Multi-scale and multi-dimensional modelling of music structure using polytopic graphs

Corentin Louboutin

UNIVERSITE BRETAGNE LOIRE MATHSTIC

UNIVERSITÉ DE RENNES 1

THÈSE DE DOCTORAT DE

L'UNIVERSITE DE RENNES 1
COMUE **U**NIVERSITE **B**RETAGNE **L**OIRE

Ecole Doctorale N°601
Mathèmatique et Sciences et Technologies de l'Information et de la Communication
Spécialité : *Informatique*

Par

« Corentin LOUBOUTIN »

« Modélisation Multi-Echelle et Multi-Dimensionnelle de la Structure Musicale par Graphes Polytopiques »

«Multi-Scale and Multi-Dimensional Modelling of Musical Structure with Polytopic Graphs»

Thèse présentée et soutenue à UNIVERSITÉ DE RENNES 1, IRISA (UMR 6074), le 13 mars 2019
Unité de recherche : 6074 (IRISA)

Rapporteurs avant soutenance :
Florence Levé, Maître de conférence, UFR des Sciences Université de Picardie Jules Verne/ MIS, Amiens
Gérard Assayag, Directeur de recherche, IRCAM/CNRS-STMS, Paris

Composition du jury :

Président :	Emmanuel Vincent, Directeur de recherche, INRIA, Nancy Grand-Est
Examinateurs :	Florence Levé, Maître de conférence, UFR des Sciences Université de Picardie Jules Verne/ MIS, Amiens
	Gérard Assayag, Directeur de recherche, IRCAM/CNRS-STMS, Paris
	Matthew Davies , Senior Researcher, INESC TEC, Porto
	Geraint Wiggins, Professor of Computational Creativity, Vrije Universiteit Brussel / Queen Mary University of London
	Frédéric Bimbot, Directeur de recherche CNRS, IRISA (UMR 6074)
Dir. de thèse :	Frédéric Bimbot, Directeur de recherche CNRS, IRISA (UMR 6074)

Résumé

Le développement massif de données numériques a permis l'essor du domaine de la Recherche d'Information Musicale (MIR). Les travaux au sein de cette communauté visent à développer des systèmes informatiques capables de classifier, recommander, analyser, composer des morceaux de musique. Afin de réaliser ces tâches, les méthodes actuelles se basent sur des modèles d'observation de l'inter-dépendance des éléments musicaux locaux tels que les notes ou les accords.

Si certains de ces modèles prennent en compte la structure des morceaux à analyser, très peu rendent compte de cette organisation à des échelles intermédiaires telle que la phrase musicale ou au-delà. Il s'avère donc nécessaire de modéliser l'organisation interconnectant les éléments musicaux formant des segments, à savoir décrire comment ces éléments sont reliés les uns aux autres à différentes échelles afin de former un tout.

Qu'est-ce qu'une section musical ? Plusieurs échelles coexistent dans la musique : au niveau élémentaire, il y a les éléments de base tels que les notes, les accords. Ensemble, ces éléments forment des rythmes, des mélodies, des enchaînements qui une fois assemblés forment des motifs qui pourront être répétés, modifiés ou mélangés avec d'autres motifs pour à leur tour former des phrases et des sections.

Dans le cas de la pop, ces sections correspondent souvent à des couplets, des refrains, des ponts, ... La durée typique d'une section est d'environ 15 secondes[1], mais cette durée peut être allongée ou raccourcie en fonction du tempo. Le but est alors de comprendre quelles sont les relations importantes entre les notes, les motifs, les phrases au sein d'une section. L'hypothèse centrale de ce travail est que la structure d'une section musical s'appuie sur un système d'implication dont l'attente est plus ou moins fortement infirmée.

Comment crée-t-on un système d'implication et quelles sont les principales structures qui en découlent ? Généralisant le principe d'implication développé par certaines théories musicologiques (Narmour en étant la figure de proue), le modèle Système&Contraste (S&C) a été conçu afin de définir les propriétés d'organisation permettant de caractériser les sections en tant qu'unités structurées autonomes au sein d'un morceau. La structure d'uns section est alors déterminée par les relations de similarité, d'analogie ou de transformation entre les éléments de base qui le constituent.

[1] Conformément à l'échelle de base pour la description de la structure musicale ("form") telle que définie par Bob Snyder concernant les « trois niveaux d'expérience musicale » [Snyder and Snyder, 2000]

Une formalisation simple du modèle S&C permet de rendre compte de la structure d'une section formée de quatre éléments : face à la succession des trois premiers éléments et leurs similarités, il est possible de créer un système logique simple suscitant une attente. Le contraste apparaît alors comme la surprise que peut apporter le dernier élément vis-à-vis de cette attente et contribue à la délimitation de la section musical en marquant la conclusion du schéma d'implication.

Prenons un exemple hors du registre musical. Si l'on considère la suite, 1-2-3-7, les trois premiers éléments 1-2-3, créent une attente (4) du fait de la relation simple qui les relie (+1). Cependant, la dernière observation (7) vient contraster avec cette attente et crée une surprise. Celle-ci peut varier en intensité, suivant le nombre de propriétés intervenant. Par exemple 7 sera moins contrastif que **A**, qui surprend par de nombreux aspects (lettre, taille, gras). Dans le modèle S&C, appliqué à la musique, le principe est le même, transposé aux propriétés musicales.

Toutefois, pour la musique, l'attente est beaucoup plus compliquée à décrire du fait de la coexistence de nombreuses dimensions musicales : chaque événement de base possède un nombre important de caractéristiques (rythme, mélodie, harmonie, timbre, paroles...) et décrire les relations pour toutes ces dimensions est un enjeu considérable. Une des contributions de ma thèse consiste à proposer et à tester des formalismes capables de décrire les relations entre accords, rythmes et mélodies.

En outre, la création de l'attente dans le cas d'un système de quatre éléments n'est pas toujours le résultat d'un processus linéaire incrémental. Du fait de la multitude de dimensions pouvant intervenir, la création de l'attente peut résulter de relations qui ne sont pas systématiquement séquentielles. Ainsi, s'intéresser à la relation entre le premier et le troisième élément du système peut considérablement simplifier le processus cognitif permettant de visualiser l'attente. Par exemple, si l'on considère la section représentée en Figure 1, en considérant une projection purement logique des trois premiers éléments, le lecteur pourrait attendre une mélodie identique au deuxième élément du fait de la relation d'identité entre la première mélodie et la troisième. Ainsi, la description d'une section à l'aide du modèle S&C, du fait des relations qu'il fait intervenir, n'est plus linéaire mais matricielle.

Le modèle S&C permet de rendre compte de la structure de très nombreuses sections musicales. Il est particulièrement approprié dans le cas de la description de l'attente et de la surprise pour les séquences du type A-B-A'-C ou A-A'-B-C. Ces structures sont très récurrentes dans la musique pop (par exemple le refrain de Waka Waka de Shakira ou de la Macarena), le jazz (Beautiful Love), ou même la musique classique (avec par exemple la forme antécédent-conséquent).

Cette modélisation permet par ailleurs de décrire les relations à plusieurs échelles au sein d'une section, ce qui a donné lieu à la formalisation du modèle polytopique décrit ci-après.

Comment les procédés structurels s'établissent-ils sur plusieurs échelles simultanément ? Si de nombreuses sections musicales peuvent se décomposer en quatre éléments (carrure), il est nécessaire, pour considérer l'échelle où les éléments de base

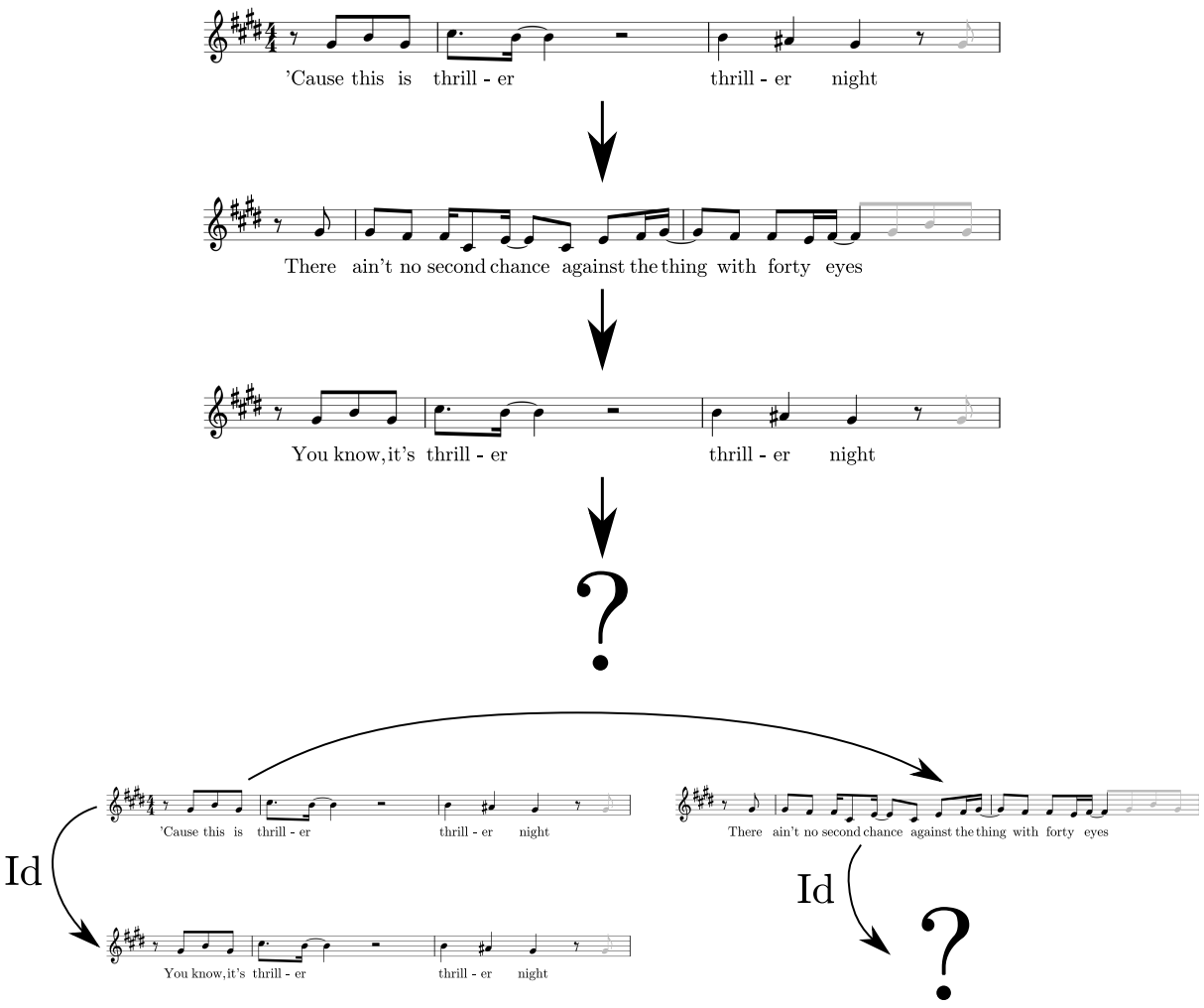


Figure 1: **Michael Jackson - Thriller** (comp.: Rod Temmperton) Thriller, EMI 1982. Timing: 2'26-2'40. *"Thriller", pp. 25-26, published by Rodsongs (PRS), 1982*

apparaissent, de le subdiviser en seize. Une partie de mes travaux de thèse s'est donc concentrée sur la généralisation du modèle S&C pour que celui-ci puisse décrire la structure d'une section à partir des relations entre ces seize éléments. Cette généralisation s'est faite par l'utilisation du modèle S&C à plusieurs échelles : PGLR (Polytopic Graph of Latent Relations).

Dans le PGLR, l'organisation entre les éléments de base est décrite à l'aide de plusieurs sous-systèmes de quatre éléments, qui sont assemblés pour former un autre système à une échelle supérieure. L'utilisation de plusieurs échelles associée à l'utilisation du modèle S&C permet alors d'organiser les éléments en les plaçant sur un polytope, notamment un n-cube (carré, cube, tesseract, ...), ce qui rend possible la construction d'un graphe de dépendances entre les éléments de base formant la section.

En effet, l'utilisation du modèle S&C permet de passer d'une description linéaire du contenu musical à une description matricielle. Ainsi, le graphe de dépendances associé à un S&C s'inscrit dans un carré. C'est l'association de ces carrés à plusieurs échelles qui induit un n-cube. Pour décrire un section de seize éléments, quatre sous-systèmes de

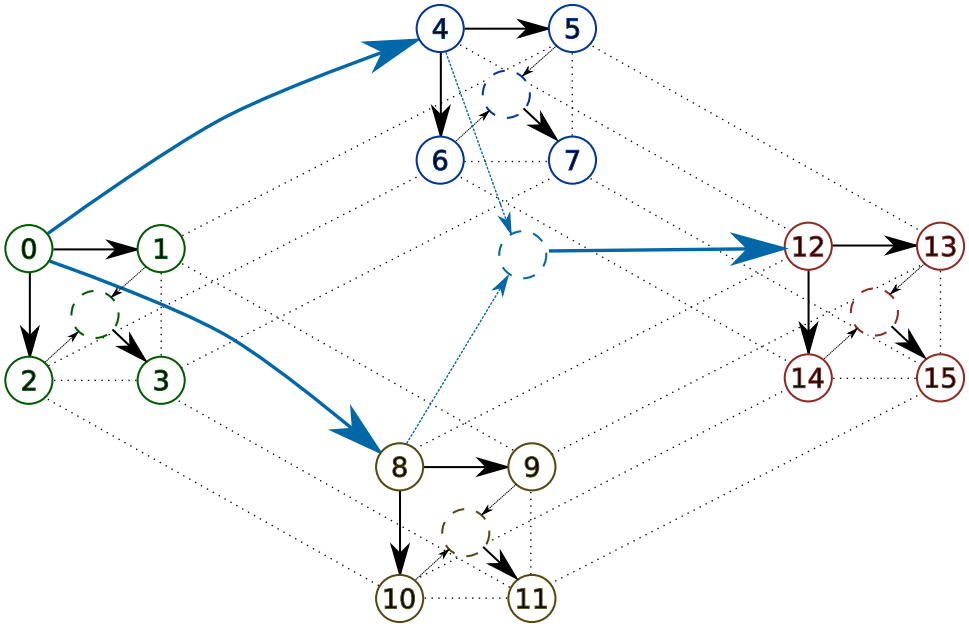


Figure 2: Graphe de dépendances induit par la description d'un segment de 16 éléments à l'aide d'un PGLR associé au modèle S&C.

quatre éléments chacun sont assemblés en considérant le système formé par les quatre premiers éléments de chaque sous-système. La Figure 2 donne une illustration du graphe de dépendances correspondant à une telle description.

Ces concepts se relient à des principes de la théorie de l'information tels que la « complexité » et permettent de générer de la musique. Du fait de la géométrie du polytope, un élément est susceptible d'avoir plusieurs dépendances. Ou, pris sous un autre angle, du fait de la multiplicité du nombre de carrés dans un tesseract, il existe plusieurs manières de choisir les quatre sous-systèmes. Cela a pour conséquence de multiplier le nombre de descriptions possibles de la structure d'une section. De même, il existe de nombreuses manières de représenter une relation entre deux éléments. Il faut donc faire un choix, c'est à dire une optimisation ! Ici, le critère utilisé pour guider cette optimisation est issu de la théorie de l'information : la complexité.

L'idée latente à ce concept est que la meilleure description est celle qui est la plus simple. Une description est simple s'il suffit de peu d'information pour la représenter. Ainsi, pour chaque possibilité de structure de dépendance et pour chaque relation, on associe un coût. Il suffit ensuite de choisir la description qui minimise la somme de tous ces coûts.

Le concept de complexité se rattache aux principes de longueur minimale de description ou complexité de Kolmogorov, qui désignent la taille du programme le plus simple capable de générer une section musicale (au sein d'une famille P de programmes prédéfinis). Par conséquent, en recherchant la description la plus simple d'une section, on cherche aussi à trouver un programme capable de générer cette section.

Étude expérimentale, résultats Pour évaluer la pertinence du modèle et sa capacité à décrire efficacement la structure de sections musicales, nous avons comparé ses performances à un modèle séquentiel. Cette comparaison s'est principalement faite sur une tâche : la prédiction de sections. Le corpus utilisé pour cette tâche est issu du jeu de données RWC POP (pour Real World Computing Popular Music Database), couramment utilisé dans le domaine du MIR du fait du grand nombre d'informations qui y sont stoquées. De ces données ont été extraites des séquences de 16 accords ainsi que les mélodies durant 4, 8 ou 16 mesures. La mesure utilisée pour l'évaluation des différents modèles est une mesure statistique, dérivée de la vraisemblance, la perplexité. Cette mesure permet d'évaluer le nombre moyen de branchements possibles à partir d'un élément. Ainsi, plus ce nombre est faible, plus le modèle associé est capable de prédire efficacement une section.

Que ce soit pour la prédiction de séquences d'accords, ou la prédiction de segments rythmiques et mélodiques, le modèle polytopique s'avère toujours plus performant que le modèle séquentiel. De plus, le modèle combinant les multiples possibilités de graphes de dépendances permet d'améliorer drastiquement les performances du modèle polytopique. L'utilisation du modèle S&C et notamment de la modélisation de l'attente sous la forme d'un élément virtuel est particulièrement intéressant pour la prédiction de séquences d'accords.

Discussions et ouvertures Cette thèse formalise et généralise les concepts latents du modèle Système & Contraste. Le modèle présenté, Polytopic Graph of Latent Relations, permet de rendre compte de dépendances à plusieurs échelles simultanément et utilise ces relations entre les éléments sur plusieurs dimensions pour caractériser une attente. Le contraste entre cette attente et les éléments réellement observés génère ensuite une surprise.

Ce modèle aura permis de mettre en évidence des mécanismes structurels importants, qui utilisés à bon escient seront susceptibles d'améliorer grandement les performances sur des tâches telles que la prédiction ou la génération de contenu musical. Toutefois, il est de nombreux aspects du modèle qui peuvent encore être développés. En effet, dans ce manuscrit, le modèle S&C n'est généralisé qu'aux sections pouvant être découpées en seize éléments. Il pourrait être intéressant par la suite d'élargir le champ de recherche pour encapsuler les sections dont les tailles ne seraient pas aussi simples. De plus, si un certain nombre de formalismes ont été proposés pour décrire les relations entre les éléments musicaux (accords, rythmes, mélodies), il peut être intéressant de développer ces formalismes pour intégrer d'autres dimensions, ou incorporer des connaissances musicales permettant d'interpréter plus facilement les descriptions faites à l'aide du modèle.

Le modèle Système & Contrast est un modèle encore très récent mais qui montre un très grand potentiel et qui apportera très certainement beaucoup, à la fois au domaine du MIR et à la musicologie.

Contents

1 Introduction **13**

1.1 Context and Focus 13

1.2 Outline of the thesis 16

2 Music Structure Analysis **19**

2.1 Structure and Scale 19

2.2 Information and Expectation 23

2.3 System & Contrast Model 25

2.3.1 Intuitive and General Presentation 25

2.3.2 Musical Examples 29

3 Multi-scale System & Contrast **35**

3.1 Square Formalisation 35

3.2 Multi-scale Analysis 36

3.2.1 Polytopic Graph of Latent Relations 37

3.2.2 Primer Preserving Permutations 38

3.3 Polytopic Model Specifications 41

3.3.1 Antecedent Functions 41

3.3.2 Relations 41

3.3.3 Sequential Model 42

3.3.4 Tree Systemic Model 43

3.3.5 Static S&C Model 44

3.3.6 Dynamic S&C Model 46

3.3.7 Relational Static S&C Model 47

4 Musical Data and Evaluation Method **49**

4.1 Examples of Multi-Scale Descriptions 49

4.2 Corpus Creation . . . 53
4.2.1 RWC POP . . . 53
4.2.2 Chord Annotation . . . 54
4.2.3 Creation of the New Corpus . . . 56
4.2.4 Test Corpus . . . 57
4.3 Evaluation Methodology . . . 59

5 Application to Chord Sequences **63**
5.1 Context and Practical Considerations . . . 63
5.1.1 Harmony . . . 63
5.1.2 Chord Representations . . . 64
5.2 Relation Formalisms . . . 65
5.2.1 Triad Circles . . . 65
5.2.2 Optimal Transport . . . 67
5.2.3 Musicologically-constrained Optimal Transport . . . 69
5.2.4 Multi-scale Generalisation . . . 72
5.3 Results . . . 72
5.3.1 Benefit of Multi-Scale Organisations . . . 73
5.3.2 Impact of the Representation . . . 75
5.3.3 Role and Importance of the Virtual Chord . . . 76
5.3.4 Relational Static S&C Model Performances . . . 78
5.3.5 Additional Observations and Considerations . . . 79

6 Rhythm and Melody **83**
6.1 Context and Practical Considerations . . . 83
6.2 Relation Formalisms . . . 86
6.2.1 Rhythmic Relations . . . 86
6.2.2 Melody Relations . . . 91
6.3 Time Alignment Optimisation . . . 93
6.4 Results . . . 96
6.4.1 Rhythm Modelling . . . 96
6.4.2 Melody modelling . . . 104
6.4.3 Summary . . . 108

7 Conclusion and Perspectives **109**

7.1 Contributions of this work 109

7.2 Extensions and perspectives 111

7.2.1 Improving the model 111

7.2.2 Music Cognition 113

7.2.3 Music Generation 113

7.2.4 Broadening the scope 115

7.2.5 Final Words 116

Chapter 1

Introduction

1.1 Context and Focus

It is quite common sense that listeners do not perceive music only as a mere sequence of sounds, nor do composers conceive their works as such. Music is essentially the result of patterns of which inner organisation and mutual relationships participate to the overall *structure* of the musical content, at different time-scales simultaneously.

However, what exactly is music structure remains an open scientific question. Considering the definition given by the Merriam-Webster dictionary, structure is "*the aggregate of elements of an entity in their relationships to each other*". Indeed, the structure of a whole is highly determined by the *relation* between its elements.

One of the most frequent attributes of music is that it contains *redundancies*, in the form of repetitions or similarities. These redundancies are organised in such a way that they create *expectations*. And ultimately these expectations may or may not be denied to create a *surprise*. Indeed, the expectation arising at some point in time from listening to a musical passage is built on the feeling that "something is going to happen". The surprise is then "*the feeling caused by something unexpected or unusual*" (Merriam-Webster) that actually happens then, typically by a more or less strong denial of the expectation created upstream.

Therefore, defining the structure of a musical group requires to describe the organisation in time of the elements that constitute that group, their relations to one another, and how these relations create a flow of information. The present work assumes that redundancy, expectation and surprise are factors that are essential to describe this flow of information, and therefore the structure. Movements, sections, phrases, motives, are groups at various scales. In this thesis, the focus is put on the structure of *sections* (i.e., a special type of segment at a particular general temporal scale that also happens to be a group).

Multiple questions then arise concerning the framework in which the actual structure of a large-scale musical object can be described:

- *What are the elements that constitute the musical object?* Both for audio and sym-

bolic representations of music, there are multiple elements that can be used to describe what happens at the elementary level of musical events: Mel Frequency Cepstral Coefficients (MFCC), onset time, tristimulus harmonic energy ratio, sharpness, spread, skewness, etc. as audio descriptors, and notes, chords, rhythms, metric, instruments, lyrics, etc. for symbolic representations. As both a *multi-dimensional* and *multi-scale* process, how can we connect and articulate these low-level musical elements with the structural description of a large-scale musical object?

- *What are the essential dependencies between the elements forming the large-scale musical object?* Potentially, all the elements constituting a musical object may be considered to have relationships with one another. However, some relations may have more importance than others, from a structural point of view. Therefore, if we assume that there is an underlying graph structure behind the relationships within a musical object, there is a need to define a subset of these relationships, i.e. a graph *topology*, as representing the primary dependencies on which rests the core of the structural description.

- *How do we describe the relationship between two elements?* Defining the topology of the dependency graph between the elements is not sufficient, as it provides no information about the actual relationships between the elements. As a consequence, there is a need to find some framework that can be used to describe formally the musical or perceptive *relations* between two musical elements. Are these elements identical? Similar? If so, in which manner? Are they totally different? Are they more different or less different than two other elements? What properties should these relations possess and how does this impact the structural organisation of the larger musical object?

