FACULDADE DE ENGENHARIA DA UNIVERSIDADE DO PORTO

CoDi: Leveraging Compatibility and Diversity in Computational Mashup Creation from Large Loop Collections

Gonçalo Nuno Botelho Amaral Rolão Bernardo

DISSERTAÇÃO

Mestrado Integrado em Engenharia Informática e Computação

Supervisor: Gilberto Bernardes, PhD

July 21, 2021

CoDi: Leveraging Compatibility and Diversity in Computational Mashup Creation from Large Loop Collections

Gonçalo Nuno Botelho Amaral Rolão Bernardo

Mestrado Integrado em Engenharia Informática e Computação

July 21, 2021

Abstract

Music is a pervasive media in retail spaces, despite their considerable licensing costs and time-consuming creative production methods. Free licensed music can ultimately reduce costs and enhance the diversity of the musical contents. To leverage the generation of free music at scale with minimum costs, intelligent music algorithms have been advanced, so that retail industries can obtain full rights; stream it globally and disseminate to large markets. However, the multidimensional and subjective nature of music is yet to be fully understood and systematized.

In this dissertation, we advance a novel approach to music generation with *quality* assessment and stylistic *diversity* from the recombination of pre-recorded public domain loops, which proliferate in web musical repositories. Recombination will be tackled at both the vertical and horizontal dimensions of musical structure. Loops are audio files with varied timbral qualities (e.g., lead, comping, and drums), and so a formal structure of vertical layers allows for cohesive harmonic fluctuations between instruments, and horizontal layers assure melodic transitions throughout the generation.

To regulate quality we propose metrics for rhythmic, harmonic, and timbral compatibility holding minimal post-processing of the audio content. To ensure diversity across multiple stylistic conditions, we adopt evolutionary diversity optimization. Musical audio will be defined by a reduced feature vector in a continuous descriptor space. The Artificial Immune Systems opt-aiNet will be adopted to optimize the search for optimal compatibility with multiple diverse solutions.

To this end, we adopt the Artificial Immune System (AIS) opt-aiNet algorithm to efficiently compute a population of compatible and diverse mashups from loop recombinations. Optimal mashups result from local minima in a feature space that objectively represents harmonic and rhythmic compatibility. We implemented our model as a prototype named CoDi. We conducted an objective evaluation of three algorithmic models of AIS, Genetic Algorithm and Brute-Force tackling three dimensions: loop recombination compatibility, mashup diversity, and computational model efficiency – which the latter reports a 93% efficiency comparing AIS with BF. Furthermore, a subjective evaluation through a perceptual experiment was employed to determine the relationship between estimated user enjoyment as *pleasantness* as the musicological metrics of CoDi. Listening test results have proven to be significantly correlated to the values of the evaluation function for both individual and continuous mashups, employing a statistical analysis through linear regression. We propose a functional prototype for automatic generation of music as recombinations of loops, at scale.

Keywords: Music Information Retrieval, Music Similarity, Functional Languages, Generative Music, Artificial Immune System, Audio Compatibility, Audio Similarity

Resumo

A música é um meio de comunicação social omnipresente em espaços comerciais, apesar dos seus consideráveis custos de licenciamento e métodos de produção criativa demorados. A música gratuita licenciada pode, em última análise, reduzir os custos e aumentar a diversidade dos conteúdos musicais. Para potenciar a geração de música livre à escala com custos mínimos, foram avançados algoritmos musicais inteligentes, de modo a que as indústrias retalhistas possam obter plenos direitos; difundi-la globalmente e divulgá-la em grandes mercados. No entanto, a natureza multidimensional e subjetiva da música ainda não foi totalmente compreendida e sistematizada.

Nesta dissertação, avançamos uma nova abordagem à geração de música com avaliação de qualidade e diversidade estilística a partir da recombinação de loops de domínio público pré-gravados, que proliferam nos repositórios musicais da web. A recombinação será abordada tanto nas dimensões vertical como horizontal da estrutura musical. Os loops são ficheiros áudio com qualidades tímbricas variadas (por exemplo vozes, acompanhamentos, e percussão). Assim, uma estrutura formal de camadas verticais permite flutuações harmónicas coesas entre instrumentos, e as camadas horizontais asseguram transições melódicas ao longo da geração.

Para regular a qualidade, propomos métricas de compatibilidade rítmica, harmónica e timbral, mantendo um pós-processamento mínimo do conteúdo áudio. Para assegurar a diversidade através de múltiplas condições estilísticas, adotamos uma otimização evolutiva da diversidade. O áudio musical será definido por um vetor de características reduzido num espaço contínuo de descritores. A opção Artificial Immune Systems opt-aiNet será adoptada para otimizar a procura de compatibilidade com múltiplas soluções.

Para tal, adotamos o algoritmo opt-aiNet, do Sistema Imunitário Artificial (AIS), para calcular eficazmente uma população de soluções como mashups compatíveis e diversos a partir de recombinações de loops. Os mashups ideais resultam de mínimos locais num espaço de características que objetivamente representa compatibilidade harmónica e rítmica. Implementamos o nosso modelo como um protótipo chamado CoDi. Realizámos uma avaliação objectiva de três modelos algorítmicos de AIS, Algoritmo Genético e Brute-Force abordando três dimensões: compatibilidade de recombinação de loops, diversidade de mashup, e eficiência do modelo computacional – no qual este reporta um valor de 93% de eficiência com AIS e comparativamente a abordagens BF. Além disso, foi utilizada uma avaliação subjective através de uma experiência perceptual para determinar a relação entre o grau estimado de agradabilidade do utilizador e as métrica musicológicas pertencentes às funções de avaliação do CoDi. Os resultados dos testes de escuta provaram estar significativamente correlacionados com os valores da função de avaliação para ambos mashups individuais ou contínuos, recorrendo a uma análise estatística através de uma regressão linear. Propõe-se um modelo funcional para a geração automática de música através de recombinações

de loops, à escala.

Keywords: Recuperação de Informação Musical, Semelhança Musical, Línguas Funcionais, Música Generativa, Sistema Imunitário Artificial, Compatibilidade Áudio, Similaridade Áudio

Acknowledgements

The feeling of conclusion and retribution after writting this dissertation is beyond whatever feeling I can explain. These words are from myself as a thankful student, a friend, and a person of family. First, I'd thank myself for knowing what it takes and carrying through every single time.

I want to express my gratitude to professor Gilberto for the mentorship, the constant advice, the incentives, and the critical viewpoints of my work and for what is, generically, Sound and Music Computing – an area I would've never found had it not been for hours of searching.

Coming from outside of the portuguese mainland implicated getting a safe space of friendship, support, and chosen family. I want to give my deepest thankfulness to Joana - for giving me the utmost love in friendship, support and professional independence to follow what we want, when we want. To Simão – for being the friend I needed and the roomate I never asked for. To Yaguas – for being my proud friend and showing me that we need to live unapologetically. To Teresa – always ready to welcome myself as myself, and nothing more. Now I see a brighter future. To Ganso – who is the representation of what are working family values, chosen family, and love. To Rosinha – always ready to live life at its fullest, sharing that with friends, and coloring peoples' lives with such a generous energy. To Colar – for putting a smile on my face every day, and being my emotional support system at any moment of the day or night. To Alexa, Xica and Xana – and everyone really – for listening to my frustrations every day and every night. To Road – for every single moment possible. Finally, to dearest Quina – who is the personification of the light at the end of the tunnel, and for being a one-person family who gave me the opportunities, the smiles, the unapologetical academic construct of being yourself and loving your closest friends.

Home is where your heart is, and wherever I am, these people have travelled with me in my heart since ever. Wherever I am, I am in my Home when I think of them. To Xico – who showed me life outside the bubble, and making sure that we know what friendship and chosen family is, along with the sacrifices we must take. To Inês Almeida Peixoto – who is the pillar which glues every connection I have. I would not be me, if not for you. I love you, and miss you. To Mena, Inês, Ana, Isabel M., and everyone of my Theatre family – which are my true first friends. To André and Mário – my mentors for anything in life. To Vitória – who came to Porto with me, and knows the true meaning of coming from the outside.

I want to personally thank my mother Lorina – the embodiment of resilience – for every gesture of love, kindness, and sacrifice for a brighter future. My brother João – the embodiment of purity – for showing me the greatness in what we do. My cousins Júlia and Carolina – the embodiment of independence – who showed me that we must appreciate, by ourselves, the rainbow after the storm. My godmother Elisabete, godfather Pedro, and Duarte – the embodiment of family. My grandmother Madalena – the embodiment of motherhood – for always making me feel loved.

Gonçalo Nuno Botelho Amaral Rolão Bernardo

"I want to create a musical DNA,
a technology through which your own personal tastes
can be reflected in the sound waves that you are listening to."

Alexey Kochetkov, Mubert

Contents

List of Figures

List of Tables

Abbreviations

ACF	*Auto-Correlation Function*
AIS	*Artificial Immune System*
BF	*Brute Force*
BS	*Beat Spectrum*
DFT	*Discrete Fourier Transformation*
DTW	*Dynamic Periodicity Warping*
DPW	*Dynamic Periodicity Warping*
EANN	*Evolutionary Artificial Neural Networks*
EDM	*Electronic Dance Music*
FP	*Fluctuation Pattern*
GA	*Genetic Algorithms*
IOI	*Inter-Onset-Intervals*
MFCC	*Mel-Frequency Spectrum Coefficients*
MIR	*Music Information Retrieval*
OP	*Onset Pattern*
RH	*Rhythm Histogram*
RP	*Rhythm Pattern*
SC	*Spectral Centroid*
TIV	*Tonal Interval Vector*

Chapter 1

Introduction

This chapter introduces the topic of this dissertation, computational music mashup creation from a large loop collection, and their relevancy under the current problems and challenges in Sound and Music Computing. Section 1.1 clarifies the context of the work. Section 1.2 details the underlying motivation. Section 1.3 exposes the methodologies employed, and the problem underlying systems for Computational Music Mashup Creation whilst motivating the work for this dissertation with Section 1.4 referring to questions developed through the development. Section 1.5 describes our logical objective and technological processes of which the generative system is objectively prepared to solve, with measurable output quality and stylistic diversity. Section 1.6 lists the objectives and the main problem addressed. Finally, Section 1.7 defines the structure of the dissertation and Section 1.8 includes a peer-reviewed publication related to the current work.

1.1 Context

The automatic creation of musical content using computational technology became popular by the mid-20th century. Particular emphasis has been devoted to the generation of Western tonal music [4] that contributed to the evolution of computer music as means of programming, consequently addressing a Sound and Music Computing (SMC) field. SMC inherits artistic, scientific and technological developments into areas such as artificial cognition, neurosciences and interactive design. By combining these methodologies it aims at understanding, modelling and generating

sound and music through computational approaches. For that reason, the computational retrieval of music information data was possible to advance through Music Information Retrieval (MIR) and Information Systems. Studies within MIR have grown to develop systems capable of assisting in music production in such an efforts to elevate understanding and knowledge defined as what is within music. A great deal of work has been made to build methods of collecting high-level information through music signals, resorted to as content-based approaches [27].

A new area of operation within MIR environment is in the field of Creative-MIR. Some of the key aims of the Creative-MIR are to open up new possibilities for music production, engagement and modulation, which is enabled by the capacity to analyze and perceive music signals comprehensively. Creative-MIR concerns the content-based processing of signals, especially that of automated music compositions [49]. Lately, these mechanisms are expected to become more prevalent and significant, consequently positioning researchers to face the difficulties of the Creative-MIR. Technological advances in Sound and Music Computing and Creative MIR have been building the groundwork for breakthroughs in the large-scale and personalized free music recombination to support artists and consumers. Nevertheless, the capabilities of musicological algorithms are limited to the knowledge gathered nowadays. Several groups and conferences dedicated to Musical Engineering were possible to make a transition from offline to online tools assisting sound and computing processes of auditory data. Notable regards for conferences are EvoMUSART[1], ISMIR[2] or Sound and Music Computing network[3].

Music mashup creation is a composition process greatly associated with the multiple genres of electronic dance music, involving the recombination of pre-recorded musical audio [82]. The computational modeling of music mashups has been pursued in academic and industry environments in light of digital music's significant growth and the increased interest of lay-users in the mashup creation practice. In the latter scenario, computational mashup creation overcomes the need for advanced knowledge on music theory, practice, and digital signal processing. The computational modeling of music mashup features two foremost challenges: 1) the retrieval of compatible audio from a dataset – either the search for several optimal matches or the search of musical audio matches to a target query –, and 2) audio transformations that 'force' the audio to synchronize at some attribute level. This dissertation focuses primarily on the underlying methods of the former processing approach, which is commonly referred to as content-based retrieval within Creative-MIR. Its application to music mashup creation has been recently identified as one of the grand challenges of the community [41].

The capabilities of storing large amounts of audio collections progressed from minimal and personal levels with artists producing their own content, to large scale musical repositories on the Web. These collection particularly include various licenses with variable degree of access from copyleft free music to highly restricted copyrighted and paid content [2]. The public domain's

[1] International Conference on Computational Intelligence in Music, Sound, Art and Design - http://www.evostar.org/2021/evomusart/

[2] International Society for Music Information Retrieval - https://www.ismir.net/

[3] http://www.smcnetwork.org/

online resources stores large amounts of prerecorded musical audio. Loops, i.e., audio files containing different perceptual qualities, instruments, and durations are a particular case of interest here. Although having access to this data proliferated through the musical web environment is easier nowadays, working with smaller and known datasets can relieve setbacks in implementing a framework. Rather, searching for the exact musicological data from audio assets can be time exhausting.

Music is a media element present in retail spaces with growing licensing costs and time-consuming production of generative processes. As the notion of free music came to solve the concern of pricing and variation of the generated musical content, it additionally enhanced processes with new capabilities of dealing with increasing amounts of data in the algorithmic creation of music, in comparison to initial researches of computer analysis technologies of musical data [86, 23]. Recent frameworks offer effective tools for the creative assisting of musical resources with users. This content is consequently possible to get streamed globally and target upon larger markets. However, the multidimensional composition of music is not fully structured to handle complex assessments of networks containing musicological attributes, for which, in the midst of the current environment, is the reason for writing this dissertation.

1.2 Motivation

This area of Algorithmic Compositions for musicological advances has grown extensively through different branches within the formal structure of what describes the musical *corpora*, thanks to progressive MIR techniques. Behind the generation of music, frameworks are created considering technological architectures with Similarity and Compatibility metrics of Harmony, Rhythm, Timbre, and Organizational Structure. The evolution of musical repositories on the web has undoubtedly supported the implementations of these generative systems.

One of the recent concerns is reviews in algorithms oriented to optimize diversity from the auditory data or the structural failure to ensure correctly established metrics. Additionally, the growing use of loops and quality of both timbre and execution of the framework are made for artists to sell their content for the application [1] and not for the sole purpose of computer-assisted production of different auditory results in computational music generation. That is the motivation behind this dissertation: developing content on a large scale inside a formal structure created from horizontal and vertical segments, ensuring that the harmonic quality remains in the instruments' variance and the generative system's harmony. Current applications are limited because of brute force use inside the musical sequences, asking musicians for audio sequences for each genre, and producing no automatic generation of stylistic diversity, but rather an automatic generation of music.

1.3 Problem Statement

Loop-based music development has become popular in retail spaces as the distribution costs tend to be expensive as well as the generative composition being time-consuming. The problem of cost and variety of the created musical material can be solved by free music, or generative music, which also strengthened processes with new capabilities to deal with increasing quantities of data in the algorithmic production of music. Recent frameworks offer advanced methods for the creative assistance of users with musical resources. Therefore, this data can be streamed worldwide and aimed at larger markets. The multidimensional nature of music, however, is not completely tackled in existing solutions, thus motivating this dissertation.

The advantage of these generative models has undeniably facilitated the development of musical libraries on the web. The ability to store large amounts of audio collections progressed until larger musical web repositories. Furthermore, these models were created completely free of charge. The online resources of the public domain store large quantities of pre-recorded loops, and while it is simpler to reach these criteria via the musical network environment, dealing with smaller and established datasets will mitigate setbacks in applying a structure.

Latest researches include analysis of algorithms aimed at maximizing auditory data diversity or institutional inability to ensure properly defined metrics. Because of the use of brute force within the musical sequences, asking musicians for audio sequences for each genre and creating no automatic generation of stylistic variation, but only an automatic generation of sound, makes existing implementations quite limited.

1.4 Research Questions

We advance 3 research questions (RQs) which directed this investigation.

RQ1 What are the technological problems for architectures driving Generative Music?

> According to the state of the art in computational mixing of musical content, what are its main concerns? Are there specific consequences for adopting certain methods or musicological metrics?

RQ2 Which methodologies have been implemented to aid the generation of stylistic diversity and quality of music?

> How do we succeed in improving the mixes towards user's preferences? What are the structural solutions for accounting the multidimensional nature of music?

RQ3 Can optimization algorithms promote compatibility and diversity in musical audio recombination?

> What are the mixing strategies in mind? What are some expected technical difficulties for the generation of the musical content in a diverse solution space promoting compatibility?

1.5 Methodology for Organizational Framework of Music Generation

A novel approach is proposed for the production of music, with quality assessment and stylistic diversity from the recombination of pre-recorded, public domain loops that proliferate in repositories of web music. Both the vertical and horizontal aspects of the harmonic system can be tackled by recombination. Audio loops are samples with differing timbral features, therefore making a formal vertical layer arrangement enabling bigger harmonic variations between instruments to be coherent, and horizontal layers maintain melodic transitions within the generation.

To regulate quality we propose metrics for rhythmic, harmonic, and timbral compatibility holding minimal post-processing of the audio content. To ensure diversity across multiple stylistic conditions, we adopt evolutionary multimodal optimization. Musical audio will be defined by a reduced feature vector in a continuous descriptor space.

Figure 1.1 depicts a diagram CoDi's architecture model.

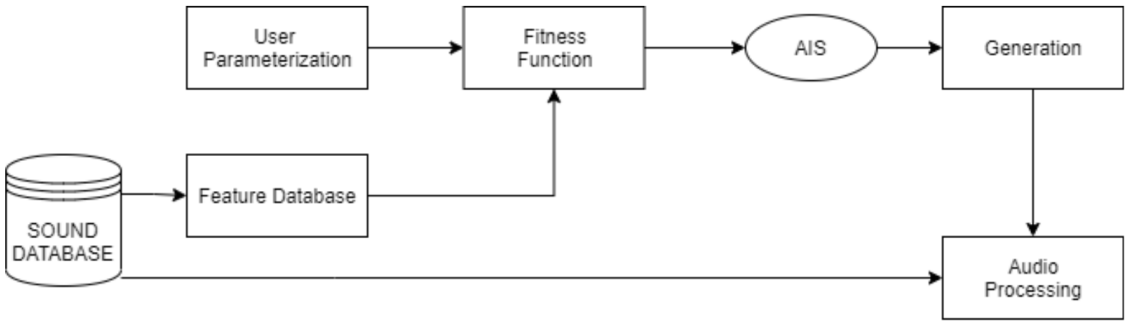

Figure 1.1: Proposed system framework for music generation with loop recombination [3])

The architecture of this application is a loop-based, generative system which is gathering auditory information from a dataset coming from musical repositories. Thus, for each of the source's audio track of that dataset, four descriptors of the musical *corpora* will be configurated to capture harmonic, rhythmic, and timbre attributes.

Computational methodologies for computing audio features and capturing the musical audio compatibility of the above-mentioned attributes are based upon previous literature surveyed in Chapter 2. We are applying a multidimensional framework capturing horizontal and vertical layers of musical structure. An Artificial Immune System is adapted to generate several auditory combinations and continuations of tracks.

1.6 Objectives

The aim regarding the dissertation is the search for a framework capable of supplying users with concise and pleasant generation of music, pervasive for many retailing industries. Upon this work, the focus is to target the development of a system for generating automatic musical tracks at scale,

with a recombination of loops in both the vertical and horizontal dimensions of musical structure. Therefore, the corresponding research literature allows for definitions of specific objectives such as:

- The development of a functional prototype for automatic generation of music as recombinations of loops at scale that are adapted for the two dimensions of the musicological nature.
- To implement musical understanding, computationally, as compatibility metrics for musical attributes. These structures help defining musical style in audio files along several combinations possible.
- Define a computational method for optimized search in large musical loop spaces. This will offer an optimization framework adapted to process large audio files in a continuous space of descriptors.

This work adopts a multimodal optimization algorithm artificial immune system (AIS) for searching compatible and diverse loop mashups from large loop datasets. Compatibility is driven by two main objective criteria, harmonic compatibility [75], and rhythmic compatibility [56], broadly following the metrics proposed in Mixmash [60]. Diversity accounts for the thorough and concurrent exploration for optimal matches across the entire search space. Conversely to existing systems driven by computationally expensive brute-force (BF) search methods (e.g., [27] and [60]), we aim to provide a computer-aided tool that enables a fluid (efficient) workflow on a large user-defined loop dataset while promoting a diverse set of optimal mashups. The model was implemented in Python as a prototype application named CoDi.

1.7 Document Structure

The remainder of the dissertation is assembled as follows:

Chapter 2 is a literature analysis on the State-of-The-Art components for this research on Computational Music Mashup Creation, in regards to generative architectures of musical composition in loop-based environments, and recent methods of similarity and compatibility in defined musicological attributes within Harmony, Rhythm, Timbre in a multimodal Formal Structure. An agenda on evolutionary optimization system for musical audio recombination is addressed. Further investigations for each of its sections are described as well as further developed work on these metrics through means of computation. Following Chapter 3 details the overview of the system, namely the audio features adopted, feature extraction, and the optimization search. Practical work can be found on Chapters 4 outlines the evaluation procedure of CoDi and the results, respectively. In Chapter 5, a description is given, in significant detail, for the conclusion of the developed implementations and research, with expected output of results and understudy of their consequences in the technical areas, and future work involved within the discussion.

1.8 Publications

The initial development of research for this dissertation affirmed a deep case study analysis for measures of the four musical modules previously mentioned. During that time of studying and research, a submission for a collaborative paper from Cocharro et al. [21] was possible, and accepted for one of Springer's editions, named:

- Cocharro, Diogo & Bernardes, Gilberto & Bernardo, Gonçalo & Lemos, Cláudio. Revisiting Rhythmic Representations and Similarity. In: Luisa Castilho et al. (Eds.) Perspectives on music and musicology. Current Research in Systematic Musicology. Springer (2021)

Furthermore along the investigation, additional case studies were conducted, mainly in leveraging for compatibility and diversity in Computational Music Mashup Creation, such as:

- Bernardo, Gonçalo & Bernardes, Gilberto. Leveraging Compatibility and Diversity in Computational MusicMashup Creation (2021), submitted for AudioMostly.

An exhaustive literature on state-of-the-art systems in Computational Music Mashup creation while addressing a tentative agenda on affirming for Audio Compatibility in music production:

- Bernardo, Gonçalo & Bernardes, Gilberto. Musical Audio Compatibility Retrieval: Towards Computer-aided Music Production (2021), submitted for International Symposium on Computer Music Multidisciplinary Research (CMMR).

Finally, a statistical analysis undertaking efficiency of the algorithmic models which are the AIS, GA, and BF approaches:

- Bernardo, Gonçalo & Bernardes, Gilberto. Multimodal Optimization for Music Mashup Creation (2021), submitted for Doctoral Congress of Engineering.

Chapter 2

Generative Strategies for Musical Audio Recombination: a Literature Review

The technological advances of sample-based audio recombination resides in two fulcral elements: representation, and perceptual distances at a given attribute level, commonly addressed as similarity and compatibility. Systems for computational musical audio recombination have three approaches of brute-force (or rule-based), Neural Network algorithms, or optimization strategies for accelerating the process of generation. Academic research focused on promoting sample-based music creation have become popular alongside the increasing interest. According to the proposed objectives of the dissertation, there are four lines of action for musical structural parameters such as Harmony, Rhythm, Timbre and Formal Structure. In this chapter, a description for a literary review on generative architectures acting as instruments of sample recombination in Section 2.1, as well as musicological strategies for compatibility approaches inside the domain of audio, and what constitutes the four descriptors mentioned from Section 2.2 through Section 2.5. A majority of the methodologies are dependent of structural descriptors, and so, raises a need for intermediate representations minimizing the structure onto the four modules of musical compatibility and similarity.

2.1 Generative Systems for Loop-Based Methodologies

In this section, a review on the extensive literature of computational models in the context of music generation through sample-based data. Further ahead in respective sections, a description

is attained to Genetic Algorithms, Artificial Neural Networks, and Artificial Immune Systems, paving new Evolutionary models for Computational Music Mashup Creation.

