary of the PCC is presented in Table 4.3.One can conclude, from all data presented before in this section, that the developed HC method did not find a suitable time window to calculate HC. Within the time intervals analysed, of *20s* and *5s*, the result did not fully capture HC. Moving on to the values' analysis, according to Table 4.2, a strong value of PCC is observed when using the cosine distance metric applied at the whole mashup, the PCC being *0.59*. Also, using a time window of *5s* located at the end of a mashup, both the Euclidean and cosine distances PCC values are considered strong, *0.68* and *0.73* respectively. A moderate correlation is observed in the excerpt of *20s* from the middle of the mashups when using entropy as a metric, being the respective PCC value *-0.32*. The PCC value also corresponds to a moderate correlation in the case of a *5s* time window and Euclidean distance at the beginning of the mashups, PCC value being *0.44*, and for the same time window but located at the middle of the mashups, this time regarding the cosine distance, having a PCC of *0.37*. Even though these results seem promising, having a good result in an excerpt of *5s* would imply also having a good, or even better, result for an excerpt of *20s*, which can not be confirmed. This could be explained, in the first case, by the fact that the beginning of songs is usually less complex and the music components could have a sparser presence, thus less content being analysed. As far as the middle of the mashup is concerned, it could be the case that the excerpt was randomly taken from an interval in which the music phrases align better as a coincidence or there is some kind of harmonic plateau. So, as a consequence, the temporal location for that window could not be determined either for the same reason. To sum up,

the best results were associated with the entire mashups.

Regarding the metrics, entropy did not capture the HC as expected and it did not provide an improvement to the results when comparing to existing metrics as the Euclidean and cosine distances.

Chapter 5

Conclusion

This chapter serves the purpose of reviewing the objectives stated at the beginning of the document, and also the purpose of summing up some general ideas and takeaways from the developed work.

The main purpose of this dissertation was to determine whether it is possible or not to determine a value for HC given a different time window and its location in a temporal scale. Given the results of the previous section 4, the algorithm did not excel, but there might be some external circumstances that had something to do with it, for instance the rhythmic dimension.

The comparison between each metric in each time interval and location proved that the Euclidean distance is still the most stable metric when analysing HC, meaning that neither the cosine distance nor the entropy proved to have better results. Entropy not being a better metric when compared to the Euclidean distance could be a consequence of harmony being more constant during a time interval, whereas rhythm "typically" has greater changes across a musical structure, which could indicate that applying entropy to the horizontal dimension of a mashup as a metric might have a better outcome.

The results of the listening test were essential to obtain the data's ground truth. Even though the results generally aligned to what was expected, its comparison with the metrics' results showed some incoherence, i.e. having a time window of *5s* presenting good metrics' values when the perceptual test showed a not so good evaluation (despite being evaluated in different scales).

The study of a possible time window showed that, according to this method, it is not possible to reduce the time period being analysed, contrasting to what was initially thought. Even though the results appear to more satisfying in the middle of a mashup, the PCC showed that they were not statically significant. To determine a time location was not possible either as a result of a temporal window not having been defined either.

5.1 Limitations and Future Work

One of the main limitations of this work was directly related to the rhythmic dimension of a song, which ended up not being explored. Even though the initial work and method both accounted for

it as far as an algorithm is concerned, building the listening test proved to be much more complicated than predicted - finding two different songs which phrases are perfectly aligned for their whole duration was not possible. In order to fully assess and compare the perceptual evaluation and the metrics, it had been planed to conduct a listening test in which the mashups would have the maximum Harmonic Compatibility value and the rhythmic alignment would be tested as perfectly aligned (phrases should always coincide), medium aligned (same *tempo*) and not aligned (completely disregard the rhythm). Finding the perfect samples proved to be very complicated because of the phrase's alignment. Adding this component to the method could have a major impact on the entropy results.

References

[1] Gilberto Bernardes, Diogo Cocharro, Marcelo Caetano, Carlos Guedes, and Matthew Davies. A multi-level tonal interval space for modelling pitch relatedness and musical consonance. *Journal of New Music Research*, 45:1–14, 05 2016. doi:10.1080/09298215.2016.1182192.

[2] Chuan-Lung Lee, Yin-Tzu Lin, Zun-Ren Yao, Feng-Yi Lee, and Ja-Ling Wu. Automatic mashup creation by considering both vertical and horizontal mashabilities. In *ISMIR*, 2015.