In this work, we consider the large-scale musical object as being a symbolic representation of a *section*, considering that sections are musical segments which form macroscopic constituents of the global piece. In pop songs, which will be the main focus of the present work, sections usually correspond to a segment such as a chorus or a verse, lasting approximately 15 seconds and exhibiting a clear beginning and end [Snyder and Snyder, 2000]. As we deal with symbolic music data, we will consider that basic low-level elements are notes, with a pitch and an onset, themselves forming aggregate of notes, or chords at the next level. The aim is then to describe the dependencies between these elements, their relationships with one another, and how these relationships are organised as regards the structure of the whole section.

The musical content observed at a given instant t within a music section obviously tends to share privileged relationships with its immediate past, hence the sequential perception of the music flow and the "natural" inclination towards modelling the music flow with chain-rule dependencies. But music content at instant t also relates with distant events which have occurred in the longer term past, especially at instants which are metrically homologous to t, in previous bars, motifs, phrases, etc. This is particularly evident in strongly "patterned" music, such as pop music, where recurrence and regularity play a central role in the design of cyclic musical repetitions, anticipations and surprises. But it is also discernible in a number of other music genres, which rely abundantly on all

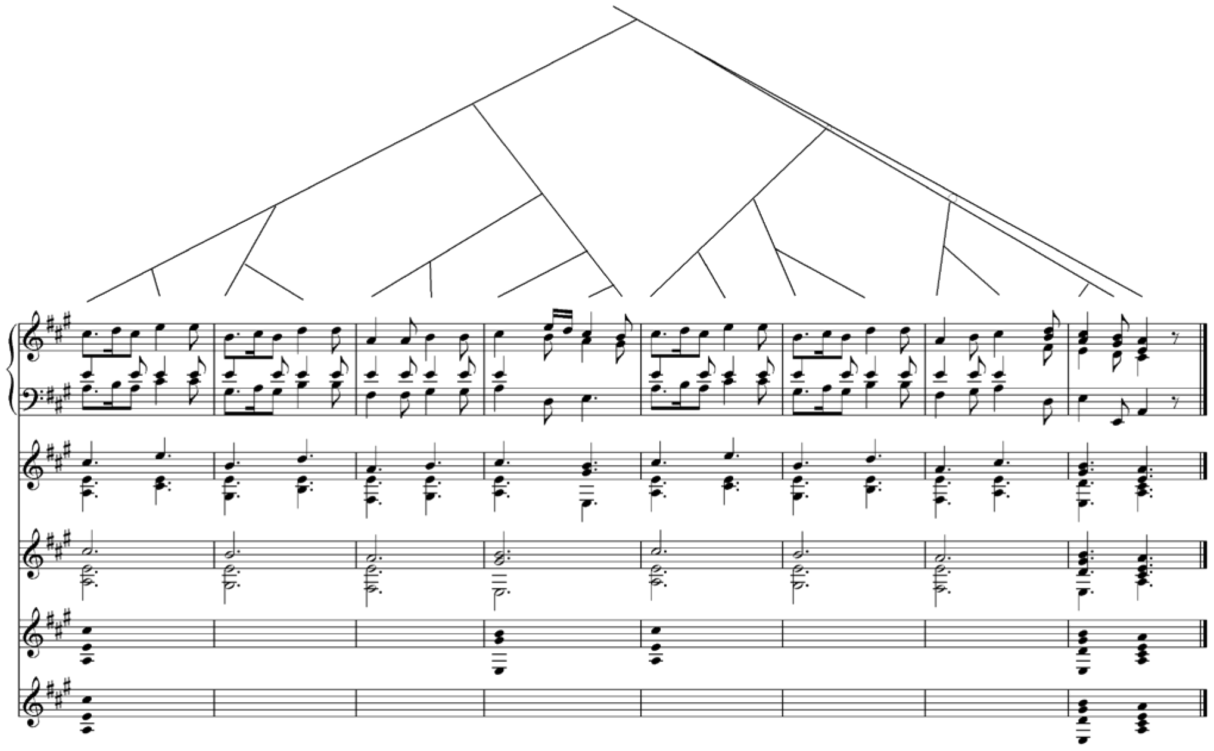


Figure 1.1: Example of structure description of the theme of the first movement of Mozart's piano sonata in A major, K.331, using Lerdahl and Jackendoff's model.

sorts of similarities, progressions, expectations and denials, simultaneously at different scales.

To overcome the limitations of purely sequential models in music content descriptions, hierarchical models are often resorted to, in order to provide a representation framework for the grouping structure of a musical passage.

The most famous hierarchical approach is undoubtedly the Generative Theory of Tonal Music (GTTM) by Lerdahl and Jackendoff [Lerdahl and Jackendoff, 1985], which has been for many years a source of inspiration for a wide variety of work in music structure modelisation. However, hierarchical approaches such as GTTM rely axiomatically on an adjacency hypothesis, under which the grouping of elements into a higher level object is only possible *via* neighbouring units. This can be seen on Figure 1.1 by the fact that none of the branches in the GTTM tree cross another one. The reference to "formal syntactic theories of sequential structure in music" recently made by Pearce and Rohrmeier [Pearce and Rohrmeier, 2018] shows that this hypothesis of a sequential structure of the music remains strongly rooted in the MIR community.

The work presented here is based on a non-sequential approach, called *Polytopic Graph of Latent Relations* (PGLR). Under this scheme, relationships between musical elements within a musical section are assumed to be developing predominantly between homologous elements within the metrical grid at different scales simultaneously. This approach generalises to the multi-scale case the System&Contrast framework which aims at describing the logical system of expectation within a section and the surprise resulting

from that expectation, as a 2×2 square matrix (i.e. a particular case of 2-dimensional polygon structure).

For regular sections of 2^n events, the PGLR lives on a n-dimensional cube (square, cube, tesseract, etc...), n being the number of scales considered simultaneously in the multi-scale model. By extension, the PGLR can be generalised to other regular (or slightly irregular) n-polytopes.

Each vertex in the polytope corresponds to a low-scale musical element, each edge represents a relationship between two vertices and each face forms an elementary system of relationships. In addition, the last vertex in each elementary system can be viewed as the denied realisation of a (virtual) expected element, itself resulting from the implication triggered by the combination of former relationships within the system. It is important to understand that a system may be formed from non-adjacent elements in the polytope, implying a non-sequential description of the structure.

The estimation of the PGLR structure of a musical section can then be obtained computationally as the joint estimation of:

1. the description of the polytope (as a more or less regular n-polytope)

2. the nesting configuration of the graph over the polytope, reflecting the flow of dependencies and interactions as elementary implication systems within the musical section (this flow being assumed to be time-wise causal - but not necessarily sequential).

3. the set of relations between the nodes of the graph, with potentially multiple possibilities which need to be disambiguated (hence the "latent" nature of the relations, as they are not actually observed).

The aim of the PGLR model is to both describe the time dependencies between the elements of a section and to model the logical expectation and surprise that can be built on the observation and perception of the similarities and differences between elements with strong relationships. The description of the structure can be related to a compression problem in the sense that it tries to compact the information of a section by inferring structural information. As such, the PGLR formalism can be seen both as some sort of cognitive model and as a compression scheme.

1.2 Outline of the thesis

In Chapter 2, we connect the work of this thesis to other works that have been done by the Music Information Retrieval (MIR) community. This first chapter introduces the main concepts that are at the base of this study, such as music structure, expectation or the Minimum Description Length (MDL) principle, along with the different approaches focusing on these concepts that exist in the MIR domain. This chapter also serves as an introduction to the System&Contrast (S&C) model.

Chapter 3 formalises the S&C model and presents its generalisation to multi-scale structure description: the PGLR model. This chapter develops different computational approaches that implement the PGLR model as well as the corresponding algorithmic designs. It introduces the key concepts of *antecedent function* and (latent) *relation*.

The subsequent chapter, Chapter 4, provides some examples of descriptions and analyses that can be done using the PGLR model presented in Chapter 3, as well as details about the creation of a corpus used for quantitative evaluations. Moreover, it defines the performance measure that is used for these evaluations, namely the *perplexity*, an objective measure to compare the prediction (and compression) ability of the different models.

Generalising the proposed concept to multiple musical dimensions, Chapter 5 and Chapter 6, provide different formalisms for the description of relationships between harmonic, rhythmic or melodic data. For each dimension, extensive results are given to describe the behaviour of PGLR models.

Finally, Chapter 7 summarises the contributions of this thesis and indicates a number of directions which could be explored to further improve and exploit the present work.

Chapter 2

Music Structure Analysis

2.1 Structure and Scale

The domain of Musical Information Retrieval (MIR) is vast, and its community aims at providing tools and computing systems which are able to recommend [Celma, 2010, Van den Oord et al., 2013, Flexer and Stevens, 2018], classify songs by genre and geographical origin [McKay et al., 2010, Conklin, 2013, Velarde et al., 2018], or compose and generate new musical pieces [Conklin, 2003, Bresson et al., 2010, Nika et al., 2016, Pachet et al., 2017]... All these tasks (and many more) are not easy to solve, as musical pieces can be described using a multitude of features, such as notes, chords, patterns or various properties of an audio file, as well as lyrics, or even features related to their usage, such as purchase statistics, or number of listeners. Therefore to achieve these tasks, there is a need for models that can integrate and organise the information contained and conveyed by music pieces at many levels.

Considering both the audio and the symbolic domains in MIR, researchers begin to have a good understanding of how the various dimensions involved in music can be used to improve MIR results, and structural information is often mentioned as being of great interest to describe music content. However, there is no general definition of what exactly is the *structure* of music, i.e. what is the exact nature of the information which can be used to describe the organisation of a musical content. Therefore, the awareness of the difficulty to give a clear definition or formalisation of music structure has grown in the MIR community, to try to bridge the gap with the musicological conceptions of music structure that are usually described in treaties and books about musical form, mostly focused on classical music (for example the sonata form [Hepokoski and Darcy, 2006] or the rondo form [Clercx, 1935]...)

While there are a number of models that aim at describing the structure of a global piece, there are only a few ones dedicated to the modelling of the musical structure of a *section* [Snyder and Snyder, 2000], i.e. the structural unit which lies just at the level under the global structure of the piece — a section can also be referred as a phrase [Jusczyk and Krumhansl, 1993], a sentence [Schoenberg et al., 1967, Caplin, 1998] or a period [Schoenberg et al., 1967, Caplin, 1998, Monelle, 2014]. In the present work,

sections will generally correspond to a passage that is 8-bar long and that can be considered as having a rather well-defined beginning and end. For example, a chorus, a verse or a bridge in typical pop-music songs, which makes it possible to loop or to delete (when remixing a song, for instance) [Bimbot et al., 2010].

If the notion of section is very common in the MIR community (enough at least to create a sub-domain which focuses on the segmentation of music pieces to find such sections), it is interesting to note that most of the segmentation methods are based on sequential approaches, in the sense that they describe the structure of a section using adjacency relationships between elements [Pearce and Rohrmeier, 2018]. But in most cases, they do not use explicit syntactic or procedural descriptions of the structural information of the sections. Analysers tend to consider the structure of a section as being a set of properties, some of them being predominant [Sloboda, 1991, Smith and Chew, 2013], but most of the segmentation methods achieve a clustering of the song in multiple sections which can be then characterised, manually or not, with some labels.

Considering the segmentation or "chunking" [Wiering et al., 2009] problem which is also referred to as structure discovery problem, one can distinguish two main families of algorithms: the *bottom-up methods* which consist in grouping lower-scale elements until they form a segment with good properties, or the *top-down methods* which consist in splitting the musical piece into successive slices. The first family includes methods such as the one named Temporal Gestalt units, designed by Tenney et al. [Tenney and Polansky, 1980] and implemented afterwards by Eerola et al. [Eerola and Toiviainen, 2004], which consists in considering a distance between each pair of successive notes of a melody and creating a boundary once the sum of the distances inside a group of notes reaches a threshold. This method is particularly interesting because it works at two scales: at the basic scale, sets of notes form a "clang" [Tenney and Polansky, 1980], and then, lists of clangs form a segments (sections or sequences). However the process of combining notes in clangs and then clangs together is still a sequential process especially since the two processes are not done simultaneously but sequentially. Some improvements of this method were proposed by considering other measures to compute the distance between two notes [Cambouropoulos, 2001, Cambouropoulos, 2006, Temperley, 2004], or derivations where chunking methods improve segment boundaries detection by analysing the behaviour of the notes at these boundaries [Chang et al., 2004, Ferrand et al., 2003].

Within the family of top-down methods, probabilistic methods such as the one used by [Bod, 2002] and [Juhász, 2004] dominate the field. These are based on Markovian models which are used to detect the boundaries of the segment in a piece given some probabilistic rules that are learned on an annotated corpus.

Among these probabilistic methods the information dynamics of music (IDyOM) model [Abdallah and Plumbley, 2009, Pearce and Wiggins, 2012, Sears et al., 2018] occupies a central role. Based on a multiple viewpoint model [Conklin and Witten, 1995] of music data, they are based on statistical n-grams to model the flow of information in music. An entropic criterion is used to infer segment boundaries following the idea that sections borders happen where expectations are low [Narmour, 1992] i.e. the uncertainty of the prediction is high.

Another method, designed by Ferrand et al. [Ferrand et al., 2003] consists in using a

sliding window of finite memory to compute a density score over a melody based on the intervals encountered if they also appear in the memory window. Pattern matching methods, both in audio [Maddage et al., 2004] and in symbolic analysis [Cambouropoulos, 2006], must also be mentioned. The latter, for example, considers the repeated sequences inside a piece for computing a score over the piece, given the degree of overlapping, duration and frequency of repeated patterns. Then, they use the local maxima to detect segments boundaries. The interest of this method is that it provides a bit more information than the boundaries of the segments, as it can be used to tell which are the segments that are similar. However, even these approaches are mainly based on a sequential point of view of the musical content, and do not provide much insight on the structure of the segment itself. In fact, these pattern-matching algorithms are used to compute a score for each possible onset at every point on the time scale, and then, the segment boundaries are detected by finding the local optima of this score.

As an alternative, there also exist a number of hierarchical or multi-scale models for musical structure. That is the case for the segmentation method presented by Tenney et al. [Tenney and Polansky, 1980], and if we focus on studies dedicated to the description of the harmonic structure, it is possible to find a larger family of methods that are based on a hierarchical description of the musical content.

In fact, as the harmony is considered as one of the most important musical dimension in classical music, many musicologists have been working on systems or methods to describe the structure of chord progressions inside a piece. Using the tools developed in Linguistic Sciences and in the MIR domain, these methods build some sort of grammatical models to describe the structure of a piece. For example, de Haas et al. [de Haas et al., 2009, de Haas et al., 2011, De Haas et al., 2013] as well as Rohrmeier et al. [Rohrmeier, 2007, Rohrmeier, 2011] who use a grammar model based on standard musical progression to describe the structure of a chord sequence. The structure can then be represented as a hierarchical graph, such as the tree shown on Figure 2.1. Using the same principle, Steedman [Steedman, 1996] also builds a grammar to describe the structure of a large variety of chord progressions that would occur in blues music. Recently Deguernel et al. [Déguernel et al., 2017] used a similar model to describe the harmonic structure of a piece that is then applied to machine improvisation.

The problem with such methods is that they are usually based on rules taken from standard musicology, which tend to be very specific to a given music genre ("classical" music, blues, jazz) and/or over-focused to some particular set of conventions. Moreover, the rules used in the grammar are strongly related to musicological descriptions of harmony, and are therefore not prone to be generalised to other musical dimensions. For example, it would be rather difficult to describe a rhythmic section on the basis of its evolution from the dominant to the tonic !

Recently, Guichaoua [Guichaoua, 2017], has introduced a new segmentation method for the analysis of chord sequences. This approach considers different structural descriptions for sections of various lengths. The structure model behind his works is based on the construction of systems of sub-sequences. To build these systems, non-sequential dependencies between chords are exploited to encode binarily the compliance (vs non-compliance) of their logical relationship. As is the case for the present work, the model

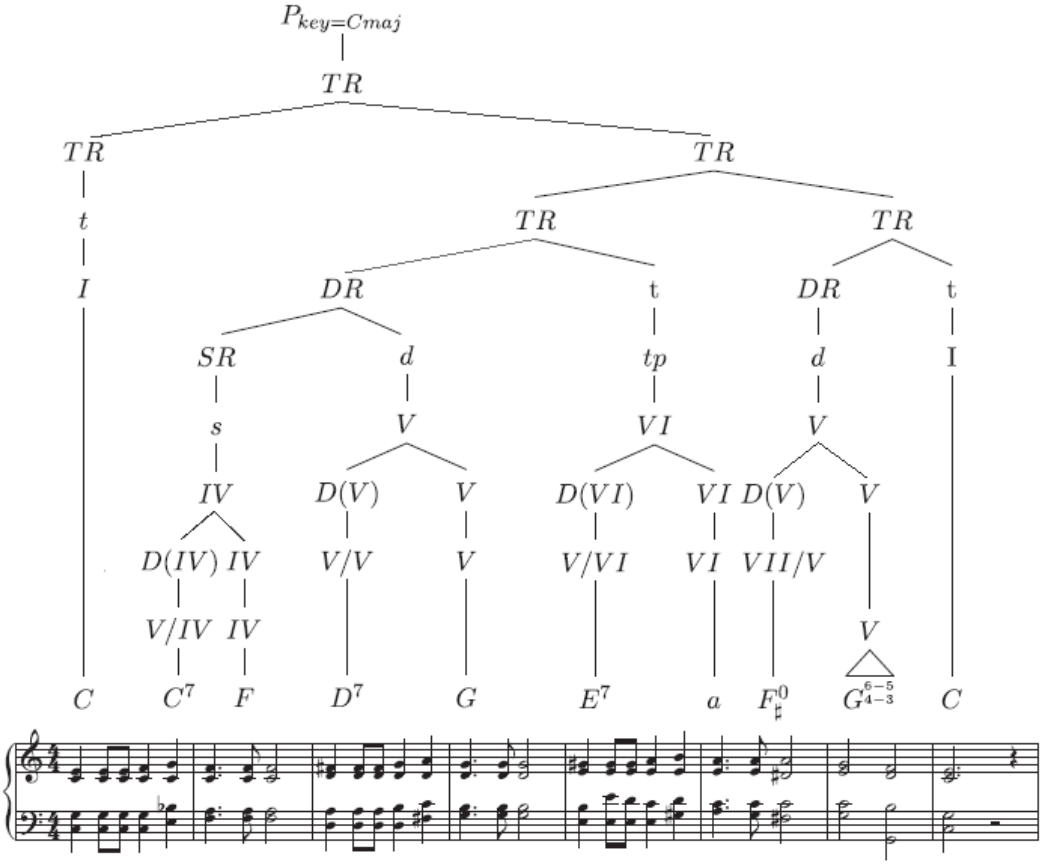


Figure 2.1: Rohrmeier's hierarchical description of the structure of a chord sequence using a syntactic grammar based on the tonal function and degree of each chord.

used to describe these systems of chords, is the System&Contrast model that will be described further in Section 2.3.

Guichaoua's algorithm aims at splitting the whole piece in segments that can be described using this framework. As there may be multiple possible descriptions for a section or multiple possibilities of segmentation of the whole piece, for each section and associated description model, a complexity score is computed. These scores are then summed to compute a complexity description score for the whole song. Following Ockham's Razor principle, the segmentation that is ultimately chosen is the segmentation with minimal complexity score. As a by-product of this algorithm, it is possible to describe the structure of each section using the models associated with complexity score calculation.

Stemming from the same research group, the work presented in this thesis is closely related to Guichaoua's approach: it shares the conception of multi-dimensional relationships as an essential aspect of music structure, a fundamental property of the polytopic model. In this work, we extend the concept in several directions: (i) more sophisticated and powerful relationships between musical elements, forming graphs of latent relations and (ii) generalisation and experimentation of the approach to new musical dimensions (namely rhythm and melody).

2.2 Information and Expectation

Along with the notion of structure which plays a growing role in the MIR domain, the interest of the community towards information theory has also greatly increased over the past few years. This is directly related to the rapidly expanding volume of data provided for the MIR evaluation tasks, for which a lot of modelling techniques are based on learning schemes.

In fact, most of the methods that are used to characterise or capture knowledge from large datasets are based on statistical methods and probabilistic models following the principles stated by Meyer [Meyer, 1956, Meyer, 1957]: practised listeners, composers and performers musical style may be "regarded as a complex system of probabilities". The methods following this point of view include Markov Models or n-grams[1] as well as neural networks and deep learning[2]. These approaches are closely related to Shannon's information theory [Shannon, 2001] according to which an observed data can be viewed as the output of a communication system and the encoding cost of the variable is considered as highly-dependent on its emission probability. As it is virtually impossible to know the real probability density function for that variable, it is generally estimated on the basis of the frequencies of its realisations (in a given training set).

In the MIR domain, probabilistic methods thus tend to estimate probability distributions of musical events, such as notes, rhythm, chords, or audio features by considering the frequency of such events inside a corpus. What changes across models is essentially that they consider different types of musical events and compute and combine statistical estimations in different ways.

Reflecting the fact that, in the Information Theory domain, Shannon's communication principles can be challenged by Kolmogorov's complexity theory [Kolmogorov, 1965, Chaitin, 1966, Vitányi and Li, 2000], there is a similar "contention" in the automatic music analysis field [Meredith, 2012a, Bimbot et al., 2016].

In fact, considering Shannon's theory, a musical object would be the result of the realisation of a random variable following a probabilistic distribution which can be estimated using frequencies of musical objects that already exist. In such case, the relevance of the structure of a musical piece would be based on the frequency of observing comparable structures in the training corpus used for estimating the probabilities.

On the other hand, following the point of view associated with the Kolmogorov's complexity theory, a musical piece can be considered as the output of a *program*. In that case, the "best" description of the structure of the musical piece can be defined as the shortest program that is able to generate the piece as result.

As the Kolmogorov complexity is a rather abstract concept, its value can not be computed in the general case and it must be estimated/approximated. One framework to do this is *minimum description length* (MDL), that is (for a music piece) the size of the shortest program, restricted to a specific subset of programs, that takes no input (or

[1] [Conklin and Witten, 1995, Shao et al., 2004, Noland and Sandler, 2006, Abdallah and Plumbley, 2009, Pearce and Wiggins, 2012, Lin and Zhang, 2018]

[2] [Li et al., 2010, Grill and Schlüter, 2015, Calvo-Zaragoza et al., 2016, Lattner et al., 2017]

including the specification of its input variables) and returns the piece as its only output. Equivalently, approaching the Kolmogorov complexity can be understood as the quantity of information strictly needed to compress the musical piece to obtain a description that is shorter than its *in extenso* description.

The reason the MDL principle has sparked more and more interest in the MIR community is because it becomes more and more obvious that compressing a musical object and retrieving information from it form a one and single problem. In fact, the more information is known about the object, the greater the compression capacity, and (if the class of compression programs is well-chosen) the inverse may be true too. Based on such viewpoints, and following the idea that the shorter the description is, the more structural regularity may emerge from the object, Meredith et al. [Meredith, 2012b, Meredith, 2013a, Meredith, 2018] directly associated the notion of music analysis with the search of a shortest description of its organisation and designed algorithms of compression based on the detection of regularities and redundancies [Meredith, 2013b]. They also used existing compression algorithms from other fields to apply them to MIR tasks such as classification or pattern recognition with the goal to investigate if compression techniques may be useful to retrieve information [Louboutin and Meredith, 2016].

In parallel, Mavromatis et al. [Mavromatis, 2009], also used the MDL principle - but in a framework that is more related to Shannon's theory - as the aim to select the simplest statistical model (such as a hidden Markov model) that would best fit the statistical properties of the corpus.

Therefore Shannon's and Kolmogorov's approaches are not contradictory, as both can be used to compress the information present in the *in extenso* representation of music pieces, one by using entropic encoding and the other by exploiting structural redundancies and regularities. Both are based on a different (yet complementary) point of view, and as a consequence, depending on the "information" on which they are based, MIR approaches may borrow from one or the other concept. Probabilistic methods consist in finding model that can generate a musical flow with similar probabilities to those observed in the training data (or derived from expert-knowledge), whereas complexity-based methods aim at providing models that give short descriptions of music pieces which can potentially be interpreted as the underlying organisation of the musical content.