2.1.1 Genetic Algorithms - GA

Through Genetic Algorithms (GAs), evolutionary models have built recent methodologies to restrict specific knowledge to the issue's domain [59, 20]. Beginning with the generation of a random population constituent of chromosomes as symbols of binary data (gene pool) computationally designated to act like candidates, each gets calculated a fitness score from a function significant to a rule-based system that examines the chromosome's capabilities of solving the issue. Generally, candidates with scores higher than a threshold end the procedure and are considered fit to fulfill their purpose. On the contrary, the new population is estimated by either:

- Selection/Reproduction: Ranking chromosomes through a candidacy of fitness values.
- Crossover: Segments of randomly matched chromosomes have intertwined values within a crossing site.
- Mutation: Segments of the individual chromosome are changed through positioning by randomness, or when a segment is discarded.

The first use of Genetic algorithms for music generation was documented by Horner [48], using a technique of thematic bridging for the development of the new melodic sequences. Two approaches to help construct this concept were suggested in an early study [87]: automated loop extraction and aided loop selection, laying out the basis for the musical sector. Further use cases aiding in musical composition have a common situation when implementing their computational measures, whether loop-based [81, 80], or not. In possession of large-scale assets, a concern raises in maintaining a navigation that is optimized and efficient within the musical space. Additionally, the choice of loops are just as important for the process. With the aid of MIR techniques, loop selection and extraction made possible an ease of search within the database. Past studies on Compatibility Estimation for:

- Loop Extraction: Shi and Mysore [81] implemented a system made to segment a loop from direct contact, by using compatibility measures of harmony, timbre and energy [87, 81]. Additionally, Smith et al. [85, 84] proposed to analyse a pattern of repetition to extract loops.
- Loop Estimation: Kitahara et al. [53] affirmed a concept of a manual input level of excitement from the user, however limiting usability and compatibility. Had the user found a loop to his liking, it would have a chance to be musically incompatible.

In 1989, Dr. David Goldberg defined GA as natural selections and genetics, through mechanical methodologies. Nowadays, the possibilities of deriving and mixing different artificial intelligence algorithms are increasingly clearer, as this research explains ahead the Darwinian synthesis

of using Neural Networks with GA, and, additionally, the symbolic and computational use cases of Artificial Immune Systems, both implemented for musicological means.

2.1.2 Artificial Neural Networks - ANN

Evolutionary algorithmic models combined with Artificial Neural Networks (ANNs) have been proposed as EANNs, or Evolutionary Artificial Neural Networks. An advantage from this conjoining is the search optimization [50, 96] capability within recombinations of increasing amounts of data as implementations below will refer to.

A structural analysis of ANN's constituents are:

- Neuron: unit/element of artificial computing.
- Architecture: connections and corresponding patterns.
- Learning Phase: Procedure of training for neural networks.

One of the major dissimilarities between ANNs and GANNs mentioned is the process of storing relevant data [59]. Whereas systems based on Genetic Algorithms have data that is capable of predicting solutions, Artificial Neural Networks include possible extra information that may not have any use for the domain. Instead, ANNs assign low weights to the irrelevant data [88]. In Chen et al, [20] a technological implementation of ANN for Loop Compatibility allowed models based on Convolutional Neural Networks (CNN) evaluating representations within time and frequency, and additionally, Siamese Neural Networks (SNN) processing individual compatibilities for two segmented audio loops. By conducting a user test of combinations by the model, CNN performed better than both SNN and the rule-based systems. Despite not implemented through the use of Neural Networks, compatibility estimation systems such as MixMash and AutoMashupper [60, 27], explained in the Harmonic section of this dissertation, are mainly rule-based and may be augmented by machine learning as they do not make decisions for the user, but rather aid that choice.

2.1.3 Evolutionary Optimisation for Musical Audio Recombination - Artificial Immune Systems (AIS)

Evolutionary algorithms are a class of artificial intelligence methods greatly motivated by optimization processes inspired by natural phenomena, such as natural selection, species migration, bird swarms, human culture, and ant colonies [83]. Evolutionary optimization algorithms can be defined by two main criteria: modality (unimodality and multimodality) and the number of objective criteria to optimize (single- and multi-objective). The modality denotes optimization strategies that seek solutions for one local optimum (unimodal) or multiple local optima (multimodal) in a single run of the algorithm across multiple iterations. Multimodal evolutionary algorithms usually account for the population diversity, resulting from a comprehensive exploration of the search space [97].

Single and multiobjective optimization differ in terms of the objective search strategy applied. Single objective finds optimal solutions to a single objective function, whereas multi-objective accounts for problems with conflicting objectives with no single optimal solution [79]. In our work, we concentrate on multimodal and single-objective search optimization, particularly the AIS opt-aiNet algorithm.

Decastro [28] presented the AIS *opt-aiNet* algorithm to solve multimodal function optimization problems. The algorithm can evolve a population of cells towards a set of optimal and diverse solutions to a problem. It employs immunological concepts of clonal proliferation, mutation, and repression to establish a network of inhibitors in the immune system network. In other words, opt-aiNet integrates local and global search to find optimal solutions while maintaining their diversity. Furthermore, the algorithm presents two additional important features: the automatic determination of the population size and a defined convergence criterion. [17] adopted the opt-aiNet to the problem of computer-aided orchestration, i.e., the search for instrumental combinations that match a reference sound by combining instrumental note samples. They showed the primacy of the method in promoting diverse solutions with optimal quality.

For an Artificial Immune System (AIS), automated search spaces can be aided by detecting segmented combinations of music files from datasets meeting a desired specification. In contrast from previously mentioned traditional GA, the AIS is worthy of seeking several variations as strong candidate options, whilst ensuring variety. An assessment of the empirical and subjective variety of ten candidate solutions discovered by the AIS was carried out in Abreu [3], compared with 10 individual GA runs and fifteen Computer-Aided Musical Orchestration algorithms in addition. In order to acquire separate variated arrangements, the compositions found by the AIS displayed compatibility and variability analogous to executing GA several cycles.

In addition to Abreu [3], an artificial immune system (AIS) called opt-aiNet would be implemented to check for variations of segmented instrumental sounds minimizing the distances to a target sound that is embodied in a fitness method, as a similar strategy in this dissertation. Inside a fitness function, timbral compatibility is structured so the stronger combinations have better convergence of higher fitness values, since the goal is to obtain variations that equate to the limits of the fitness function, as seen in Figure 2.1. This fitness function is represented by several peaks, layers, and the dark spots reflect values of individual sound fragments, or combinations. The extensive search for all possible combinations includes heuristics to reach a resolution in much less time, or perhaps to suffer from heavy computational and time-consuming algorithms. Therefore, the method through AIS works best when speaking about musical production and where different solutions to the same problem are needed.

De Castro and Timmis [28] have developed an artificial immune system (AIS) designated *opt-aiNet* for multidimensional optimization algorithms, presenting a variety of potential solutions as an optimum fitness value. It employs immunological concepts of clonal proliferation, mutation and repression to establish a network of inhibitors in the immune system. Opt-aiNet integrates local and global search to find and achieve optimal fitness scores at the same time as maintaining the variety of solutions, in both dimensions of horizontal and vertical space of musicology. It is

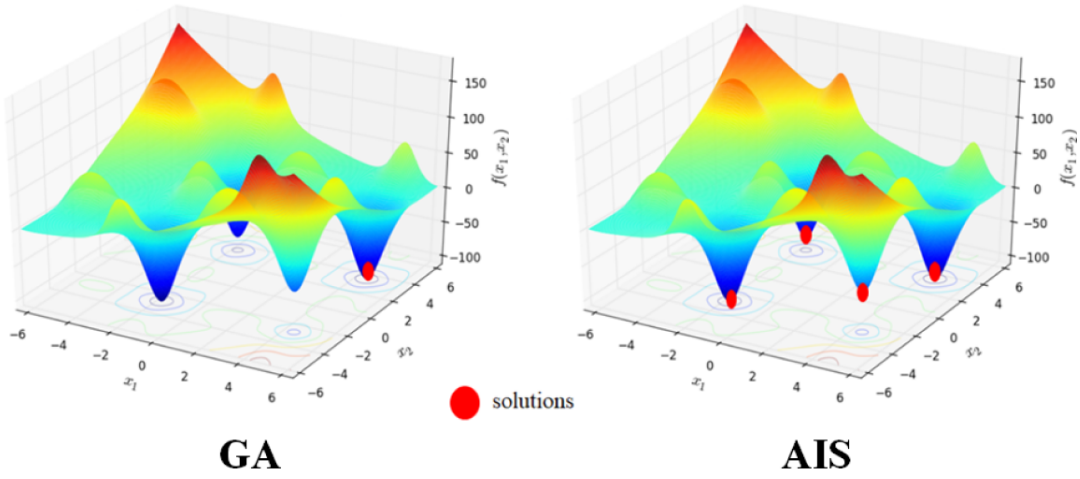

Figure 2.1: 3D Representation of Recombination solutions from AIS vs Genetic Algorithm

capable of finding sets of good candidates who are distinct from each other, but considered equal as solutions for the optimization. Abreu [3] implements its fitness utilizing such attributes, in order to estimate fitness through compatibility between the audio configurations in the dataset and the intended musical output. As mentioned above, AIS is often implemented to scan for variations that approximate the pattern of the target sound, named orchestrations. Every orchestration is a set of sounds differing in length that belong to their musical repository. An additional component in their model was the phase vocoder design component utilized to track or condense each audio fragment to the average temporal length, ensuring simultaneous ending and begining times, as they are played together.

Another example of use cases for these algorithms are the implementation of Opt-aiNet by the named system ChordAIS [65] assisting its users in generating realistic and functional chord progressions. However, this is through a dimension of symbolic domain, not audio processing. By procedures of optimization within the encodings of the objective function for progressions and attractiveness of musical properties of Harmonicity, technological possibilities for Opt-aiNet are to concurrently locate and output several multimodal solutions that are considered optimal, resulting in multiple candidates of high quality that can be applied to the user's desire. It is based upon more abstract knowledge of musicological principles that allow a stronger manipulation of descriptors.

Validation is carried out by various experiments and hearing tests determining the auditory consistency of the candidate chords proposed by ChordAIS [65]. These outputs were rated by most listeners as stronger candidates for progression than the chords discarded by the system. When comparing with two related architectures, ConChord (Gilberto et. al [8]) and ChordGA, which use a regular GA instead of Opt-aiNet, consumer evaluation revealed a preference for ChordAIS over ChordGA and ConChord. According to the findings, by recommending good-quality candidates in all the keys evaluated, ChordAIS was considered capable of assisting users in the generation of tonal chord progressions.

An analysis review for evolutionary computation inside the musicological space notes that architectures based on neural networks typically feed from the signal itself [66, 96, 20]. Whereas other computational systems, for this regard, need intermediate representations that minimize the subjective nature of the musical structure which are further ahead segmented and explained into 4 sections of metrics: Harmony, Rhythm, Timbre and Formal Structure.

2.1.3.1 Search Optimization and Fitness Function within AIS

As we're presenting a search space according to distances, our optimization process seeks to minimize the distance between the continuous flow of solutions within the evaluation function, as they are considered better for auditory perceptual senses. Recent literatures, however, adopted maximization of similarity. As it is executed by an existing loop extraction algorithm, the main concern is ensuring that both quality and stylistic diversity are shown in compositional results of the framework. Furthermore the Artificial Immune System opt-aiNet is adopted to optimize the search for optimal compatibility with multiple diverse solutions within minimum locals.

2.1.4 AIS vs GA

Carpentier et al. [19, 18, 89, 90] , because of their ability to discover and manipulate the search space, use GAs to conduct this study of search optimization methodologies. The mechanics behind its exploration is essential for searching for new potential solution space regions, or fitness function peaks, in order to strengthen the value of the existing candidate solutions. Nevertheless, the regular GA, as seen in Figure 2.1, faces a shortage of variety upon convergence, resulting with only a single solution leading to a single peak generated by the GA. The stochastic design of the detection algorithm will not mean that the optimal solution is discovered, sometimes trapped in local optimum. Performing the GA several cycles to the same specifications also leads to the subsequent peaks of the fitness function observed in dissimilar solutions.

2.2 Harmonic Compatibility

Harmony is a fundamental principle in Western tonal music as it relates to how sounds are vertically aligned and looks to minimize concepts like dissonance. Measures for Harmonic Compatibility are defined as the capability for musical audios to generate audio of enjoyable quality, when merged [74]. These methods are commonly expressed as two dimensions of horizontal (alternatively known as sequential compatibility) and vertical (also described as simultaneous) compatibility of musicological structure [54, 44], in which the underlying principle is to automatically quantify the degree of harmonic proximity of two or more musical audio samples. When mixing audio tracks, output possibilities are either harmonically compatible – making pleasant music – or have a very dissonant auditory return.

Harmony compatibility in audio-content-based processing, notably in musical audio recombination, has been researched under 'harmonic mixing' [10, 60]. Moreover, Bernardes et al. [10]

stress the importance of concurrent hierarchical dimensions in harmonic mixing at the level of the key for large-scale structure (mostly enforcing the horizontal dimension) and dissonance and perceptual relatedness or harmonic distance at the small-scale structure.

We identified three main categories of harmonic mixing methods: 1) key affinity, 2) chroma-based similarity and in related (or enhanced) tonal pitch spaces, and 3) psychoacoustic (dissonance) metrics. Their main difference relies on the type of representation adopted, from which typical metrics (e.g., Euclidean or angular distances) compute their similarity.

The methodologies for searching quality audio candidates in the harmonic compatibility of generative music are following:

- Key Affinity prominent procedures methods used in commercial applications [95].
- Spectral Similarity methods to find the spectral candidate with similar characteristics and patterns as the target track.
- Dissonance/Consonance based methodologies searching for the highest values of consonance between merged musical audios [71]. These are perceptual calculations for the quality in the mixing of tracks, as they can both output successful or badly perceptual tones.

Similarities in the spectral matching of audios are helpful to maintain technical correctness in the structural harmonicity of the auditory mixing. If the user's purpose is to test the sole power of attraction, measures of dissonance/consonance computations give perceptual data of the entire output. The ideal candidate is the one most compatible and of similar pattern and features, compared to the auditory source.

Approaches recurring Dissonance and Consonance have higher possibilities of varied outputs due to their evaluation of *pleasantness* [15, 24] coming from the mix, as the most compatible and suitable candidate is a conjunction of root note and its corresponding octave.

2.2.1 Key Affinity

The affinity between musical keys is defined by distances across major and minor keys in the double circular representation shown in Fig. 2.2, known within the DJ community as the 'Camelot Wheel'. This method favors large-scale harmonic similarity by promoting major-minor modes mixes and intervals of fifth relations across musical keys and poorly captures small-scale harmonic similarity [9].

2.2.2 Spectral Similarity

Methods from Spectral similarity do not prioritize defined frequencies, but are rather focused on the harmonicity of mixings, taking into account higher levels of auditory representation such as chromagrams or spectograms for pitch. Studies for in-depth harmonic compatibility have undergone solutions for matching Chromagram values of different windows, with definitions of

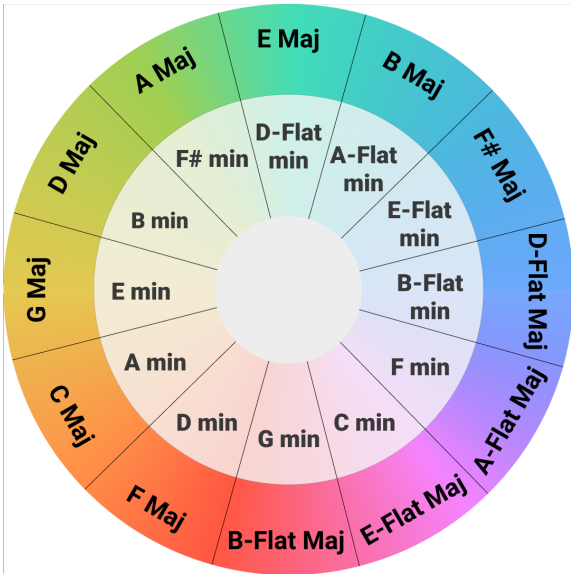

Figure 2.2: Camelot Wheel representation of major and minor keys, representing the perceptual proximity of keys, commonly adopted in the context of harmonic mixing.

harmonic-based metrics in musical tracks both at small and large-scale whilst making new assessments to perceptual relatedness within the Chroma space such as key affinity and vectors of similar chroma [27]. Alternatively, transforming the auditory space into tonal vectors through DFT correlation, or analysing windows of similarity with dissonance-based approaching. Beat-synchronous chromagrams over pre-defined tempo and key ranges were possible for gathering harmonic compatibility [27]. Ultimately, these are complying with the environment of Western harmonicity and tonality [60].

2.2.2.1 Chroma Matching

Tonal information within musical audio files is computationally possible to be represented as Chroma vectors, in which each segmented excerpt of the track is put into one of its corresponding twelve dimensions of pitch classes (i.e., from C through B pitch class).

The concept of Chroma matching through vectors built groundwork for important case studies in mashup creation, recognition of chords [52] or for magnifying techniques of MIR [64]. Notable regards for this dissertation are MixMash and AutoMashupper, both frameworks to aid users in the creation of mixings through different audio files. A base for Compatibility Estimation is the implemented research of Davies et al [26, 27] with AutoMashupper on which sections of audio tracks are conjoined or intermittently switched at different phases through algorithmic composition. To ensure optimal sequential compatibility, Chroma vectors between 2D convolutions of the songs are integrated with beat-synchronous data of the musical files. However, both applications of Mixmash and AutoMashupper have a general concern of lacking a vertical dimension of simultaneous compatibility according to the musical structure.

This system was generalized by Lee et al. [54] and suggested that two components could be more compatible if they have additional quantities of harmonic instability. There are two drawbacks of AutoMashupper, which treats the harmonic and rhythmic resemblance as part of mashability, While it is theoretically possible for two music segments to correspond, While having various harmonies and rhythms, varied results are .

A measure of Spectral Balance came to popular rise, in recent times, after the implementation of Automashupper, where developers segment regions dependent of the beat offsets of the tracks. By a concept of perceptual loudness, they find the flattest spectrum coming from the beat offset and create a mathematical conception of mashability.

2.2.2.2 Tonal Interval Space - TIS

Music theory specifies meanings of consonance and dissonance through intervals between the notes in the musical scale. This information is valuable for determining both the perceptual human listening and the existence of tonal intervals. This human perception of musical sections does not prioritize methodologies using Chroma vectors as they are listed through proximity of frequency in ascending order, rather than proximity of better harmoniously sound pitch classes, as Euler's model of Tonnetz states and is observable in Figure 2.3.

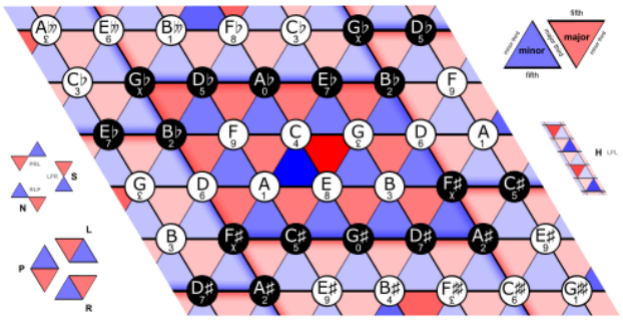

Figure 2.3: Euler's Model for Tonnetz states (Source from Wikipedia's public domain.)

The Tonal Interval Vector measure, or TIV, previously implemented in recent researches [60, 11, 8], reformulates the possibility of Chroma-matching vectors and models of Harmonic Compatibility, with finding commonality between the intervals of pitch classes, as observed in Figure 2.4. This research allows the prioritization of less dissonant mixes, between audio segments, through the calculation of similarity in the dissonance of musical tracks.

Fernandez [74] conducted an online survey running several algorithms of Harmonic Compatibility on a ranking of worth, according to additional standard errors. Results were that the implementation of TIV had better performance within the metrics established for harmonic compatibility, stated in Figure 2.5.

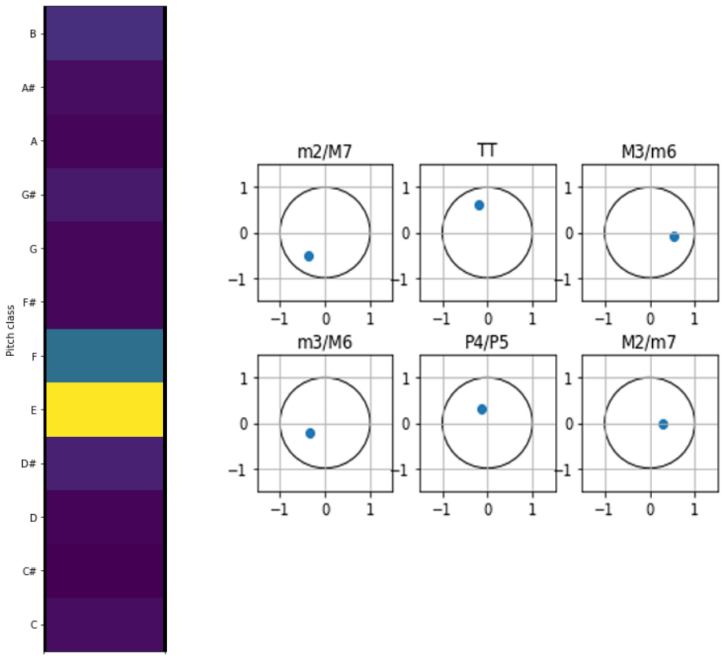

Figure 2.4: TIV representation of Chroma Vector [7]

2.2.3 Dissonance/Consonance

Methods taking the phenomenon of dissonance have recently taken a conception of simultaneous consonance, in which a perceptual salience is noted from musical notes playing simultaneously, and represented in the spectrum. As referred in the theory of musical environment, consonant combinations are perceptually known to appeal users, rather than dissonant sets taking less enjoyable combinations of tones [44].

This section presents algorithmic methods taking representations that are driven by perceptions of Interference and Harmonicity.

2.2.3.1 Pure Dyads and Roughness

According to the terminology of chords, classifications can vary according to different dimensions. Terms such as dyad or triad are composed of, respectively, a collection of two or three notes played simultaneously. As such, the concept of pure dyads are of pairs characterized by two pure-tones concurrently playing [44].

Interference of pure tones are possible through masking and beating, although studies are limited regarding consonance through masking, in contrast with beating. The impression that is beating comes from a pair of frequencies that are close within the spectrum, and consequently produce an auditory sense of amplitude modulation. Accordingly recognized as perceptually unpleasant, the beating from the sum of all frequencies is designated as Roughness.

The groundwork for more approachable models using this measure came from Plomp and Levelt [71], where dyads are composed of an average coming from a set of tones and the divergence

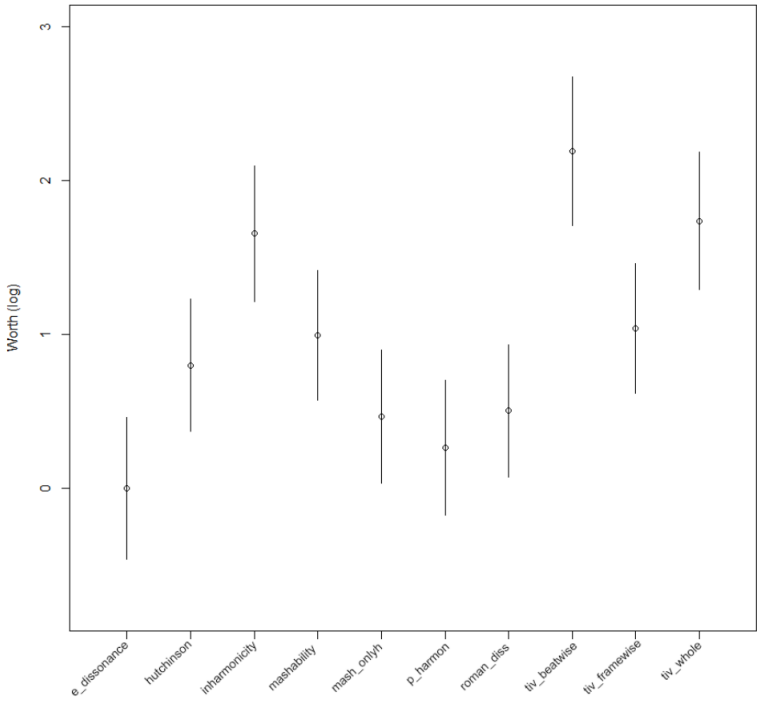

Figure 2.5: Harmonic compatibility algorithm worth with their standard errors [74].)

length between their harmonics. A concept of Critical Bandwith was defined as the frequency difference of two pure tones in which the sense of "Roughness" fades and smoother results are possible. Results were this methodology of perceptual roughness is dependent of the dyad difference in frequency, and the Critical Bandwith. Studies further developing this model were from Gebhardt et al. [38] implementing a mixing methodology, as shown in Figure 2.6, capable of diminishing levels of roughness with a model of Pitch commonality from Parncutt and Strasburguer [69]; and integrated with a roughness-based architecture in the works of Hutchinson and Knopoff [51]. To achieve the greatest value of compatibility in a sequential dimension, a Spectral Modeling tool[1] helps the current model to extrapolate values of roughness of two musical segments within a range of one octave composed of both 48 ascending pitch shifts, as well as descending ones too. Later on, Maffei [58] implemented a similar model minimizing levels of roughness by decreasing the highest candidates of frequencies contributing for the perceptual Interference of roughness and dissonance.