[3] Gilberto Bernardes, Matthew Davies, and Carlos Guedes. *A Hierarchical Harmonic Mixing Method*, pages 151–170. 11 2018. doi:10.1007/978-3-030-01692-0_11.

[4] Miguel Pérez Fernández. Harmonic compatibility for loops in electronic music, 2020.

[5] Gilberto Bernardes, Diogo Cocharro, Carlos Guedes, and Matthew Davies. Harmony generation driven by a perceptually motivated tonal interval space. *Computers in Entertainment*, 14, 12 2016. doi:10.1145/2991145.

[6] Mashup2. https://mashup.mixedinkey.com, note = "[Online; accessed 20-February-2022]", 2022.

[7] Native Instruments. Traktor DJ 2. https://www.native-instruments.com/en/products/traktor/dj-software/traktor-dj-2/, note = "[Online; accessed 21-February-2022]", 2022.

[8] Catarina Maçãs, Ana Rodrigues, Gilberto Bernardes, and Penousal Machado. Mixmash: An assistive tool for music mashup creation from large music collections. *International Journal of Art, Culture and Design Technologies*, 8:20–40, 07 2019. doi:10.4018/IJACDT.2019070102.

[9] Matthew Davies, Philippe Hamel, Kazuyoshi Yoshii, and Masataka Goto. Automashupper: Automatic creation of multi-song music mashups. *Audio, Speech, and Language Processing, IEEE/ACM Transactions on*, 22:1726–1737, 12 2014. doi:10.1109/TASLP.2014.2347135.

[10] Peter Harrison and Marcus Pearce. Simultaneous consonance in music perception and composition. *Psychological Review*, 127, 12 2019. doi:10.1037/rev0000169.

[11] Native Instruments. Mixxx DJ your way. https://mixxx.org, note = "[Online; accessed 21-February-2022]".

[12] Tim Pohle, Dominik Schnitzer, Markus Schedl, Peter Knees, and Gerhard Widmer. On rhythm and general music similarity. pages 525–530, 01 2009.

[13] Felix Dobrowohl, Andrew Milne, and Roger Dean. Timbre preferences in the context of mixing music. *Applied Sciences*, 9:1695, 04 2019. doi:10.3390/app9081695.

[14] Roman Gebhardt, Matthew Davies, and Bernhard Seeber. Harmonic mixing based on roughness and pitch commonality. 12 2015.

[15] Numpy. URL: https://numpy.org/.

[16] Scipy. URL: https://scipy.org/.

[17] Librosa. URL: https://librosa.org/doc/latest/index.html.

[18] Dmitry Bogdanov, Nicolas Wack, Emilia Gómez, Sankalp Gulati, Perfecto Herrera, Oscar Mayor, Gerard Roma, Justin Salamon, José Zapata, and Xavier Serra. Essentia: An open-source library for sound and music analysis. MM '13, New York, NY, USA, 2013. Association for Computing Machinery. URL: https://doi.org/10.1145/2502081.2502229, doi:10.1145/2502081.2502229.

[19] soundfile. URL: https://pysoundfile.readthedocs.io/en/latest/index.html#module-soundfile.

[20] Hpcp. URL: https://essentia.upf.edu/reference/std_HPCP.html.

[21] Gregory Cox. On the relationship between entropy and meaning in music: An exploration with recurrent neural networks. 08 2010.

[22] Emmanuel Amiot. Entropy of fourier coefficients of periodic musical objects. *Journal of Mathematics and Music*, 15:1–12, 07 2020. doi:10.1080/17459737.2020.1777592.

[23] Roman Gebhardt, Matthew Davies, and Bernhard Seeber. Psychoacoustic approaches for harmonic music mixing. *Applied Sciences*, 6:123, 05 2016. doi:10.3390/app6050123.

[24] Free stock music for your youtube videos or multimedia projects - 100% free. URL: https://www.free-stock-music.com/.

[25] Diemo Schwarz, Guillaume Lemaître, Mitsuko Aramaki, and Richard Kronland-Martinet. Effects of Test Duration in Subjective Listening Tests. In Hans Timmermans, editor, *International Computer Music Conference (ICMC)*, pages 515–519, Utrecht, Netherlands, September 2016. Hans Timmermans, HKU University of the Arts Utrecht, HKU Music and Technology. URL: https://hal.archives-ouvertes.fr/hal-01427340.

[26] Audacityteam. Home, Aug 2022. URL: https://www.audacityteam.org/.

[27] Charles Zaiontz. Intraclass correlation, 2019. URL: https://www.real-statistics.com/reliability/interrater-reliability/intraclass-correlation/.