The work that is presented in this thesis is mostly based on the second type of approaches, in the sense that our aim is to describe the structure of a section using a model which gives an explanation of the inner organisation of musical sections. This is achieved by inferring simple (latent) relations between specific pairs of elements composing the section, under the hypothesis that the more redundancy or regularity is present within the section, the simpler the relations are.

Another hypothesis underlying the work presented here is that, by formalising the principle of *expectation*, it is possible to reduce the quantity of information necessary to describe a musical section. This notion of *expectation* (expectedness or expectancy) is present in a lot of studies both in MIR and standard musicology[3]. It is alternatively

[3][Duerksen, 1972, Lerdahl and Jackendoff, 1985, Schmuckler, 1989, Narmour, 2000, Huron, 2006, Ockelford, 2006, Abdallah and Plumbley, 2009, Pearce and Wiggins, 2012, Tillmann et al., 2014, Loy, 2017, Agres et al., 2018]

referred to as implication[4], surprise[5] or anticipation[6]. Here, we want to draw the attention of the reader on the fact that *expectation* has different meanings in all these studies: sometimes it is is a cognitive projection made by the brain such as for Tillmann et al. [Tillmann et al., 2014], sometimes it can be the deduction that a predictive model can trigger based on music descriptor statistical evaluation [Farbood, 2006], or sometimes it is understood as a step-by-step process directly applied to the analysis of the music content [Lerdahl and Jackendoff, 1985, Narmour, 1992]. However all these points of view have in common that they tend to model the expectation as the ability to predict the future.

The work presented in this thesis is strongly based on the System&Contrast (S&C) model [Bimbot et al., 2012a] which has recently been introduced as a generalisation of Narmour's *Implication-Realization* model (proof given in [Bimbot et al., 2016]). The principle of the S&C model is to give a compact and simple description of the structural organisation of basic elements in a musical section, considering that expectation (in the sense we use it) is a way to reduce the information necessary to encode surprise in the musical flow. The next section provides a detailed presentation to the System&Contrast model.

2.3 System & Contrast Model

Approaching structure as *"the arrangement of - and relations between - the parts or elements of something complex"* (Oxford dictionary), and given that a musical section may be considered as "something complex", the System&Contrast model bases its principles on the fact that the structure of a section can be described using a *system* of relations between the elements that compose it. In other words, a musical segment is composed of elements (with multiple properties) which can be described on the basis of their mutual relationships to form a strong "logical" unit, that is the *section*.

In this section, we review briefly the System&Contrast (S&C) model, following almost the same structure as the one used by Bimbot et al. [Bimbot et al., 2016], but sometimes with different examples and illustrations.

2.3.1 Intuitive and General Presentation

The basic principle behind the S&C model is that, given a set of three objects (which may or may not be musical objects), it is often possible to characterise an expectation on a fourth object based on the relations between the first three objects. For example, given the sequence of three objects "A", "B" and "C", it is possible to characterise the expectation of the reader as "D". And then, "A-B-C-D" forms a strong logical system. However for objects that have more than one property, or for sequences of objects that are not generated by the iteration of a single relation, it appears that a matrix representation

[4][Narmour, 1989, Bharucha, 1987]
[5][Abdallah and Plumbley, 2009, Tillmann et al., 2014]
[6][Lévy, 2004]

Figure 2.2: Examples of (plain) square systems.

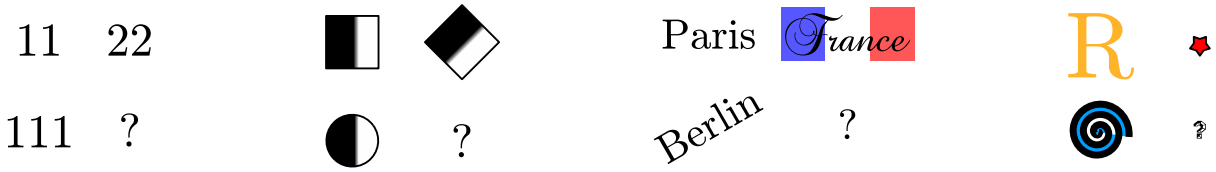


Figure 2.3: Examples of incomplete square systems.

is beneficial to the characterisation of the logical relations, via the formation of a system.

Figure 2.2 shows some examples of such *square systems* i.e. systems of 4 elements, which correspond to a very common form in music. Here, the relation that would be considered useful to *describe* the observations are quite simple. In the first system, B is expected at the bottom right position as it is the letter following A in the alphabet like 2 is following in the digit set. In the second system, one expects a U after the sequence r, s, t, i and as S is capitalised compared to r, and as t is small, one expects also a capitalised letter as the fourth element. The third system involves a rich set of properties such as font, orientation, animal and ecosystem which are easy to explain. The fourth example illustrates that objects may have multiple properties, some of them having no obvious relation and which therefore, do not participate in the description of the system. Here, the only systemic properties are the shape (circle for all elements), and the size (larger on the left, smaller on the right). Therefore, one would expect a circle smaller than the third one and bigger than the second one with any kind of colouring motif.

A fundamental property of a square system is its redundancy, in the sense that the fourth element can be easily guessed from the rest of the system. In fact, Figure 2.3 presents some systems where the fourth element has been replaced by a question mark. It is easy to *deduce* some properties of the fourth element by observing the relations between the first three elements. For the first example, one would expect the number 222, a circle with the filled sector rotated for the second, “Germany” with the flag of that country as background for the third, and, in the last example, almost anything different from a spiral provided it is small. Note that the question mark for this last example could be used as a possible object to fill the system in the sense that it has a size corresponding to the systemic expectation.

Considering now a complete set of four elements, and assuming that the first three elements form a system, it can be very interesting to compare the *expectation* created by the system with the actual fourth element; and to describe how it deviates from that expectation and to what extent. The fact that we are able to characterise the expectation based only on the observation of the first three elements makes it possible to characterise the fourth element by is difference with the expectation. This results in the concept of

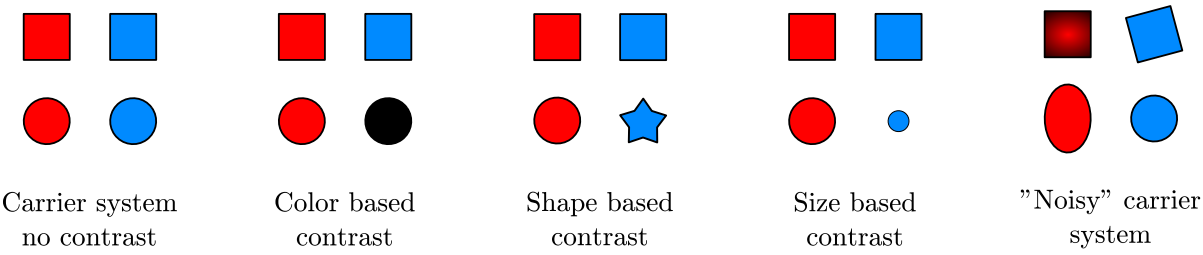


Figure 2.4: Examples of square S&Cs.

contrast.

In summary, a square *System&Contrast* (S&C) can be defined as a system of four elements, where the last element may have some properties that create some conflict with the properties that are expected by analogical inference in fourth position, on the basis of the first three elements. The *contrast* can be assimilated to some surprise triggered by the system and the contradiction (i.e. denial) of the logic implied by its first three elements.

Some basic examples of S&Cs, based on simple properties, are given on Figure 2.4. As it is the case for the fourth example, the contrast may not directly contradict the expected properties but can also involve properties that are not useful to describe the rest of the system. In such a case, the contrast is a surprise in the sense that it contradicts the logical framework used to describe the expectation from the observation of the first three elements. Conversely, it may also happen some properties of the first three elements do not form any particular expectation, if these properties happen to be unrelated. This is the case in the fifth example in Figure 2.4, where texture, orientation and deformation appear as not to be showing any systemic behaviour.

In the S&C approach, sequences are therefore represented in a matrix arrangement (instead of a purely sequential chain) based on relations between the first element (say x_0), called the *primer*, and its neighbours in the 2×2 matrix, via latent relations (say f and g). In fact, to create an expectation, it would also be possible to resort to a single iterated relation, f, but the model would have less potential : with two relations on potentially different musical dimensions, and between non-contiguous elements, the S&C model offers greater latitude to create sophisticated expectations, and combine them at several scales. The fact that the S&C model introduces non-sequential dependencies constitutes in fact one of the most important hypothesis of the model: the structure of a musical section relies on relations that are not necessarily sequential. This is illustrated by the elementary graph of dependencies within a S&C, as depicted on Figure 2.5.

With the notations of Figure 2.5, with f and g denoting the relations, and x_0 the primer, we have:

$$\begin{cases} x_1 = f(x_0) \\ and \\ x_2 = g(x_0) \end{cases} \tag{2.1}$$

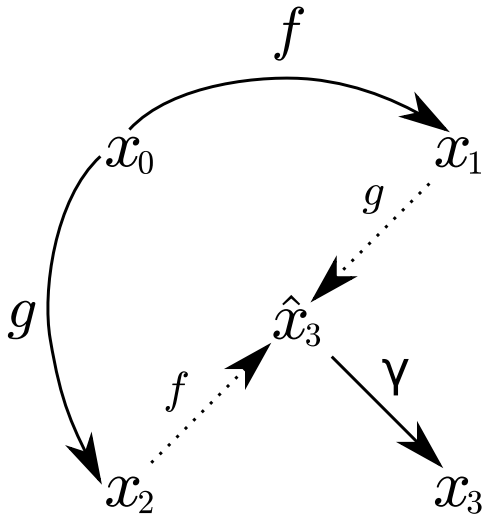


Figure 2.5: Elementary graph of dependencies within a S&C.

Then, the expectation can be represented by a *virtual element*, $\hat{x_3}$, equal to $f(g(x_0))$ (or $g(f(x_0))$) meaning that $\hat{x_3}$ is to x_2 what x_1 is to x_0 (and also that $\hat{x_3}$ is to x_1 what x_2 is to x_0). In the rest of this work, we will only consider commutative relations, but in the case where the relations were not, a convention can be set for the creation of the expectation, for instance, $f(g(x_0))$. Function γ is then the contrast function, which describes the difference between the actual observation x_3 and the virtual element representing the expectation, $\hat{x_3}$.

Note that the S&C model enforces a causality principle in the sense that the direction of the relationships between elements within the system is assumed to be in accordance with the order in which these elements occur in the unfolded sequence. As a consequence, the contrast is always in final position, although, in some cases, it would also be possible to explain the system on the basis of a disparity affecting another element than the last one.

As illustrated in the examples above, f and g may only apply to a subset of the properties of the primer. They can be understood as (more or less complex) transformations of the primer. We call the quadruplet (x_0, f, g, γ) the S&C description of the sequence $X = x_0x_1x_2x_3$. Such a description may be considered as the "genetic program" of the system [Bimbot et al., 2016]— echoing Narmour's genetic "code" [Narmour, 1989]— or at least, a "generative program", as we will later elaborate on.

A S&C description requires the choice of a logic space to describe the relations inside a system. This can be viewed as an optimisation process that can readily be formulated as a minimum description length (MDL) problem. In fact, considering the S&C description as a program generating the sequence X, the MDL approach can be understood as a compression scheme over this description. The aim then, is to find the space of relations for which the description of f, g and γ are the shortest.

As it is not the focus of this work, we do not discuss here the case when there are more or less than four elements in a S&C. The matter has been formalised by Bimbot et al. [Bimbot et al., 2012b, Bimbot et al., 2016] and solved, to some extent, by the im-

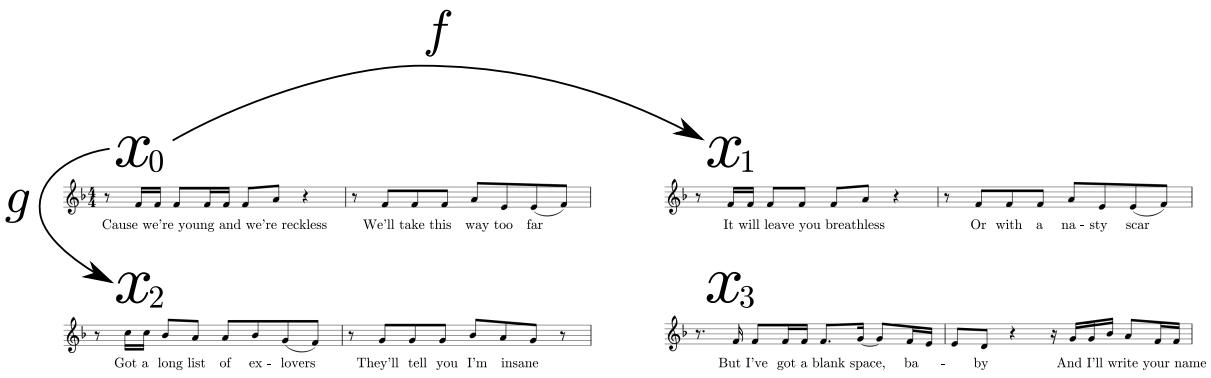


Figure 2.6: **Taylor Swift - Blank Space** (comp.: Taylor Swift, Max Martin, Shellback) Blank Space, Big Machine Records/Republic Records 2010. Timing: 1'08-1'28. *Transcribed by the author*

plementation proposed by Guichaoua [Guichaoua, 2017]. Still, it is important to keep in mind that the S&C model can also be used to describe the structure of sections that have three, five or six (or even seven) elements, while sections with more or less elements can be handled by changing the time-scale.

2.3.2 Musical Examples

Music generally appears as a *sequential* and *dynamic* presentation of acoustic or symbolic objects. However, an important hypothesis underlying the application of the S&C model to music is that dependency relations between discrete objects inside a musical section may be considered as *matricial*. Analysing a musical passage therefore involves implicit operations of delinearisation of the musical flow and of discretisation of its properties.

Rather than re-using former illustrations of descriptions and analysis of musical sections with the S&C model, such as those presented by Bimbot et al. [Bimbot et al., 2012b, Deruty et al., 2013, Bimbot et al., 2016], we present a few new examples. As we focus only on square S&Cs, each section is decomposed in four elements, x_0, x_1, x_2 and x_3.

The first example shown on Figure 2.6, is a part of the chorus of a famous pop song (which has been viewed more than 2.3 billions times on YouTube). This section follows a *aabc* structure which is one of the most common pattern in music. To analyse the section using the S&C model, we divided it in four 2-bar elements. The relation f between the first two is almost identity, in fact, there is a sixteenth note which does not appear in the second element, but all the rest is identical except for the lyrics which have only in common the rhyme at their middle-end and end. The relation g between the primer and the third element, creates a strong change in the melody, as it indeed drastically modifies the melody contour in terms of pitch but also the rhythm by adding notes on the last beat of the first bar and removing two other onsets.

Given these two relations, one would logically expect some repetition of x_2 with a very small change of the rhythm and lyrics that would rhyme with the middle and end of x_2. However, x_3 contrasts with this expectation by showing a drastic change both in rhythm and pitch, and discrepancy in the rhyme: "lovers" vs "baby". In this case,

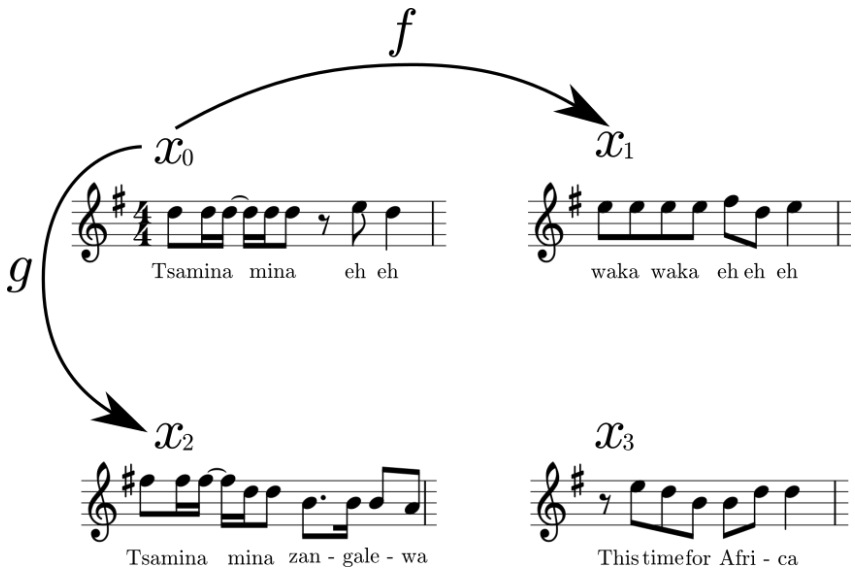


Figure 2.7: **Shakira - Waka waka (This Time for Africa)** (comp.: Zangalewa, arr.: Shakira, John Hill) Waka waka (This Time for Africa)(The Official 2010 FIFA World Cup (TM) Song), Epic Records/Sony Music Entertainment 2010. Timing: 1'03-1'11. *Transcribed by the author.*

almost all musical dimensions follow a *aabc*. However, whereas for some dimensions such as the pitch contour or the lyrics, the discrepancy between a and b is large, for other musical dimensions (such as for the rhythm) the inner structure could be encoded as $aa'a''c$. Some other sections in the song, such as the first part of the same verse, shows a similar organisation but with a weaker contrast.

The second example is also a well-known pop song, and its S&C description is depicted on Figure 2.7. This section has a $aba'c$ form, which is extremely common in music. Here, relations f and g are both different from identity in almost every dimensions that could be reasonably considered. Function f marks a strong change of the rhythm on the first part of the 2-bar element and a small change of the pitch contour as it is almost a transposition of the primer's pitch. There is also a rhyme between lyrics, with the repetition of the "eh eh eh". On the other hand, g strongly changes the second part of the primer to increase the pitch contour, the rhythm density and the lyrics. The contrast is not excessive in term of rhythm, as the rhythmic pattern of x_3 is almost the same as that of x_2. However, some changes appear on the first half of the element, where g creates the expectation of a change in the second part. The pitch contour is a bit different of what could have been expected but, it is in the lyrics that the contrast is the most drastic. In fact, while all the first sentences were in Fang language, the last one is in English which creates a drastic change of sonority (and cultural background). Still, the rhyme is consistent with the expectation (but considering the rest of the lyrics, one would have probably expected "Zangalewa" instead of Africa!) And for the first part, the most straightforward syllables would have been "waka waka" instead of "This time for". Therefore, the whole form, $aba'c$, can be seen as the conjunction of all the patterns that are obtained by considering separately the musical dimensions: $aba'b'$ (rhythm), $aa'bc$ (melody), *abac* (first part of lyrics), *aabc* (second part of the lyrics), *aabb* (rhymes), and... *aaab* (lyrics language)! In

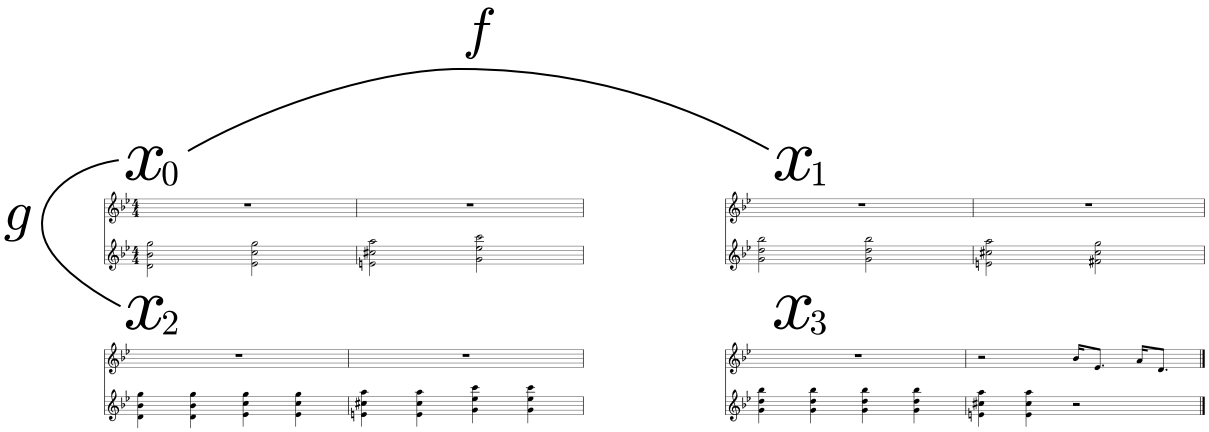


Figure 2.8: **Smokey Joe & The Kid - Smokid All Star (Feat. Waahli, NON Genetic, Pigeon John, ASM, Youthstar, Blake Worrell, Chill Bump, Dj Netik)** (comp.: Smokey Joe & The Kid) Smokid All Star, Benzaï Lab/Believe Music 2016. Timing: 0'35-0'56. *Transcribed by the author*

other words, these multiple musical dimensions are all redundant within the section, but in different ways.

The next example shown on Figure 2.8 is taken from the accompaniment of a recent rap-song. It also exhibits an $aba'c$ form. Here f changes the harmony and g transforms the rhythm by doubling each chord. The only contrast here is what happens at the end of x_3: the organ stops to let the trumpets and other wind instruments mark a strongly different rhythm with notes that induce two different chords on the same strong beat. Without this change on the last strong beat, x_3 would have been exactly what would be logically expected. In fact, stopping the accompaniment to create a contrast is a very common strategy in some types of pop music (this process was actually also used for the two first examples).

In Figure 2.9, which is the chorus of "People Are Strange" by the famous band *The Doors*, the structure obtained after a delinearisation of the content is a bit different than for the previous examples. Here, the first three elements of the S&C, which have actually a lot in common (such as the lyrics "When you're strange" at the beginning), all last three bars. There is therefore a direct contrast based on the element duration, as x_3 lasts only one and a half bar (counting the anacrusis). In fact, the local meter, the number of beats by element and the position of the musical material w.r.t. the beat are musical dimensions that are frequently used to create surprise.

Another observation that can be made from this example, is that the guitar part, which governs the harmony, has no other deviation with the expectation, except for its length. Indeed, f keeps the guitar identical and g changes it to a stagnant line. Therefore, in the absence of contrast, one may expect a stagnant line for x_3, which is actually the case.

As we view it, the structure of the global section is somehow the reunion of the structural patterns emerging for its multiple musical-dimensions, hence the need for a multi-dimensional approach of the music structure analysis (as originally sketched

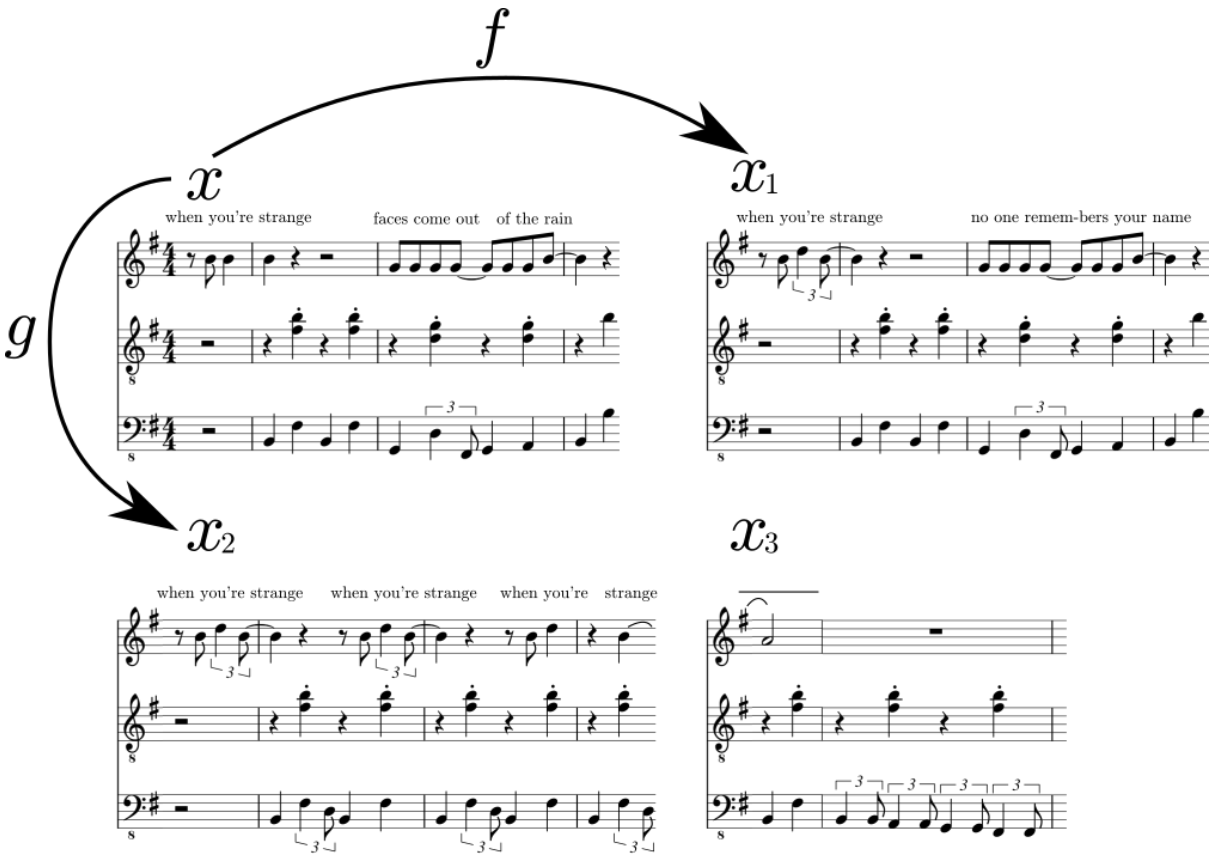


Figure 2.9: **The Doors - People Are Strange** (comp.: Jim Morrison, Robbie Krieger) People Are Strange, Elektra 1967. Timing: 0'19-0'40. *Transcribed by ear and corrected by the author*

in [Peeters and Deruty, 2009]). As noted by Smith et al. [Smith and Chew, 2013], the listener certainly focuses on different dimensions at different moments in a song and may locally perceive a predominant structural role of the melody for one section and of the harmony for another one. But we believe that a proper description of the structure of a musical section must account for the patterns formed by the multiple musical dimensions and their congruence in reinforcing the consistency of its structural construction.