2.2.3.2 Psychoacoustic Metrics for Harmonic Compatibility

Systems based upon the notion of Periodicity have in consideration a pattern recurrence rate in time and physical properties of a sound wave such as frequency, therefore making possible the spectrum transformation into integers that are multiples of the fundamental frequency. Studies that regard this methodology are limited to higher debates of measures such as an auto-correlation

[1] https://github.com/MTG/sms-tools

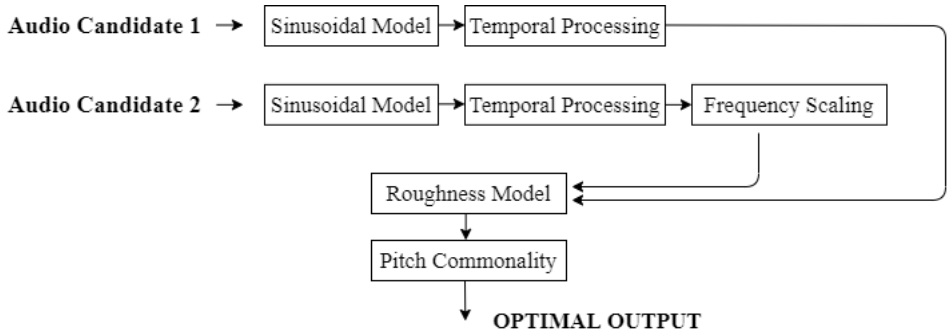

Figure 2.6: Model of Pitch Commonality [38, 39])

transformation within the dimensions of time, and pattern similarity within the frequencies of the spectrum [14, 44]. Thus, new forms of achieving consonance are also possible.

Peaks within the spectrum allow a perceptual inference of frequencies such as the fundamental or any corresponding harmonics. Thus, even if they are not physically present in the mixing, users go through this phenomenon of Virtual Pitch in order to sense the tones. A model was achieved by calculating overtones, masking, loudness, and the high points of pure tones. The inference of musical chords are possible through an algorithm of Pitch commonality, in which the power of each of the twelve pitch classes mentioned are perceived.

Like stated, Gebhardt's work [38] aided the process of musical mixing. While getting hold of the Virtual Pitch algorithm, he made possible to determine, for both target and source tracks, the perceiving power of triads composed of the three highest perceptual frequencies, available for keys of Minor and Major scale [63]. The inter-relationship between the excerpts is a methodology for consonance, as it is calculated from the 2 pairs of triads through the pitch commonality algorithm [69]. For Harmonic Compatibility, the final results concluded the algorithm of pitch commonality integrated within roughness [69] performed worse than models disregarding components of masking or gain levels [38]. The model of spectral pattern matching returning the best output considers a divergence in the harmonic processing of the spectrum [45]. This methodology returns a probabilistic distribution and with the aid of Kullback-Leibler divergence, ensures a constant value of 0 if no pitches are salient.

2.3 Rhythmic Compatibility

Symbolic representations of computational approaches provide degrees of associated abstractions that are transparent and aligned within Western music's notation formality. The complexity behind audio processing and rhythm compatibility has posed harsher limits in the compilation evidences so far. One of the common use cases is note onset detection [6], not freely accessible, but rather implicit in audio recordings. In this section, a thorough literature on the audio domain representations underpinning computational methods for audio compatibility adopted in content-based generation. As such, a general viewpoint is shown in Figure 2.7

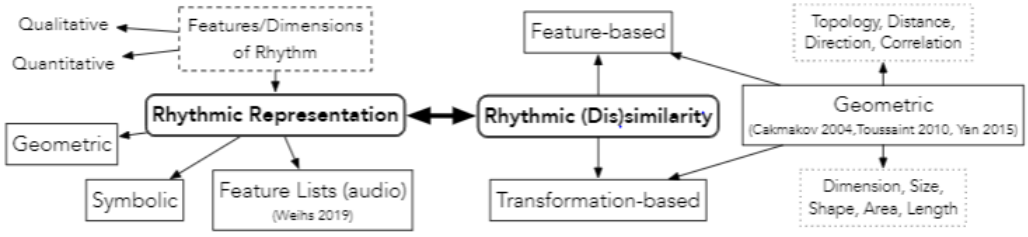

Figure 2.7: Dimensions of Rhythm (Dis)similarity [21])

Early computational mashup creation systems focused on rhythmic-only features related to the temporal arrangement between two or more musical tracks [10, 43]. Lee et al. [54] concentrated on rhythmic matching serving tempo as an input parameter for the system and employing beat matching by stretching the beats through phase vocoder. Today, this strategy proliferate in commercial software such as Tracktor, [2] Mashup 2, [3] and Mixxx. [4] Davies et al. [27] perform beat and downbeat tracking to align the tracks, based on two onset detection functions – kick drum and snare drum – while assuming a constant tempo and time signature across the entire duration of a musical track.

Rhythmic representations from musical audio can be organized into three abstraction levels: low, mid, and high. For this writing, low and mid dimensions are better focused for the matter of rhythmic representation and compatibility through audio computations. Roads [76] describes low-level classification as the retrieval of feature lists from which it is capable of segmenting individual events. The classification of the middle level as transcription-based is equal to the level of data as abstract representations. Finally, depending on the application, high-level definition as the composition or style analysis. As mentioned from developing a previous publication, this is following similar notations of Cocharro et al. [21]

2.3.1 Low-level Rhythmic Description of Compatibility

In low-level rhythmic definitions of audio, time series that reveal the volume of change of the content of an audio signal, over time, are given a fundamental focus. The difference between consecutive descriptor values of positive, negative derivatives and total shifts in the musical audio material, is usually adopted to compute a novelty function [6].

Representative examples of baseline descriptors for the novelty feature computation are energy and spectral-based representation. The primary application of novelty features is to detect abrupt shifts in the musical audio sound, which typically signify the onset of notes. Different details may be inferred based on the baseline function implemented. In order to catch improvements to the audio signal that align with rhythmic variation, embracing sole energy information may not be enough. In the frequency domain, a note transition over a series of conjoined notes can be best

[2] https://www.native-instruments.com/en/catalog/traktor/, last access on 20 April 2021.
[3] https://mashup.mixedinkey.com/, last access on 20 April, 2021.
[4] https://mixxx.org/, last access on 20 April, 2021.

recorded. Similar to the energy-based innovation, this latter of spectral flux approach operates from variations between consecutive spectral vectors.

2.3.2 Mid-level Rhythmic Description of Compatibility

A popular rhythmic representation of musical audio at the mid-level is a systematic series of string from beginning of onset times. Onsets may be interpreted from novelty functions as local maximums. Inter-Onset-Intervals (IOIs) could be further determined from the resulting series, supplying higher-level knowledge about the length of the periodicity of the rhythmic structure. The salience of feature periodicity has, however, been more commonly discussed as continuous functions, as several representations of the periodicity function were suggested. Their key distinction is based on the type of information being represented like event locations, and their interpretation as both sequence of events and histograms.

In the auto-correlation function (ACF) as states Dixon [30], limited bandwidth copies of the signals are implemented from the amplitude envelope to detect periodicities in each band. To approximate tempo, meter, and the periodicity distribution to predict compositions, relations between periodicities are then used.

The beat spectrum [33, 5, 67], as a result of time lags, reflects rhythmic periodicity. The beat spectrum can be calculated from diagonal sums of pair distances in a self-similarity matrix [33] after signal parameterization (e.g., brightness, loudness and MFCC). The measurable signal will uncover the periodic structures from the peaks in the musical audio. Repetitive music can have high peaks in the repetition times of the beat spectrum.

The Rhythmic or Fluctuation Patterns(FP) [57] define rhythm in various frequency bands as the modulation of loudness. The effect is a time-invariant 24 critical Bark bands interpretation that captures the audio signal's repeating patterns (i.e. periodicities) and can demonstrate the rhythmic structure of the critical bands. Pohle et al. [72] extend this concept of FP, as shown in Figure 2.8 in a semitone space by an onset-based representation of rhythmic patterns, and its representation of the matrix, called Onset Patterns, is normally minimized using the DCT [72] that dismiss higher order coefficients.

From the FP algorithm, the rhythmic histogram description of musical tracks through all bands can be derived, consequently anotating the magnitude of each modulation frequency bin [56]. In an improved ACF version of a discrete transform signal, Tzanetakis and Cook [94] suggest a related beat histogram guided from peaks.

Since in periodicity representations of musical rhythm [46] the temporal component is partly disregarded, Holzapfel and Stylianou [47] follow the scale transformation as a special case of the Mellin transformation, to ACF, allowing the rhythmic representation invariant of tempo shifts, accounting for identical rhythms in various tempos. Tempo estimation and beat tracking, commonly noted as an unified value derived from the periodicity function detailed above, are two additional and fundamental mid-level rhythmic representations of musical audio [29, 42, 30].

Established state-of-the-art methods operate well among clear beats through musical materials, such as in commercial musical genres. Expressive tempo fluctuations and non-percussive

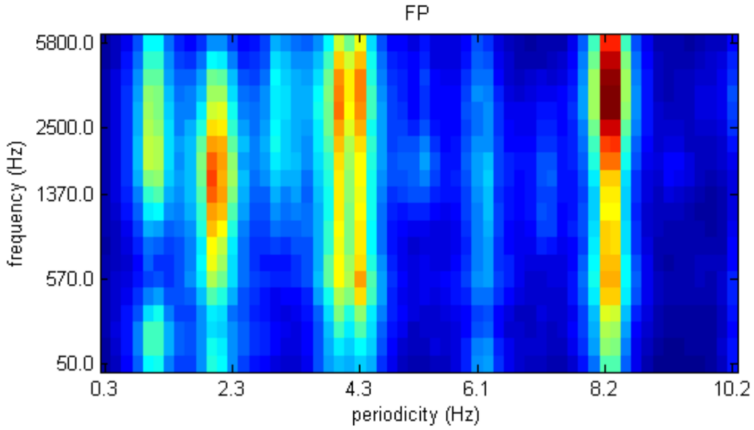

Figure 2.8: Fluctuation Pattern of measured periodicities from a musical source [72]

audio content are rather quite difficult [62], and investigations for end-to-end algorithms have tackled tempo detection and beat tracking, leveraging deep learning methods [16], displaying and improved outputs, including non-percussive musical content [40].

Relevant to the context of our work is rhythmic compatibility computation adopting continuous periodicity functions, such as those shown in Fig. 2.9. Three representative examples are the beat spectrum, rhythmic or fluctuation patterns, and rhythmic histograms.

2.3.3 Feature-Based Methodologies for Audio Compatibility

To understand processes of compatibility, questions of what is measured and how it is measured [91] are brought to concern. There are distinguishable two broad approaches to rhythm compatibility:

- Feature-based methods, which compare the number of common traits;
- Transformation-based methods, which measure how much effort is required to transform one pattern into another [92].

For the current development, the dissertation is mainly focused upon common methodologies for feature-based transformations (e.g. Euclidean Distance). As it is used in [29] for auditory compatibility computation, it can also be implemented to classify related rhythmic patterns as groups in clustering algorithms.

For the calculation of rhythmic similarities across musical audio, Foote et al. [34, 36, 35] follow the cosine distance around the beat continuum and the Fourier beat spectrum coefficients. Using rhythmic histogram, FP, and Onset Patterns [56, 70, 13], related metrics were introduced to new computational systems. Supporting the Euclidean distance is among the most common cases of the Minkowski distance with broad rhythmic compatibility formulations and can be generalized to just about any rhythmic data representations, representing the distance between the magnitude of the vectors.

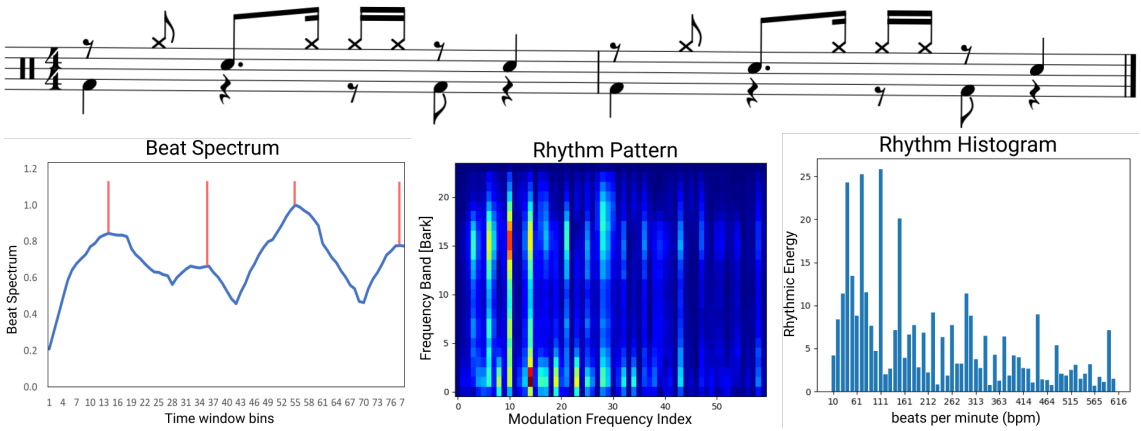

Figure 2.9: Rhythmic representation of a snare drum and closed hi-hat rhythm (represented in ascending order in the top image). The three bottom images represent rhythmic periodicity functions (from left to right): beat spectrum, rhythm patterns, and rhythmic histogram.

Before the distance estimation, the dynamic time warping (DTW) algorithm is used to account for the estimation of the shortest distance through musical rhythms, typically applicable to audio-driven models that have stronger granularity in the temporal resolution and express micro-timing irregularities. By order to align them in the best possible strategy to reduce the Euclidean distance, the DTW calculates the minimal cost throughout sequences [70]. Dynamic periodicity warping (DPW) is a related technique yet was proposed to compensate for tempo variations in the computation of their compatibility by rhythmic audio representation [46].

2.4 Timbre Compatibility

Perceptual dimensions of sound bearing timbre or analyzing source signal classifications have gained scientific awareness, as a timbral field of musical studies leverages uncertainties for researchers left to investigate. Timbre is characterized as every aspect of the sound independent of loudness coming from the mix's individual parts. A sounds' timbral nature is defined as a musical piece's character independent of pitch and intensity. Timbre is one of the most elusive musical audio attributes. It has been defined as "the psychoacoustician's multidimensional wastebasket category for everything that cannot be qualified as pitch or loudness" [61]. Unpacking the above definition for the context of automatic mixing of musical audio, we can frame the idea of timbre compatibility as a metric capturing musical structure attributes independent of pitch and loudness.

Studies upon timbral manipulation methodologies are directed towards analysis in musical sections of human emotional responses or specific genre attributes with identification algorithms. When judging a songs' piece, we are perceptually judging the timbre of that corresponding section. Recent studies [31, 32] conducted different timbral manipulations impacting the overall perceptual sound quality. Regarding timbral compatibility measures, a framework for assisting in creating musical mashup creation [60] compute the cosine distance between Mel-frequency

spectrum coefficients (MFCC) as a timbral similarity metric to support users in selecting musical audio samples with varying degrees of timbral resemblance via a graph-based visualization of large archive.

Inspired by the key role of equalization in audio mixing, Davies et al. [27] proposed a measure of spectral balance to promote an equal energy distribution across the low, mid and high spectral regions of a mix spectra. Mathematically, they apply the concept of spectral flatness to capture the balance of the resulting mix.

Rocha [77] identifies polyphonic patterns that have been empirically chosen by a limited number of features to characterize timbre in the context of EDM. While specifying three kinds of properties, it collects the most important dimensions of polyphonic textures for contrast with others. Their analytical structure consists of three evaluations by means of MFCCs, Spectral Flatness, and Dirtiness, which are further converted into a feature vector representing timbre of an auditory fragment. Similarity between two separate sources is defined by measuring the Euclidean distance between the two corresponding vectors, and the similarity rating model produced positive results. As there is no ground reality for both literary computational knowledge and database for timbre compatibility, there was no model evaluation inside the built architecture. However, it calculates a similarity ranking focusing on various timbral features and equal weighting. A feature vector was developed [77] to define a specific timbre, its most innovative feature being 'dirtiness,' that reasons for the harsh texture that is typical of certain forms of EDM.

Computer-aided orchestration is another relevant application dealing with timbral compatibility to approximate a target query sound – the task includes assessing the perceptual compatibility of multiple instrument note combinations to the query timbre and the compatibility of the instrument note mix. In Abreu et al. [17], the timbral dimensions considered are fundamental frequency, pitch, and amplitude of the spectral peaks, loudness, spectral centroid, and spectral spread. The artificial immune system opt-aiNet is adopted for an efficient optimization search across large musical instrumental note corpus, considering both quality and diversity in the resulting orchestration mixes.

Following attributes from [31, 32], the next subsections express the relevance for recent methodologies adopting timbre similary.

2.4.1 Spectral Slope Manipulation

This concept is linearly correlated to the spectral centroid audio descriptor, an independent variable α is defined as the variance of partials' amplitudes acting as a lowpass filter. Consequently, upper partials' amplitudes can get intensified and perceptually brighter.

$$A_k = \frac{1}{k^{\alpha}} \tag{2.1}$$

Symbol A as amplitude of partial and k as order of partial where $k = \frac{F_k}{F_1} = 1, 2, ...$

- If α=1, the partials have an amplitude proportional to the inverse of their harmonic number (1, 1/2, 1/3, ...) approximating a sawtooth oscillation.
- If α=0, all partials and F1 have equal amplitude, approximating a pulse stream.

2.4.2 Odd to Even Ratio

Within this function, a variable β is controlling the even-number partials' gain level by multiplying with values between 0 and 1. It is densely correlated to the auditory descriptor named OER. Researches within waveshapes affirmed square waves are related to distortion, which appends odd harmonics. It is a conventional method for continuous waveshaping variation.

- If α=1 and β=0, the resulting sound would have no even-number harmonics and is similar to a square-wave oscillation.
- If β=1 approaches a similarity to a sawtooth oscillation. Negative values give similar outputs with inversed phases. Amounts progressing a threshold of one emphasize the oscillation an octave higher.

2.4.3 Inharmonicity

This concept transforms the frequency of partials based on its corresponding partial number, with a constant variable δ in which the ratio of partials is extended or contracted through the whole spectrum, exponentially. It is linked to the auditory descriptor that is Inharmonicity, and, consequentially, raises concerns regarding the phase positions of harmonics are raised.

$$F_k = F_{k-1} + F_1 \delta^k \tag{2.2}$$

Symbol F as frequency of partial and k as order of partial.

- $\delta > 1$ generates stretched partials
- $\delta < 1$ in compressed partials
- $\delta = 1$ is a perfect harmonic sound

Several instruments feature large amounts of distinctly inharmonic partials, and no computational limits for partial detuning have been found. Instruments like the piano are distinguished by the detuned nature of the partials coming from the struck strings.

2.4.4 Distortion

Signal clipping of the output at an established amplitude in order to generate further partials. Although not directly associated with an audio descriptor, it can influence multiple features inside a spectral flux, spectral centroid, spectral flatness, inharmonicity, or spectral slope. Hence, it is a

natural outcome of any physical system without limitless bandwidth. Henceforward the increasing interest in the subject for the auditory system [55].

After a normalization process of the source signal, a constant variable γ is multiplied, and consequently, every value greater than one is set to one, and every signal smaller than –1 is set to –1.

Dependent on the input signal, it can accentuate the signal's properties, making the sound brighter and raise the amplitude of odd harmonics. If the signal presents a common inharmonic partial situation, static signals can become dynamic with the output beating between the original sound's inharmonic partials and the distortion's new partials. Dobrowohl additionally considered the concept of plucking - a comb-filter effect [32].

Conclusions for the research allowed a better understanding of how one sawtooth waveshape can implicate different outputs and user likings. To calculate deviations between the influence of the effects and the user desires, a Perception Threshold value is assigned to each effect. The highest value for timbral change was achieved within the Inharmonicity effect compared with the other manipulations' capabilities.

2.5 Compatibility of Formal Structure

The advancement and accessibility of technology and audio editing techniques have minimized entry barriers for mashup production. Music sequencers constituent of loop-based systems, such as Ableton Live, for instance, lead the user to suit the rhythms and push the audio sampling keys. As more than just a result, despite structured musical training, systems producing mashups are now most frequently produced by music lovers. Users will need to focus on their own backgrounds and musical learning for the above resources to find suitable music clips to be blended together. If the amount of streaming music already available rapidly increases, ill become time-consuming and labour intensive to obtain adequate clips.

Automatic schemes have been suggested by several prior studies to build mashups. Noted, these approaches [26, 27] concentrated only on the vertical appropriateness of the selected music segment, because they always deemed the adequacy of the music sections to be layered, which was described as the word 'mashability'. AutoMashupper [26, 27] seems to pertain as the initial research that included a detailed calculation analysis to identify suitable music fragments and an automated mashup generation schematic to be layered together. In order to produce the final mashup, samples from other tracks with both the maximum mashability upon its grounds of chromagram compatibility, rhythmic compatibility, and spectral balance, have been overlaid with every section mostly with corresponding segment throughout the main song. This architecture is however supported by future work [25], which live input recording is known as the base track in Davies [25], and the accompanying music parts are overlapped on the input signal. Tsuzuki et al. [93] concentrated on allowing users to overlap vocals from various performers all along the same accompanied music.

Chuan-Lung Lee et al. [54] implemented a mechanism developed to produce mashup compositions dynamically by taking into account both the vertically and horizontally mashabilities. Two new variables are also regarded and explored inside the vertical mashability of "harmonic change control" and "volume weighting". Subjective assessments reveals that the auditory enjoyment of the generated mashups would be increased relative to that of the existing equivalents created in Davies [27]by keeping many such variables into consideration. In addition, pleasantness can gain varying degrees of hearing satisfaction in mashups by combining with the horizontal mashability. As a direct consequence, the first background track listeners choose to derive from and some necessary structure metrics (such as the amount of background sections N and the quantity of lead segments per background division M) given a collection of multitrack songs with formal component names, the framework can then produce an automatic mashup. Investigations into these dual dimensions of Nature in Music, as Figure 2.10 denotes, assumes that multitrack songs can include at least two types of background and leads, in the sorted manner. This presumption is rational since multitrack songs can effectively be retrieved via mashup websites[5].

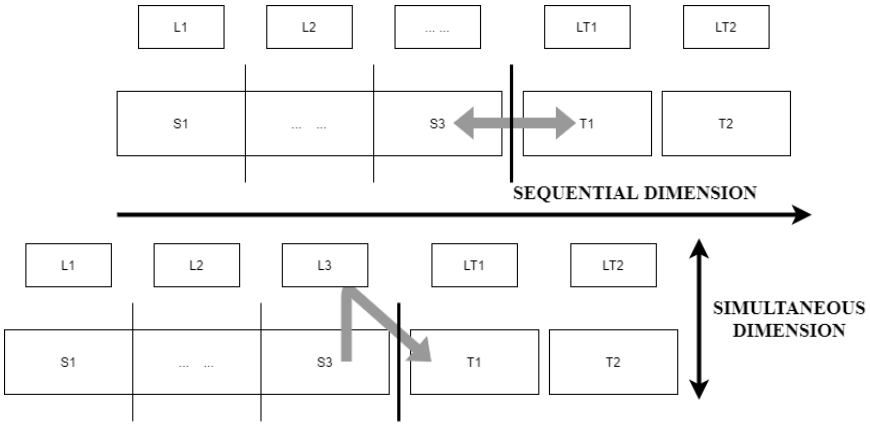

Figure 2.10: Horizontal and Vertical Multidimensions of Musical Space for Mashup Creation [54])

Initially, the framework developed by Lee would collect audio information in the preprocessing step and pre-compute vertically and horizontally mashability for each potential combination of units in the music package. Thus, as per user-specified structure variables such as the amount of leads per context segment M, it specifies which auditory section should be placed in the resulting mashup arrangement and from where. Additionally, how certain the fragments are converted to create the subsequent mashup generation. The two phases, vertical and horizontal mashup processes, would be done through iterative manners until the necessary duration of the consumer is met by the resulting mashup. Firstly, the mashup generation stage will adjust the speed, loudness, and pitch of each unit to the appropriate values. Next, to create the final mashup album, the units would be matched and concatenated.

[5]Mixter - http://ccmixter.org

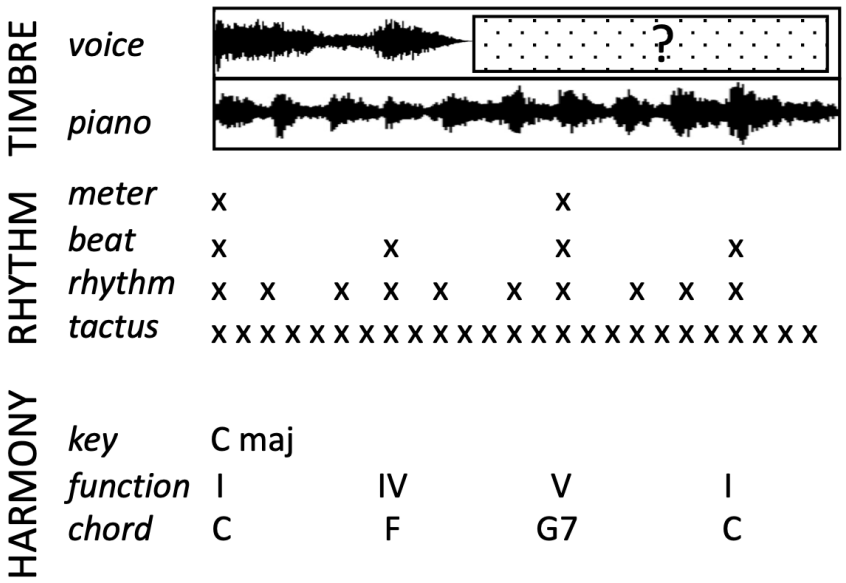

Figure 2.11: Overview of the multiple attributes and hierarchies considered in musical audio compatibility underlying the composition process.

In producing mashups, it gives much greater versatility than just the structures suggested in previous experiments, as it takes into consideration the recently enacted vertical mashability metrics. A case study test indicates that there has been a statistically meaningful increase in user retention.

2.5.1 What Does Compatibility Entail?

Fig. 2.11 shows a typical computer-aided music production search problem from large musical audio archives. It aims at finding a vocal musical audio sample that 'fits' the existing project context, i.e., a previous vocal track excerpt and an accompaniment piano track. First and foremost, the journey typically commences by inspecting large-scale 'meta-data' attributes, such as the duration, key, style, or instrumentation, to convey some unity across the musical structure at higher-level hierarchies. One aims to find musical audio samples that belong to the same category (e.g., to belong to the key of C major and the binary meter of Fig. 2.11). However, finding samples that match high-level hierarchies does not guarantee a good 'fit' or compatibility with the existing context. For example, within binary meters, one can have a myriad of inner accents (e.g., 6/8 or 2/4 meters), whose manifestation can be inspected at lower structural levels.

Musical audio compatibility, with a tool-set of representations and metrics, may be of use here to account for the multiple hierarchical levels of interaction across musical audio tracks. Metrics must adhere to multi-objective and multi-hierarchy constraint optimization levels, which can identify and distinguish, for example, varying degrees of syncopation at the low-level tactus grid, while equally fitting the high temporal metrical organization. Furthermore, cognitive aware models and methods ought to leverage varying degrees of novelty and familiarity to a given context, understood here as both the specific audio content from a particular work and the social-cultural

and subjective dimensions involved in the composition process. The latter can be made explicit as constraints defined manually by the user in a human-in-the-loop approach [99].

Harmonic compatibility must entail representations and metrics for large-scale compatibility at the level of the tonal plan or key changes of the musical work. Mid-scale considerations of harmonic function or tension (namely in the context of Western tonal music) ought to promote a structural *intent* of the harmonic sequences. At the small-scale greater inspection of dissonance concepts and pitch affinity at the level of the musical surface. For example, lets consider the compatibility between the chord of C major and two other candidate chords of A Maj and A min. Roughly, their combination presents only slightly different degrees of dissonance; however, A Maj may not 'fit' the large-scale organization of the musical structure composition as A min, considering a key of C major.

Rhythmic compatibility ought to equally consider the large- and mid-scale accents across the phrase structures. Downbeat alignment plays a vital role in detecting larger structural organizations and the saliency of the beat or pulse structures within the metrical grid inspects the mid-level compatibility. A relevant example of the latter mid-level compatibility is the distinction between binary and ternary organization within equal metrical duration (e.g., distinguishing 3/4 and 6/8 meters). Finally, lower level or rhythmic manifestation, micro-timing deviations below the beat level are important interactions to consider to accommodate the same rhythmic 'feeling' manifestation and avoid, for example, straight and swing feeling clashes, which can cause disruptive rhythmic dissonance effects.

Finally, timbre compatibility is the most elusive attribute in grasping the multi-dimensional perceptual qualities, namely at the instrumentation level and the spectral attributes the multiple musical audios occupy. At the large-scale structure, coherent timbral content per musical layer is typically pursued. At the small scale, a balanced interaction across the spectral energy space each layer occupies can be accounted for, which can equally account for the linear progression of individual tracks, an important dimension for the harmonic and melodic articulation of voice leading.

2.5.2 Prototypical Formal Structure of Musical forms

As musical genres increasingly get more complex, representations of both notation and musical form must remain coherent in the structural network, as analysed in a study conducting identification of four categories of musical forms most known to the nature that is music. These are also represented with capital lettering which affirm the segments corresponding to specific musical structures.

- Strophic: characterized as AA, AAA or AA' comprising one main theme, repeated with or without minor alterations.

- Binary: consisted as AB or ABAB with two main alternating tunes which ends during the second theme.

- Ternary: represented as ABC, or ABA'. While having three main tunes, or two main themes plus an initial theme re-insertion presenting a possibility of slight variations.
- Rondo: represented as ABACA, ABACABA or ABACADAEA, it is also characterized as a tune that is recurring and alternated with others contrasting themes.

Strophic form [78, 22] is amongst the most prominent forms of music, often related to as verse, more often used in popular music, folk tunes, or songs which are verse-based. Due to its repetitiveness, it is considered highly simple in regards to other remaining types, and is normally designed with a AAA structure typically with a length of 8 to 16 measures. Most misconceptions fell on two dichotomies of strophic/round confusion and binary/ternary ambiguity. As observed in Wu [98] and Figure 2.12, modifications of the repetition theme in (a) result in visualization differences that can be easily mistaken with the depiction of the alternating theme, like case (b). Appendix A.1 shows a brief list of common musical forms integrated in musical space and formal structuring.

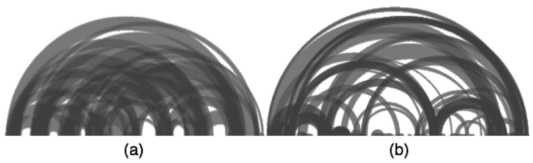

Figure 2.12: Visualization comparison of respectively (a) Strophic form and (b) Rondo form [98])

2.6 Summary

Generative architectures surrounding the loop-based methodologies and aid of musical composition, like Mixmash or Automashupper, can be limited in terms of product scaling and development, which is the reason of suggesting generative applications based on compatibility metrics of the environment of Music. Although Neural Loop Combiner seems to be a close approach to optimized searches in musical space, the dataset available is quite scarse and would affirm long terms of producing specific information for that domain. At this moment, algorithmic perspectives for search optimization seem to point that AIS implement an optimized architecture that allows working through descriptive spaces in large loop scales without storing unnecessary data.

Chapters before this segment have described a literature review on generative music systems based on methodologies of audio-based loop recombinations, whilst describing a musicological environment of attributes defining musical quality and stylistic diversity. These metrics, as explained, can be sectioned into four layers of audio compatibility for Harmony, Rhythm, Timbre, and Formal (multimodal) Structure. Each component presents their specific computational domain capable of, with the help of latest technologies for Creative-MIR, achieving limitless signal analysis with heavy loads of auditory data information coming from datasets annotating the necessary attributes.

Chapter 3

CoDi

In this chapter, for Section 3.1, we address the implementation of CoDi – found in GitHub repository accessed in `shorturl.at/cCFHV` with all code and results – following a multimodal architecture for an effective search space for diversity and compatibility. Section 3.2 follows an agenda taking the procedure that was developing the feature extraction of the musicological metrics within the CoDi model and how that generates a feature dataset for evaluation. Section 3.3 mentions the properties within the computed evaluation functions assessing the audio files for the vertical mashup generation. Section 3.4 denotes the computed evaluation functions assessing the audio files for the horizontal mashup generation. Section 3.5 denotes the computational explaining of the developed opt-AiNet algorithm within CoDi, upbringing the optimized search space mentioned in chapters above. Finally, Section 3.6 covers representation when executing CoDi and the process enveloped in its implementation, and Section 3.7 mentions a summarized procedure of the chapter.

3.1 Overview

In CoDi, we adopt the opt-aiNet to promote the computational efficient search for musical mashups resulting from the recombination of musical audio loops, [1] understood in the opt-aiNet as network cells. Mashups result from the combination loops $l_i \in S$, where $i = \{1,...,L\}$ is the index of the

[1] An audio loop is a short segment of audio, e.g., a measure of a drum beat, which is created to be repeated over time [37].

loop in the dataset S, with a total number of L musical loops. A mashup p is then a combination of musical loops.

Each loop l_i is represented in a feature space, where distances equate to their compatibility. The smaller the distance, the greater their compatibility. Optimal compatible mashups result from minimizing an evaluation function E_p in the feature space. Diversity is guaranteed by pursuing a combination of local and global search in exploiting the feature space, resulting from iterative clonal mutation and selection of mashup candidates in the search space (see Fig. 3.1).

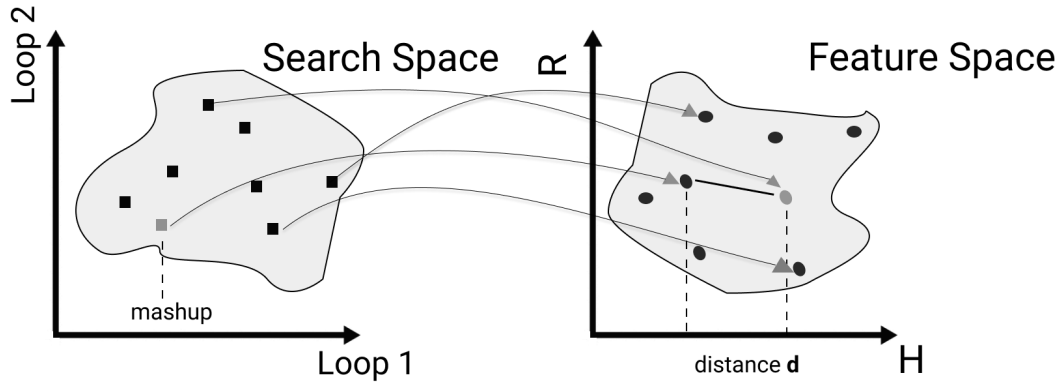

Figure 3.1: Representation of both search space (left) and correlated feature space (right) taking harmonic and rhythmic measurements within the loop dataset.

Fig 3.2 shows the architecture of CoDi. A user-defined dataset S with musical audio loops l_i is the collection of musical audio adopted in the mashups. Feature extraction algorithm define harmonic $T(k)$ and rhythmic $r(b)$ representations for each dataset loop l_i, which are stored into a feature dataset. The AIS opt-aiNet is then adopted to search for multiple compatible and diverse mashups by evolving a random initial population. Finally, a set of optimal mashups result from overlaying the mashup component loops.

3.2 Feature Extraction and Dataset

The feature extraction module is responsible for creating two vector representations that capture the harmonic, rhythmic and timbral content of each musical audio loop l_i.

We adopt TIV, $T(k)$, as a representation for the harmonic content of an audio loop. $T(k)$ is a 12-dimensional vector computed as the DFT of a chroma vector $c(m)$, such that:

$$\begin{aligned} T(k) &= w_a(k) \sum_{m=0}^{M-1} \bar{c}(m) e^{\frac{-j2\pi km}{M}} \quad , \\ k \in Z \quad &\text{with } \bar{c}(m) = \frac{c(m)}{\sum_{m=0}^{M-1} c(m)} \quad , \end{aligned} \tag{3.1}$$

where $M = 12$ is the dimension of the input vector and $w_a(k) = 3, 8, 11.5, 15, 14.5, 7.5$ are weights derived from empirical ratings of dyads consonance used to adjust the contribution of each

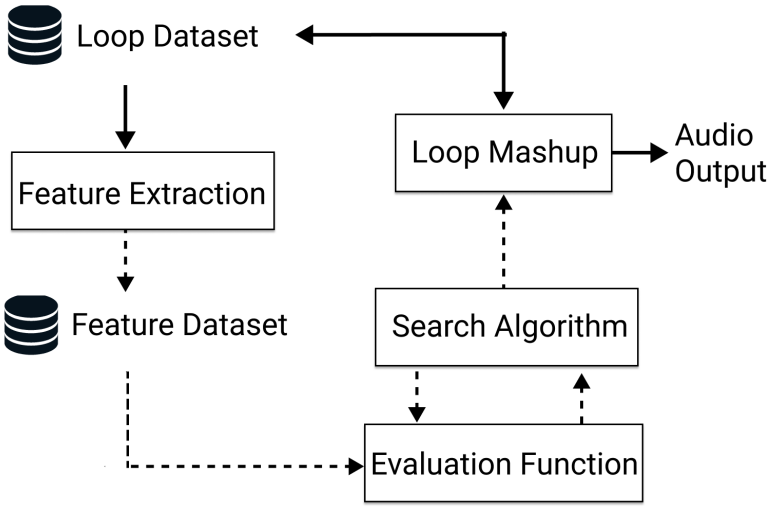

Figure 3.2: Multiple modules considered in CoDi underlying our AIS computational composition process. Rectangular blocks are processing functions. Solid and dashed arrows denote audio or control flow of information between processing modules, respectively.

dimension k of the DFT space. We set k to $1 \leq k \leq 6$ for $T(k)$, since the remaining coefficients are symmetric. $T(k)$ uses $\bar{c}(m)$ which is $c(m)$ normalised by the DC component to allow the representation and comparison of different hierarchical levels of tonal pitch.

Following [10, 11], we adopt the $T(k)$ space to compute the harmonic compatibility H between two given loops l_1 and l_2 using Eq. 3.2, which combines the dissonance D and perceptual distance P metrics shown in Eq. 3.3 and 3.4, respectively. The lower the values of H, the higher the degree of harmonic compatibility between two audio loops l_i. [74] has shown that the harmonic compatibility H indicator perceptually captures human judgments of pleasantness to a higher degree than remaining harmonic compatibility metrics.

$$H_{l_1,l_2} = D_{l_1,l_2} \cdot P_{l_1,l_2} \tag{3.2}$$

$$D_{l_1,l_2} = 1 - \frac{a_1 T_1(k) + a_2 T_2(k)}{a_1 + a_2 w_a(k)} \tag{3.3}$$

where a_1 and a_2 are the amplitudes of $T_1(k)$ and $T_2(k)$, respectively.

$$P_{l_1,l_2} = \sqrt{\sum_{k=1}^{6} |T_1(k) - T_2(k)|^2} \tag{3.4}$$

A rhythmic histogram $r(b)$ [56], where $b = 60$ bins, is adopted to represent the rhythmic content of a musical loop as amplitude modulations. The representation derives from rhythmic patterns [68], a matrix representation of fluctuations in different frequencies on critical bands of the human's listening range. Their fundamental difference is that rhythmic histogram $r(b)$ accumulates all frequency bands onto a single bin, resulting in a vector of 60 frequency modulation

bins in the [0,600] BPM range [56]. The motivation to adopt rhythmic histograms $r(b)$ instead of the most common rhythmic patterns representation is to minimize pitch or spectral differences in the compatibility computation, namely in light of the typical single-instrument nature of musical loops used in production settings [68, 67].

To compute a rhythmic histogram $r(b)$, we adopt a two-stage extraction process. First, we group the frequency bands by loudness sensation, using a short-time Fourier transform. The resulting spectral representation is then converted into a time-invariant 24 critical Bark bands modulation frequency spectrum by applying a second Fourier transform. High amplitudes values in the rhythmic histogram $r(b)$ denote a recurrent period in the musical audio. For example, Figure 3.3 shows the rhythmic histogram with four predominant peaks at multiples of the tempo.

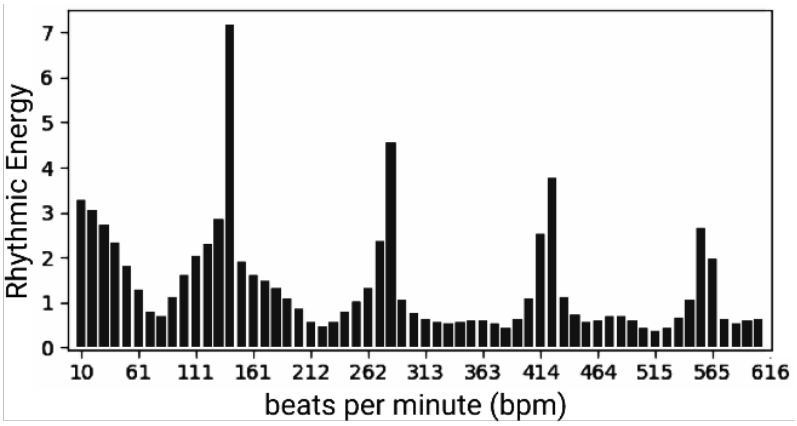

Figure 3.3: Rhythmic periodicity function for an audio loop including leads and brass within rhythmic pulse.

The distance between rhythmic histograms $r_1(b)$ and $r_2(b)$ from two musical loops l_1 and l_2 is computed as their angular distance, such that:

$$R_{l_1,l_2} = \arccos \frac{r_1(b) \cdot r_2(b)}{|r_1(b)||r_2(b)|} \quad . \tag{3.5}$$

The MFCC is based upon the raw Cepstrum mathematical equation and leverages a filtering process into the magnitude spectrum through sets of overlapping filters in triangular shape, based upon the mel scale as Fig 3.4

The magnitude spectrum is consequently filtered, transformed onto the logarithmic scale, and further modified to Mel-frequency Cepstral Coefficients when going through a Discrete Cosine Transform. Equation 3.6 raises the general form for transforming frequency onto the mel scale, and the distance of two MFCCS $mfcc_1$ and $mfcc_2$, from two musical loops l_1 and l_2, is computed through an angular distance in Equation 3.7, such that:

$$mfcc_{l_1} = 2595 * log_{10}(1 + \frac{frequency_{l1}}{700}) \quad , \tag{3.6}$$

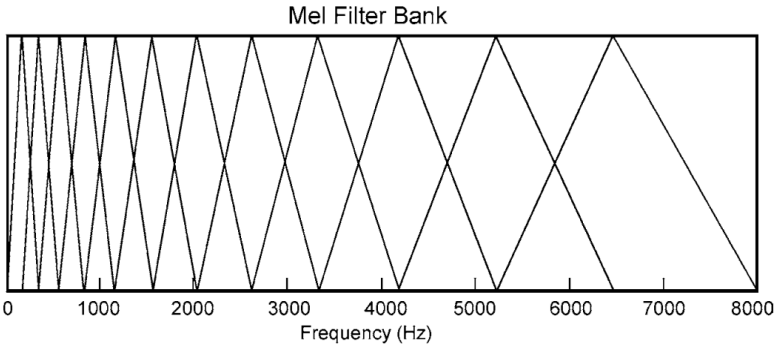

Figure 3.4: Representation of mel scale.

$$M_{l_1,l_2} = \arccos \frac{mfcc_1 \cdot mfcc_2}{|mfcc_1||mfcc_2|} \quad . \quad , \tag{3.7}$$

To ensure a human-in-the-loop approach within CoDi, we assigned weights to each of the featured metrics within the extraction process, as w in ranges of [0,1] so that users can infer between more – or less – harmonic compatibility ($w_H = 0.45$), rhythmic compatibility ($w_R = 0.45$), spectral balance ($w_F = 0.05$), or continuous generation with MFCC ($w_M = 0.05$).

3.3 Vertical Evaluation Function

To withhold compatibility in a vertical dimension within a diverse space of solutions, AIS opt-aiNet assesses a population of musical loop mashups p at each iteration. The population evolves across multiple iterations by minimizing an evaluation function E_p. To compute an objective evaluation value per mashup p, we first define H_p and R_p as the sum of all pairwise distances across the component loops l_i in the mashup p. A total of $u(u-1)/2$ pairwise values per harmonic and rhythmic representation are summed using Eq. 3.2 and 3.5, respectively, where u is the total number of loops l_i in a mashup p. Then, we apply Eq. 3.8 to combine the resulting harmonic H_p and rhythmic R_p compatibility metrics linearly. Furthermore, a high penalty is applied to mashups p that include repeating loops l_i, such that:

$$E_p = w_H H_p + w_R R_p + w_F F_p \quad , \tag{3.8}$$

where $F_p = 0$ if no l_i duplicate loops are found in p and $F_p = 0.5$, otherwise.

3.4 Horizontal Evaluation Function

In order to sustain for compatibility and continuity in a horizontal dimension within a diverse space of solutions, a new evaluation function assesses the same population of musical loop mashups p at each iteration. Since we are dealing with more than one run of the model CoDi algorithm (not the number of iterations within the search algorithm), the function must search for the population

of selected mashups in the previous population, or iteration, found in $p-1$. Just as the vertical dimension, we entail for minimizing an evaluation function E_p. To compute this new objective evaluation for continuous generation, we retain the mashup p, as well as the computed H_p, R_p, and F_p as the sum of all pairwise distances across the component loops l_i in the mashup p. After retaining that information, we measure the distances of the current population, and the previous solutions found in $p-1$ – raising new metrics of comparison such as $H_{p,p-1}$, $R_{p,p-1}$, $F_{p,p-1}$, and to assure timbral quality and avoid auditive discrepancies in the continuation, we address $M_{p,p-1}$

We apply Eq. 3.9 to combine the resulting harmonic – H_p and H_{p-1} – with rhythmic – R_p and R_{p-1} – as compatibility metrics of the entire population of selected mashups for generation, gathered in $H_{p,p-1}$, $R_{p,p-1}$, $F_{p,p-1}$, and $M_{p,p-1}$ with the addition of MFCC metrics – M_p and M_{p-1} – as a balanced measure of auditive continuity comparing the current and previous population. The same high penalty is applied to mashups p and that include repeating loops l_i, such that:

$$E_{p,p-1} = w_H H_p + w_R R_p + w_F F_p + w_H H_{p,p-1} + w_R R_{p,p-1} + w_F F_{p,p-1} + w_M M_{p,p-1} \quad , \qquad (3.9)$$

where $F_p = 0$ if no p duplicate loops are found in p and $F_p = 0.5$, otherwise.

The following Figure 3.5 makes a representation of the different metrics within the feature extraction for a non-context generation – or one run of the CoDi model – and context generation – with more than one run of the CoDi model. We entail that the difference between these two dimensions of non-context and context generation is the use of MFCC coefficients comparing the current and the previous population.

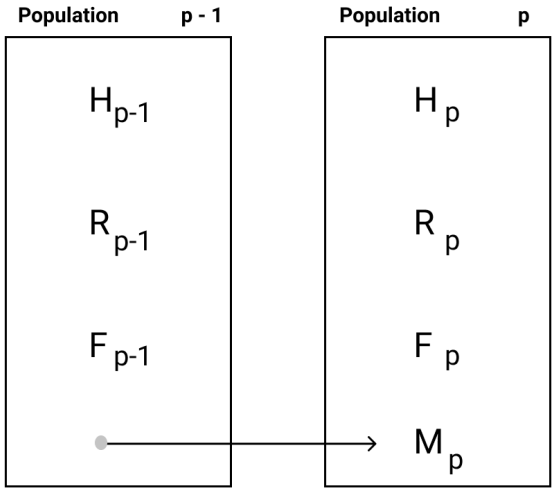

Figure 3.5: Feature Extraction representation for non-context and context generation. Metrics of harmonic, rhythmic and timbral compatibility are addressed for both dimensions.

3.5 Searching Algorithm

The immunological operations in AIS opt-aiNet – cloning, mutation, and affinity suppression – evolve an initial random population towards compatible and diverse mashups in the immune network. Maintenance of compatibility is assured by the evaluation function E_p and leveraged by cloning and mutation operators, which optimize the population of mashup candidates across multiple regions. Valleys (or local minima) in the multimodal search space indicate optimal mashup candidates p. Fig 3.6 shows a flowchart diagram of the AIS opt-aiNet algorithm used in our work.

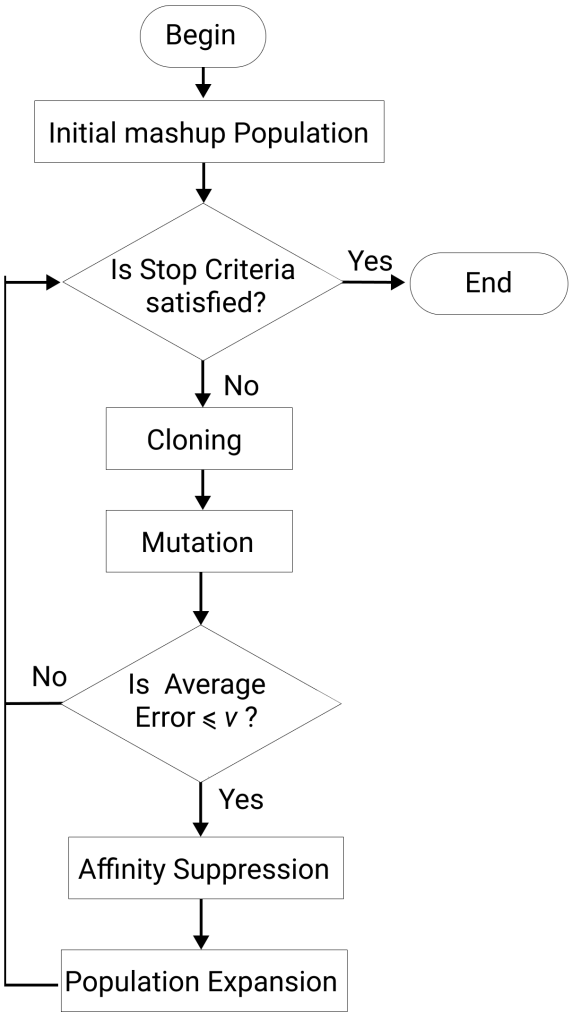

Figure 3.6: Opt-aiNet flowchart diagram.