Going back to our previous example, it is furthermore interesting to note that the last melodic element shows no anacrusis and only contains the ending note of the melody, while all other elements contained an anacrusis. This rises the problem that will be developed in Chapter 6: quite often, melodic elements are not time-aligned with their corresponding harmony-governing element or accompaniment. Here, one would consider, metrically speaking, the beginning of the section at the first beat of the first plain bar of the segment, but this is almost two beats after the actual beginning of the melody...

Our last example is taken from a heavy metal piece, "2 Minutes to Midnigth" by *Iron Maiden* (see Figure 2.10). Here the section follows an *abac* form, where each basic element lasts 4 bars. Function g is almost the identity, whereas f creates drastic changes in the system. However, the last element x_3 is really different from x_1 and creates a contrast for every instrument. But the contrast does not affect the whole section for

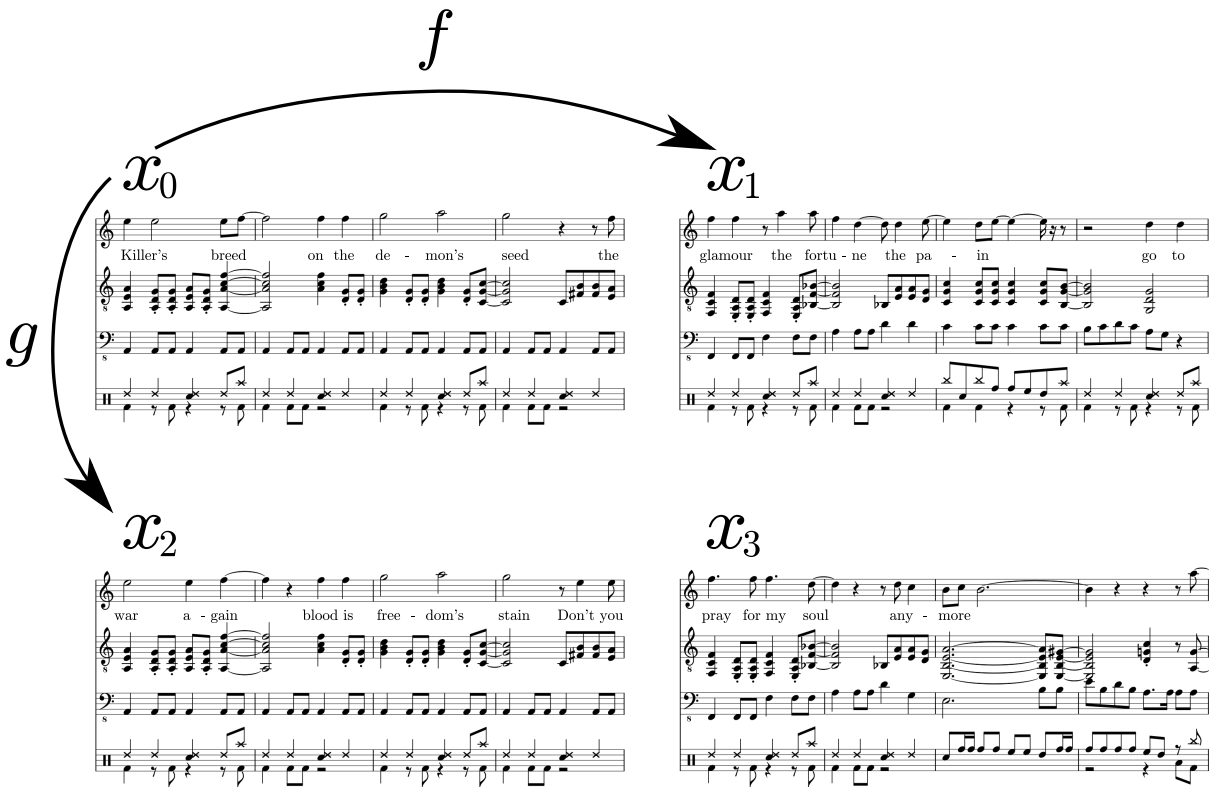


Figure 2.10: **Iron Maiden - 2 Minutes To Midnight** (comp.: Bruce Dickinson, Adrian Smith) 2 Minutes To Midnight, EMI 1984. Timing: 0'52-1'13. *Transcribed by ear*

every instrument. In fact, the guitar and the drums have contrastive parts only over the second half of x_3, the bass has a contrast which is beginning at the last beat of the second bar of x_3, while the whole melody is entirely contrastive. Therefore, for one dimension, the contrast can be considered at the scale of the whole system. However, by considering other dimensions, it would be more effective to describe the contrast at smaller scales. Note incidentally that the same observation can be made for the previous examples, such as, Figures 2.6, 2.7, and 2.8.

All these examples, along with those that were presented in the previous studies, confirm the relevance of using compression schemes to describe musical structure. Moreover, the analysis of these musical sections highlights the multi-dimensional property of music and the need for modelling the structural information at several layers of information simultaneously.

But not only is music structure multi-dimensional, in the sense that it affects several musical properties, it is also developed over multiple time-scales simultaneously. It is therefore important to investigate how these scales articulate with one to another. However, these two fundamental properties of the music may be addressed independently, even if in practical use, they are musically correlated. The aim of the following study is to formalise these processes in the context of a computational approach.

Chapter 3

Multi-scale System & Contrast

This chapter is dedicated to the formalisation of the principles underlying the implementation of the System&Contrast model into a multi-scale framework: the Polytopic Graph of Latent Relations. In particular, we investigate how 6 specific configurations (called *Primer Preserving Permutations* or *PPP*) can be defined and used to characterise the inner organisation of musical sections at different scales simultaneously.

Given that a wide majority of musical sections in pop-music are composed of 8 bars (or 16 strong beat[1]), this chapter will focus only on the description of these sections, assuming they are subdivided in 4 and 16 elements, i.e. square systems and square systems of square systems. However, this hypothesis is not considered as too severely restraining, as former works by Deruty et al. [Deruty et al., 2013], Bimbot et al. [Bimbot et al., 2016] and Guichaoua [Guichaoua, 2017] have shown that less regular structures could be approached as deformations of square ones.

3.1 Square Formalisation

The principle behind the S&C model is that a sequence of musical elements can be described by using non sequential dependencies between these elements. This new organisation creates a strong process of expectation (via logical implications). Applied to a sequence of four elements, $(x_i)_{0\leq i\leq 3}$, it can be understood as the geometrical arrangement of the elements into a square matrix:

$$X = \begin{bmatrix} x_0 & x_1 \\ x_2 & x_3 \end{bmatrix} \tag{3.1}$$

Based on this arrangement, two relations f and g can be assumed, which relate the *primer*, that is the first element of the sequence, x_0, and its neighbours in X:

$$\begin{cases} x_1 = f(x_0) \\ x_2 = g(x_0) \end{cases} \tag{3.2}$$

[1]Here, bars are in 4/4 with therefore alternation between strong beat (S) and weak beat (W): S-W-S-W.

Note that, as mentioned in the previous chapter, these two relations may apply to a subset of the properties characterising the elements of the system.

The S&C model envisions the fourth element, x_3, as being in relation with a *virtual* projected element $\hat{x_3}$ which would result from the combination of f and g, applied to x_0, namely the logical expectation that one could build, by analogical implication, from the observation of the three first elements. The disparity between $\hat{x_3}$ and the actual (observed) x_3 is modelled by a *contrast* function γ:

$$\hat{x_3} = f(g(x_0)) \tag{3.3}$$

$$x_3 = \gamma(\hat{x_3}) \tag{3.4}$$

The description of a section using the S&C model is the quadruplet (x_0, f, g, γ) which, due to the fact that these relations are often simpler to describe, can be used as a compact description of the *in extenso* representation of the section. It can be viewed as a minimal description in the sense of the Kolmogorov's complexity [Vitányi and Li, 2000] in line with several other work in MIR[2].

3.2 Multi-scale Analysis

A number of approaches in music modelling are based on hierarchical representations of the dependencies between elements to describe a piece structure or even a section structure. For example Rohrmeier and De Haas [De Haas et al., 2009, de Haas et al., 2011, Rohrmeier, 2007] use a tree-like model based on formal languages to describe structures of chord progressions. However, dependencies inside their model are formalised in terms of musicological rules assuming some tonal background, such as a dominant to tonic progression. Therefore, the structure model they use is inherently restricted to chord progressions.

Our aim is to propose another model, based on the S&C framework, which has also a hierarchical basis, but that would focus on similarities between elements at given metrical positions to infer non-sequential dependencies between the elements of a section. As the dependencies do not depend on the musicological properties of the elements, this structure model can be used with any dimension: rhythms, chords, melodies, meters, nuances...

Considering a basic element as all the musical information that is related to a given strong beat, it appears that a wide majority of sections are composed of 16 such elements. It can be less or more, but 16 elements sections is the most common case, especially in pop music. As the S&C model is designed to describe systems of four elements, there is a need to generalise its principles in order to be able to describe a full section. The generalisation that is presented in this section is called Polytopic Graphs of Latent Relations (PGLR) [Louboutin and Bimbot, 2017b]. Its main principle is to reconsider the sequential nature of musical section by relating musical elements one another, not on the basis of their contiguity, but rather taking into account the fact that they lie in homologous metrical positions. This section details how this can be formulated by viewing the musical section as a polytope.

[2][Mavromatis, 2009, Temperley, 2014, Louboutin and Meredith, 2016]

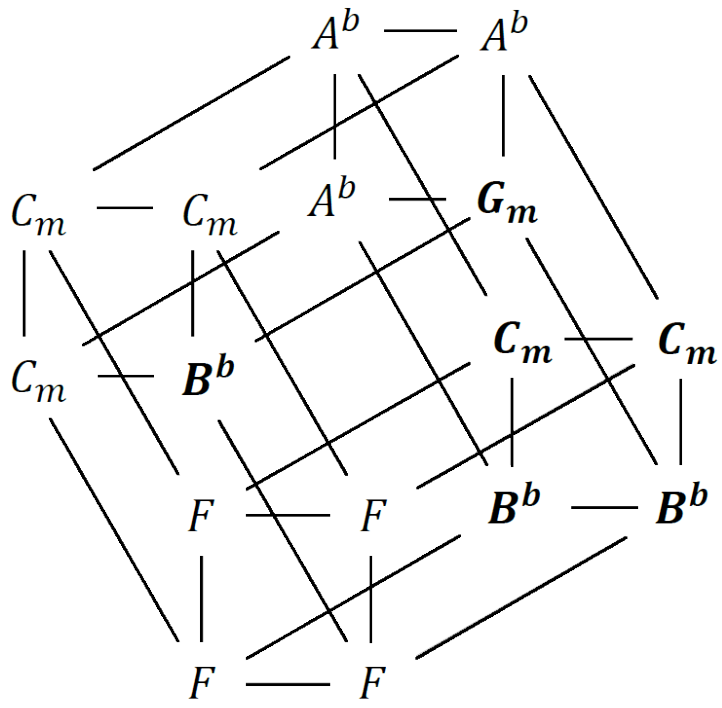


Figure 3.1: Polytopic representation of the chord sequence taken from *Master Blaster* by Stevie Wonder.

3.2.1 Polytopic Graph of Latent Relations

Indeed, elementary systems of four elements, as described in the previous section, can further be used to describe longer sequences of musical elements. In particular, sequences of 2^n elements can be arranged as an n-dimensional cube (square, cube, tesseract, etc...), and more generally speaking, on an n-polytope, n being the number of scales considered simultaneously in the multi-scale model.

Each vertex in the polytope corresponds to a musical element of the lowest scale, each edge represents a latent relationship between two vertices and each face forms an elementary system of relationships between (typically) 4 elements.

For instance, a sequence of 16 chords can be divided into four sequences of four successive chords, each of them being described as separate systems. Then, these four S&Cs, taken as elementary objects, can be related by forming an upper-scale S&C, linking the four primers of the 4 lower-scale S&Cs. Figure 3.1 represents such a description projected on a tesseract, and is illustrated in the case of the chord sequence from the chorus section of *Master Blaster* by Stevie Wonder:

$$Cm\ Cm\ Cm\ Bb \quad Ab\ Ab\ Ab\ Gm \quad F\ F\ F\ F \quad Cm\ Cm\ Bb\ Bb$$

The PGLR approach views a sequence of musical elements within a structural section as exhibiting privileged relationships with other elements located at similar metrical positions across different timescales.

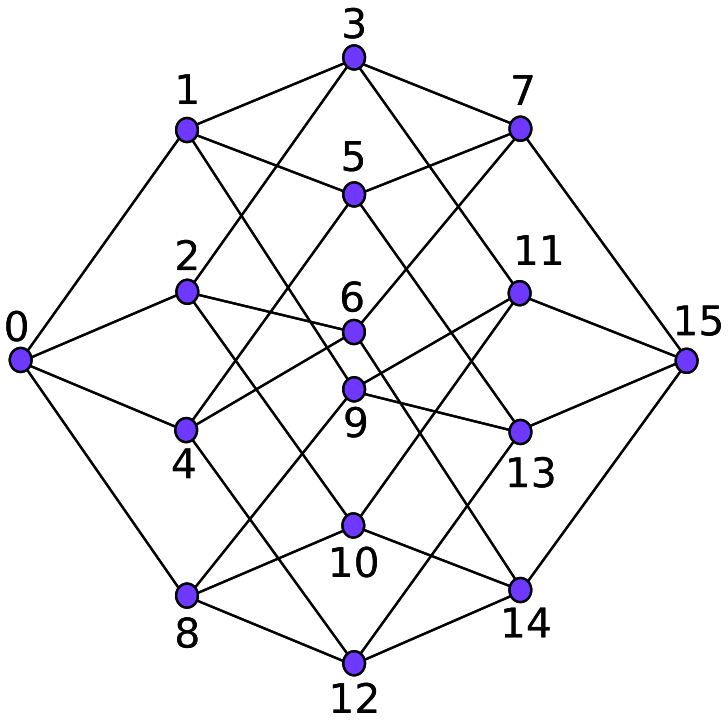


Figure 3.2: Polytopic representation of a sequence of 16 elements, where elements at the same depth are aligned vertically. The resulting partial order between vertices is causal.

As opposed to the sequential viewpoint which assumes a total order of elements along the time-line, the polytopic organisation on the tesseract leads to a partial order (illustrated on Fig. 3.2), where elements of the same depth are aligned vertically and where the fourth element of each face can be defined in reference to the virtual element resulting from the implication of the three others.

In the most general case, valid systemic organisations can be characterised by a graph of nested systems, the flow of which respects the partial ordering of Fig. 3.2. Note however that there is a possible conflict between three implications systems for elements 7, 11, 13 and 14, each possible implication corresponding to a face of the tesseract. For instance, node 7 can be viewed as resulting from 3 implication systems: $[1, 3, 5, 7]$, $[2, 3, 6, 7]$ and $[4, 5, 6, 7]$. Element 15, as it has the highest depth can be viewed as the contrastive element of 6 concurrent systems.

3.2.2 Primer Preserving Permutations

One way to handle these conflicts is to constrain the graph to preserve systemic properties at higher scales. This can be achieved by forcing lower-scale systems to be supported by parallel faces on the tesseract, while the first elements of each of the 4 lower-scale systems are used to form an upper-scale system. This approach drastically brings down the number of possible graphs to 6, which corresponds to specific permutations of the initial index sequence (see Table 3.1), termed here as Primer Preserving Permutations (PPP) [Louboutin and Bimbot, 2017a].

PPP_0	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
PPP_1	0	1	4	5	2	3	6	7	8	9	12	13	10	11	14	15
PPP_2	0	2	4	6	1	3	5	7	8	10	12	14	9	11	13	15
PPP_3	0	1	8	9	2	3	10	11	4	5	12	13	6	7	14	15
PPP_4	0	2	8	10	1	3	9	11	4	6	12	14	5	7	13	15
PPP_5	0	4	8	12	1	5	9	13	2	6	10	14	3	7	11	15

Table 3.1: List of the 6 Primer Preserving Permutations.

To illustrate a PPP, let's consider the subdivision of a sequence of 16 chords into four sub-sequences of four successive chords. Each sub-sequence can be described as a separate Lower-Scale S&C (LS): $[0, 1, 2, 3]$, $[4, 5, 6, 7]$, $[8, 9, 10, 11]$ and $[12, 13, 14, 15]$. Then, these four S&Cs can be related to one to another by forming the Upper-Scale S&C (US) $[0, 4, 8, 12]$, linking the four primers of the 4 LS. This configuration (PPP_0) turns out to be particularly economical for describing chord sequences such as the one taken from *Master Blaster*:

$$Cm\ Cm\ Cm\ Bb \quad Ab\ Ab\ Ab\ Gm \quad F\ F\ F\ F \quad Cm\ Cm\ Bb\ Bb$$

as most similarities (identities) develop between neighbouring elements.

But, if we now consider another example such as the following progression:

$$Bm\ Bm\ A\ A \quad G\ Em\ Bm\ Bm \quad Bm\ Bm\ A\ A \quad G\ Em\ Bm\ Bm$$

a different configuration appears to be more efficient to explain this sequence. In fact, grouping chords into the following 4 lower-scale S&Cs: $[0, 1, 8, 9]$, $[2, 3, 10, 11]$, $[4, 5, 12, 13]$ and $[6, 7, 14, 15]$, which are all parallel faces in the tesseract, and then relating these four systems by a upper-scale system $[0, 2, 4, 6]$ (configuration PPP_3) leads to a less complex (and therefore more economical) description of the relations between the data within the section. In fact, by doing such grouping, the number of identity relations used for the description is higher than using PPP_0[3]. Fig. 3.3 illustrates these two configurations.

Algorithm 1 recursively generates all the PPP for a sequence of size 2^n. The dimension of the n-cube corresponding to the sequence is used to build a set of vectors $\{2^i | 0 \leq i < n\}$. Then, if n is even (resp. odd), the algorithm selects two (resp. one) vectors in the set of directions (which is represented as a set of integers that are powers of 2: 1, 2, 4 and 8 for a section of 16 elements) in argument to build a face (resp. edge) from the primer. The primer and each new point are used to compute the PPPs of dimension $n-2$ (resp. $n-1$) using the the direction that are not already selected. Each sub-sequence of size 2^{n-2} with the same structure (indexed by k) is then concatenated to form a sequence of size 2^n.

[3]Some examples of non-identity relations between elements will be given in Chapter 4, formalisms for relation between chords, rhythms or melodies will be developed in Chapter 5 and 6.

Input: Primer p, set of directions used D

Function `PPP(`p`,`D`)`**:**

```
if |D| = 2 then /* D = {d_1; d_2}|d_2 > d_1                    */
    return [[p; p + d_1; p + d_2; p + d_1 + d_2]];
end
L = [];
if |D| mod 2 = 1 then
    for d ∈ D do
        D' = D − {d};
        S_1 = PPP(p,D');
        S_2 = PPP(p + d,D');
        for k = 0 to |S_1| − 1 do
            L.Append(S_1[k] ⊕ S_2[k]);
        end
    end
else
    for d_1, d_2 ∈ D^2|d_2 > d_1 do
        D' = D − {d_1; d_2};
        S_1 = PPP(p,D');
        S_2 = PPP(p + d_1,D');
        S_3 = PPP(p + d_2,D');
        S_4 = PPP(p + d_1 + d_2,D');
        for k = 0 to |S_1| − 1 do
            L.Append(S_1[k] ⊕ S_2[k] ⊕ S_3[k] ⊕ S_4[k]);
        end
    end
end
return L;
```

end

Algorithm 1: Recursive function that generate the list of PPPs for a n-cube of dimension. The initial call will then be $f(0, \{2^i | 0 \leq i < n\})$. Here $\oplus$ is used as the concatenation operation on lists.

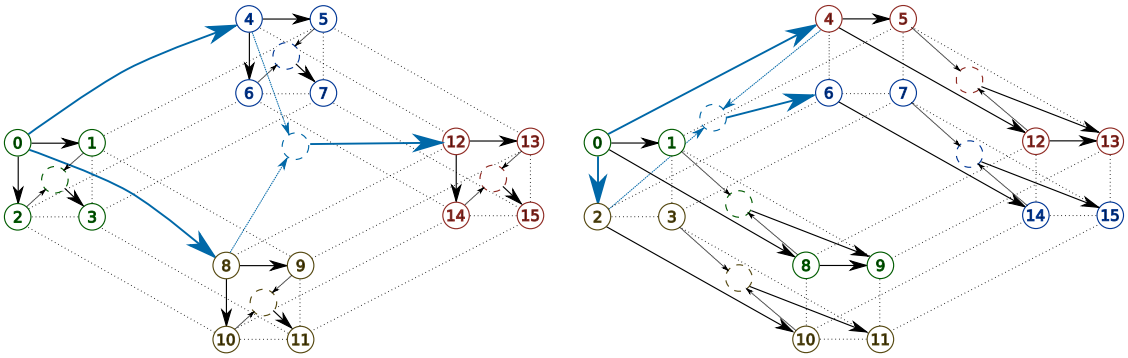


Figure 3.3: Representations of two PPP-based PGLRs on a tesseract: PPP_0 (left), PPP_3 (right). In blue, the Upper-Scale (US) S&C – in black, the 4 Lower-Scale (LS) S&Cs. Dotted nodes indicate the virtual elements ($\hat{x}$) in the implication scheme (Section 3.1).

3.3 Polytopic Model Specifications

3.3.1 Antecedent Functions

The concept of PGLR offers a new way to describe the dependencies between the elements in a section. But due to the multiplicity of possible graphs of dependencies over a polytope, there may be a lot of different models of description. This section presents a list of proposed models based on the PGLR framework. Each model is a first-order model, i.e. each element has only one antecedent in the graph. Therefore, a model of dependencies can be assimilated to its *antecedent function* i.e. the function which, for each element, returns its antecedent in the polytopic graph. The antecedent function associated with a model is denoted Φ_M.

3.3.2 Relations

In the PGLR formalism, an element can be any musical object: a rhythmic patterns, a chord, a melodic cell, particular dynamics or instrumentations, a local metric value or tempo, etc... For each type of musical dimension considered, specifying a full description of the structure of a system, requires to describe the set of relations between elements that are dependent. Whereas, for the sequential model, the antecedent function does not depend on the formalism used to describe the relations, this is not the case for the models that are based on the S&C framework.

In fact, in the S&C framework, the antecedent of the last element in a system of four elements is a virtual element. Therefore, the antecedent function requires a formulable relation of the dependencies in order to generate such a virtual element. This particular element can only be created by applying the combination of the two first relations to the primer. That is, only by formulating the relation between the three first elements, can we model the expectation.

Moreover, to ensure that it is possible to specify a virtual element for any system, the relation space, $\mathcal{R}$, and the element representation space, $\mathcal{E}$, must verify the following

properties:

1. $(\mathcal{R}, \circ)$ is a commutative group
   That is, any combination of relations is in $\mathcal{R}$, combinations can be done in any order, identity relation is in $\mathcal{R}$ and any relation has a unique inverse in $\mathcal{R}$. This property ensures that we can combine relations f and g to obtain a virtual element, and it also ensures that each relation has a symmetric, which can be used to describe the antecedent of an element $\mathcal{E}$ by a dual relation.

2. $\forall (e_0, e_1) \in \mathcal{E}, \exists! r \in \mathcal{R}, e_1 = r(e_0) \in \mathcal{R}$
   That is, for every couple of elements in $\mathcal{E}$, the relation between the two elements is in $\mathcal{R}$ and is unique. This property of uniqueness guarantees that we can describe any relation between any elements with the same relation formalism.

3. $\forall e \in \mathcal{E}, \forall r \in \mathcal{R}, r(e) \in \mathcal{E}$
   That is, any relation of $\mathcal{R}$ can be applied to every element of $\mathcal{E}$, and its image is also in $\mathcal{E}$. This property ensures that we can apply any relation to any primer to obtain a virtual element, and that the virtual element is also in $\mathcal{E}$. The benefit is that the contrast relation can also be represented with a relation living in $\mathcal{R}$.

4. $\forall e \in \mathcal{E}, \forall r \in \mathcal{R}, \exists x \in \mathcal{E}, r(x) = e$
   That is, for every relation in $\mathcal{R}$, any element of $\mathcal{E}$ has an antecedent in $\mathcal{E}$ by this relation. This property ensures that given the contrast relation and the observed contrastive element, it is possible to describe the virtual element, that is, the expectation.

5. the complexity of the relations must be measurable. That is, there exists a function, C, that can give a cost to any relation of $\mathcal{R}$. This property is needed to apply a complexity score to the relation description.

By using a set of relations that satisfy such properties, it is possible to fully describe any S&C. The interest of such a formalism is that given any element of the S&C and the three relations f, g and γ it is possible to reconstruct the full sequence of elements.

Moreover, it is important to note that the last property is necessary to give a complexity score to the description of a sequence. The complexity of description of a whole sequence of elements, X, by a model, M, can be defined as the total sum of each individual relation cost between one element, x_i and its antecedent in the graph, $\Phi_M(x_i)$:

$$C_M(X) = \sum_{i=1}^{15} C(r(\Phi_M(x_i), x_i)) \tag{3.5}$$

This cost function can be useful to chose the best PPP, as it will be discussed later.

3.3.3 Sequential Model

As opposed to models based on the PGLR framework considered in this work, it is worth noting that the *sequential model* (which we will denote as Seq), turns out to be the most

Figure 3.4: Representation of the relations used by a sequential analysis of a sequence of 16 elements.

common approach in standard (and computational) musicology. It is indeed considered as "common sense" in music analysis that an element depends essentially on what happened just before. In fact, this hypothesis is discarded in the case of polytopic models, and it is therefore important to compare the two approaches, both formally and experimentally.

In the case of first-order models (as we consider here), the sequential model relates an element with its immediate predecessor as shown on Figure 3.4. In the statistical framework, the model is called a bi-gram model.

Therefore, the antecedent function for such a graph (here, a chain) can be very easily defined by:

$$\Phi_{Seq}(x_i) = x_{i-1} \tag{3.6}$$

Φ_{Seq} being independent on the type of formalism chosen to describe the relation between elements.

From here, all models that will be presented are based on the multi-scale principle supporting the PGLR framework. As such, they will often be referred as multi-scale or polytopic models in the next chapters.

3.3.4 Tree Systemic Model

A first type of multi-scale model which we will consider in this work is what we call *tree systemic model* and denote as Sys. Tree systemic model is an implementation of the PGLR, which consists in partitioning the observed sequence in four *lower-scale* (LS) systems of four elements. But, instead of considering a virtual element, it simplifies the description of a system of four elements by describing the three last elements in relation to the first element of the system, the *primer*. It can be seen as related to the S&C framework, where the virtual element is replaced by the primer. So as to describe the global structure of the section and model the relation between each of the four systems, an *upper-scale* (US) system is used to describe the relation between the four primers of the LS systems. It can be seen as creating a new system with the sub-sequence of the section which contains only the primers of each of the LS systems. Fig. 3.5 shows the representation of the graph structure created by the tree systemic model.

Given such a graph of dependencies, it is possible to recursively define an antecedent function Φ_{Sys} for the Tree Systemic model:

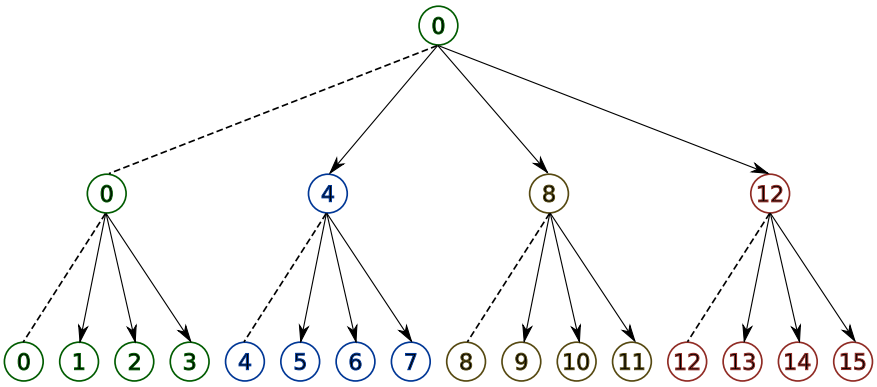


Figure 3.5: Representation of the relations used by a multi-scale analysis of a sequence of 16 elements in the Tree Systemic model. Here, the sub-sequence Y used in Equation 3.7 to pass from a scale to the upper one is $(0, 4, 8, 12)$.

$$\Phi_{Sys}(x_i) = \begin{cases} x_{i-1} & \text{if } i \pmod 4 = 1 \\ x_{i-2} & \text{if } i \pmod 4 = 2 \\ x_{i-3} & \text{if } i \pmod 4 = 3 \\ \Phi_{Sys}(y_{\frac{i}{4}}) \text{ in } Y = (y_l)_{0 \leq l < \frac{n}{4}} = (x_{4*j})_{0 \leq j < \frac{n}{4}} & \text{otherwise} \end{cases} \tag{3.7}$$

Here, each of the first three lines relates an element to the primer of its sub-system. The last line is used to pass from the lower scale to the upper one by considering the sub-sequence, Y, formed by the primers of each sub-system only. Y is obtained by sampling X at each index that is a multiple of 4.

Note that this function can be applied on the initial sequence but also on any permutation of the sequence. In particular, for each PPP, a new model can be defined: Sys_0, Sys_1, ..., Sys_5.

By considering all the possible descriptions corresponding to the 6 PPPs, the model that corresponds to the best PPP (according to some description cost such as the one defined by Equation 3.5) can also be defined. This "optimal" model (for sequence X) will be referred to as Sys^X in Chapters 5 and 6.

Ultimately, the antecedent function can be generalised to any section containing 2^{2p} elements, where p is the number of scales. And, by introducing a "special" system of two elements, the tree systemic model can also be generalised to the description of sequences of length 2^{2p+1}.

3.3.5 Static S&C Model

Let us now introduce the *static S&C model*, $S\&C$, which is the direct implementation of the PGLR framework. In fact, as for the previous model, it consists in describing the structure of a sequence by partitioning that sequence in 4-element systems. But here the S&C formalism is applied to the description of each system of four elements to generate

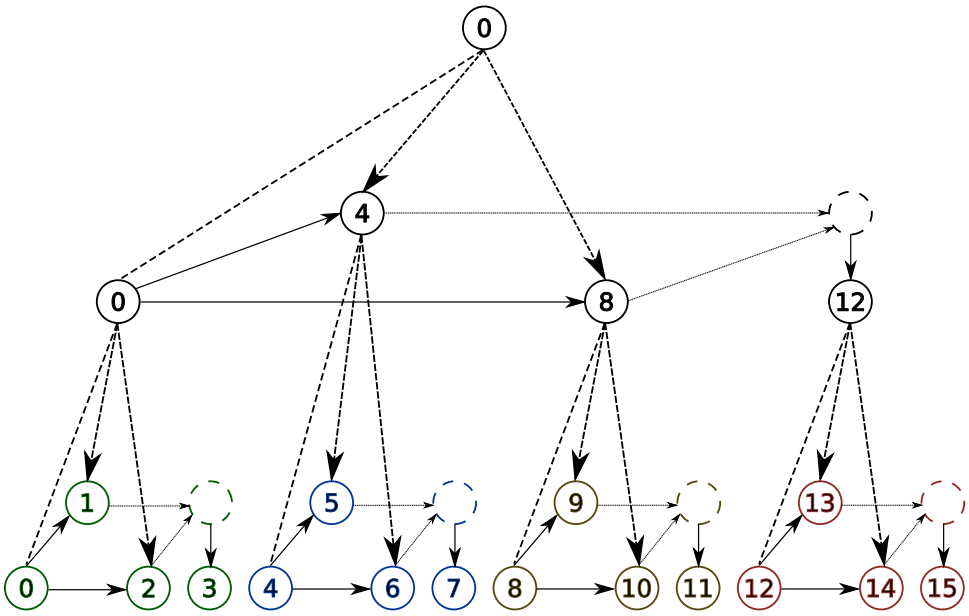


Figure 3.6: Representation as a tree of the relations resulting from a multi-scale analysis of a sequence of 16 elements, with the creation of a virtual element (that is, using the static S&C model). Here, the sub-sequence Y used in Equation 3.8 to pass from a scale to the upper one is $(0, 4, 8, 12)$.

a *virtual element* based on the first two relations. The difference with the tree systemic model is therefore that the last element of a system is described in relation to this virtual element which may not actually exist in the observed sequence, whereas the reference was the primer, in the Tree Systemic model. The connection between the lower-scale and the upper-scale one is achieved in the same way as for the tree systemic model, that is by creating a new sequence with only the primers of each lower-scale systems and describing the structure of this sub-sequence using the S&C formalism.

For sections that have more than two scales, the process can be recursively iterated: each sub-sequence is partitioned in systems of four elements and the sequence of the next scale is obtained by down-sampling the full sequence and taking only the primer. Fig. 3.6 represents the graph of relations for a sequence of 16 elements as a tree, while Fig. 3.7 represents the same graph projected on a 4-cube.

Formally, let r be a deterministic function that describes the relation between two elements. The antecedent function, $\Phi_{S\&C}$, is defined recursively as follow:

$$\Phi_{S\&C}(x_i|T) = \begin{cases} x_{i-1} & \text{if } i \pmod 4 = 1 \\ x_{i-2} & \text{if } i \pmod 4 = 2 \\ r(x_{i-3}, x_{i-2})(x_{i-1}) & \text{if } i \pmod 4 = 3 \\ \Phi_{S\&C}(y_{\frac{i}{4}}|r) \text{ in } Y = (y_l)_{0 \leq l < \frac{n}{4}} = (x_{4*j})_{0 \leq j < \frac{n}{4}} & \text{otherwise} \end{cases} \tag{3.8}$$

As for the Sys model, the function is also valid for sequences of length 2^{2p+1}. Moreover, the antecedent function can be applied on both the initial sequence and each permutation

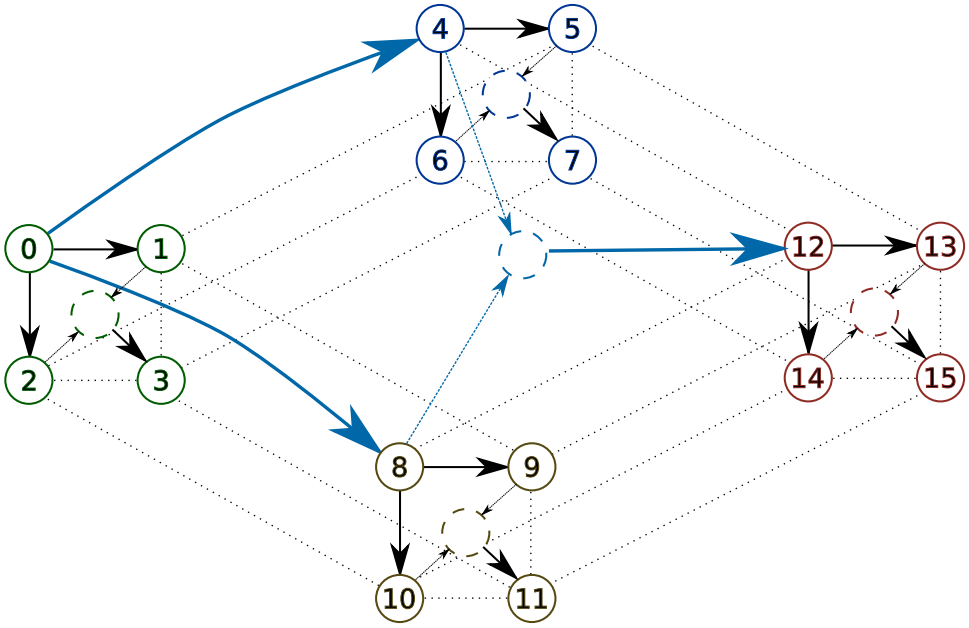


Figure 3.7: Representation on a tesseract of the relations resulting from a multi-scale analysis of a sequence of 16 elements with the creation of a virtual element (that is, using the static S&C model).

of the sequence. Therefore, for each PPP, a distinct model can be considered. This model can be indexed by its PPP number, for example: $S\&C_0$, $S\&C_1$... The model whose PPP minimises the description cost is referred to as $S\&C^X$ in Chapters 5 and 6.

3.3.6 Dynamic S&C Model

The *dynamic S&C model*, *Dyn*, also uses the PGLR formalism and the S&C concept. The difference with the static S&C model is that the number of systems of four elements used to compute the description is much larger, because there is no constraint of homology between lower-scale system. In fact, the main principle of the dynamic S&C model is that the description of a system of four elements is necessary for each element which can be seen as a contrastive element in the polytope. But an interesting thing about the polytopic representation is that for some elements, there are multiple possible systems to describe a single element which appears in a contrastive position. In such a case, the choice of the best system to be used is an optimisation problem that is explained hereafter.

Let a *choice* be a sequence of sets, each one containing four ordered indexes (in increasing order). Each set corresponds to a system used to explain an element in a metrically contrastive position. Therefore, the indexes of a set form a square (a face) in the polytopic representation and the last index corresponds to a possible contrastive element. The *choice* must contain a set for each element in a possible contrastive position in the sequence (4 sets for a sequence of 8 elements, 10 sets for a sequence of 16 elements...), N sets in the general case. The sets are ordered in the sequence in the ascending order of the contrastive index of each set, that is for $\mathcal{K} = \mathcal{K}_0 \dots \mathcal{K}_k \dots \mathcal{K}_N$ and $\mathcal{K}_k = \mathcal{K}_k^0 \dots \mathcal{K}_k^3$,

we have for all $k_1 < k_2$ and $j_1 < j_2$:

$$\mathcal{K}_{k_1}^{j_1} < \mathcal{K}_{k_1}^{j_2} \text{ and } \mathcal{K}_{k_1}^{3} < \mathcal{K}_{k_2}^{3}$$

An example of a choice for a 8-element sequence would be:

$$\{[0,1,2,3],[0,1,4,5],[0,2,4,6],[2,3,6,7]\}$$

A choice is therefore a set of faces on the tesseract, each one being used to describe one of all the elements of the section that are in a contrastive position.

For a choice $\mathcal{K}$ and an index i, we define $\mathcal{K}_{k_i^*}$ as the first set in $\mathcal{K}$ containing index i. That is for all $k < k_i^*$, $\mathcal{K}_k$ does not contain the index i.

Given a choice C, we can, a posteriori, build an antecedent function $\Phi_{S\&C_{Dyn}}$:

$$\Phi_{S\&C_{Dyn}}(x_i|\mathcal{K}) = \Phi_{S\&C}(y_l) \text{ in } y = (x_{\mathcal{K}_{k_i^*}^j})_{0\leq j<4} \text{ with } l \text{ subject to } y_l = x_i \tag{3.9}$$

When there are more than one possible system that can lead to an explanation of a contrastive element, a cost is associated with each system, based on its relation costs. Following the MDL principle, the system that is used for the description is the one with the minimal cost.

For example, looking at Figure 3.2, which gives another point of view of the polytopic representation of a sequence of elements, it appears that nodes 7, 11, 13, 14 are contrastive in three different implication systems and 15 in 6 implication systems. Therefore, there exists $3^4 * 6 = 486$ distinct choices for a sequence of 16 elements.

3.3.7 Relational Static S&C Model

The *relational static S&C model* is just an improvement of the static S&C model. Its main aim is to use the relation redundancies in an S&C description to create an even more simple description. That is, achieving a better compression by considering the redundancies between the S&C description of all the lower-scale systems.

To fully describe a section of 16 elements, $X = x_0x_1\ldots x_{15}$, a static S&C model, associated with any PPP, requires:

- the first element, x_0, which is the primer of the upper-scale S&C;
- the upper-scale relations, F, G and Γ, which describe the relations between the primers of the lower-scale S&Cs;
- the set of lower-scale relations, $(f_i, g_i, \gamma_i)_{0\leq i<4}$.

Under the hypothesis that there exists a relation formalism to describe the relations between relations, it is then possible to describe the lower-scale systems with three S&Cs

of relations instead of four S&Cs. In fact, it is possible to create an S&C to describe the relations between all f_i relations, and the same for g_i and γ_i.

The new description becomes:

- the first element, x_0, which is the primer of the upper-scale S&C;
- the upper-scale relations, F, G and Γ, which describe the relations between the primers of all the lower-scale S&Cs;
- $(f_0, F_f, G_f, \Gamma_f)$, the S&C to describe the relations between all "f" relations of the lower-scale S&C in the previous description.
- $(g_0, F_g, G_g, \Gamma_g)$
- $(\gamma_0, F_\gamma, G_\gamma, \Gamma_\gamma)$

For this model, the antecedent function of an element is the same as the one defined for the static S&C model. The only difference is the way the relations are encoded and the way the evaluation can be done.

In this chapter, we presented the detailed formalism of possible implementations of the PGLR framework, a multi-scale generalisation of the S&C model. The next chapter introduces the experimental data and methodology that have been used to compare these various possible schemes in different situations.

Chapter 4

Musical Data and Evaluation Method

In this chapter we develop how the S&C model (in its static form, defined in the previous chapter, see Section 3.3.5) can be used to manually analyse and describe a musical section. Then we present the data that are used in our experiments in computational structure analysis using PGLRs and the evaluation methodology.

4.1 Examples of Multi-Scale Descriptions

As the static S&C model generalises the 2×2 matricial S&C model to a multi-scale model, it is possible to use it to analyse and describe the main dependencies between the basic musical units at different scales simultaneously.

Considering again the examples from Section 2.3.2, it is now possible to improve the analysis of their musical content on the basis of smaller lower-scale elements. In fact, each of the 4 elements in the passage from "Smokid All Star" can be split in four smaller element (see Figure 4.1). Then, except from the last one, each element contains only one chord which is played either once, or twice.

Let us now focus on the "primer preserving permutation" PPP_0 (as defined in Chapter 3). PPP_0 is, so to say, the less disruptive polytopic configuration, as it follows more closely the time-line of the music). We will therefore consider a S&C description of each set of four consecutive elements.

Considering a purely chromatic logic, the first four elements (chords Gm, Cm, A and Cm again) form an S&C where function f (from Gm to Cm) changes two notes, by displacing the lower one by one semitone (d to e_{flat}) and the middle one by two semitones (b_{flat} to c), while keeping the third one unchanged. From similar considerations, function g (from Gm to A) changes all three notes of the two chords, upping them (from bottom to top) by two, three and two semitones respectively.

By combining these displacements for each note, one could logically expect in fourth position a chord containing the three notes (f, e_{flat} and a) (a $F7^{(no5)}$ chord).

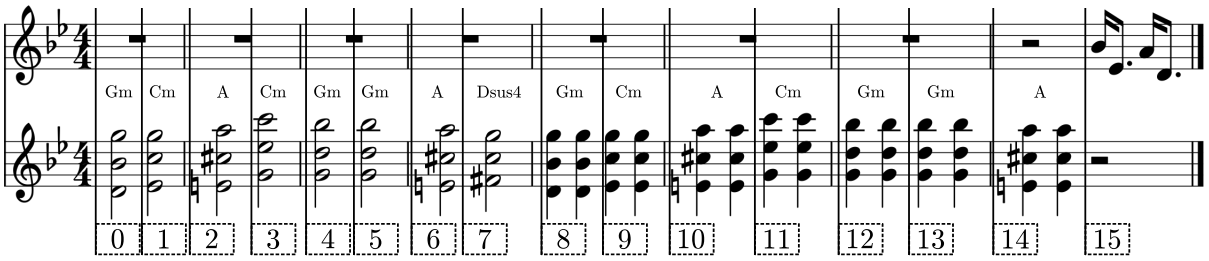


Figure 4.1: **Smokey Joe & The Kid - Smokid All Star (Feat. Waahli, NON Genetic, Pigeon John, ASM, Youthstar, Blake Worrell, Chill Bump, Dj Netik)** (comp.: Smokey Joe & The Kid) Smokid All Star, Benzaï Lab/Believe Music 2016. Timing: 0'35-0'56. *Transcribed by the author*

"Logically" is used here in the sense of strictly "mechanical" chromatic displacements. Listeners may not expect an $F7$ chord as the most "natural" continuation of Gm Cm A as it is more related to the Bb key than to the Gm key which may create a surprise when played after a A chord. Indeed, the logical implication which provides the most *compressible* element here may result (and generally does result) in a different element from the one that would be considered as the most *likely* one (which is in fact a very interesting feature).

Ultimately, the contrast function γ then describes the discrepancy between the implied chord, here $F7^{(no5)}$, and the Cm chord, actually observed: $+2$, 0 and $+2$ semi-tones, from bottom to top.

Here, it must be noted that a different "logic" may lead to a different expectation. For instance, a different way of pairing the notes between the first three chords in the system, could lead to D (instead of $F7^{(no5)}$) as the most rational counterpart of A, and the contrast function would be different.

Considering now the second group of four elements, as chords 4 and 5 are identical (Gm), the logical implication according to the 2×2 S&C model would be a repetition of chord 6 in position 7 (therefore an A chord, whichever logic is considered). The contrast lies in the $Dsus4$.

The third group of four elements $(8 - 11)$ follows the same chord pattern as $0 - 3$ and the last one $(12 - 15)$ has the same implication than the second group $(4 - 7)$ but the contrast is radically different and "surprising", as two diads appear on trumpets and wind instruments which were silent until then, while the chords played by the organ stop.

Looking now at the upper scale, we have a more global S&C which links those four S&Cs. The contrast inside this higher scale S&C is here the composition of contrasts from the second and the last group (which condenses in elements 7 and 15).