The AIS opt-aiNet algorithm starts by instantiating a random population of mashups p. The initial number of mashups in the population (i.e., population size or the number of network cells) is not relevant. The algorithm includes mechanics for the automatic adjustment of the population size via affinity suppression and population expansion. Cloning is responsible for creating a number N_c of offspring cells per mashup in the population (or network cell) that are identical copies of their parent cell. Each clone includes the parent and its N_c offspring. The offspring clones undergo an

operation of somatic mutation to become variations of its parent. In other words, mutation asserts if a given loop l_i in a mashup p is changed. A probability of a given loop l_i within a mashup p to be mutated is inversely proportional to the mashup p evaluation value. Following [17], we adopt Eq. 3.10 represented in Figure 3.7 to define the mutation probability of a given loop l_i within a mashup p.

$$\chi = \exp(-\gamma \hat{E}) \quad , \tag{3.10}$$

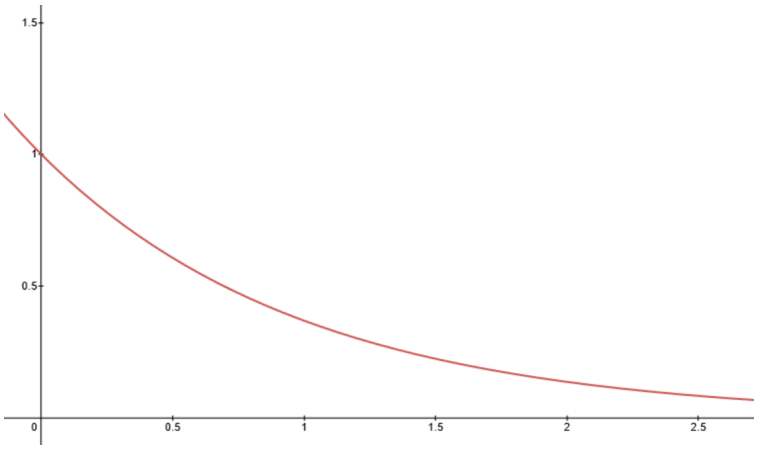

Figure 3.7: Function of mutation probability given in a loop within the population.

where $\gamma = 1.2$ is a constant and $\hat{E}$ is the normalized evaluation value to the [0,1] range from the corresponding mashup p (or to the cell undergoing mutation). For each of the loops l_i in the mashup p, a random decimal value in the [0,1] range will determine its mutation onto a different audio loop index i. If the decimal value is $\leq \chi$, another loop from the dataset S is randomly fetched, or, in other words, its index i is replaced by a random index number in the $[1, L]$ range.

A clonal selection performs an elitist optimization of the population to retain the best-ranked mashups per clone. To this end, all clone mashups are evaluated using Eq. ,3.8 and the mashup with the smallest evaluation value E_p per clone is retained in the population. Then, the population's average fitness is computed to assess if the local region optimization of the mashups has stabilized. To compute the stabilization of the population, we compute the average error of the population evaluation using Eq. 3.11, which computes the modulo of the ratio between the average evaluation values of the previous iteration to the current iteration subtracted from unity. If the average error is $\leq v$ where $v = 0.001$, the population is said to have stabilized, and the algorithm continues to affinity suppression and population expansion. If the condition does not hold, a new iteration with a clonal selection of the population is performed.

$$\text{Average Error} = \left| 1 - \frac{\text{Average of Old Evaluation}}{\text{Average of Evaluation}} \right|, \tag{3.11}$$

To maintain diversity, AIS opt-aiNet adopts the suppression operator to excludes mashups p with high affinity, or below a given distance threshold t in the feature space. Pairwise mashup

distances in the feature space are computed as the angular distance A of the weighted combination of harmonic $T(k)$ vectors and the linear combination of rhythmic $r(b)$ vectors using Eq. 3.12 and 3.13, respectively. Prior to the angular distance computation each mashup p is represented by the concatenated $(T_c(k), r_p(b))$ vector.

$$T_p(k) = \sum_{i=1} a_{l_i} T_{l_i}(k) \tag{3.12}$$

$$r_p(b) = \sum_{i=1} r_{l_i}(b) \tag{3.13}$$

Hence, suppression excludes mashups p (candidate mashups) within a given radial range, retaining the local optima mashups that minimize the value of E_p in Eq. 3.8 in multiple regions of the search space. By excluding similar mashups p from the immune network, we ensure diversity in the population. The remaining mashups in the immune network after suppression are referred to as memory cells.

The AIS opt-aiNet includes two stopping criteria conditions. Whenever one condition is met, the iterative method is stopped, and the population is output. The system's output is the total number of mashups in the population ranked by their evaluation E_p value in ascending order, i.e., from the best to worse ranked local optima. The stopping criteria include the user-defined maximum number of interactions u or once the number of memory cells in a population has stabilized over two consecutive iterations. If the number of memory cells does not stabilize, a percentage $d = 40\%$ of random network cells is appended to the population to expand the immune network capacity to explore the space further.

Finally, once the algorithm outputs a population of local optima mashups, each mashup p is synthesized by overlapping its component audio loops l_i, retrieved from the loop dataset S given their index i.

3.6 Representation and Execution

Before running the model, users must gather two datasets of harmonic and rhythmic loops. When running CoDi, the user specify the number of runs of the model, as this number will determine if the mashup generation comes with context – values higher than 1 – or not – value of 1.

After the search algorithm finishes, the solution dataset is presented to the user while asking if the content should be played, or not. If the user accepts playing the entire run of the model, auditory information of the entire mashup is represented – such as the name and ID of each loop gathered proportionalyl to the number of layer (in our case, two of harmonic content and one of rhythmic content) – while it is playing. When finished, or when the user decides to not play the mashup content, users must choose the index of the mashup of which they desire whether for compositional purposes, or pure enjoyment of *pleasantness*. Figure 3.8 represents the three mentioned phases of CoDi, from top to bottom.

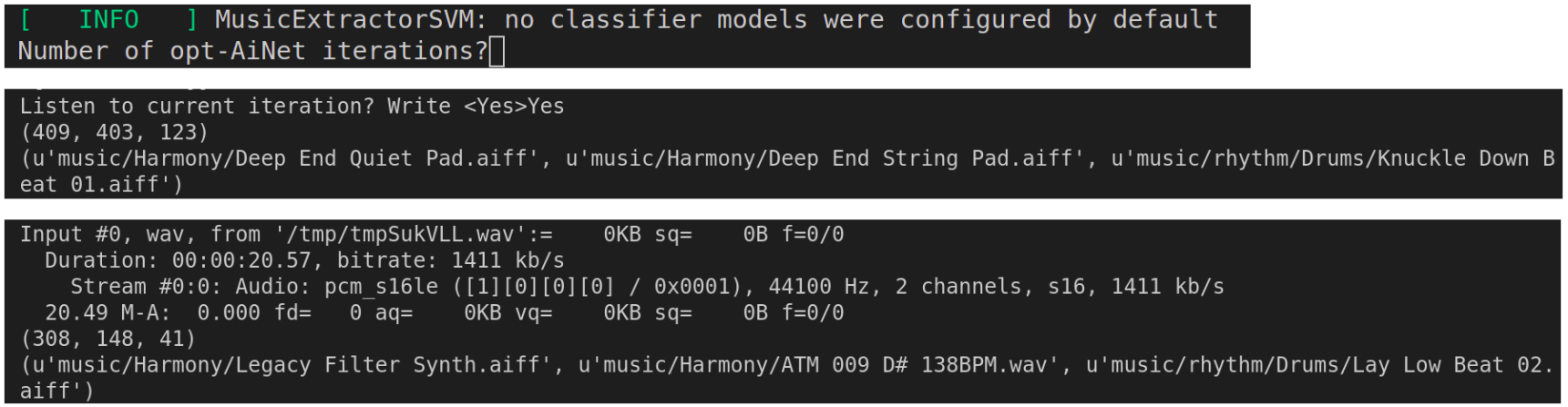

Figure 3.8: Three distinct phases of running CoDi. Top figure shows the initial question for initial number of iteration. Middle figure denotes the user accepting the output of the generated mashups, as the model shows the path of each selected loop for the generation. Lowest figure represents auditory information of the mashup playing.

3.7 Summary

In this chapter, we propose CoDi, adopting the opt-aiNet to promote the computational efficient search for musical mashups resulting from the recombination of musical audio loops. This model follows a multimodal architecture for an effective search space for diversity and compatibility with a procedure developing a feature extraction from musicological metrics within the CoDi model, and how that generates a feature dataset for evaluation.

We develop evaluation functions regarding music generation for vertical dimension – or non-context generation – and horizontal dimension – or context generation. CoDi's computational explaining of the developed opt-AiNet algorithm is additionally leveraged, upbringing the optimized search space.

Chapter 4

Evaluation

The current chapter is based on the research paper [12] which the author co-authored. The contents used are solely from the author.

As shown in previous studies [27, 39], compatibility between musical loops in a mashup is fundamental to user enjoyment. However, diversity in mashup creation is equally important in promoting multiple solutions from which users can select, taking into account their personal preferences. Therefore, an application for assisting users in mashup creation should provide multiple and perceptually different solutions. Section 4.1 mentions an initial developed phase of the dissertation where three models based on AIS, GA, and BF-based approaches searching for compatibility, diversity, and an efficient computational performance.

In this context, we adopted an intermediate evaluation procedure, with objective measures to evaluate the 1) compatibility, 2) diversity, and 3) computational performance of the developed GA and BF models compared to an initial developed AIS model in Pure Data, miXmash-AIS, with two layers of auditive content. All systems use the same feature space, which results from the combined harmonic H and rhythmic R representations. More importantly, the same evaluation function in Eq. 3.8 to assess the compatibility of a mashup E_p. Similar to miXmash-AIS, CoDi returns mashups ordered by fitness value.

A dataset of 171 drum intrumental drum loops were added to a dataset of 551 hip-hop instrumental loops from Apple Loops which are commonly distributed with proprietary Apple Digital Audio Workstation software, such as Garageband and Logic [1] was adopted as our dataset S. Included loops range between five to 24 seconds and feature diverse tempo (or bpm) and multiple

[1] https://support.apple.com/guide/logicpro/apple-loops-in-logic-pro-lgcp734a05f6/mac, last access on 10 May 2021.

instruments within a large set of spectral regions, roughly in the [40,10000] Hz range. We have defined CoDi's AIS parameters to withstand an *initial population* of 30 cells and a value of 300 *max iterations*. The number of *clone generations* – for each network cell – is set to 15, and respectively with an *affinity threshold* of $t = 0.1$, as mentioned in Chapter 3.

4.1 Implemented Models for Objective Evaluation

Following the development of a research paper regarding the best choice of algorithms for an optimized musical space, a development of three distinct algorithmic models were developed in Pure Data, as mentioned. These models are based on the opt-AiNet AIS algorithm, as well as a GA approach, and a BF-based approach. These models were algorithmic comparisons with the same evaluation function analysing harmonic and rhythmic compatibility and diversity within a large dataset of loops.

The motivation to pursue such a model is to:

- Tackle current scalability limitations in state-of-the-art (brute force) models.
- Enforce compatibility, i.e., recombination quality, of audio loops
- Create a pool of diverse solutions that can accommodate personal user preferences or promote different musical styles.

A proposition was made while developoing miXmash-AIS, a multimodal music mashup optimization system for loop recombination at scale. It adopts the AIS opt-aiNet algorithm to leverage compatible and diverse mashups while addressing the scalability issues in existing state-of-the-art BF solutions for computational music mashup creation. In promoting a diverse set of optimal mashups, the system can account for personal preferences and different stylistic traits. The conducted evaluation compared the proposed system to a standard Genetic Algorithm (GA) and a Brute Force (BF) approach. Figure 4.1 shows a visual representation of the Pure Data environment of miXmash-AIS.

An objective comparison of AIS opt-aiNet to a standard GA and BF approaches in the task under study denotes the primacy of the AIS opt-aiNet in finding local and global optimal mashups, closely matching the compatibility values of the BF approach.

4.2 Quantitative Evaluation - Intermediate Experiment

4.2.1 Evaluating Compatibility

To objectively assess and compare the compatibility in all models under evaluation, we average the evaluation function values E_p of the 10 best-ranked mashups, thus providing an average indicator of the model compatibility. The smaller the average compatibility value, the better it complies with

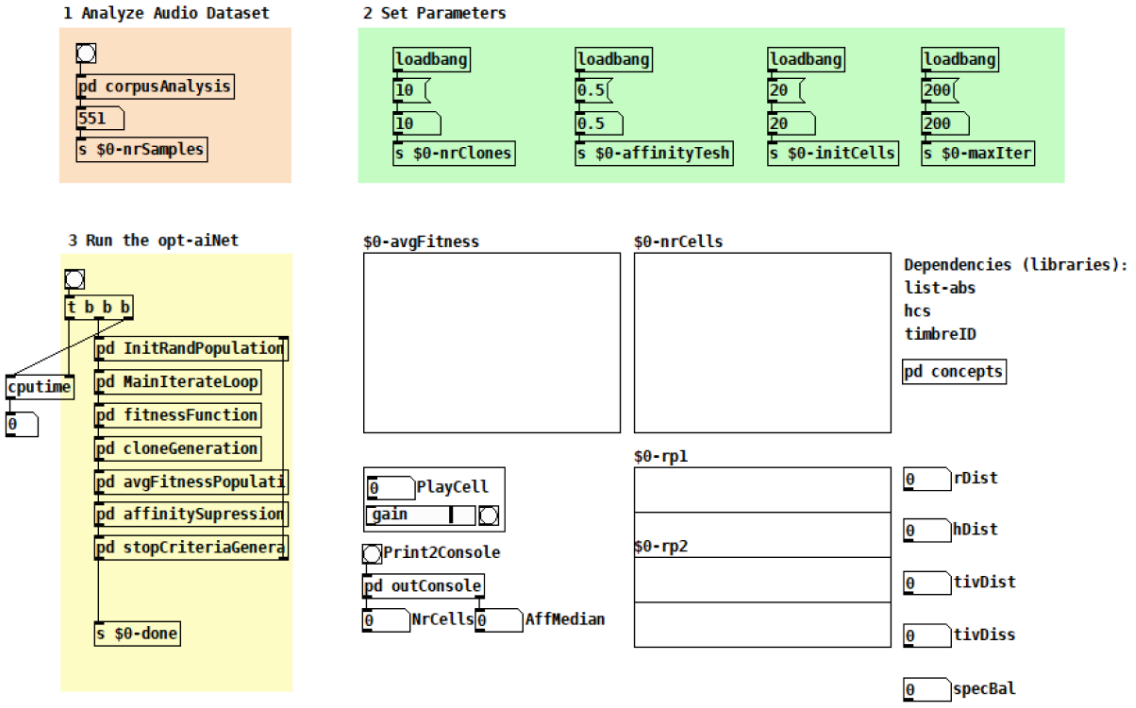

Figure 4.1: Pure Data environment of the developed miXmash-AIS model

the objective criteria we aim to minimize, i.e., harmonic H and rhythmic R compatibility, and no repeating (overlapped) loops F. Furthermore, in the AIS models, we ran the algorithms 10 times to capture the diversion in optimization convergence mechanics across multiple runs. Diversion is the optimization is expected to be more clearly noticeable in the results of the GA. The AIS opt-aiNet algorithm also does not guarantee similar results with each run. However, the affinity suppression and population expansion in AIS opt-aiNet algorithm minimize this behavior.

4.2.2 Evaluating Diversity

To objectively measure diversity in the three models under evaluation, we inspect distance relations across mashups in their feature space, i.e., within the same space as the affinity is calculated. Mashup locations in the feature space reflect perceptual relations amongst the regions in which they are found. Therefore, it is possible to associate diversity in the feature space with the conception of diversity along the perceptual dimensions, correlated with adopted harmonic ad rhythmic features. We propose using the average distance across all unique pairwise mashups from the 10 best-ranked evaluation set in each model. A total of 45 distance values are averaged per model.

4.2.3 Evaluating Performance

The computational performance of the models is instrumental to the task of computational mashup creation at scale due to the combinatorial explosion of the loop recombination number. Depending on the size of the loop dataset L, defined *a priori* by the user by selecting the dataset, it affects the

complexity of the problem – e.g. a BF approach in the dataset under consideration, which includes 551 loops, results in 151525 unique combinations for mashups with two overlapping loops.

The associated computational runtime cost of each iteration for AIS under evaluation can be defined as $\mathcal{O}(LV)$, where L is the current population size, and V is the length of the combined rhythmic and harmonic representation vectors. The affinity suppression in AIS has an additional computational cost of $\mathcal{O}(L2V)$. The BF approach does not feature multiple iterations, and its computational cost can be defined as $\mathcal{O}(L^2V)$. The population in CoDi's and miXmash-AIS algorithm is defined by the user (previous systems [17, 65] adopt values in the [10,30] range), whereas both systems stand on value. These costs indicate greater computational gains when adopting an approach with GA than AIS, whose affinity suppression adds complexity.

On the research conducting the AIS, GA, and BF versions of miXmash, the AIS and GA suggested substantial gains compared to the BF. However, the former algorithms are dependent on their ability to converge. Therefore, to assess the performance of the models in the real-case scenario of Apple Loop collection recombination, we computed, for both CoDi and miXmash-AIS, the average CPU usage in milliseconds (ms) over 10 runs. Furthermore, we equally report the number of interactions and the number of population cell count at convergence, which is particularly relevant for AIS due to the dynamic behavior of its population number adaptation.

4.3 Qualitative Evaluation - Perceptual Test

In order to perceptually assess the vertical and horizontal mashups developed by the CoDi model and its evaluation function, a conducted online listening test was built for each multimodal structure. The principal aim of this experiment if to study and control the relationship between the user enjoyment of the mashup solutions outputted and their capability of mashability. We want to measure musical quality and sense of *pleasantness* of CoDi from harmonic and rhythmic feel.

To theoretically explore this experiment, users are ideally listening and evaluating each mashup creation - whether an individual mashup (vertical) or a continuous generation of two or more connected mashups (horizontal) from the same initial solution. However, there are some impractical issues regarding that approach. Initially, it is difficult to ask listeners to judge their level of enjoyment of a mashup with several layers of harmonic mixing. Furthermore, creating several mashup combinations can create a long and tiring experience for user participants. Third, the estimated mashability is, through empirical observation, connected to the likelihood of very high harmonic and rhythmic compatibility. Thus, it is not trivial to meaningfully relate the concept of mashability between long mashups with different sections of individual mashup combinations within the same generation.

Nevertheless, an ordered output of evaluation values, or mashability within each individual mashup, is denoted. Hence, in order to prevent the issues referred above, we conducted two different sections for the listening experiment studying the multimodal approach of a singular mashup, or a continuous mashup of two of more generations from the same initial solution.

4.3.1 Listening Experiment

To create the mashup dataset employed in the experiment, a total of 20 mashups were randomly picked – 10 for vertical loop combinations, and 10 for continuous horizontal mashups (which means the initial seconds are the same for every output). Each output comes from a three-layer mix of audio files from the large dataset of 722 (171 rhythmic plus 551 harmonic) instrumental loops. For each mashup generated, we calculate and rank the level of mashability from a normalized range of [0,1], and picked the total 10 vertical and horizontal outputs from low, middle, and high values.

The purpose is to estimate a representative scenario using CoDi for within the human-in-the-loop systems and Creative-MIR, where the best result is an automated music mashup creation result given by CoDi, where lower ranked results represent the best recommendations minimizing the cost.

The experiment is conducted as follows: participants are asked to listen to each randomized mashup that is presented to them – each mashup has three layers of auditory mixing – with the possibility of replaying the listening audio file at any instance. Next, users conducting the experiment must rate their degree of *pleasantness* in a rating from 1-10 which denotes the enjoyment while listening to the mashup and, efficiently, how successful the mixing of the audio files was. This happens for each section in the multimodal approach – vertical and horizontal. As we have two different evaluation functions for CoDi, the horizontal evaluation section takes 2 full iterations of CoDi and presents 10 mashup generations from the initial mashup, taking the same initial 18 seconds. In this case, we are evaluating the function through the degree of *pleasantness* and *continuity* of the mashup generation from CoDi's model.

In total, a set of 100 participants – musicians and non-musicians – were recruited to take the listening test. The duration of the questionnaire, in the QuestionPro platform, took a total to 5-10 minutes for participants. As mashups are appreciated independently of any musical training, this was not a criterion for selecting participants, however the concern was to explain to participants what the term "loop", "pleasantness", and “music mashup” meant. To prevent order effects, the archive of the 20 total vertical and horizontal mashup generations were presented in a distinct random order for each participant.

Section 4.4 gives specific attention to the results of the listening test. In the post completion of the listening test, some participants gave their general opinion about CoDi, its generated mashups and how even the worst results can sound pleasant for different users.

In Appendix B, a list of figures showing the developed perceptual listening survey, and the procedure along the completion of it, from start to end.

4.3.2 Correlation Analysis - Pearson Correlation and R-Squared

To investigate the results of the listening test we assess the mean ratings that each participant addressed per mashup generation. Regression analysis is a form of infering statistics in models of

linear regression. Firstly, to gather the total values of the 100 participants and ascertain the average of each rating for the generated mashup.

Pearson's correlation coefficient is a test measuring the statistical relationship between two continuous variables. Known as the best method of measuring the association between variables of interest, it is based on the method of covariance giving information about the magnitude as well as the direction of the relationship between them. Cases must be independent to each other and variables must be linearly related. An assertion of this information can be generated through a scatterplot checking if it possesses a relatively straight line. *P-values* help determine if the relationship in the sample is also existing in larger populations. For each independent variable, the *p-value* tests the null hypothesis that the variable has no correlation with the dependent variable. If no correlation is found, there is no existing association between the changes in the independent variable and the shifts in the dependent variable – insufficient evidence that an effect in the population is made.

If the *p-value* for a variable is less than the significance level, the information data provides evidence that is enough to reject the null hypothesis for the entire population – essentially, the hypothesis favors that there is a non-zero correlation where if independent variable changes, there is an association with the variation in the dependent variable in the population. Nevertheless, if high p-values are found, insufficient evidence is present within the gathered data to conclude any existing correlation.

The coeffiecient of the regression is gathered from the sign and value between each independent and dependent variable. Positive coefficients explain that as values from the independent variable increase, the dependent variable' mean increases. Negative coefficients suggest that if the independent variable increases, the dependent variable tends to decrease.

The following items assert the degree of correlation coming from the Pearson coefficient correlation values, which can be grasped through the linear slope:

- Perfect - If the correlation is near ± 1. As a variable increases, the other variable also increases (if positive) or decrease (if negative).
- High degree - If the coefficient value lies between ± 0.50 and ± 1.
- Moderate degree: If the value lies between ± 0.30 and ± 0.49.
- Low degree: When the value lies below ± 0.29.
- No correlation: Value is zero.

R-squared – or coefficient of determination – evaluates scattered data points along the fitted regression line. For the same dataset, higher R-squared values represent smaller differences between the observed data and the fitted values.

The following items assert the meaning behind the values of the coefficient of determination:

- 0 - a model that does not explain any of the variation in the response variable around its mean. The mean of the dependent variable predicts the dependent variable as well as the regression model.
- 1 - a model that explains all the variation in the response variable around the mean.

The higher the coefficient of determination, the better the regression model fits the values coming from the experiment. Nevertheless, a good model can have a low value, and contrarily, a biased model can have a high value. This is why we address the randomization procedure of the mashups inside the survey, for each participant.

4.4 Results

In this section we address the results for both experiments made throughout the development phase – an intermediate study on implementend algorithmic models in Pure Data and their capabilities of searching through large archive and finding optimal solutions, and finally, a listening experiment reaching 100 participants, this time for the CoDi model developed and its generated mashups both vertically and horizontally.

4.4.1 Intermediate Experiment Results

The following tables denote the intermediate experiment taking in account the study upon the best algorithmic model searching optimization in a large archive of audio file. Tables 4.1, 4.2, and 4.3 present the results for the dimensions under evaluation – compatibility, diversity, computational performance – for each AIS, GA, and BF models. A boxplot was generated for each dimension to be evaluated, correspondingly. The two models of AIS and GA ran 10 times to account for their variability in local optima convergence. From each run of the algorithms, results account for the 10 best-ranked mashups. The BF model performs the same at every run of the algorithm, since it computes all possible loop dataset combinations – meaning the algorithm was necessary to run only once.