Note however that, in this configuration of the static S&C model, the upper system does not relate the whole lower scale S&Cs but only their primers, i.e. the first (0), the fifth (4), the (doubled) ninth (8) and the (doubled) thirteenth (12) chords. These are all Gm chords and they form a non-contrastive system (as the last doubled chord is exactly what one would logically expect from the observation of the three first chords). This upper-scale system is very simple to encode. But as a counterpart to this, it is quite

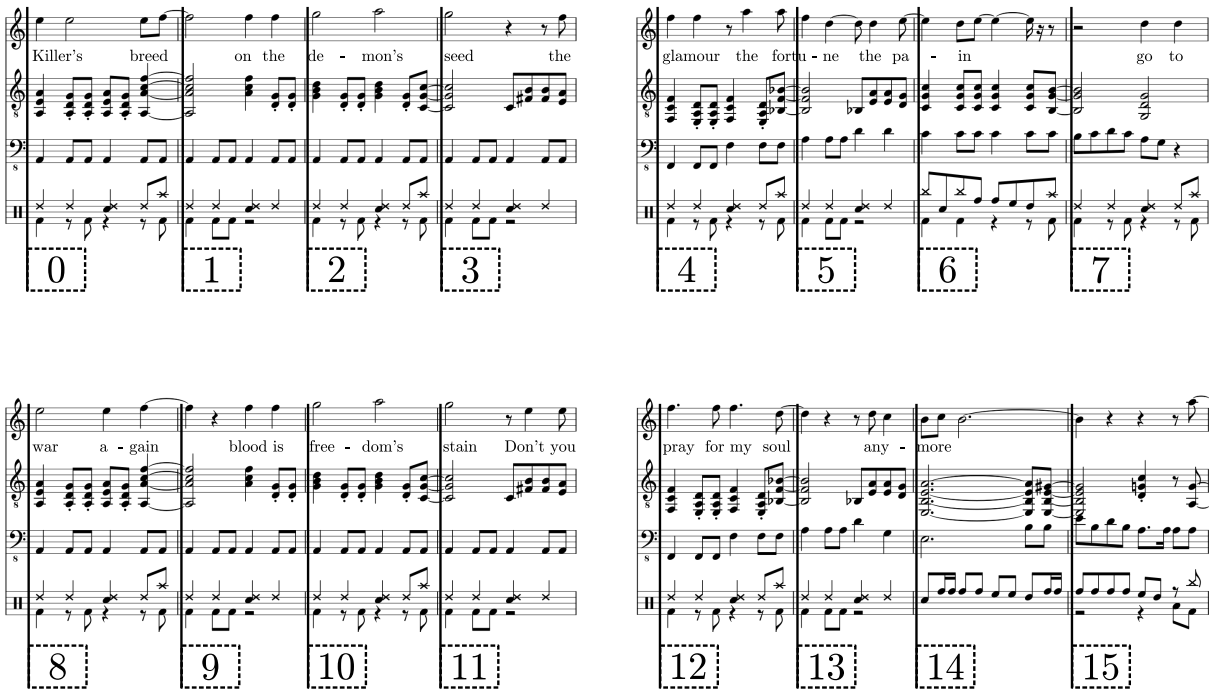


Figure 4.2: **Iron Maiden - 2 Minutes To Midnight** (comp.: Bruce Dickinson, Adrian Smith) 2 Minutes To Midnight, EMI 1984. Timing: 0'52-1'13. *Transcribed by ear*

complex to encode all the displacements and contrasts in the lower-scale S&Cs.

Considering another PPP, the picture is different. For example PPP_3, which relates $[0, 1, 8, 9]$, $[2, 3, 10, 11]$, $[4, 5, 12, 13]$ and $[6, 7, 14, 15]$ as lower scale S&Cs and $[0, 2, 4, 6]$ as the upper-scale one. In that case, the upper-scale is (almost) non-contrastive - one could expect some inversion of the A chord. At the lower scale, only the last S&C shows a real contrast. Therefore, the representation of the chord sequence using PPP_3 requires less information than PPP_0 to encode all the S&Cs.

If we consider now another one of the examples presented earlier (Figure 4.2), using PPP_3 would also reduce considerably the amount of information to encode the global structure, as only the upper- and the two last lower-scale S&Cs would be considered as contrastive. The third lower-scale system is almost only contrastive for the melody, as there is only one note which is contrastive for the bass. The guitar and drums are not contrastive at all. Therefore, considering each instrument separately, they may result in different types of description as the contrast vary in intensity for each instrument and system. However, it is very interesting to see that for all instruments, the PPP resulting in the simplest description is PPP_3, leading in a common structural description.

Moreover, by considering the last lower-scale S&C of PPP_3, $[6, 7, 14, 15]$, one may experience some difficulties to easily describe the logical or even the musical expectation for any of the instruments. Here, there is no obvious explanatory space of relations between the elements. This is one of the challenges in the S&C conception (and not an easy one, neither for a human being, nor for a computer). In fact, one of the contributions of this work is to explore possible relation spaces for different types of musical dimensions, as described in the next two chapters.

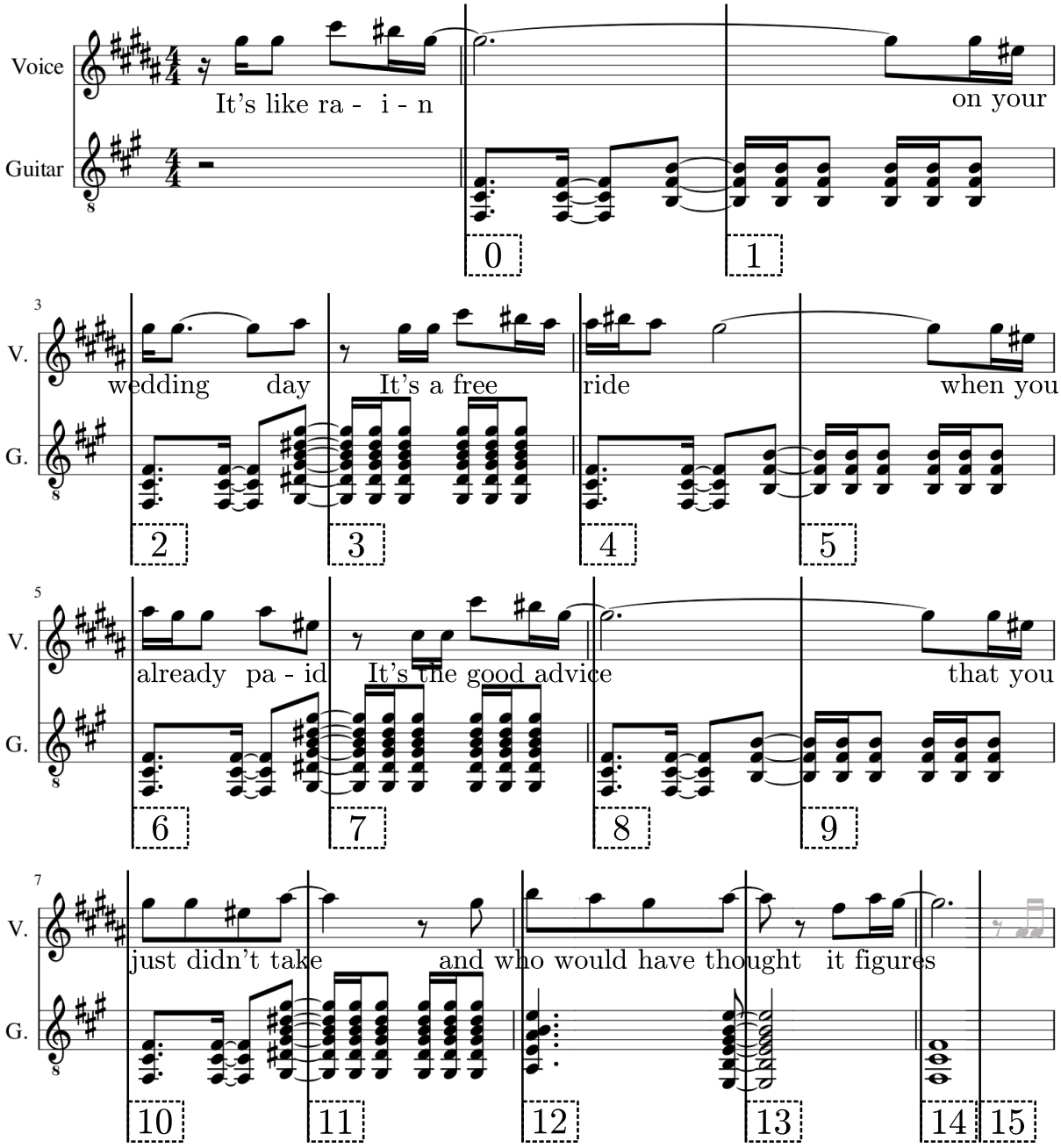


Figure 4.3: **Alanis Morissette - Ironic** (comp.: Glen Ballard, Alanis Morissette) Ironic, Maverick 1995. Timing: 0'42-1,06. *Transcribed by ear*

But let us consider now a third and last example, presented on Figure 4.3. This is a passage from one of the most famous Canadian pop-songs from the 90's. Here, the global form of the section is *aaab* and PPP_3 is (as it is often the case), the best PPP to efficiently describe the section's structure. However (and this will be discussed in Chapter 6), this example shows how the presence of anacruses creates a phase shift between the melodic motifs and their accompaniment. The incidence of that phenomenon tends to intensify at smaller scales, and needs to be accounted for (and dealt with) in the analysis process. Here, the beginning of the sentence "It's like rain" is clearly part of the

melody associated with the section, and it is particularly relevant to "take it onboard" to construct the systemic description of the whole section. But as it is sang in anacrusis, the notes are outside the segment's metrical boundaries as opposed to the other instruments. And corresponding motifs in the rest of the section are also ahead of phase. We show in Section 6.3 how this problem can be handled algorithmically.

The three previous examples illustrate the process that lies behind the algorithms that are presented in the forthcoming sections. They also point out some of the difficulties that need to be overcome. In particular, it reveals how important the data representation is, for finding a good space of relations.

4.2 Corpus Creation

For implementing, testing and evaluating algorithmic implementations of the PGLR model, appropriate music data are absolutely necessary.

This section presents the work that was done to exploit the RWC POP corpus with the goal to build a simplified corpus containing musical information, in an adequate representation, so as to be used for quantitative experiments.

4.2.1 RWC POP

The Real World Computing (RWC) Music Database has been designed and created for scientific use by Goto et al. [Goto et al., 2002]. The main goal behind their work was to create a database that could be publicly used by researchers and at a very small cost (virtually for free, just incurring the duplication and shipping costs).

A first advantage of such a database is to make possible to evaluate, on the same corpus, different methods on the same task. Performances can be compared and diagnostics of differences between them can be achieved in a much easier way.

A second objective was to create a consequent database that could be used in any scientific contexts without creating any copyright complications. For example, the use of some musical samples from commercial music at a conference may be an obstacle to the diffusion as a video on the web.

A third goal for creating this corpus was to provide a tool for researchers of the MIR community that would greatly increase their capacities to conduct some statistical studies. In fact, at the time, the lack of large music databases available was seen as an obstacle to research progress and the provision of RWC. This may be debatable today, where some MIR applications are required to handle collections of songs that are countable in millions of items... Nevertheless, the RWC database still offers an valuable benchmark for a number of MIR tasks and since its creation, the RWC database has been used in many occasions for a wide variety of tasks, such as declipping [Gaultier et al., 2017], music annotation [Bittner et al., 2014], music classification [Homburg et al., 2005], beat detection [Durand and Essid, 2016], structure analysis [Peeters and Bisot, 2014], etc...

The RWC database is distributed under a "research" license. It originally contains 215

songs and is divided into four sub-sets, RWC Classical, RWC Jazz, Royalty-Free Music Database and the last one, RWC Popular (RWC POP), which is the subset used in this work. The database was then enlarged to include two more categories RWC Genres, which contains 100 songs and Musical Instrument Sound Database (50 songs).

RWC POP consists in 100 songs, for which an audio (WAV) version and a standard MIDI file, transcribed by ear, are provided. Over this collection, 20 songs have lyrics in English and were composed in the style of American hits from the 80's. The other 80 songs have Japanese lyrics and were written in the 90's Japanese style. Many composers, arrangers and writers have been called for to compose these songs so as to ensure a significant variety of style or influences over the collection.

4.2.2 Chord Annotation

As Guichaoua [Guichaoua, 2017] worked on chord progression analysis, a first need for his experiments (as well as ours) was to incorporate harmony-related information to the corpus, that is providing an annotation of chords which could be used along with the S&C model. The interest of the harmonic dimension is that it is often used in standard musicology to describe the structure of a musical phrase [De Haas et al., 2009] or even a piece [Hepokoski and Darcy, 2006].

The description of the structure of a chord sequence using the PGLR model may be very useful to characterise the structure of a section. In fact, chords and/or harmony are often (yet not always) an essential musical dimension to inform music structure. However, given a musical section (in audio, midi or even as a score), it can be hard to find the associated chord sequence, as such information is derived indirectly from the data. The problem of chord annotation is indeed a major subject of interest in the MIR community and the representation of the chord progressions may differ w.r.t. the method used to infer the chord sequence associated with a section.

Several types of representation of the chord content of a musical section co-exist:

- Chord progression, where only the information relative to the successive chords is kept. In such a case, chords are considered as following each other without any duration or onset information. That is each chord of the sequence is annotated with its name (a symbol from a list of predefined possibilities) and each chord has the same importance, independently of its duration or onset.

- Chord segmentation, it is the most commonly used annotation for audio files. It consists in annotating a recording with, at each chord change, the name of the new chord, its onset (in seconds) and its duration (in seconds). In such a representation, a chord that has a long duration may have more importance than one with a short duration. However, as the annotation is based on a recording, it is annotated in seconds and it may be hard to associate a precise musical time with the temporal duration of the chord. In fact, the problem of alignment between midi files and audio files, which consists in associating musical duration to temporal duration, is also a problem of interest in the MIR community [Hu et al., 2003, Raffel and Ellis, 2016].

- Chord grids, which consist in representing the harmony of a section using a schematic point of view. A discrete scale of description is chosen (often a bar or a beat) to assign to each musical temporal unit, a chord name corresponding to the principal chord being played (or perceived) at this time. This representation combines the previous representation with a discrete representation that is musically relevant. However, if the scale is too large, some harmonic information may be missing in the final chord sequence. On the opposite, if the sampling is too frequent, it may create a great amount of redundancy.

For our study, the chord grid is the most suitable representation, as the PGLR model is mostly designed to describe the relations and dependencies between events that are taken at specific metrical positions (each beat, strong beat or bar). For chord representation, we decided to use a scale that is related to the tempo. Generally it is set to a beat, but for some fast tempos, the scale may be set to a half-beat while for slow tempos, the scale is set to two beats duration. This representation may create a lot of redundancies, but it is precise enough to guarantee that models may infer some structure over a section. Diminishing further the scale could be used to detect some very small changes of chords, but it would drastically increase the computation time of the analysis of a section whereas the number of cases where the precision would be useful is not sufficient to justify such an augmentation of the computation time.

Using some semi-automatic method, Cho et al. [Cho, 2011] obtained a chord annotation of the songs from RWC POP. But its main problem is that the result is provided together with an audio segmentation representation. That is, each chord duration and onset is given in seconds instead of musical unit. Fortunately, Goto et al. [Goto, 2006] also gave a beat annotation for the songs of the dataset. Then, by taking at each beat one of the chord played or sustained on the beat, it is possible to get a chord grid representation. However, as for some beats, there may be two chords played or as a change may occur off-beat, it may create some ambiguity. This is why the 100 RWC-Pop data were post-processed manually [Guichaoua and Bimbot, 2018] to ensure a better consistency of the annotated data with the metrical information, yielding to three versions of annotation, with different levels of manual intervention. In our experiments, we use the last one, i.e. the "cleanest" annotation.

For each song, the annotation is represented as a CSV file where each line represents a section, characterised by a (semiotic) label, the length of the section (in beats) and the sequence of chords grouped by cells of four beats. Some additional conventions were used to simplify the annotation:

- if there is only one chord in a cell, the chord lasts four beats;
- if two chords in a cell are separated by a comma, then each chord is considered as lasting for two beats;
- if two chords separated by a comma are also in parenthesis, then these two chords should be considered to be each lasting for one beat;
- if a chord is replaced with a N, it means that the corresponding time interval is filled with silence or non-pitched content;

	A	B	C	D	E	F	G	H	I	J	K
1	H	8	A&9	A&9							
2	I	32	C#m7	F#m7	Abm7	AM7,(AM7,B)	C#m7	Abm7	F#m7	D6,Ab7	
3	A	32	C#m7	C#m7	F#m7	F#m7	Abm7	Abm7	AM7	Ab7	
4	B	32	C#m7	C#m7	F#m7	F#m7	Abm7	C#m7	DM7	Ab7	
5	P	36	AM7	B6	Ab7d9*	C#m7	F#9*	F#9*	F#m7	Abm7,F#m7	Em6,D6
6	C1	32	E	Abm	Bm6	C#m7	F#m7	Abm7	Am7	B7s4	
7	C2	32	E	Abm	Bm6	C#m7	F#m7	Abm7	Am7	B7s4	
8	J	8	E	D6,Ab7							
9	A	32	C#m7	C#m7	F#m7	F#m7	Abm7	Abm7	AM7	Ab7	
10	B	32	C#m7	C#m7	F#m7	F#m7	Abm7	C#m7	DM7	Ab7	
11	P	36	AM7	B6	Ab7d9*	C#m7	F#9*	F#9*	F#m7	Abm7,F#m7	Em6,D6
12	C1	32	E	Abm	Bm6	C#m7	F#m7	Abm7	Am7	B7s4	
13	C2	32	E	Abm	Bm6	C#m7	F#m7	Abm7	Am7	B7s4	
14	I	32	C#m7	F#m7	Abm7	AM7,(AM7,B)	C#m7	Abm7	F#m7	D6,Ab7	
15	X/A	32	C#m7	C#m7	F#m7	F#m7	Abm7	Abm7	AM7	Ab7	
16	Y/B	36	C#m7	C#m7	F#m7	F#m7	Abm7	C#m7	DM7	Abm7,F#m7	Em6,D6
17	C1	32	E	Abm	Bm6	C#m7	F#m7	Abm7	Am7	B7s4	
18	C2	32	E	Abm	Bm6	C#m7	F#m7	Abm7	Am7	B7s4	
19	I	32	C#m7	F#m7	Abm7	AM7,(AM7,B)	C#m7	Abm7	F#m7	D6,Ab7	
20	K	24	AM7	AM7	A	A	A	A			

Figure 4.4: CSV file that gives the harmonic description of the seventh song from RWC POP.

- if a chord is replaced by a "%", it means then the corresponding duration is deleted (for example, for ternary bars).
- if a cell information is between brackets, it means that the corresponding beats are also to be considered as the beginning of the next section, even though they can also be used to describe the end of the current section. This happens when two sections overlap.

For example a cell containing $C, (C, Dm)$ represents the sequence C, C, C, Dm, whereas if the cell contains $(Am, N), (E, \%)$ then it means that the cell contains only three beats and the corresponding sequence is Am, N, E (where N indicates that the second beat is silent). The CSV file for song RWC POP 07 is represented on Figure 4.4.

4.2.3 Creation of the New Corpus

As working on chord sequences was only a part of the objectives of this thesis, we worked on building a corpus which also contains other musical dimensions, namely melody and rhythm. Fortunately, the RWC POP database contains a midi file for every song as well as the corresponding audio file. Therefore, extracting the melodic line of each song was manageable.

However, the segmentation of the piece in small segments was initially made with audio files, and, unless calling for an alignment algorithm that may give some inaccurate results (or doing it manually!), there was no segmentation of the midi files. Analysing each melodic section would then have been impossible. But, as some segmentation information was present together with the chord annotations (and as the precision scale used for the

chord sampling is known), it was possible, by aligning a melody with its chord annotation to determine section boundaries. This global alignment was done manually (which did not take not too long), using both audio files and midi files.

By condensing all the melodies, chord sequences and segmentation annotations in one XML file for each, we created a new corpus which is a simplified (and reduced) version of the RWC POP corpus. In this format, the corpus can very easily be parsed automatically, using the *music21* Python library.

An excerpt of such a file is presented on Figure 4.5.

4.2.4 Test Corpus

Once all songs were annotated, we built different sub-corpora containing only the sections that needed to be used as input for the algorithm, to provide a structural description of the section. As our current implementation of the S&C model can only be used on sections of 16 elements, we had to select sections of length 16 or 32 in the corpus.

In this study, we present experiments that consider each dimension separately. Some perspectives on multi-dimensional structure description where all dimensions may be analysed together will be mentioned in Section 7.2, but they were not actually tested in the scope of this work. Therefore three parallel corpora have been created : one for the harmonic analysis of the structure, one for the rhythmic information and one for the melodic content.

In the harmonic case, when two sections have the same label, only one is kept in the corpus (section labels differ at least in terms of subscripts or superscripts as soon as their harmonic content is different in one position). For rhythmic experiments, all sections of proper size have been kept, as there may be some variations from one instance of a section to another, even if they have identical labels.

Moreover, as the midi file for song 56 was not exploitable, but the corresponding chords had been annotated, the sections of this song are present in the chord corpus but not in the melody one. This results in total in a small difference of number of sections in each corpus : the chord corpus contains 727 sequences whereas the melodic corpus (used both for melodic and rhythmic experiments) contains 791 sections.

Finally, for the chord corpus, as most of the sequences contain not 16 but 32 chords, we sub-sampled these sequences by taking one chord out of two (the one played on strong beats). This can be seen as changing the scale precision for the chord grid of the corresponding section.

Ultimately, in the cases when a chord sequence contains a N, i.e. a void chord (silence or non-harmonic beat), the N is replaced with the immediate (non-N) predecessor[1]. This choice of completion can be justified by the fact that, even after a silence, the last chord played is usually still perceived by the listener.

[1]Note that no sequence of the corpus that have a real chord begins with a void chord. In such a case, another completion rule should have been added.

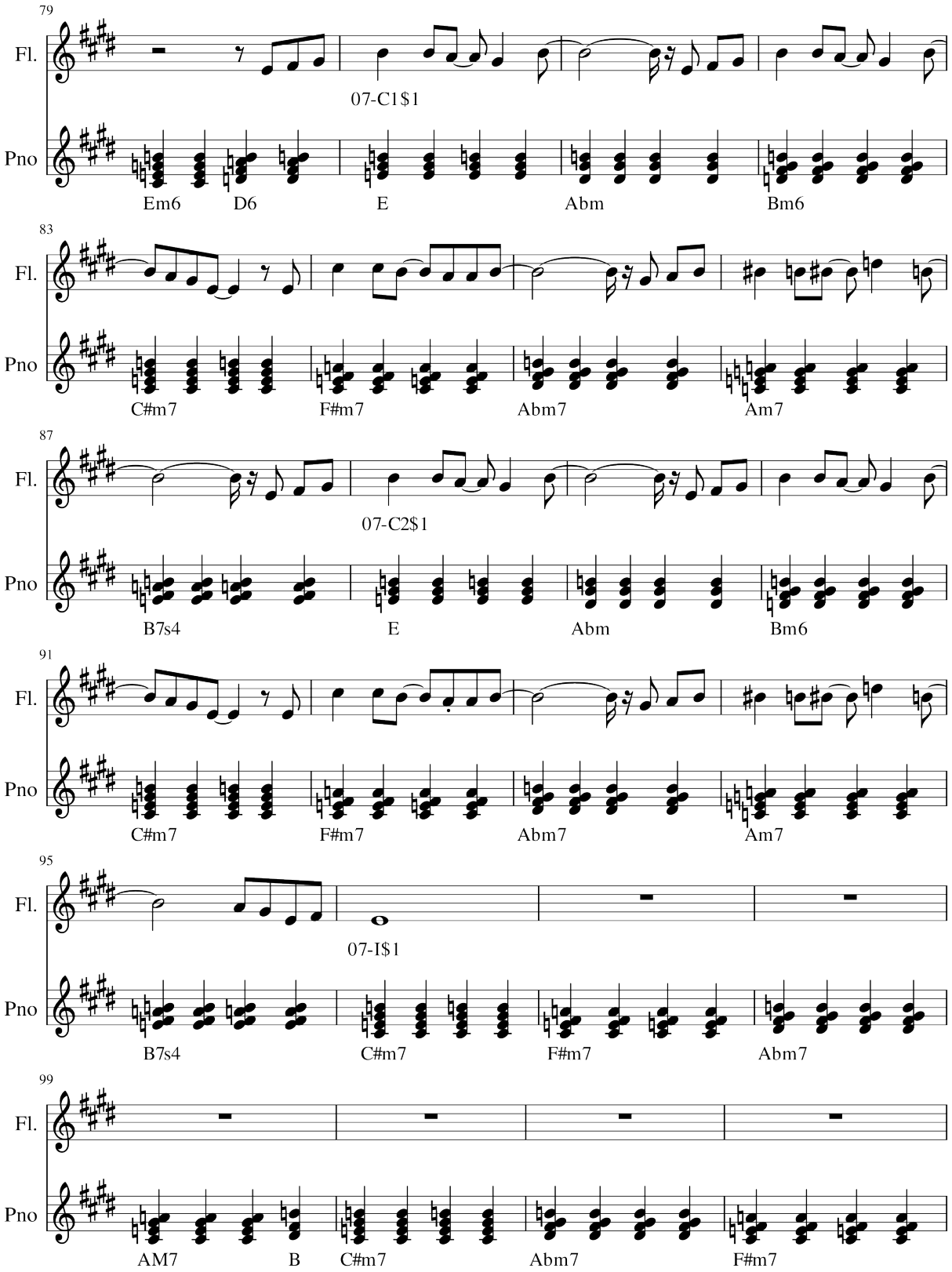


Figure 4.5: Part of the XML file, opened with musescore, that gives both the harmonic, melodic and segmentation description of the seventh song from RWC POP.