By comparing the average compatibility values across the AIS (1.368) and GA (2.694) models, AIS outperforms the standard GA model with a significant reduction of the compatibility value, finding local minima in the search space with smaller – and therefore, optimal – compatibility values. These results reinforce the importance of the multimodal search in guaranteeing a comprehensive search across the search space, which typically guarantees enhanced access to global and local optima. Conversely, the population of the GA algorithm typically converges to the same region and does not guarantee to converge to local optima. The average compatibility of the BF approach (.774) is lower than the AIS (and GA), thus presenting a set of more compatible mashups in its 10 best-ranked solutions. However, we must account the AIS is enforcing the exclusion of mashup solution in the same region of the feature space, which denotes perceptually similar mashups. Therefore, it can be excluding compatible and perceptually similar mashups, as it only retains the optimal mashup in a surrounding affinity region. The lower median affinity value of the

Table 4.1: AIS opt-aiNet objective evaluation.

Run Count (#)	Iteration Count	CPU Time (ms)	Cell Count	Average Compatibility	Median Affinity
1	35	2953	11	1.368	1.640
2	30	2593	12	1.528	1.639
3	60	4547	12	1.353	1.667
4	40	3250	13	1.556	1.655
5	75	6469	14	1.228	1.638
6	140	9749	14	1.451	1.613
7	80	7328	18	1.323	1.676
8	160	11984	12	1.259	1.602
9	95	8484	14	1.370	1.582
10	140	9625	12	1.247	1.529
Total Average	-	6708	-	1.368	1.624

Table 4.2: Genetic algorithm objective evaluation.

Run Count (#)	Iteration Count	CPU Time (ms)	Cell Count	Average Compatibility	Median Affinity
1	200	1578	20	2.921	0.441
2	200	1422	20	2.403	0.356
3	200	1313	20	3.072	0.509
4	200	1422	20	2.537	0.226
5	200	1141	20	2.644	0.278
6	200	1000	20	2.647	0.103
7	200	1453	20	2.610	0.458
8	200	1891	20	2.568	0.398
9	200	1266	20	2.780	0.745
10	200	1797	20	2.756	0.865
Total Average	-	1428	-	2.694	0.438

Table 4.3: Brute force objective evaluation.

Run Count (#)	Iteration Count	CPU Time (ms)	Cell Count	Average Compatibility	Median Affinity
1	-	151523	-	0.774	1.494

BA compared to the AIS reinforces this assertion, as it indicates the found 10 best-ranked mashups in the BF have smaller diversity. Figure 4.2 expresses a boxplot for the three algorithmic models, in which this dimension is visually justified with the data collected.

The median affinity for the three AIS (1.624), GA (.4379), and BF (1.494) denote a clear advantage of the AIS in promoting diversity in the 10 best-ranked mashups as it finds mashups whose global distances in the feature space are more spread then the remaining models. The GA

performs very poorly in terms of diversity as it typically converges most solutions towards a unique local optimal region in the search space. Equal to compatibility, Figure 4.3 shows a boxplot for the diversity dimension and the consequential behaviour of the three AIS, GA and BF models.

By comparing the computational efficiency of the three models given by the CPU time to complete a set of mashup solutions (i.e., the set of optima mashup solutions in AIS or GA and the full set of pairwise comparison in BF), some gains are in the average CPU time of GA (1428 ms) and AIS algorithm (6698 ms) compared to the CPU time of the BF (151523 ms). The GA could even be further optimized as no stopping criteria have been defined. The observation that the evaluation function E_p needs several iterations to improve the population in the latter stages of convergence. If no diversity is required, GA clearly stands as the best optimization strategy due to its efficiency. The BF approach presents an obvious high computational cost in such a combinatorial explosion problem. By inspecting the diverse iteration count in the AIS opt-aiNet and the resulting average compatibility values in each run, we can denote the capacity of stopping criteria in the algorithm to assess optimal convergence conditions.Figure 4.4 shows a boxplot for the CPU Performance of the three AIS, GA and BF models.

Giving the final regards to this experiment, the opt-aiNet AIS would be an appropriate choice for the algorithmic model implementation of CoDi. While the GA stands as the most efficient algorithm, its poor results in terms of compatibility reinforce the primacy of the AIS opt-aiNet in efficiently finding optimal compatible loop mashups. Furthermore, the AIS opt-aiNet showed to promote a diverse mashup population, outperforming both GA or BF approaches. The AIS opt-aiNet promotes diversity to a greater degree than the GA and the BF approach. Finally, GA and AIS have significant computational performance gains compared to the BF approach.

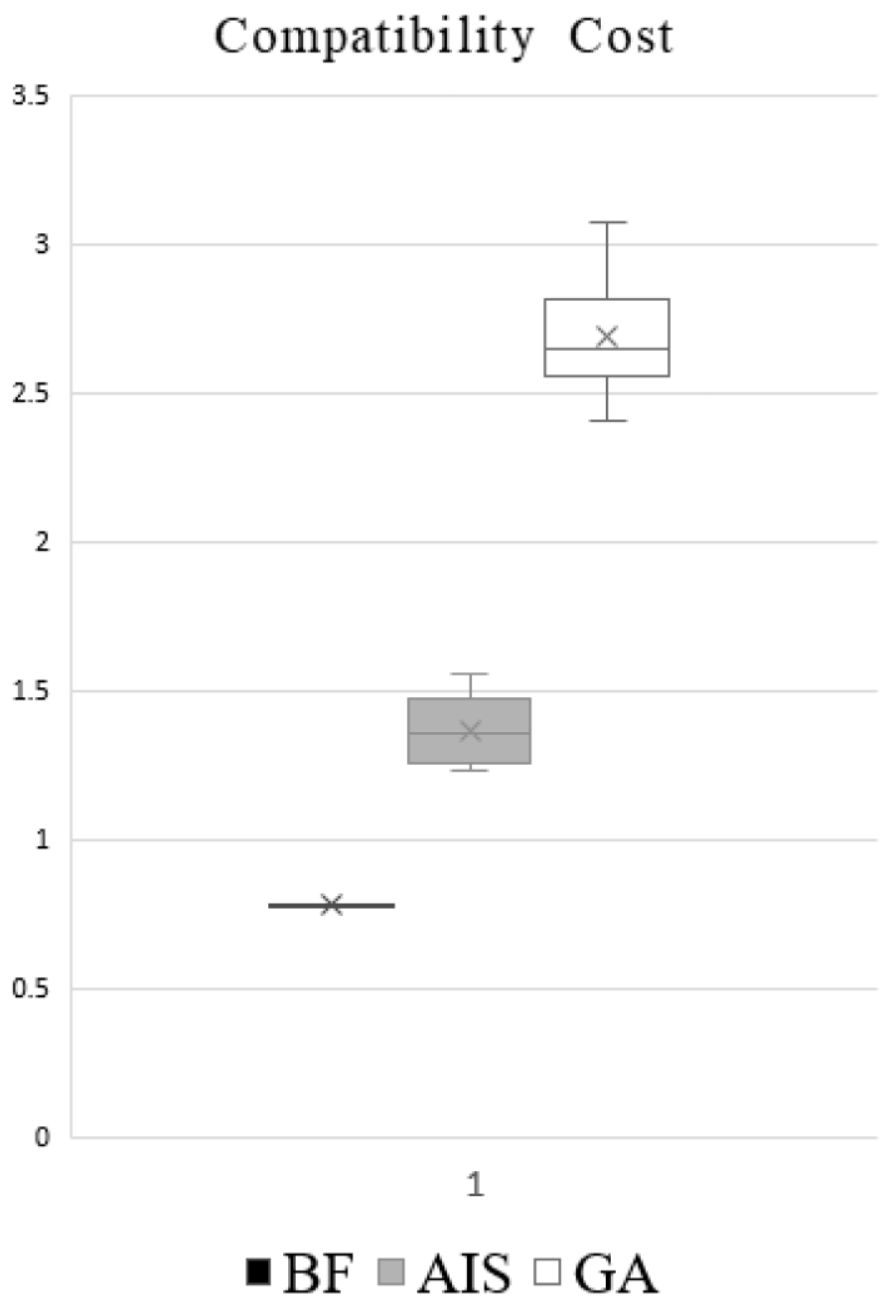

Figure 4.2: Boxplot Overview of the dimension related to Compatibility in the Objective Evaluation

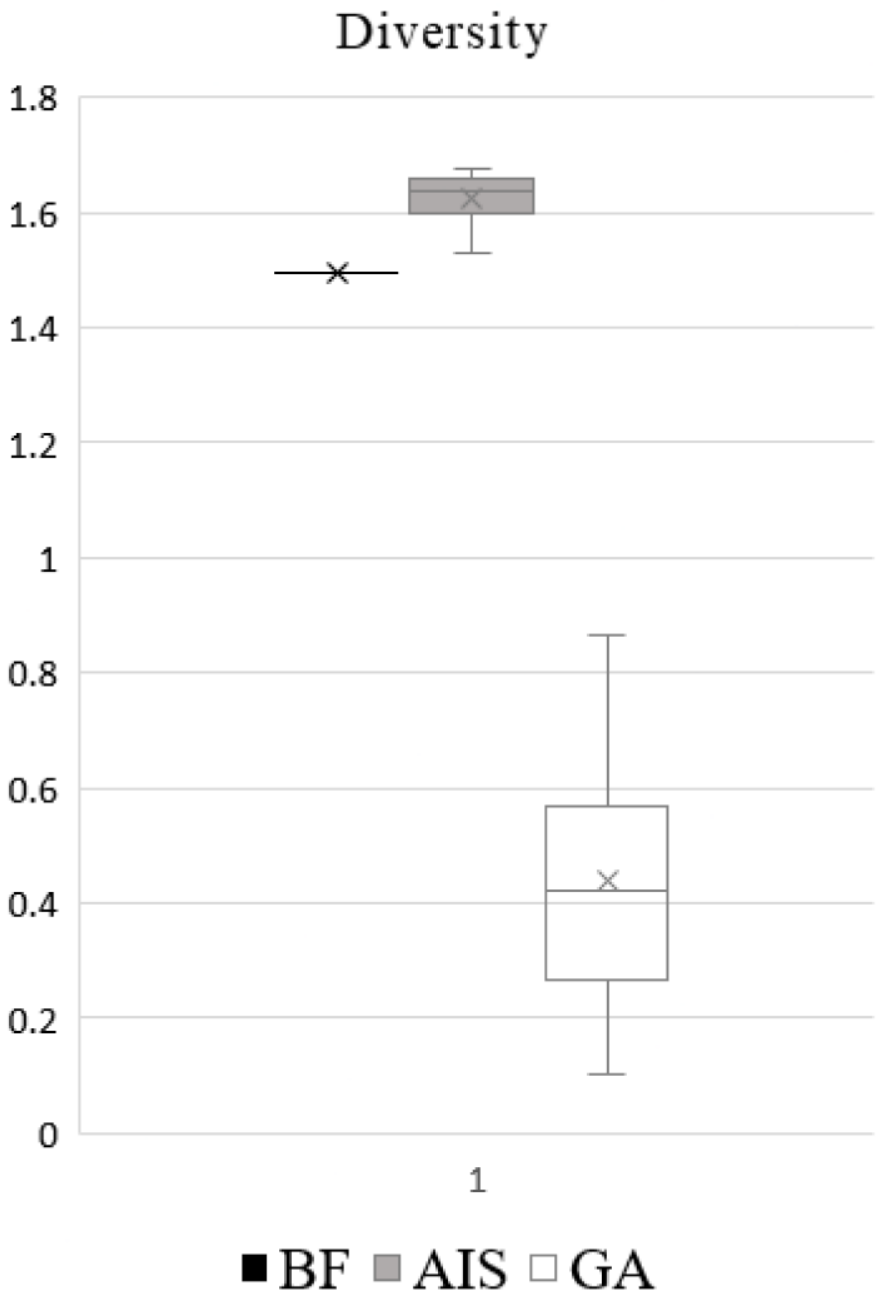

Figure 4.3: Boxplot Overview of the dimension related to Diversity in the Objective Evaluation

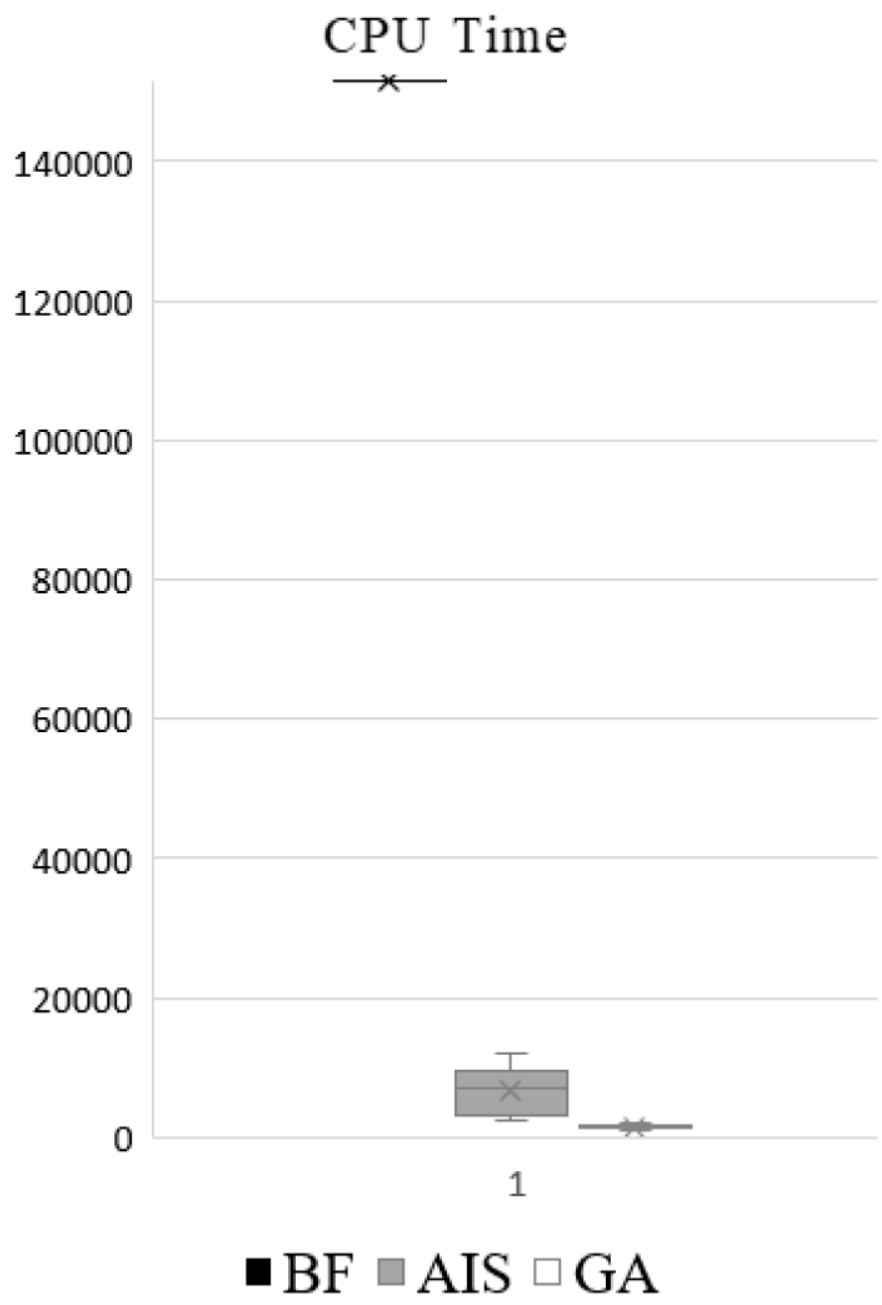

Figure 4.4: Boxplot Overview of the dimension related to CPU Performance in the Objective Evaluation

4.4.2 Listening Experiment Results

In order to assess the effectiveness of the evaluation functions for both vertical and horizontal dimensions, we raised the development of a listening experiment survey reaching 100 participants found in Annex B. After gathering all survey information, we employed an objective assessment on the correlation degree between the values of the evaluation function from CoDi, and the mean of the perceptual rankings made by the participants.

It followed through a qualitative online listening perceptual tests on 100 users – both non-musicians and musicians – measuring the value of *pleasantness* , which ensure quality is guaranteed at first hand. Secondly, to infer the possibility to compare the solutions in the implemented musical space. Initial developed models in Pure Data [73], mentioned further in Chapter 3, allowed for a study on three models based on AIS, GA, and Brute-force approach comparing diversity, compatibility, and computational performance.

4.4.2.1 Vertical Dimension

Starting from the Vertical dimension section of the perceptual test – with an evaluation function of harmonic, rhythmic, and spectral balance – Figures 4.5 denotes a scatter plot taking in account each of the 10 mashups' ID, and the mean average of all the 100 submissions and ranking values for that corresponding mashup. Note that we submitted each mashup from the best minimized value to the worst, which entails for the perceptual degree of *pleasantness*. For participants, the mashup IDs were random so that no external manipulation gets across the results, and is handled.

The results have shown that since we are minimizing the cost values and searching for diverse and compatible audio, we observe that if the evaluation function values increase, the values of *pleasantness* decrease, and vice-versa. The CoDi model generates mashups which are not, in fact, completely inaudible, and looking at the graphics at first hand, the linear correlation is accordingly variating with the levels of harmonic mixing. Additionally, by calculating the standard deviation and keeping the mean average value for each mashup, we employ for an error graph in Fig 4.6 showing the discrepancy of values for different users, as some may enjoy mashups with worse values of evaluation and overall harmonic mixing.

The measurement of the statistical study with the Pearson Correlation and the Coefficient of Determination, R-squared, is as observed in Figure 4.7 in which the representation of the linear regression confirms the proportional correlation when the evaluation values increase – affecting the harmonic mixing – and the survey rating diminish too. The information gathered makes notes of high values of r-squared (approx. 0.982) in which the CoDi is a reference model that explains all the variation in the response variable around the mean. As mentioned, the higher the coefficient of determination, the better the regression model fits the values coming from the experiment. In regards to the *p-value* of approximately zero, CoDi's hypothesis makes a favorable assumption that there is a non-zero correlation where if the independent variable of its objective evaluation changes, there is an association with the variation in the dependent variable that is in the perceptual experiment filled by the population. Furthermore on the coefficient regression, an observed

negative coefficient suggests that if the independent variable increases, the dependent variable tends to decrease – which is the case for the computational mashup creation model that is CoDi.

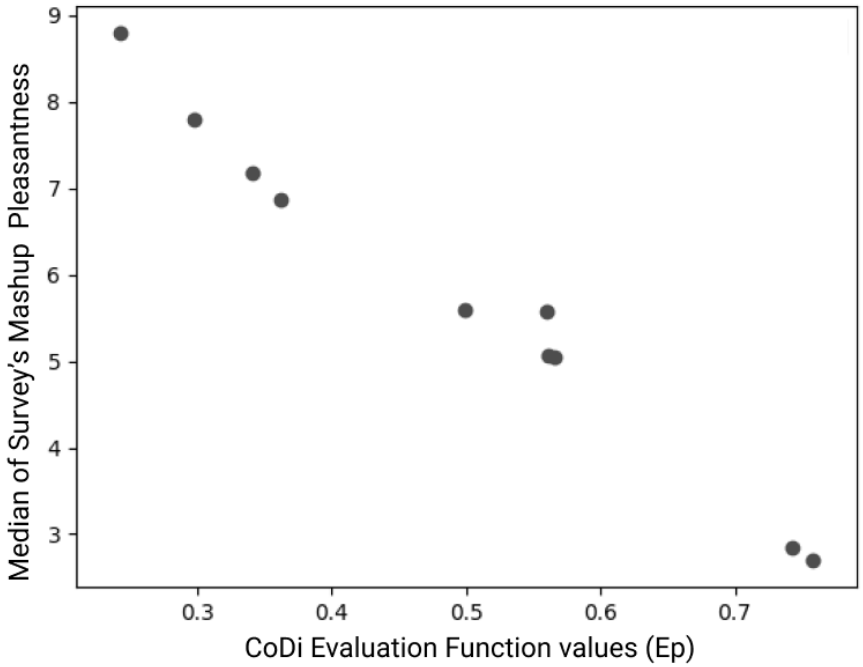

Figure 4.5: Scatter plot of the original data gathered from the Vertical dimension of the perceptual test survey.

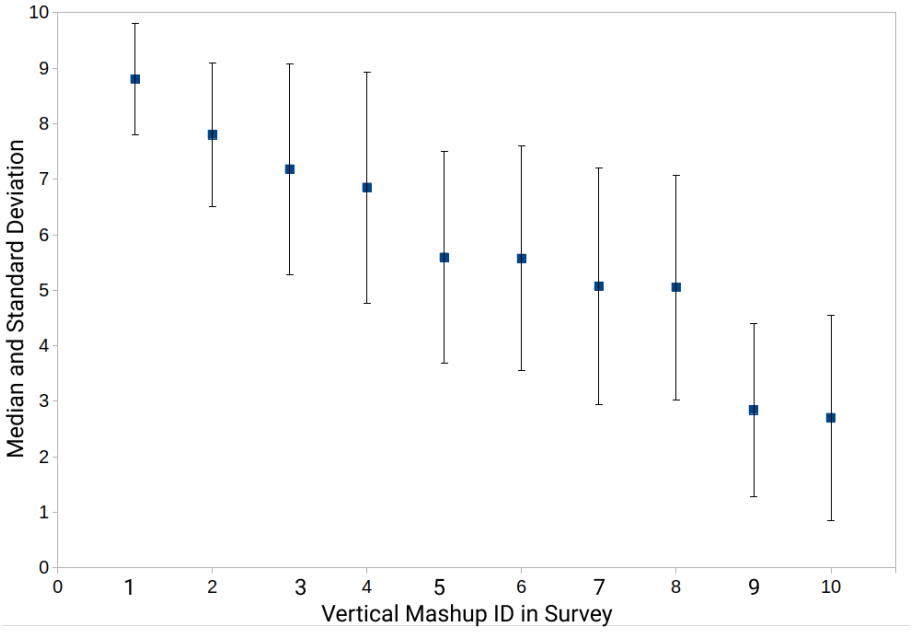

Figure 4.6: Error graph of the data with standard deviation calculation, gathered from the Vertical dimension section of the survey.

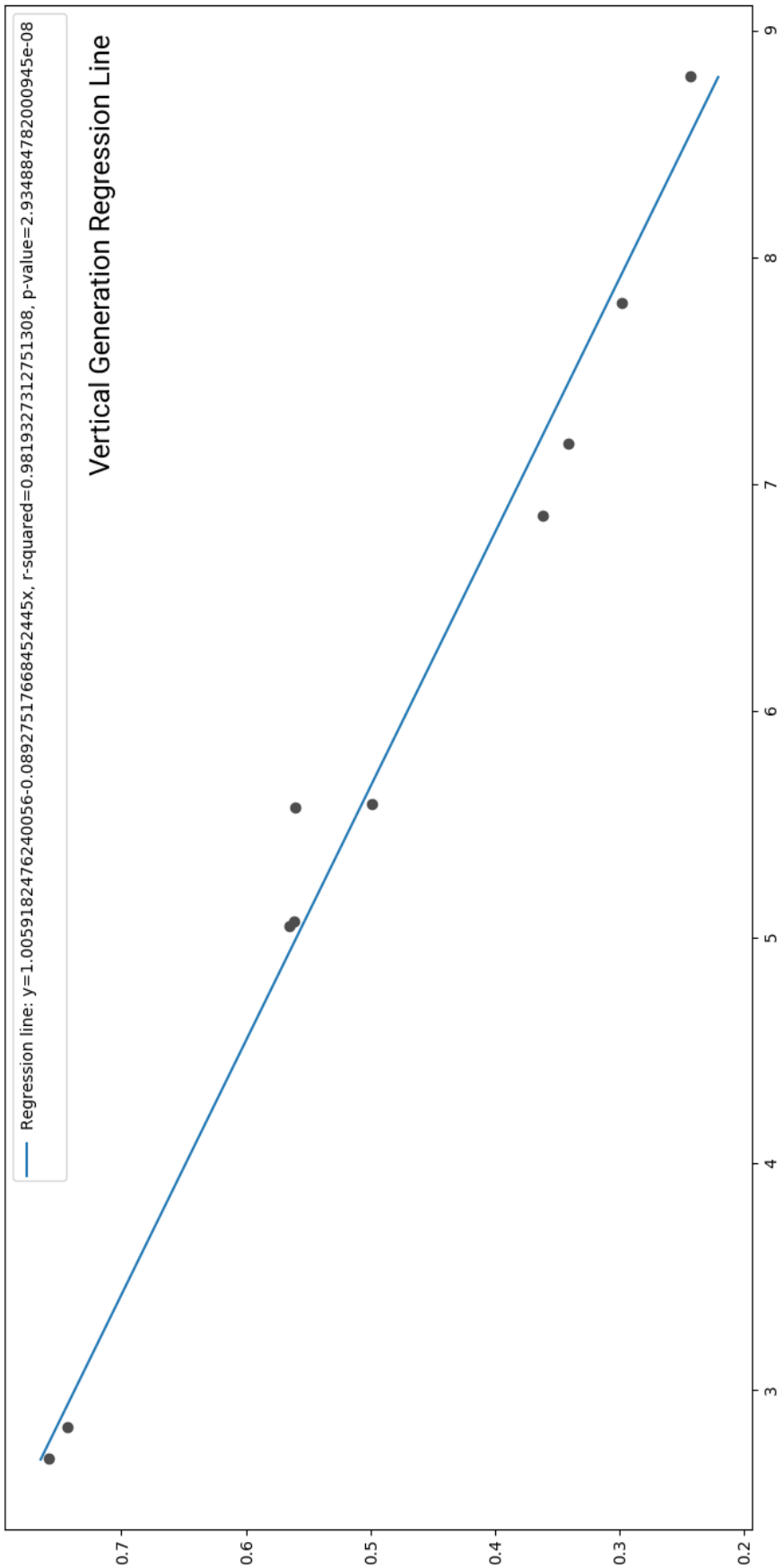

Figure 4.7: Plotted linear regression of the data gathered from the Vertical dimension section of the survey and the values of the CoDi model.