4.3 Evaluation Methodology

As there exists no "ground-truth" as of the actual structure of a music section, we considered that the different models could be compared as regards their prediction ability.

Under this approach, performance for each model is obtained by calculating a *perplexity* [Jelinek et al., 1977, Brown et al., 1992] B^*, derived from the amount of *entropy* H^* in the data that remains unexplained by the model. This is estimated by measuring how well an unseen sequence, $X = x_0 \dots x_{n-1}$, can be predicted by the model. The more powerful the model, the lower the perplexity. Alternative measures such as predictive information rate or other measures defined by Abdallah et al. [Abdallah and Plumbley, 2009] could have been used here. However, as these other measures are functions of time, they are more suited for segmentation or event detection than for prediction evaluation[2].

Given a model M, the computation of perplexity requires the definition of a probability mass function for all observable events which underlie the model. In fact, the ability of M to predict a sequence X can be assimilated to the probability, $P_M(X)$, that sequence X can be generated by model M.

Using the chain rule, this probability can be defined as:

$$P_M(X) = P_M(x_0 \dots x_{n-1}) \tag{4.1}$$
$$= P_M(x_0)P_M(x_1|x_0)P_M(x_2|x_1, x_0) \dots P_M(x_{n-1}|x_{n-2}, \dots, x_0) \tag{4.2}$$

When using Markov first-order approximation :

$$P_M(x_i|x_{i-1} \dots x_0) \approx P_M(x_i|x_{i-1}) \tag{4.3}$$

But here, even if all models that are considered are first-order models, the dependencies of the events are not necessarily sequential. We consider that an element may depend more strongly on another element than its direct predecessor.

This can be achieved with the antecedent function of a model, Φ_M, which defines which event is used as an antecedent in a PGLR. Therefore, the first-order approximation used to estimate $P_M(x_i|x_{i-1} \dots x_0)$ is, in all cases:

$$P_M(x_i|x_{i-1} \dots x_0) \approx P_M(x_i|\Phi_M(x_i)) \tag{4.4}$$

Note that, when using the sequential model for which $\Phi(x_i) = x_{i-1}$, the approximation is exactly the same as the first-order Markov chain.

To specify further the probability value $P_M(x_i|\Phi(x_i))$, we assume in this work that it can be approximated as:

$$P_M(x_i|\Phi(x_i)) \approx P_M(r(\Phi(x_i), x_i)) \tag{4.5}$$

where $r(\Phi(x_i), x_i)$ is the (forward) relation which turns $\Phi(x_i)$ into x_i (i.e. that transforms the antecedent into the current element). Note that therefore:

$$r(\Phi(x_i), x_i)(\Phi(x_i)) = x_i \tag{4.6}$$

[2]Moreover some of these measures are closely related to the perplexity (resp. entropy) and the cross-perplexity (resp. cross-entropy).

Ultimately, for each set of relations $\mathcal{S}$, the probability mass function of the relations is discrete and finite. We estimate the probability $P_M(r)$ of a relation r to appear in a description based on model M as:

$$P_M(r) = \frac{1 + Occ(d, \mathcal{C}_M)}{|\mathcal{S}| + \sum_{s \in \mathcal{S}} Occ(s, \mathcal{C}_M)} \tag{4.7}$$

where $\mathcal{C}_M$ is the training corpus associated with the model M (that contains all the relations used for the descriptions of each section within the training corpus) and $Occ(d, \mathcal{C}_M)$ is the number of occurrences of relation r observed in $\mathcal{C}_M$. Note that, rather than being the exact frequency of occurrences of r, we use a back-off technique in order to avoid null probability values.

For our experiments, we use a 2-fold cross-validation strategy to estimate probabilities and compute entropy (and then perplexity) scores. That is, we split the corpus into the even indexed songs and the odd indexed songs. The measured entropy is obtained as the average between the two partial entropies calculated by (i) learning probabilities on even songs and compute the entropy on odd songs and (ii) vice-versa.

We hypothesise an *a priori* uniformity in the distribution of the first element of each sequence and therefore, considering a finite number E of possible elements, the initial probability is estimated as $P_M(x_0) = \frac{1}{E}$ (this preserves comparability between the models).

Using these probabilities estimation, the entropy of model M can be computed as:

$$H^*(M) = -\sum_{k=0}^{z-1} P_M(r_k) \log_2 P_M(r_k) \tag{4.8}$$

This entropy quantifies, as an average number of bits per relation, the quantity of information conveyed by each relation. The more frequent is a relation, the less information it carries. Conversely, a relation that has a low probability value carries more surprise and then more information. The entropy of the probability mass function reflects the minimal number of bits required to describe each relation within the relation space.

Based on the entropy, it is ultimately possible to compute the *perplexity* as:

$$B^* = 2^{H^*} \tag{4.9}$$

B^* can be interpreted as a branching factor, that is, the average equivalent number of distinct relations between two elements, if all relations are equiprobable. It measures the compression capacity of the model and is smaller for models which capture more information in the data.

In this work, we consider specifically the *cross-perplexity* $\hat{B}$, which is obtained from the *negative log likelihood* (NLL) $\hat{H}$, computed on a test-set. In that case, the capacity of the model to catch relevant information from an unseen musical section is measured by means of a cross-entropy score, which quantifies the ability of the model to predict unknown sequences from a similar (yet different) population.

For a given model M and a sequence $X = x_0 \dots x_{n-1}$, $\hat{H}_M$ is defined as:

$$\hat{H}_M(X) = -\frac{1}{n} \sum_{i=0}^{n-1} \log_2 P_M(x_i|\Phi(x_i)) \tag{4.10}$$

with the convention $P(x_0|\Phi(x_0)) = 1/|\mathcal{S}|$.

In that context, the cross-perplexity $\hat{B} = 2^{\hat{H}}$ can be understood as an estimation of the average (per symbol) branching factor in predicting the sequence knowing its structure, on the basis of probabilities learnt on *other* sequences.

Additionally, for the models $S\&C_X$ and Sys_X, we also compute the total entropy $\hat{H}_{Mtot}(X) = \hat{H}_M(X) + Q_M$, which takes into account the number of bits needed to encode the optimal configuration of the PPP (1 among 6) for each sequence of 16 chords, namely:

$$Q = \log_2(6)/16 \approx 0.16 \text{ bits/symbol} \tag{4.11}$$

For the dynamic model (Dyn), as there are more distinct possible graphs (486), the encoding cost is larger (0.56bits/symbol). This term is equal to 0 for all models that have a unique distinct possible graph (Seq, Sys_i, $S\&C_i$).

The next two chapters introduce and study several relation formalisms for three musical dimensions: chords, rhythms and melodies. The different models presented in Chapter 3 are compared to one another, using the perplexity measure defined in this chapter. This is the occasion to investigate the impact of various properties of the different models and formalisms and their relevance for the analysis of the structure across each musical dimension.

Chapter 5

Application to Chord Sequences

The main focus of this chapter is to describe relation formalisms that were investigated along with the PGLR approach for modelling chord sequences, specifically. Firstly we identify a few contextual aspects of this particular task, then we present multiple formalisms to describe the relations between chords. Finally, the formalisms are combined with the different models presented in Chapter 3 and evaluated on a prediction task.

5.1 Context and Practical Considerations

5.1.1 Harmony

In standard musicology, the study of harmony (i.e. abstract mechanisms that govern melody and chord progressions), undeniably occupies a central place. As time passes, the interest for this domain does not decrease. As a result, many formalisms for chord progressions description have been designed, such as [Piston, 1948, Schoenberg, 1983, Cohn, 2011].

The plurality of formalisms for the description of chord progressions has the advantage to provide plenty of representations and relation formalisms between chords. Describing the relation between two chords as a movement in a space with specific properties is indeed a well-established principle. It results in well-defined vision of the description of chords and relations between them, which can be used, in musicology, to describe compositional principles, cultural idiosyncrases, as well as in automatic music processing, for the design of algorithmic methods.

However most of the approaches to analyse chord progressions consist in describing the relation between chords that are sequentially successive or within adjacent groups of chords. In fact, to describe the structure of a chord progression, a usual approach consists in using the tonal function of the chords [Bigand and Parncutt, 1999, De Haas et al., 2009, de Haas et al., 2011, Déguernel et al., 2017] and then describe the relation between chords as a functional progression. For these models the adjacency between chords and groups of chord is very important. Here, as the polytopic models consider some non-sequential relations, it partly departs from standard musicological assumptions.

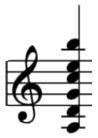
	Representation		
Full Name	Musical	PCS	TR
Am^{11}		$\{9, 0, 4, 7, 11, 2\}$	Am

Figure 5.1: Musical, Pitch Class Set (PCS) and Triad Reduction (TR) representations of the chord Am^{11}.

Considering the PGLR approach, the aim is to use existing concepts in the musicological and MIR communities, to describe in a formal and quantifiable way the relations between chords. In fact, to be able to compare the different models, it is important to choose some formalisms for describing relations that can be used with different types of models and that can describe both the relation between adjacent chords and long-term related chords.

Moreover, if we refer back to Equation 3.5, the formalisms must be provided with a criterion on which it is possible to base the computation of a complexity measure from the relation between two chords. Another important point is that, to be used with the S&C model, the formalisms of these relations must not only describe the relation between two chords but also provide a way to characterise the virtual element. Therefore, some adjustments had to be made to adapt the concepts of the musicological and MIR domain to the PGLR analysis of the structure of chord sequences.

5.1.2 Chord Representations

Speaking as generally as possible, a chord, in music, is a set of notes (here represented as "pitches") that are heard as if sounding simultaneously. However, in tonal western music, chords are more generally conceived as set of *pitch classes* supporting the local harmonic ground plan of the music. In particular, chords play a strong role in the accompaniment of the melody in pop songs.

The most frequently encountered chords are triads (i.e. sets of three pitch classes), with a predominance of major and minor triads. More sophisticated chords contain combinations of 4 pitch classes or even more.

Chords can be represented in various ways. In this chapter, we consider two types of representations that are most commonly used in MIR: (*i*) the complete set of pitch classes forming the chord (PCS description) and (*ii*) the tabular notation of the major or minor triadic reduction of the chord (TR description). Assuming 4 or 5 pitch classes per chord, this leads to potentially several hundreds of different PCS descriptions (much less in practice), but only 24 distinct TRs.

For example, following *i*, the chord Am^{11} (see Figure 5.1), would be represented by the set: $\{9, 0, 4, 7, 11, 2\}$, while its reduction using *ii* would simply be Am (or $\{9, 0, 4\}$).

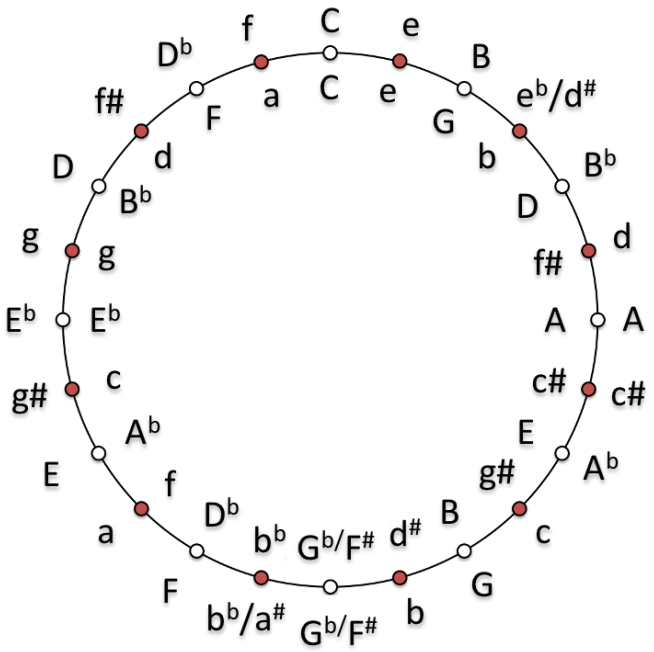


Figure 5.2: Triad circle of phase-shifts (outside) and triad circle of thirds (inside).

5.2 Relation Formalisms

As mentioned earlier, a number of formalisms exist to describe chord relations, either in classical musicology (through chromatic relations or via displacements over the circle of fifths, which is the circle of notes obtained by transposing a note to its fifth at each step) or in the framework of more recent theories, in particular Wietzmann regions [Weitzmann and Saslaw, 2004] or neo-Riemannian theory [Cohn, 2011]. Following a similar point of view, Tymockzo [Tymoczko, 2006, Tymoczko, 2008] also proposed a model based on combinations of chromatic and scalar transpositions.

Depending on the formalism under consideration, the property of uniqueness of the relation between two chords may or may not be satisfied.

5.2.1 Triad Circles

This subsection focuses on the relation formalisms that can be used along with the TR representation of chords.

We call *triad circle* any circular arrangement of triads aimed at reflecting some proximity relationship between triads along its circumference. As there are a lot of possible criteria to define proximity between chords, we decided to consider the most common one which is used in general musicology, that is the criterion of voice-leading proximity. Or, more precisely speaking, a simplified version of it.

Two chords are considered to be close to each other if (a) they share some notes and (b) if one can obtain the second chord by moving by a small number of semitones the notes of the first chord. That is, two chords are close if their (matched) notes only differ

by a small number of semitones.

In this study, only two circles will be considered: the *circle of thirds* and the *circle of phase-shifts* (see Figure 5.2).

The circle of thirds is formed by alternating major and minor triads where the neighbouring triads share two common pitch classes and where the third pitch class is away of 1 or 2 semitones from the other third pitch class. This circle can be generated by alternating the two neo-Riemanian transformations: R (stands for Relative) which change Am to C by moving the fundamental down by two semitones, and L (stands for Leading-tone exchange) which change C to Em by moving the fundamental one semitone down. The interest of such a circle is that chords that belong to a same tonality are grouped together on a sector of the circle.

The circle of phase-shifts consists of a chord progression which results from a minimal displacement on the 3-5 phase torus of triads as defined in [Amiot, 2013]. The 3-5 phase torus of triads is 2D torus defined in $\mathbb{C}^2$ by the two parametric equations:

$$|a_3| = 2.236 \text{ and } |a_5| = 1.93185 \tag{5.1}$$

Where a_3 and a_5 are the third and fifth coefficients of the Fourier transform of the triads. The coordinates in this parametric space are $(\arg(a_3), \arg(a_5))$.

The circle is obtained after applying the rotation $(-5\pi/6, -\pi/2)$ to the coordinate of each major triad, which is equivalent to transposing each major triad by half a semitone in the pitch space. Moreover, when unfolding the torus (by reversing the Fourier transform) it appears that the direct path linking two adjacent major triads comes close to a unique minor triad. For example, the closest minor triad between C and $C\#$ arc is Fm. Therefore, it is possible to insert, on this basis, minor triads in the circle of major triads and create a regular progression in the 3-5 phase space. The final circle is shown on Figure 5.2.

As the 3-5 torus of phase-shifts is a particular representation of the Tonnetz, this new circle (triad circle of phase-shifts) is, like the circle of thirds, a circular representation of a path in this specific chord space.

These two representations provide a way to express relationships between two TRs– in a unique way– as the angular displacement around the circle. Moreover, it is easy to prove that such a space of representations and relations verifies the properties defined in Section 3.3.2, i.e. that it is a commutative group which can be used to describe any relation between two chords, such that any relation can be applied to any chord to create another chord in the same representation space and such that all relations are invertible.

In fact, if considering the triads over the circle as integers modulo 24, the relation between two chord c_1 and c_2 simply becomes $c_2 - c_1 \pmod{24}$ which is always an integer. Applying a relation to a chord corresponds to an addition modulo 24, which provides an integer also lying in $[\![0; 23]\!]$. Moreover, for every chord c, and every relation r there is a unique antecedent of c by r which simply is $c - r \pmod{24}$. This resides in the fact that $\mathbb{Z}_{24}$ is a cyclic group.

The two circles of triads under consideration in our study are particularly interesting in the sense that they may be easily used to interpret the modelling results, as they are

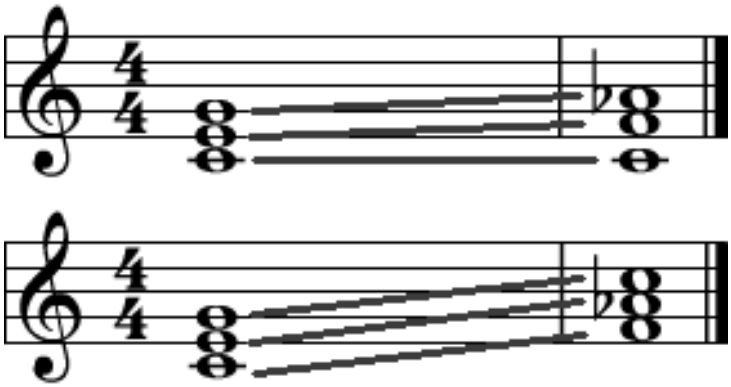


Figure 5.3: Transports $\{(0,0),(4,5),(7,8)\}$ and $\{(0,5),(4,8),(7,0)\}$ between C and Fm.

well-known by the musicological and MIR communities. But in order to further support their relevance, we also consider in our forthcoming experiments, random circles, i.e. circles on which the 24 triads are placed randomly.

Following Equation 3.5, the description cost of a relation based on triad circle formalisms can be computed as the absolute value of the rotation on the circle to move from one triad to the other one. Each individual cost can then be added into a global cost for the whole sequence using Equation 3.5.

5.2.2 Optimal Transport

This subsection focuses specifically on the description of relations between two chords represented using the PCSs. Here we consider only chords that have at least one note. Note that the triad reduction of a chord can also be represented using PCSs, as the set of three notes representing the triad reduction.

If two chords X and Y are represented as sets of pitch classes x_i and y_j, the set of *transports* between X and Y can be defined as:

$$T = \{t_k = (x_{i_k}, y_{j_k}) \mid x_{i_k} \in X,\ y_{j_k} \in Y\} \tag{5.2}$$

that is, pairs of notes across the two chords indexed by an integer k which represents a virtual mapping between their respective pitch classes. This is a simplified model that can be used to represent "voices" in chord sequences. In this study, we consider only complete transports, i.e. each note is associated to at least one voice.

Formally speaking, a transport T between X and Y is said to be complete if and only if it satisfies:

$$\begin{cases} \forall x \in X,\ \exists t \in T,\ \exists y \in Y\ /\ t = (x,y) \\ and \\ \forall y \in Y,\ \exists t \in T,\ \exists x \in X\ /\ t = (x,y) \end{cases} \tag{5.3}$$

The optimality of a transport between two chords is defined using the distance asso-

ciated with the pitch class displacement of the transport:

$$|T| = \sum_{(x,y)\in T} |d(x,y)| \tag{5.4}$$

The term $d(x,y)$ is a displacement in the pitch class space, $|d(x,y)|$ is the corresponding distance. In our work, we use two types of distances:

- the *chromatic distance* (or smoothness) [Lewin, 1998, Straus, 2003, Cohn, 2011], which is the shortest displacement in semitones from pitch class x to pitch class y. It is defined as:

  $$d(x,y) = ((y-x+5) \pmod{12}) - 5 \tag{5.5}$$

  In Fig.5.3, the first transport is minimal for chromatic distance (its cost being equal to 2).

- the *harmonic distance*, where the displacement is considered on the circle of fifths instead of the chromatic scale. It can be defined as:

  $$d(x,y) = (7(y-x)+5) \pmod{12}) - 5 \tag{5.6}$$

  In Fig.5.3 the second transport is minimal for harmonic proximity (its cost being equal to 6).

It is important to note that the transport cost can be used as a relation cost in reference to Equation 3.5.

The algorithm used to find the optimal transport between two chords is the most naive one. It tests all possible complete transports between the two chords (that is $n!$ possible cases if n is the number of pitch classes in X and Y) and chooses the one with the minimal cost. But as the two chords may have different sizes, the algorithm considers every possible duplication of notes of the smaller chord, so that it has the same number of notes than the larger chord. Therefore the total number of possibilities of complete transport, $CT(X,Y)$, between X of size $|X|$ and Y of size $|Y|$ is defined by:

$$CT(X,Y) = (\min(|X|,|Y|))^{||Y|-|X||} * (\max(|X|,|Y|))! \tag{5.7}$$

The fact that the algorithm that computes the optimal transport between two chords adapts the number of notes of the smallest chord so that it matches the number of notes of the largest one, has some consequences. As in a S&C description, the graph of latent relations is connected, i.e. every chord is linked to any of the other chord by a path in the graph. Therefore, to optimise all transports in a sequence, it is necessary to adapt the size of *all* chords in that sequence so that it corresponds to the one of the largest chord of the sequence.

Fortunately, in practice $|X| = |Y|$ or $||Y|-|X|| = 1$ and the maximum number of notes in the largest chord of the sequence is 4 which makes the optimisation of the transport between two chords reasonably tractable. This is the reason why we could afford to use a candid algorithm, and did not consider more sophisticated ones, as proposed or discussed

in [Kantorovitch, 1958, Villani, 2003]: these are indeed much harder to implement, as they are designed for transport between objects of a much higher size such as images (see for instance [Ferradans et al., 2013]).

The initial optimisation process that we investigated consisted in finding the optimal transports within the system made of f, g and γ such that the sum of the three transport costs is minimal. This meant testing all possible combinations of relations, as γ depends on the conjunction of f and g. However, with such an approach, the contrast was bound to influence the description of the system. After considering this, we felt that this situation was not desirable, for two reasons: (a) because this would mean that the contrast could have a non-causal influence on the choice of the description of f and g and (b) because two sections with the same first three elements and a different contrast could end up with totally different PGLR models.

Therefore we simplified the modelling assumption (and therefore the optimisation process), by focusing on the goal to first minimise the transports corresponding to the systemic relations, f and g an then only, to find subsequently the most economical description for the contrast, keeping f and g fixed.

By doing so, we also significantly reduced the computation time needed to optimise a given sub-system. Another advantage of such a choice is that, the optimisation of f and g can be carried out independently. In fact, as the notes of the primer (x_0) are fixed, to find the best f (resp g), we just need to explore all permutations of notes in x_1 (resp x_2). The choice of a permutation of x_1 (resp x_2) does not affect the transport cost between the primer and any permutation of x_2 (resp x_1) and vice-versa. The computation time for the optimisation of the description of a sequence of n chords falls down from $\mathcal{O}(\Theta^{n-1})$ to $\mathcal{O}((n-1)*\Theta)$, where Θ is the computation time of a single transport optimisation as defined by Equation 5.7.

When used with a sequential model, minimal transport can be viewed as some sort of voice-leading analysis [Cohn, 2011]. However, here, chords are taken regularly on a metrical grid. As a consequence, optimal transport may be used to describe the relation between two identical chords whereas voice-leading analysis only considers progression between distinct successive chords.

5.2.3 Musicologically-constrained Optimal Transport

One limitation that can be objected when considering the optimal transport formalism as described above is that the virtual chord created to model the logical expectation of a system may be outside of the tonality of the section; or it can correspond to a chord that is hard to explain with standard musicology.

For example, if we consider the chord sequence shown on Figure 5.4, $C\ F\ G\ C$, the virtual chord obtained by using the basic optimal transport algorithm is the $B7$ chord without its fifth. such a chord is not common for a progression the first three chords of which are clearly perceived as belonging to the C-major tonality. A way to explain such chord could be to consider it as a passage chord which may or may not really occur in the sequence, that is a chord which makes a link between the system and the contrast

but is "invisible". In this sense, it could play a similar role to the one of the augmented triad in Cohn's theory [Cohn, 2011], that is, it could be envisioned as describing a virtual path between the system and the contrast.

However, we also studied a relation formalism based on optimal transport which would create some virtual chord with a more straightforward interpretability in terms of standard musicology[1]. We therefore considered an extension of the concept of optimal transport enabling the incorporation of musicological constraints. The main principle behind this *musicologically-constrained optimal transport* is that a note displacement may depend on the musicological function of the departure note in the chord. This means that it could be more relevant to take into account the function of each note in the chord for estimating and applying the displacements corresponding to f and g.

For example, let us consider chord progression $C\,F\,G\,C$. In the case of (unconstrained) optimal transport, the optimal transports are $f = (0, 1, 2)$ and $g = (-1, -2, 0)$, then $g \circ f = (-1, -1, 2)$, applied to the primer $C = [0, 4, 7]$ it results in the creation of the virtual element $B7 = [11, 3, 9]$.