4.4.2.2 Horizontal Dimension

As the evaluation function for the Horizontal Dimension takes more of musicological metrics – with capabilities of MFCC – while helping the continuation of the mashup generation, another section of the perceptual experiment was conducted, evaluating the degree of *pleasantness* and, additionally, *continuity* as a form of a coherent continuous variation and generation of the mashup, for 2 full iterations of CoDi. Figure 4.8 also informs a scatter plot with each of the 10 mashups' ID, and the mean average of all the 100 submissions and ranking values for that corresponding mashup. Equal to the vertical evaluation, each mashup was submitted from the best minimized value to the worst, which entails for the perceptual degree of *pleasantness* and *continuity*. Mashup IDs were random so that no external manipulation affects results.

The results have shown that as the horizontal evaluation function's values increase, the values of *pleasantness* and *continuity* decrease, and vice-versa. The continuous generation of the CoDi model produces mashups which are not, in fact, completely inaudible following more than 1 iteration, and looking at the graphics at first hand, the linear correlation is accordingly variating with the levels of harmonic mixing. Additionally, by calculating the standard deviation and keeping the mean average value for each mashup, we employ for an error graph in Fig 4.9 showing the discrepancy of values for different users, as some may enjoy mashups with worse values of evaluation and overall harmonic mixing.

As observed in Figure 4.10, the values of the Coefficient of Determination, R-squared, are equally high as the vertical dimension (approx. 0.913) – in which the representation of the linear regression confirms the proportional correlation when the evaluation values increase – affecting the harmonic mixing and continuity – and the survey rating diminish too. The information gathered makes notes of high values of r-squared in which the CoDi is a reference model that explains all the variation in the response variable around the mean. As mentioned, the higher the coefficient of determination, the better the regression model fits the values coming from the experiment.

CoDi's *p-value* of approximately zero in the Horizontal Dimension makes CoDi's hypothesis favorable to a non-zero correlation – if the independent variable of its objective horizontal evaluation changes, there is an association with the variation in the dependent variable that is in the perceptual experiment filled by the population. Furthermore on the coefficient regression, an observed negative coefficient suggests that if the independent variable increases, the dependent variable decreases too – which is the case for the computational mashup creation model that is CoDi.

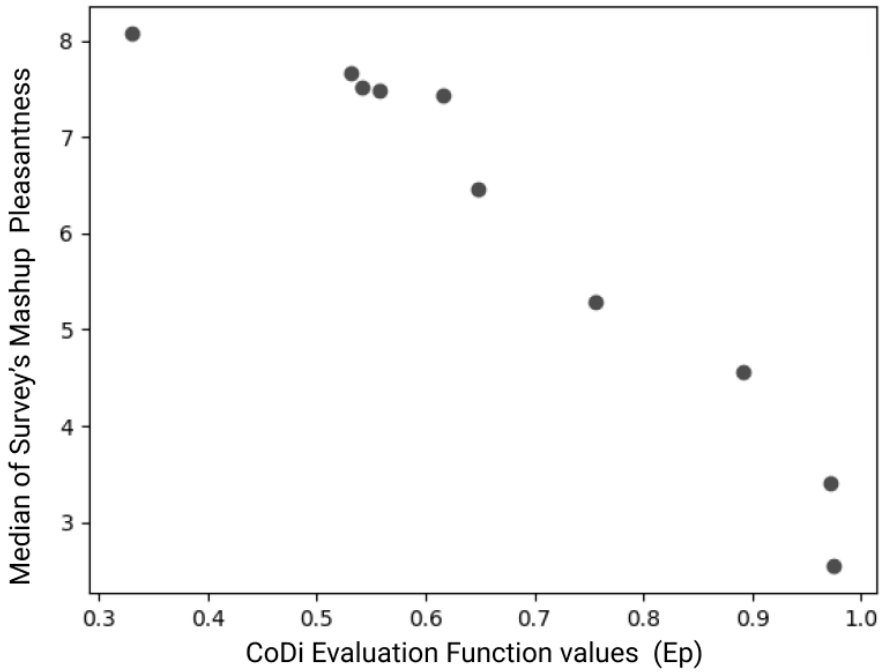

Figure 4.8: Scatter plot of the original data gathered from the Horizontal dimension of the perceptual test survey.

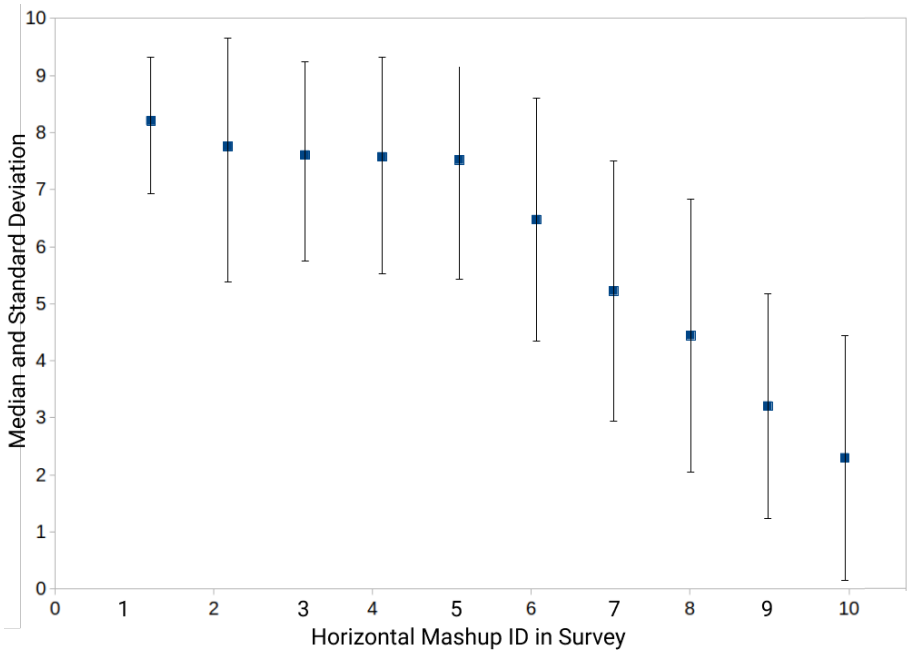

Figure 4.9: Error graph of the data with standard deviation calculation, gathered from the Horizontal dimension section of the survey.

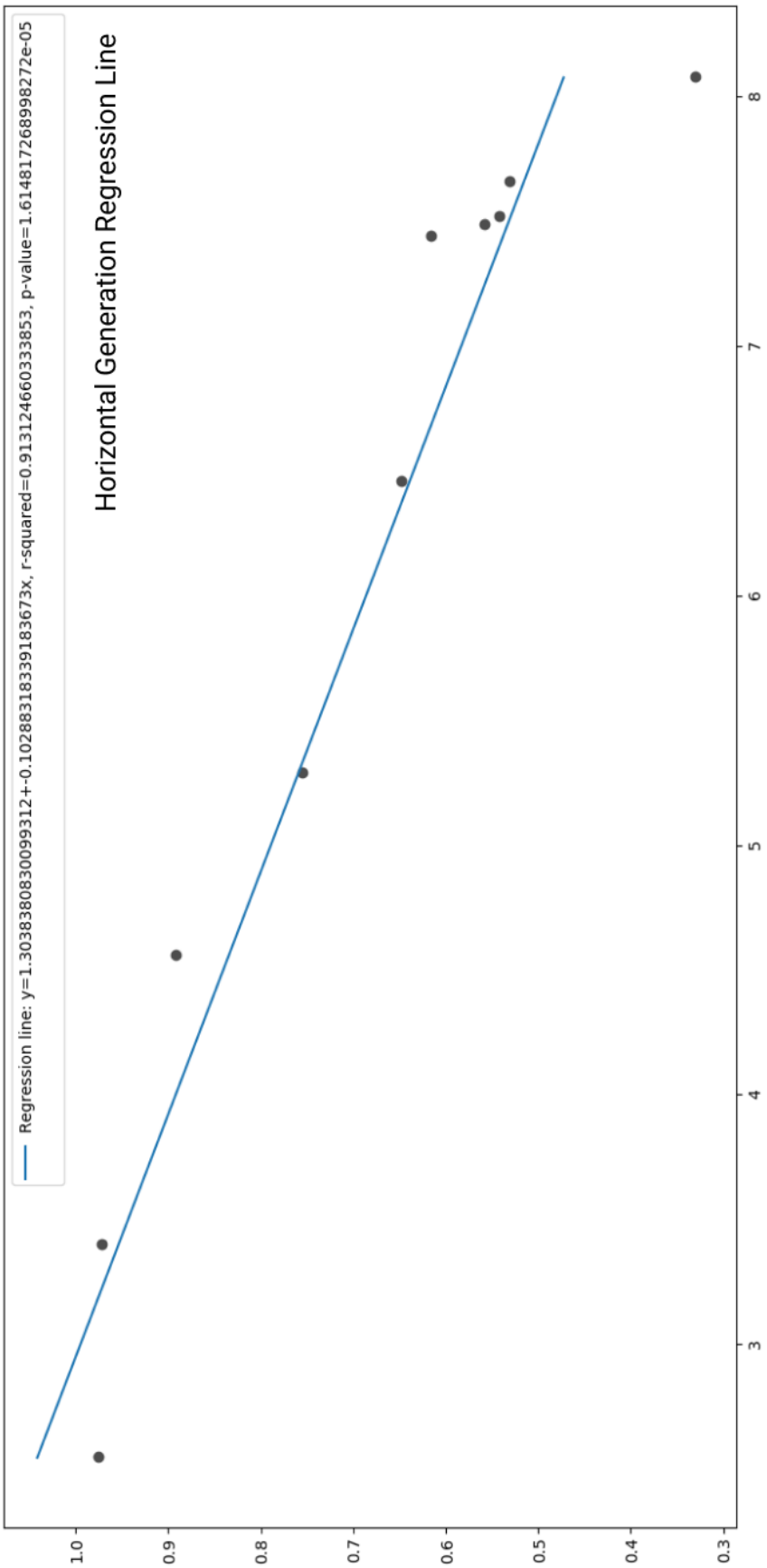

Figure 4.10: Plotted linear regression of the data gathered from the Horizontal dimension section of the survey and the values of the CoDi model.

4.5 Summary

Two evaluations of qualitative and quantitative nature make the case to leverage compatibility and diversity in large musical corpora of audio files – with metrics of harmony, rhythm and timbre – in a multimodal structure. First, an intermediate experiment and development of three algorithmic models – in Pure Data – of AIS, GA and BF approach. This assessed the best algorithm for raising an optimized search and evaluation function on compatibility, diversity, and computational performance, where AIS stands as main candidate. A final experiment with the CoDi model, developed in Python, employed a listening test reaching 100 participants, to assess the degree of *pleasantness* for the vertical dimension of the mashup generation – with only one iteration of the CoDi model – and additionally to that, the feeling of *continuity* from the horizontal dimension which took two iterations of CoDi.

Chapter 5

Conclusions

In this dissertation, we propose a functional prototype for automatic generation of music as recombinations of loops at scale. It optimizes processes with large audio files in a continuous space of descriptors, with the aid of AIS algorithm opt-aiNet. Furthermore, we raise the knowledge of computationally structured metrics for defining musical style in audio files along several combinations possible of output. We leverage an exhaustive reading on state-of-the-art solutions for the issues of compatibility, diversity and computational performance, and mention several musical systems and advantages to the use of Creative-MIR techniques for automated music generation. Additionally to illustrate the considerable prospects of signal processing in the sense of music, and for subsequent research of Creative-MIR.

We proposed CoDi, a multimodal music mashup optimization system for loop recombination at scale. It adopts the AIS opt-aiNet algorithm to leverage compatible and diverse mashups while addressing the scalability issues in existing state-of-the-art Brute-Force solutions for computational music mashup creation. In promoting a diverse set of optimal mashups, the system can account for personal preferences and different stylistic traits. An objective comparison of AIS opt-aiNet to a standard Genetic Algorithm and Brute-Force approaches in the task under study denotes the primacy of the AIS opt-aiNet in finding local and global optimal mashups, closely matching the compatibility values of the BF approach. The AIS opt-aiNet promotes diversity to a greater degree than the GA and the BF approach, such as GA and AIS have significant computational performance gains compared to the BF approach. Furthermore, a subjective evaluation through a perceptual experiment was employed to determine the relationship between estimated user enjoyment as *pleasantness* as the musicological metrics of CoDi. Results have shown high correlation between the survey submissions and the values of the evaluation function.

5.1 Contributions for Areas of Interest

With this disseration we acknowledge contributions for a scientific community, but also for the artistic community within the human-in-the-loop approach of CoDi's model. It is a supportive tool in the compositional process of musicians and non-musicians which look for diverse and compatible mashups in a multimodal optimization for context generation and non-context generation. As a computational music mashup model developed in a basis of programming informatics, we denote that this research makes improvement within the Creative-MIR dimension.

In the given literature review, we leverage that musicological metrics as means of computation have not been fully acknowledged and constructed. However, with the development of this work, we address that the present metrics of harmonic, rhythmic, and timbral content can still raise compatibility and diversity in a large-scale feature space of solutions. The feature extraction process has proven to be a good candidate to continue the development of this model and many other systems based on computational generation of music.

5.2 Discussion

While the issue of consonance has been extensively studied, its relevance to the direction of music composition and harmonic compatibility is rarely addressed. Most of prototypes are analyzed in terms of consonance/dissonance by rating chords and afterwards the generated ranking is associated with human perception tests. Driven by immunological values, opt-aiNet produces a variety of high quality options at the same time as retaining diversity. This inherent property of variety enables opt-aiNet to bring convergence to all the optimum, both global or local of the fitness function, which expresses as compositions most of which are identical to the goal but distinct from others. The AIS generates several selections when going through sample-based sounds, instead of finding a single solution that is limited by parameterizations specified a priori, extending the artistic capacity of automatic generation of music.

While the review on computational music mashup systems has almost exclusively focused on methodologies for searching and retrieving musical audio content from large musical audio archives, we must acknowledge that musical audio transformations are an important dimension to consider in the future. Existing tools and methods incorporate such transformations to a small degree and claim limitations in the resulting audio due to the artifacts introduced by these audio processing techniques (e.g., time-stretching, pitch-shifting, or spectral filtering). Existing tools acknowledge the need for multi-attribute optimization search within a pool of musical audio sample candidates, yet very little knowledge of the interaction across these attributes is known.

Curating multi-track archives to understand the underlying phenomena of musical audio recombination can promote models that veridically assess the musical layers interaction and promote better predictive models, namely by using deep learning architectures. Ultimately, advancing enhanced models of musical audio compatibility can foster the creative composition impetus across

a broader range of users beyond highly trained composers or producers and facilitate the time-consuming and demanding search across the growing number of musical audio content.

5.3 Future Work

In the future, we are planning to extend the current evaluation so it will explore the addition of timbral compatibility criteria to the evaluation function E_p without narrowing the model's capacity to promote diverse solutions. Additionally, to keep assessing for an evaluation function with added musicological metrics contemplating for continuous generation of music mashups, as we investigate further the perceptual difference that mashups that are more pleasant than others, in which they sound aesthetically better.

As opt-aiNet has proven to be a prominent candidate for an algorithm composition directed for computational music mashup models, there are several other algorithms aiding in areas such as music recomendation and music genre generation. Therefore, new features are pathways to more consistent models building onto fully functional frameworks – going from human-in-the-loop approach to the ease of any user – in which users can either generate or listen to music mashups at indefinite times.

Appendix A

Notable Music Forms

Notable Musical Forms		
Form	Structure	Further Details
Fugue	Contrapuntal form	
Strophic	A A' A" A"'	
Binary	AA', AB, AAB, ABB, AA'BB'	
Ternary	ABA, ABA'	
Rondo	ABACADA...	A:'chorus', B,C,D:'verse'
Sonata form	AB-Dev-AB	A:'principal' theme
Rondo-Sonata form	AB-AC-AB	B:'secondary' theme

Table A.1: Visualization of most notable musical forms. (Source from Wu [98])

Appendix B

Online Survey - Listening Test

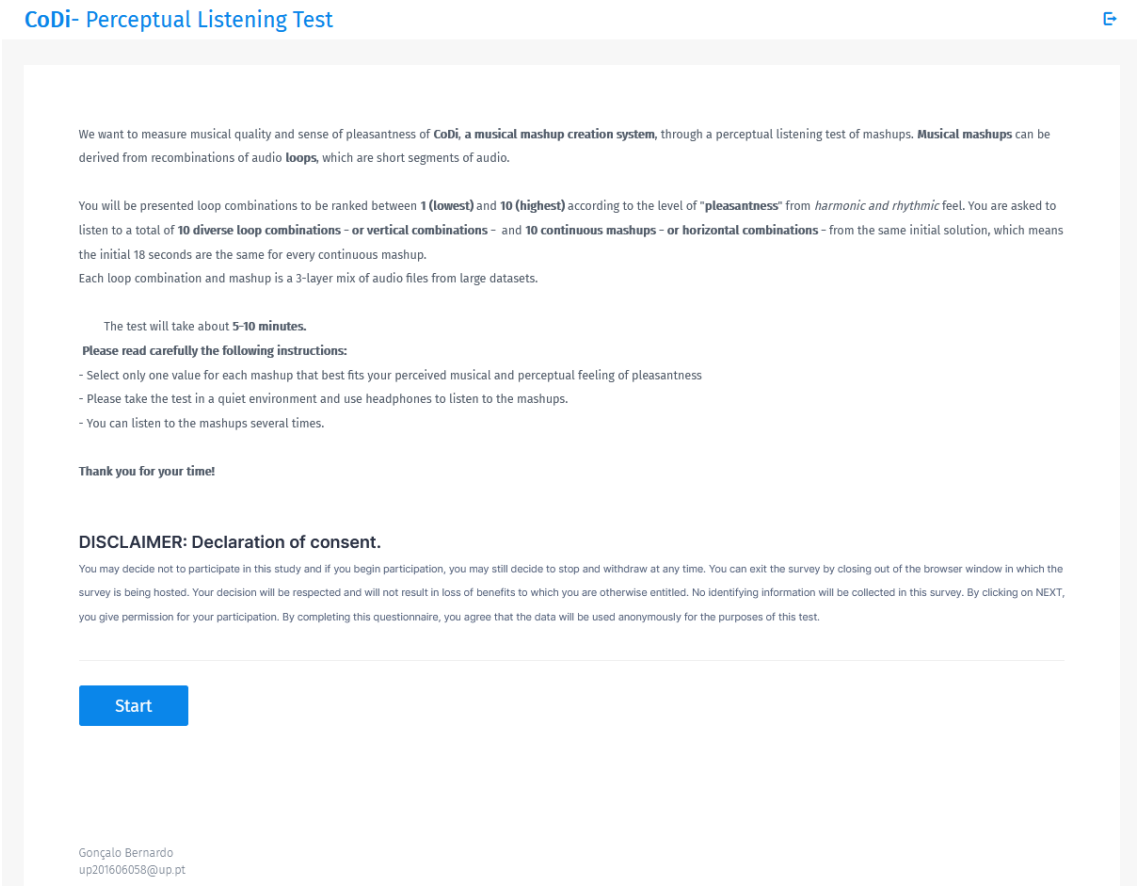

Figure B.1: Introduction of survey.

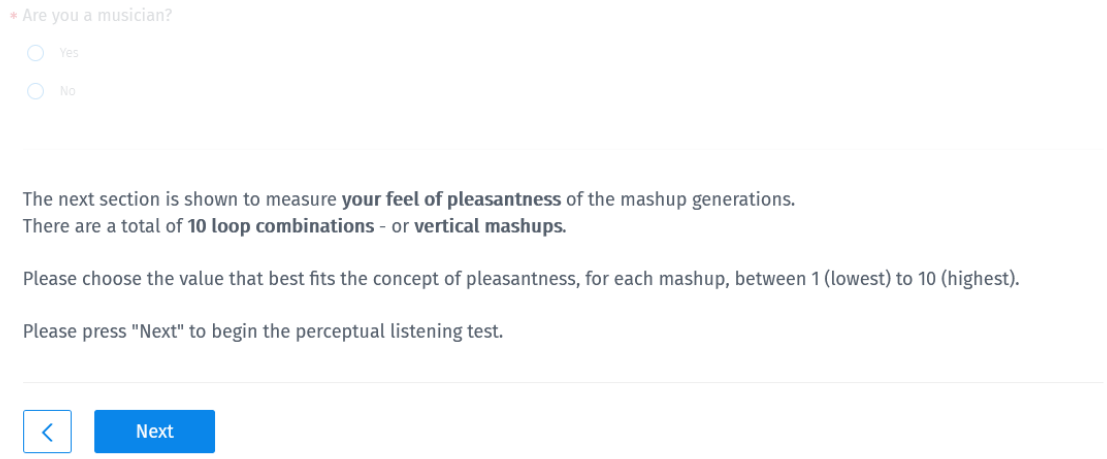

Figure B.2: Section introduction for vertical mashups.

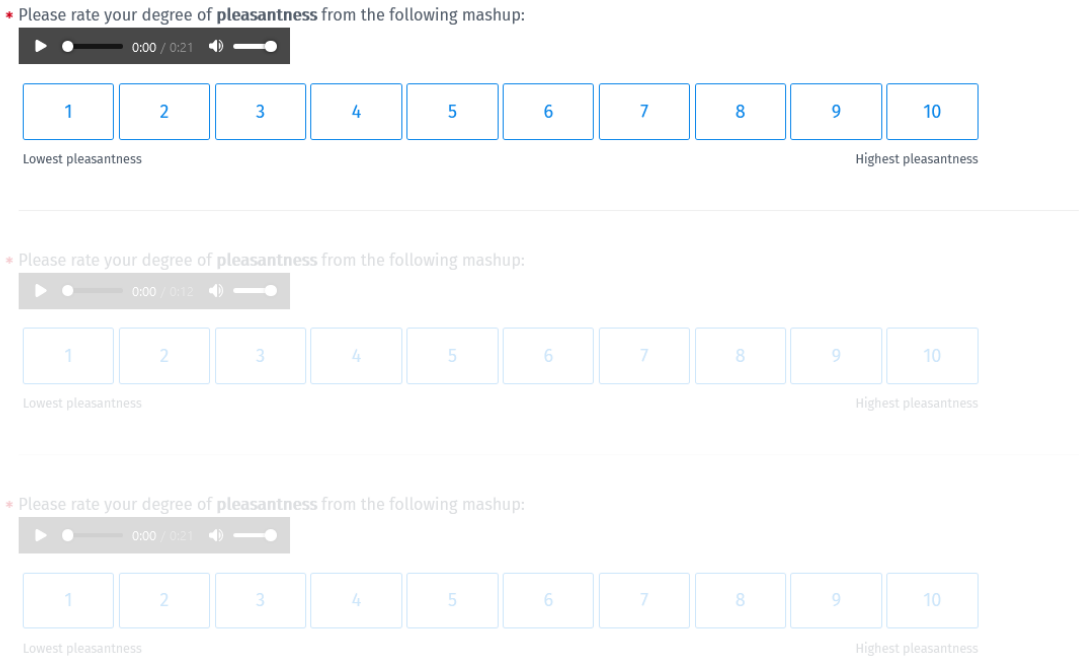

Figure B.3: Evaluation procedure of survey for vertical mashups ranking 1 (lowest) to 10 (highest) pleasantness.

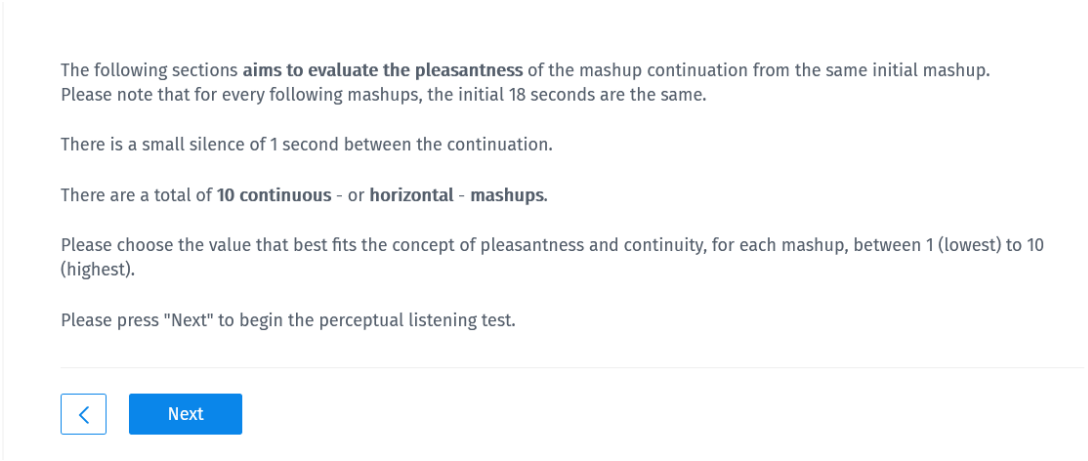

Figure B.4: Section introduction for horizontal mashups.

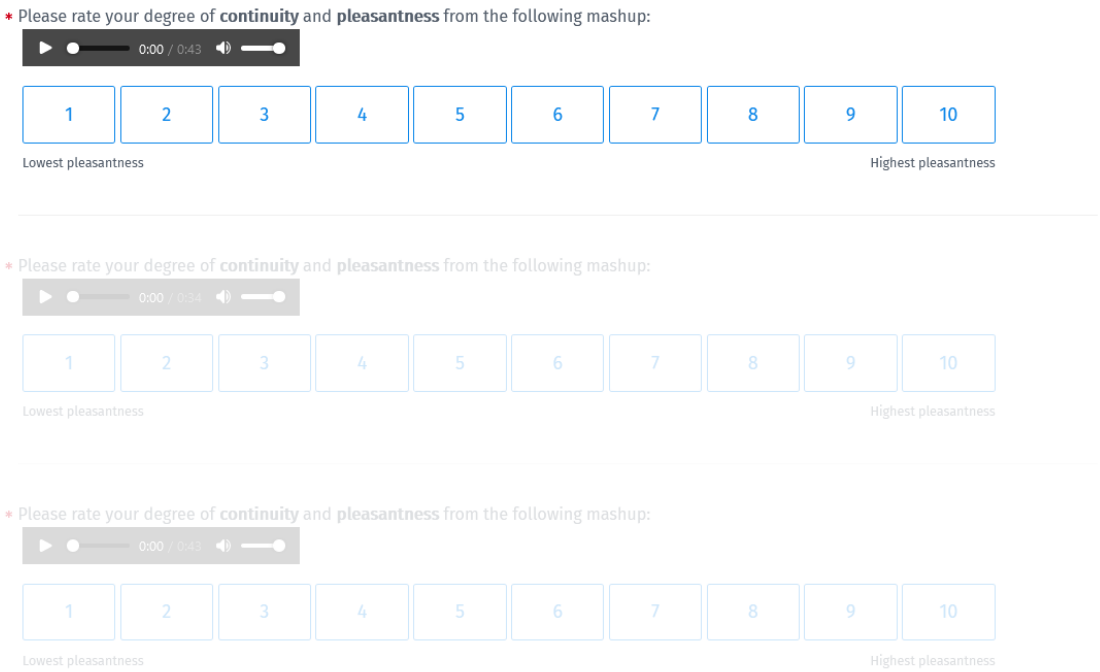

Figure B.5: Evaluation procedure of survey for horizontal mashups ranking 1 (lowest) to 10 (highest) of pleasantness and continuation.

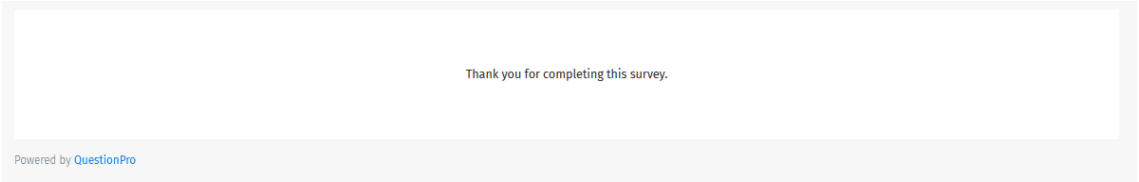

Figure B.6: End of survey.

References

[1] Mubert: Royalty-free music for app & content makers. Available at https://www.mubert.com/. Accessed: 2021-01-07.

[2] Lionbridge: CEO of AI Music Generator Mubert Wants to "Create a Musical DNA", Dec 2019. Accessed: 2021-02-01.

[3] José Abreu, Marcelo Caetano, and Rui Penha. Computer-aided musical orchestration using an artificial immune system. In Colin Johnson, Vic Ciesielski, João Correia, and Penousal Machado, editors, *Evolutionary and Biologically Inspired Music, Sound, Art and Design*, pages 1–16. Springer International Publishing, 2016.

[4] Christopher Ariza. Navigating the landscape of computer aided algorithmic composition systems: a definition, seven descriptors, and a lexicon of systems and research. In *Proceedings of the 2005 International Computer Music Conference, ICMC 2005, Barcelona, Spain, September 4-10*. Michigan Publishing, 2005.

[5] Jean-Julien Aucouturier and François Pachet. Music similarity measures: What's the use? In *ISMIR 3rd International Conference on Music Information Retrieval, Paris, France, October 13-17, Proceedings*, 2002.

[6] Juan Pablo Bello, Laurent Daudet, Samer A. Abdallah, Chris Duxbury, Mike E. Davies, and Mark B. Sandler. A tutorial on onset detection in music signals. *IEEE Transactions on Speech and Audio Processing*, 13(5-2):1035–1047, 2005.

[7] Gilberto Bernardes. Tonal Interval Space. Available at https://sites.google.com/site/tonalintervalspace/home. Accessed: 2021-01-20.

[8] Gilberto Bernardes, Diogo Cocharro, Marcelo Caetano, Carlos Guedes, and Matthew E.P. Davies. A multi-level tonal interval space for modelling pitch relatedness and musical consonance. *Journal of New Music Research*, 45(4):281–294, 2016.

[9] Gilberto Bernardes, Diogo Cocharro, Carlos Guedes, and Matthew E. P. Davies. Harmony generation driven by a perceptually motivated tonal interval space. *Comput. Entertain.*, 14:1–21, 2016.

[10] Gilberto Bernardes, Matthew EP Davies, and Carlos Guedes. A hierarchical harmonic mixing method. In *International Symposium on Computer Music Multidisciplinary Research*, pages 151–170. Springer, 2017.

[11] Gilberto Bernardes, Matthew EP Davies, and Carlos Guedes. A perceptually-motivated harmonic compatibility method for music mixing. In *13th International Symposium on Computer Music Multidisciplinary Research*, pages 104–115, 2017.

[12] Gonçalo Bernardo and Gilberto Bernardes. Musical Audio Compatibility Retrieval: Towards Computer-aided Music Production. 06 2021.

[13] Klaas Bosteels and Etienne Kerre. Fuzzy audio similarity measures based on spectrum histograms and fluctuation patterns. In *2007 International Conference on Multimedia and Ubiquitous Engineering (MUE 2007) Seoul, Korea*, volume 96, pages 361–365. IEEE Computer Society, 01 2007.

[14] Daniel Bowling, Dale Purves, and Kamraan Gill. Vocal similarity predicts the relative attraction of musical chords. *Proceedings of the National Academy of Sciences*, 115:216–221, 2017.

[15] Daniel L. Bowling and Dale Purves. A biological rationale for musical consonance. *Proceedings of the National Academy of Sciences*, 112:11155–11160, 2015.

[16] Sebastian Böck and Markus Schedl. Enhanced beat tracking with context-aware neural networks. *Proceedings of the 14th International Conference on Digital Audio Effects, DAFx 2011*, 09 2011.

[17] Marcelo Caetano, Asterios Zacharakis, Isabel Barbancho, and Lorenzo J. Tardón. Leveraging diversity in computer-aided musical orchestration with an artificial immune system for multimodal optimization. *Swarm and Evolutionary Computation*, 50:100484, 2019.

[18] Greoire Carpentier, Damien Tardieu, Jonathan Harvey, Gérard Assayag, and Emmanuel Saint-James. Predicting timbre features of instrument sound combinations: Application to automatic orchestration. *Journal of New Music Research*, 39, 03 2010.

[19] Grégoire Carpentier, Gérard Assayag, and Emmanuel Saint-James. Solving the musical orchestration problem using multiobjective constrained optimization with a genetic local search approach. *Journal of Heuristics*, 16:681–714, 10 2010.

[20] Bo-Yu Chen, Jordan B. L. Smith, and Yi-Hsuan Yang. Neural loop combiner: Neural network models for assessing the compatibility of loops. *Computing Research Repository CoRR*, 2020.

[21] Diogo Cocharro, Gilberto Bernardes, Gonçalo Bernardo, and Cláudio Lemos. Revisiting rhythmic representations and similarity. In *Luísa Castilho et al. (Eds.) Perspectives on music and musicology. Current Research in Systematic Musicology*. Springer, 2021.

[22] Rip Cohen. aaB: Strophic Design and Cognition. Available at https://www.academia.edu/9697546/aaB_Strophic_Design_and_Cognition, 2013. Accessed: 2021-02-10.

[23] Darrell Conklin and Ian Witten. Multiple viewpoint systems for music prediction. *Journal of New Music Research*, 24(1):51–73, 1995.

[24] Marion Cousineau, Josh H. McDermott, and Isabelle Peretz. The basis of musical consonance as revealed by congenital amusia. *Proceedings of the National Academy of Sciences*, 109(48):19858–19863, 2012.

[25] M. Davies, A. Stark, F. Gouyon, and M. Goto. Improvasher: A real-time mashup system for live musical input. In *14th International Conference on New Interfaces for Musical Expression, NIME, London, United Kingdom, June 30 - July 4*, pages 541–544. nime.org, 2014.

[26] Matthew E. P. Davies, Philippe Hamel, Kazuyoshi Yoshii, and Masataka Goto. Automashupper: An automatic multi-song mashup system. In *Proceedings of the 14th International Society for Music Information Retrieval Conference, ISMIR Curitiba, Brazil, November 4-8*, pages 575–580, 2013.

[27] Matthew E. P. Davies, Philippe Hamel, Kazuyoshi Yoshii, and Masataka Goto. Automashupper: automatic creation of multi-song music mashups. *IEEE ACM Trans. Audio Speech Lang. Process.*, 22(12):1726–1737, 2014.

[28] Leandro De Castro and Jon Timmis. An artificial immune network for multimodal function optimization. In *Proceedings of the 2002 CECCongress*, volume 1, pages 699–704, 06 2002.

[29] Simon Dixon, Fabien Gouyon, and Gerhard Widmer. Towards characterisation of music via rhythmic patterns. In *ISMIR5th International Conference on Music Information Retrieval, Barcelona, Spain, October 10-14*, 01 2004.

[30] Simon Dixon, Elias Pampalk, and Gerhard Widmer. Classification of dance music by periodicity patterns. In *ISMIR 4th International Conference on Music Information Retrieval, Baltimore, Maryland, USA, October 27-30*, 09 2003.

[31] Felix A. Dobrowohl, A. Milne, and R. Dean. Timbre preferences in the context of mixing music. *Applied Sciences*, 9(8), 2019.

[32] Felix A. Dobrowohl, Andrew J. Milne, and Roger T. Dean. Controlling perception thresholds for changing timbres in continuous sounds. *Organised Sound*, 24:71–84, 2019.

[33] J. Foote and S. Uchihashi. The beat spectrum: a new approach to rhythm analysis. In *IEEE International Conference on Multimedia and Expo, 2001. ICME 2001.*, pages 881–884, 2001.

[34] Jonathan Foote and Matthew Cooper. Visualizing musical structure and rhythm via self-similarity. In *Proceedings of the International Computer Music Conference*, volume 1, pages 423–430, 2001.

[35] Jonathan Foote, Matthew Cooper, and Unjung Nam. Audio retrieval by rhythmic similarity. In *Proceedings of the International Conference on Music Information Retrieval*, 2002.

[36] Jonathan Foote and Shingo Uchihashi. The beat spectrum: A new approach to rhythm analysis. volume 22, 08 2001.

[37] M. Gallagher. *The Music Tech Dictionary: A Glossary of Audio-Related Terms and Technologies*. Course Technology, 2009.

[38] R. Gebhardt, M. Davies, and B. Seeber. Harmonic mixing based on roughness and pitch commonality. In *Proceedings of the 18th Int. Conference on Digital Audio Effects (DAFx-15)*, 2015.

[39] Roman B. Gebhardt, Matthew E. P. Davies, and Bernhard U. Seeber. Psychoacoustic approaches for harmonic music mixing. *Applied Sciences*, 6(5):136, 2016.

[40] Masataka Goto. An audio-based real-time beat tracking system for music with or without drum-sounds. *Journal of New Music Research*, 30, 09 2002.

[41] Masataka Goto. Grand challenges in music information research. In *Dagstuhl Follow-Ups: Multimodal Music Processing*, volume 3, pages 217–225. Dagstuhl Publishing, 2012.

[42] Fabien Gouyon and Simon Dixon. Dance music classification: A tempo-based approach. In *Proceedings of the International Conference on Music Information Retrieval*, 2004.

[43] Garth Griffin, YE Kim, and Douglas Turnbull. Beat-sync-mash-coder: A web application for real-time creation of beat-synchronous music mashups. In *Proceedings of the International Conference on Acoustics, Speech, & Signal Processing*, pages 2–5, 2010.

[44] Peter M. C. Harrison and M. Pearce. Simultaneous consonance in music perception and composition. *Psychological Review*, 12(2):216–244, 2020.

[45] Peter M. C. Harrison and Marcus T. Pearce. An energy-based generative sequence model for testing sensory theories of western harmony. *Proceedings of the 19th International Society for Music Information Retrieval Conference, ISMIR Paris, France, September 23-27*, pages 160–167, 2018.

[46] Andre Holzapfel and Yannis Stylianou. Rhythmic similarity of music based on dynamic periodicity warping. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP March 30 - April 4, Caesars Palace, Las Vegas, Nevada, USA*, pages 2217–2220. IEEE, 03 2008.

[47] Andre Holzapfel and Yannis Stylianou. A scale transform based method for rhythmic similarity of music. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP 19-24 April Taipei, Taiwan*, pages 317–320. IEEE, 04 2009.

[48] Andrew Horner and David E. Goldberg. Genetic algorithms and computer-assisted music composition. In *Proceedings of the 4th International Conference on Genetic Algorithms, San Diego, CA, USA, July*, pages 437–441. Morgan Kaufmann, 1991.

[49] Eric J Humphrey, Douglas Turnbull, and Tom Collins. A brief review of creative mir. *Proceedings of the International Society for Music Information Retrieval Conference ISMIR Late-Breaking News and Demos*, 2013.

[50] Talib Hussain. Methods of combining neural networks and genetic algorithms, 1997.

[51] William Hutchinson and Leon Knopoff. The acoustic component of western consonance. *Interface*, 7:1–29, 1978.

[52] Nanzhu Jiang, Peter Grosche, Verena Konz, and Meinard Müller. Analyzing chroma feature types for automated chord recognition. In *AES International Conference Semantic Audio Ilmenau, Germany, July 22-24*. Audio Engineering Society, 2011.

[53] T. Kitahara, K. Iijima, Misaki Okada, Yuji Yamashita, and A. Tsuruoka. A loop sequencer that selects music loops based on the degree of excitement. In *12th Sound and Music Computing Conference (SMC2015)*, 2015.

[54] 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 *Proceedings of the 16th International Society for Music Information Retrieval Conference, ISMIR, Málaga, Spain, October 26-30*, pages 399–405, 2015.

[55] K. M. Lee, E. Skoe, N. Kraus, and Richard Ashley. Selective subcortical enhancement of musical intervals in musicians. *The Journal of neuroscience : the official journal of the Society for Neuroscience*, pages 5832—5840, 2009.

[56] Thomas Lidy and Andreas Rauber. Evaluation of feature extractors and psycho-acoustic transformations for music genre classification. In *ISMIR 6th International Conference on Music Information Retrieval, London, UK, 11-15 September*, pages 34–41, 01 2005.

[57] B. Logan and A. Salomon. A music similarity function based on signal analysis. In *IEEE International Conference on Multimedia and Expo, 2001. ICME 2001.*, pages 745–748, 2001.

[58] V. Maffei. Master thesis: Techniques for automatic dissonance suppression in harmonic mixing, Politecnico di Milano, 2016.

[59] Richa Mahajan and Gaganpreet Kaur. Neural networks using genetic algorithms. *International Journal of Computer Applications*, 77(14), 2013.

[60] C. Maçãs, A. Rodrigues, G. Bernardes, and P. Machado. Mixmash: A visualisation system for musical mashup creation. In *2018 22nd International Conference Information Visualisation (IV)*, 2018.

[61] Stephen McAdams and Albert Bregman. Hearing musical streams. *Computer Music Journal*, pages 26–60, 1979.

[62] Martin McKinney, Dirk Moelants, Matthew Davies, and Anssi Klapuri. Evaluation of audio beat tracking and music tempo extraction algorithms. *Journal of New Music Research*, 36:1–16, 03 2007.

[63] Ken'ichi Miyazaki, Yuzuru Hiraga, Mayumi Adachi, Yoshitaka Nakajima, and Minoru Tsuzaki. Virtual pitch and the classification of chords in minor and major keys. *10th International Conference on Music Perception and Cognition (ICMPC)*, 2008.

[64] M. Muller and S. Ewert. Towards timbre-invariant audio features for harmony-based music. *IEEE Transactions on Audio, Speech, and Language Processing*, 2010.

[65] María Navarro-Cáceres, Marcelo Caetano, Gilberto Bernardes, and Leandro Nunes de Castro. Chordais: An assistive system for the generation of chord progressions with an artificial immune system. *Swarm and Evolutionary Computation*, 50, 2019.

[66] Gerhard Nierhaus. Algorithmic composition: Paradigms of automated music generation. *Computer Music Journal*, pages 70–74, 2010.

[67] Elias Pampalk, Simon Dixon, and Gerhard Widmer. On the evaluation of perceptual sfimilarity measures for music. In *Proceedings of the International Conference on Digital Audio Effects*, London, UK, 2003.

[68] Elias Pampalk, Andreas Rauber, and Dieter Merkl. Content-based organization and visualization of music archives. In *Proceedings of the Tenth ACM International Conference on Multimedia*, page 570–579, New York, NY, USA, 2002.

[69] Richard Parncutt and Hans Strasburger. Applying psychoacoustics in composition: Harmonic progressions of non-harmonic sonorities. *Perspectives of New Music*, 32, 1994.

[70] Jouni Paulus and Anssi Klapuri. Measuring the similarity of rhythmic patterns. In *Proceedings of the International Conference on Music Information Retrieval*, 2002.

[71] R. Plomp and W. Levelt. Tonal consonance and critical bandwidth. *The Journal of the Acoustical Society of America*, 38 4:548–60, 1965.

[72] Tim Pohle, Dominik Schnitzer, Markus Schedl, Peter Knees, and Gerhard Widmer. On rhythm and general music similarity. In *Proceedings of the 10th International Society for Music Information Retrieval Conference, ISMIR, Kobe International Conference Center, Kobe, Japan, October 26-30*, pages 525–530. International Society for Music Information Retrieval, 2009.

[73] Miller Puckette. Pure data: another integrated computer music environment. *Proceedings of the 2005 International Computer Music Conference ICMC*, pages 37–41, 1996.

[74] Miguel Pérez Fernández. *Harmonic compatibility for loops in electronic music*. PhD thesis, Universitat Pompeu Fabra, 2020.

[75] António Ramires, G. Bernardes, M. Davies, and X. Serra. TIV.lib: an open-source library for the tonal description of musical audio. *ArXiv*, abs/2008.11529, 2020.

[76] Curtis Roads. *The Computer Music Tutorial*. The MIT Press, 1996.

[77] Bruno Rocha, Aline Honingh, and Niels Bogaards. Segmentation and timbre similarity in electronic dance music. In *Proceedings of the Sound and Music Computing Conference SMC*, pages 754–761, 2013.

[78] Joshua Ross. Types of musical forms (examples, definitions, lists). Available at https://joshuarosspiano.com/musical-forms/, May 2019. Accessed: 2021-01-15.

[79] Dragan Savic. Single-objective vs. multiobjective optimisation for integrated decision support. *Proceedings of the First Biennial Meeting of the International Environmental Modelling and Software Society*, 1:7–12, 01 2002.

[80] Prem Seetharaman and Bryan Pardo. Simultaneous separation and segmentation in layered music. In *Proceedings of the 17th International Society for Music Information Retrieval Conference, ISMIR New York City, United States, August 7-11*, pages 495–501, 2016.

[81] Zhengshan Shi and Gautham J. Mysore. Loopmaker: Automatic creation of music loops from pre-recorded music. In Regan L. Mandryk, Mark Hancock, Mark Perry, and Anna L. Cox, editors, *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. ACM, 2018.

[82] John Shiga. Copy-and-persist: The logic of mash-up culture. *Critical Studies in Media Communication*, 24(2):93–114, 2007.

[83] Dan Simon. *Evolutionary optimization algorithms*. John Wiley & Sons, 2013.

[84] Jordan B. L. Smith and Masataka Goto. Nonnegative tensor factorization for source separation of loops in audio. In *IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2018*. IEEE, 2018.

[85] Jordan B. L. Smith, Yuta Kawasaki, and Masataka Goto. Unmixer: An interface for extracting and remixing loops. In Arthur Flexer, Geoffroy Peeters, Julián Urbano, and Anja Volk, editors, *Proceedings of the 20th International Society for Music Information Retrieval Conference, ISMIR 2019*, 2019.

[86] Stephen Smoliar. Schenker: A computer aid for analysing tonal music. *SIGLASH Newsl.*, 10, 1976.

[87] Sebastian Streich and Bee Suan Ong. A music loop explorer system. In *Proceedings of the 2008 International Computer Music Conference, ICMC Belfast, Ireland, August 24-29*. Michigan Publishing, 2008.

[88] Wu W.L. Su F.C. Design and testing of a genetic algorithm neural network in the assessment of gait patterns, 2000.

[89] Damien Tardieu, Gérard Assayag, Xavier Rodet, and Emmanuel Saint-James. Imitative and generative orchestrations using pre-analysed sounds databases. In *Proceedings of the 6th Sound and Music Computing Conference SMC'06*, pages 115–122, Marseille, France, May 2006.

[90] Damien Tardieu and Xavier Rodet. An instrument timbre model for computer aided orchestration. In *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, pages 347 – 350, 11 2007.

[91] Godfried Toussaint. A comparison of rhythmic dissimilarity measures. In *5th International Conference on Music Information Retrieval ISMIR 2004, Barcelona, Spain, October 10-14*, volume 21, pages 129–149, 01 2006.

[92] Godfried T. Toussaint. *The Geometry of Musical Rhythm What Makes a "Good" Rhythm Good?* Chapman and Hall/CRC, 2nd edition, 2019.

[93] K. Tsuzuki, Tomoyasu Nakano, M. Goto, T. Yamada, and S. Makino. Unisoner: An interactive interface for derivative chorus creation from various singing voices on the web. In *Proceedings of the International Computer Music Conference*, 2014.

[94] G. Tzanetakis and P. Cook. Musical genre classification of audio signals. *IEEE Transactions on Speech and Audio Processing*, 10(5):293–302, 2002.

[95] Yakov Vorobyev. Harmonic mixing guide. https://mixedinkey.com/harmonic-mixing-guide, 2020. last accessed on 12/03/21.

[96] Rodney Waschka. Composing with genetic algorithms: Gendash. In Eduardo Reck Miranda and John Al Biles, editors, *Evolutionary Computer Music*. Springer London, 2007.

[97] Ka-Chun Wong, Chun-Ho Wu, Ricky K.P. Mok, Chengbin Peng, and Zhaolei Zhang. Evolutionary multimodal optimization using the principle of locality. *Information Sciences*, 194:138–170, 2012.

[98] H. H. Wu and J. Bello. Audio-based music visualization for music structure analysis. In *Proceedings of the 7th Sound and Music Computing Conference, SMC*, page 11. Sound and music Computing Network, 2010.

[99] Yijun Zhou, Yuki Koyama, Masataka Goto, and Takeo Igarashi. Generative melody composition with human-in-the-loop bayesian optimization. *Computing Research Repository CoRR*, 2020.