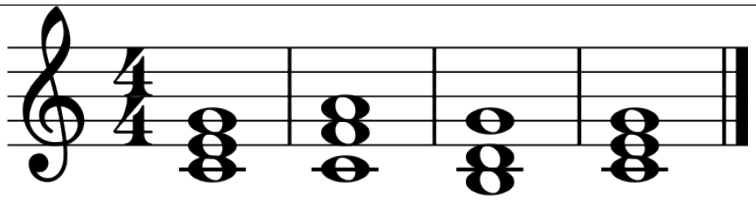
If we now consider the relations described with the musicologically-constrained framework, we have :

- $f = (root : 0, (0, 1, 2))$ (i.e. is the root of the first chord is the first note of the pc-set and the note of the pc-set are displaced by 0, 1 and two semitones following their order in the pc-set, the root stays in place, the third go up by one semitone and the fifth by two),
- $g = (root : 0, (-1, -2, 0))$ (the root is also the first note of the primer's pc-set and the transport is $(-1, -2, 0)$).
- To obtain the virtual element, we can both apply f to the image of the primer by g or apply g to the image of the primer by f (and as the two cases are equivalent, we only detail the first one). In order to apply f to $G = [11, 2, 7]$ with the root being 7 (and not 11), we first apply a circular permutation to the pc-set so that the root is placed in first place (that is the condition to apply f) yielding to the following representation : $[7, 11, 2]$. Then, by displacing the pitch classes following the transport encoded in f, we obtain the virtual element $[7, 0, 4]$ which corresponds to the C chord.

Under this approach, we generate a different virtual element, and in the case of the M-C OT, the result appears as more musicological. Figure 5.4 gives an illustration of this example.

On this example, we only considered the root function, but it is conceivable to also consider different harmonic functions such as third, fifth seventh and so on. However, considering the function of each note in a chord requires to encode each of these functions and for each chord (assuming that they exist and are clearly identifiable for all chords!). As the basic representation of the corpus is a symbolic representation of the chords, it can be easily done, especially for the root. However, if we consider the relation as a vector

[1]These experiments were carried out in the context of an informal collaboration with Mathieu Giraud

	$C\ F\ G\ C$
Initial Sequence	

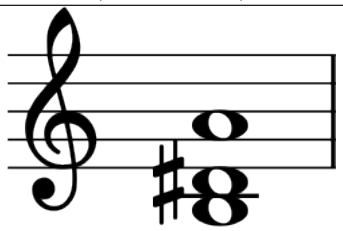
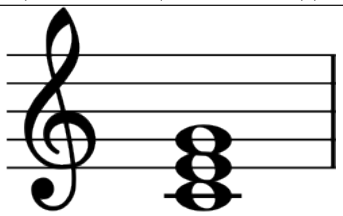
Formalism	Basic OT	Extended OT
f	$(0, 1, 2)$	$(root : 0, (0, 1, 2))$
g	$(-1, -2, 0)$	$(root : 0, (-1, -2, 0))$
Virtual Element		

Figure 5.4: Difference between the virtual chords created with basic optimal transport (basic OT) and extended optimal transport for the sequence $C\ F\ G\ C$. Here, the first virtual element is obtained by using $g \circ f = (-1, -1, 2)$ on the primer $[0, 4, 7]$ (0 is pitch class for the C note). The second is obtained by applying f on the circular reversal of $[11, 2, 7]$ where the root (7) has the index 0 in the pitch class set, that is $([7, 11, 2])$.

that describes the displacement given a note function, there can be some difficulties. In fact, there are chords without fifth or third, or chords where the root does not really appear, and applying a root displacement to a chord where the root is not played creates a conflictual situation. To handle such conflicts, for each relation with a displacement associated with a note function, a rule must be created to apply it to chords where there is no note played that have such function.

Therefore, as in the example above, we considered the root function only. And instead of encoding the displacement for each note function, we computed the basic optimal transport and used a circular permutation to ensure that the root of the third chord follows the same displacement than the root of the primer with f. In case there is no root, in the chord, the first active note above the root was chosen as a substitute for the root. For the chord sequence presented on Figure 5.4, this results exactly in applying $[0, 1, 2]$ to $[7, 11, 2]$ instead of $[11, 2, 7]$ as $[7, 11, 2]$ is obtained by circulating $[11, 2, 7]$ so that the root of G (7) is the first note of the pitch class set.

By doing so, some displacements may be applied to a note that does not have the same function than the one used to compute the transport. However, using a circular permutation that aligns roots increases the chance that a displacement is applied to a note that has the proper function.

5.2.4 Multi-scale Generalisation

One of the main benefits of formalising relations in terms of optimal transport is that the relation space is isomorphic to the chord space. In fact, if two chords can be described by PCS of n pitch classes, then the relation between the two chords will be described by a set of n intervals. And as intervals may vary from -5 to 6, once taken the interval modulo 12, we can consider each interval as a pitch class which would result in considering the relation as living in the same space as a chord.

Therefore, it is possible, by using the transport formalism, to describe systems of relation. This readily provides the possibility to implement the Relational Static S&C model, defined in Section 3.1, which aims at describing systems of relations between relations to further simplify (or compress) the description of a sequence by using higher-level redundancies.

5.3 Results

This section presents a collection of results obtained by considering the different formalisms introduced above for describing the relations between chords. The aim is to investigate the performance of the multi-scale models (that encompass both the tree systemic, the static S&C, the dynamic S&C and the relational static S&C models) and compare them to the sequential model for the description of the structure of a section based on its chord sequence representation. Moreover, a comparison between the different options of relation formalisms is also performed. This is the occasion to investigate some specificities of the S&C model such as the relevance of non-sequential modelling and the role of the virtual element.

All results are obtained using a 2-fold cross-validation scheme on the corpus of the 727 sections of RWC POP represented as sequences of 16 chords.

In our experiments, we test a number of possible combinations by associating (when relevant), different chord descriptions, representations, and relation formalisms:

- chord descriptions: full chords (up to 4 notes), triadic reductions (3 notes).
- chord representation: full pitch class set (FPCS), triadic pitch class set of 3 notes (TPCS), symbolic triadic representation (TR)
- relation formalisms: optimal transport (chromatic and harmonic) and circular relations (on circle of thirds, circle of phase-shifts and random)

We compare the sequential bi-gram model (Seq) – a very common approach in MIR as described in [Pearce, 2005] – with different configurations of multi-scale models (Sys and $S\&C$) as defined in terms of their antecedent functions, according to the formalism defined in Section 3.3. We also test the dynamic approach (Dyn).

For each multi-scale model based on the PPPs, three system optimisations are considered:

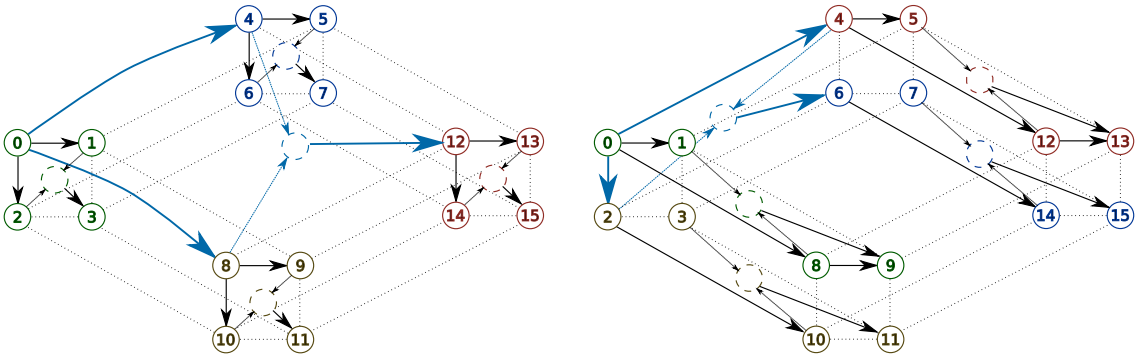


Figure 5.5: Representations of the relations used by a multi-scale analysis of a sequence of 16 events projected on a tesseract: PPP_0 (left), PPP_3 (right).

	Thirds Circle Rotation on TR	Optimal Transport				DPG
		Chromatic		Harmonic		
		on FPCS	on TPCS	on FPCS	on TPCS	
Seq	7.46	3.22	3.39	4.06	4.29	1
Sys_0	8.34	3.34	3.53	4.23	4.50	1
Sys^*	7.11	2.97	3.11	3.66	4.03	1
Sys^X	5.21	2.53	2.64	3.03	3.25	6
$S\&C_0$	6.06	2.85	3.01	3.78	3.94	1
$S\&C^*$	4.82	2.43	2.57	3.10	3.25	1
$S\&C^X$	**4.17**	**2.25**	**2.36**	**2.79**	**2.93**	6
Dyn^X	4.41	2.36	2.45	4.00	4.03	486

Table 5.1: Average cross-perplexity obtained with a 2-fold cross-validation, for the different models on RWC POP. DPG stands for distinct possible graphs.

- S_0 which corresponds to the static configuration PPP_0 (see Fig. 5.5, left);
- S^* which corresponds to the globally optimal PPP over the whole corpus which happens to be PPP_3 (see Fig. 5.5, right);
- S^X: in this case, the optimal PPP is chosen a posteriori as the one that optimises the description of the sequence X (which varies across all Xs).

Comparative results, in terms of average cross-perplexity, are provided synthetically in Table 5.1 and are commented in detail in the forthcoming sub-sections.

5.3.1 Benefit of Multi-Scale Organisations

A first observation that can be made from Table 5.1 is that the multi-scale models ($S\&C$ and Sys) globally outperform the sequential one: all cross-perplexity values are lower, ex-

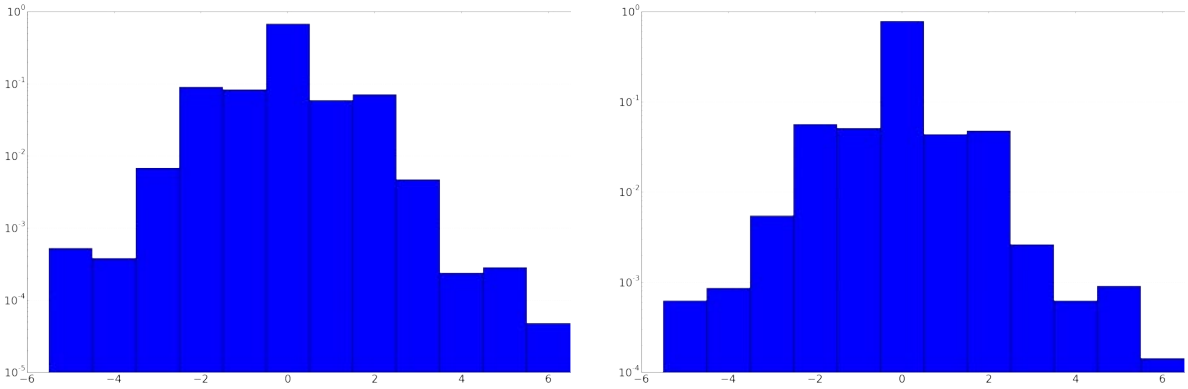


Figure 5.6: Log-distribution of the pitch displacements induced by the optimal transports computed by *Seq* (left) and $S\&C_3$ (right) models.

	Triad Circle		
	Third	Phase	Random
Seq	8.00	7.67	9.32
$S\&C_0$	6.68	6.77	7.84
$S\&C_3$	5.35	5.35	6.02
$S\&C^X$	**4.63**	**4.63**	**5.21**

Table 5.2: Average cross-perplexity obtained for the *Seq* and *S&C* models on RWC POP data with 2-fold cross-validation with the different types of triad circles.

cept for the basic Sys_0 configuration.[2]. In particular, the $S\&C^X$ model, which consists in taking the permutation that has the smallest cross-perplexity for each sequence, provides the most spectacular cross-perplexity improvement for all types of chord representations and relations, at the expense of a very limited number of Distinct Possible Graphs (DPG). It is also worth noting that each PPP used alone with the S&C dependency model has a smaller cross-perplexity than the sequential model. Note that the $S\&C^*$ configuration provides a noticeable advantage over *Seq* and $S\&C_0$ configurations.

Figure 5.6 shows that a strong reason of the good performance of the multi-scale models is that there are more identity relations implied in the S&C description of a chord sequence. That is the S&C models are more able to catch obvious long-range redundancies within chord sequences. Using logical expectation may also be contributing to the performance advantage.

The last row of the table also shows that the dynamic nesting approach is an interesting alternative as it provides cross-perplexity scores almost as favourable as $S\&C^X$. However, the *Dyn* model has a high number of distinct possible graphs, that is a high encoding cost of the model structure. Therefore, even if it can be useful for prediction, it may be less interesting in terms of compression ability.

Table 5.2 shows the cross-perplexity figures obtained with the *S&C* and *Seq* structure models for different types of triad circles. For all these relation formalisms, the sequential

[2]This result is consistent with a preliminary study made during this thesis on a small corpus of 45 chord sequences which had also indicated such trends [Louboutin and Bimbot, 2016].

	$S\&C_0$	$S\&C_1$	$S\&C_2$	$S\&C_3$	$S\&C_4$	$S\&C_5$	$S\&C^X$
Basic OT	2.85	2.75	3.53	2.43	3.05	3.18	2.25
M-C OT	3.01	2.87	3.78	2.69	3.29	3.48	2.45

Table 5.3: Average cross-perplexity obtained by the $S\&C$ models on RWC POP data, with 2-fold cross-validation using the basic (Basic OT) and musicologically-constrained optimal transport (M-C OT) formalisms.

model is outperformed by the $S\&C$ models, with again a strong advantage for $S\&C^X$ approach. $S\&C^X$ still outperforms the other models after adding the cost of the model description to the cross-perplexity score (which increases then from 4.63 to 5.18).

Therefore, the multi-scale models appear to have a definite advantage over the sequential model, as they seem to be more able to catch the redundancies in the chord sequences at multiple ranges. The simplicity of these models makes their encoding cost very small, which makes them useful for both compression and prediction tasks.

5.3.2 Impact of the Representation

The performance of triadic circle relations (TCRs) is based on a global sequence entropy while the optimal transport (OT) approach is evaluated in terms of average "per voice" entropy. In particular, the maximal branching factor of TCRs is 24 instead of 12 for OT. Therefore, the two cross-perplexity scores cannot be directly compared. However, both approaches show similar trends w.r.t. the relative model performance. This supports the hypothesis of a general benefit of the multi-scale approach rather irrespective of the way the chord information and relations are encoded.

In Table 5.1, results are also provided for optimal transport on triadic reductions (TRs) represented by PCS. Here too, the relative performance levels across models show the same trends. Note that the cross-perplexity on TRs is slightly higher because the average pitch class distance between triads tends to be larger than that between chords with 4 notes or more.

In chromatic optimal transport, the distance is computed from the set of note displacements measured on a semitone scale. We also tested a harmonic distance by considering displacements on the circle of fifths. Results in Table 5.1 show that this globally degrades the performance.

On Table 5.2, cross-perplexity values are provided for two other triad circles: the circle of phase-shifts as defined on Fig. 5.2 and a randomised circle, where triads are positioned at random. Results show that the phase circle performs quite the same as the circle of thirds, whereas the randomised circle clearly performs less well. All outperform their counterpart in the sequential model, as for all polytopic models, the identity relation is of zero cost and with higher probability.

Table 5.3 presents the results obtained when using the basic and musicologically-constrained optimal transport, presented in Section 5.2.3. For each model, the basic optimal transport formalism leads to a slightly better results. The difference is not very

US	LS_1	LS_2	LS_3	LS_4
44.4 %	8.0 %	15.7 %	19.7 %	22.4 %

Table 5.4: Proportion of sequences with contrastive US (Upper-Scale system) and LS_k (k^{th} Lower-Scale system).

high, but it must be noted that the musicologically-constrained method also requires the encoding of the root, which in term of compression requires more information (which could be seen as an issue). Nevertheless, in case we wish to interpret the result in terms of relevance of the virtual chords under musicological standards, the constrained transport may turn out to be of more interest.

5.3.3 Role and Importance of the Virtual Chord

As the virtual chord has a very special role in the S&C model, there was a need to study and evaluate its importance in the modelling scheme.

As shown in Table 5.1, the effectiveness of the virtual element in the S&C scheme is underlined by the systematic improvement observed when comparing Sys scores to $S\&C$ results. The virtual element, $\hat{x_3}$, in the S&C model, appears globally as a better antecedent for x_3, than does the primer, x_0, in Sys. Note that this improvement is present for every type of relation formalisms, and as the virtual chord may differ from a relation model to another, it implies that, due to the expectation process used by the S&C model, it is more prone to catch the structure of (relatively frequent) non-contrastive sequences such as *abab*.

However, a detailed examination of the results shows that , for about 37% of test sequences, Sys^X outperforms $S\&C^X$ using the optimal transport formalism. This can be explained by the fact that Sys^X is obviously more interesting to describe sequences with an *aaba* structure, which is (also) a rather common (contrastive) pattern in chord sequences.

To study the specific relationship between the virtual and the contrastive element in the $S\&C^X$ scheme, we also investigated on the location and the number of contrastive vs. non-contrastive elements in potentially contrastive positions defined by the PPP framework. An element in such position is said to be non-contrastive if it is the same element than the virtual chord. The following results are focusing on the triad data and the circle of third formalism.

Table 5.4 presents the distribution of actual contrasts for the Upper-Scale (US) systems and the 4 Lower-Scale (LS) systems in contrastive positions. While 44.4% of US systems are contrastive, it can also be noted that the frequency of LS contrasts (or, so to speak, the occurrence of surprises at the lower-scale span) increases with the index of the LS system (i.e., its depth in the tesseract).

Figure 5.7 depicts the proportion of sequences as a function of the number of actual contrasts observed in different contrastive positions. It can be observed that the number of contrastive Lower-Scale systems decays rapidly from over 60.4% of sequences with

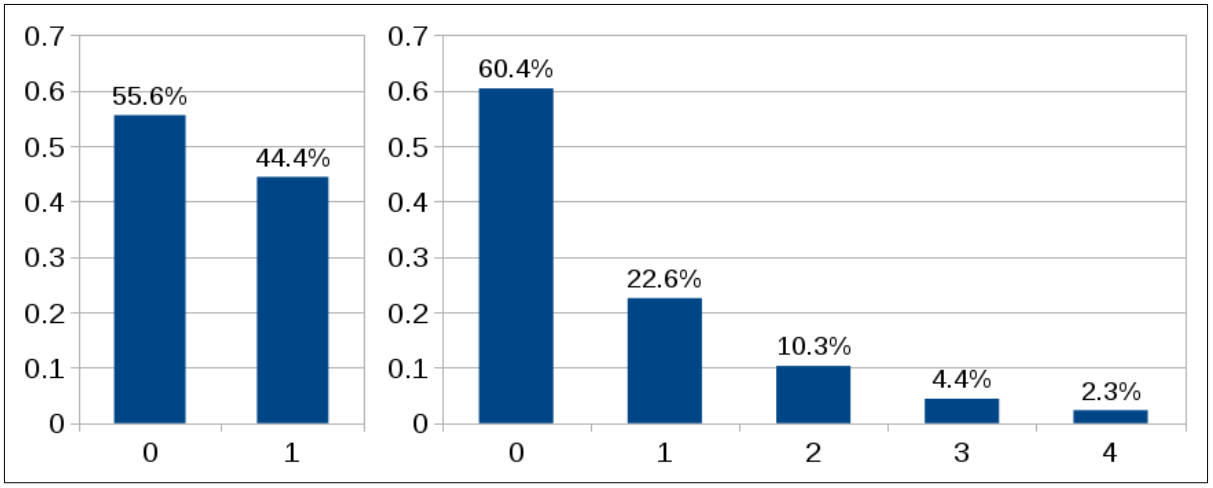


Figure 5.7: Proportion of contrastive systems within US systems (left) and the 4 LS systems (right)

	All	Diff
Systemic position	5.6	14.6
Contrastive position	3.5	18.7

Table 5.5: Perplexity of relations for systemic relations and contrastive relations, including (All) or excluding (Diff) the identity relation.

no contrastive Lower-Scale system down to only 2.3% with all four LS systems being contrastive.

It would surely be interesting to compare these profiles across different music genres and a variety of musical dimensions, in order to study possible correlations.

Table 5.5 reports on the perplexity obtained when considering separately the systemic positions and the contrastive positions. Keeping in mind that they may be specific to the corpus, results show nevertheless two very interesting trends.

Perplexity is higher in systemic positions (5.6) as opposed to contrastive positions (3.5), implying that the actual observations in contrastive positions often correspond (or are close) to the implied expectation. This can be related to the results previously presented, w.r.t. the relatively low density of actual contrasts.

However, when different from identity (column Diff), these relations show a *lower* cross-perplexity for systemic relations (14.6 vs 18.7) indicating that, when a relation is not identity, the contrast is more unpredictable and/or more distant on the circle of thirds, than it is for systemic relations.

In summary, strictly contrastive (i.e. non-identity) relations tend to be less frequent but more intense than for systemic relations. This certainly relates to the presumed role of contrasts as carrying a strong quantity of surprise. These observations may be a motivation for a differential treatment of systemic relations vs. contrastive ones.

	$RS\&C_0$	$RS\&C_1$	$RS\&C_2$	$RS\&C_3$	$RS\&C_4$	$RS\&C_5$	$RS\&C^X$
Cross-perplexities	2.87	2.93	3.01	2.87	2.95	2.97	2.69

Table 5.6: Cross-perplexities obtained using the relational static S&C model with optimal transport relation formalism.

5.3.4 Relational Static S&C Model Performances

As mentioned earlier, when using optimal transport as a relation formalism, the representation of the relations is exactly the same as the representation of chords, that is a set of integers from $\mathbb{Z}_{12}$. In fact the two spaces are isomorphic: for chords, it represents pitch classes and for relations it represents intervals. Therefore, it is possible to describe the relation between relations in the exact same way as it is done for relation between chords. This makes it possible to use the relational static S&C model presented in Section 3.3.7 with the optimal transport formalism to describe the multi-scale structure of a chord sequence.

Table 5.6 presents the cross-perplexities obtained using this model.

When compared to the static S&C model, the relational static S&C Model does not seem to provide much benefit. In fact, the cross-perplexity values obtained with the basic static S&C were 2.85 for $S\&C_0$, 2.43 for $S\&X^*$ ($S\&C_3$), and 2.25 for $S\&C^X$. Here, the values obtained with the relational static S&C model are always higher.

The reason behind these results may be explained by considering the effect of each model on the probability distributions of elements and relations. Using the static S&C model is interesting to compress a sequence of chords because it uses the note displacements instead of pitch classes to describe the full sequence. Such a process is interesting because the probability distribution of pitch classes is flat and the probability distribution of note displacements has one big peak on the identity relation and small very values for large displacements. In other words, the entropy of the relations distribution is way smaller than the entropy of the pitch classes distribution.

However, passing from the distribution of relations to the distribution of relations of relations, may reduce this effect by flattening the peak on identity relation and therefore increase the entropy. This flattening effect can be explained by the fact that, due to the large amount of identity relations, a lot of these relations may be transformed into non-identity relations of relations (in the case they are related to non-identity relation). And as there are only a few of these relations that are non-identity, only a few pair of non-identity relations lead to identity relations of relations. This results in a smaller peak of the identity for relation of relations than there was on the basic distribution of relations (between pitch classes) and therefore a higher entropy.

Some similar trends were already observed in former work, when considering intervals of intervals for melody representation and compression, when used for pattern similarity measurement, pattern discovery and music classification [Louboutin and Meredith, 2016]. There too, taking intervals of intervals instead of the basic interval representation did not increase the average compression ratio or the classification.

However, a noticeable effect of such a representation is that it decreases the difference

of performance between the different permutation models. In fact, here, the maximum difference of cross-perplexity value between two permutation models, $RS\&C_i$, is 0.24 while, for the basic $S\&C_i$ models the maximum difference is 1.68 ($S\&C_2$ having 3.68 and $S\&C_3$ having 2.85).

As a consequence, each permutation model using this new representation, $RS\&C_i$, provides a better cross-perplexity (max 3.01) than the sequential model, Seq (3.22). The fact that the use of this second step of compression improves the results of the "badly-performing" permutations can be explained using the same argument as the one used above. In fact, as for permutations performing bad using the basic static S&C model, the associated distribution of note displacements has a high entropy, i.e. it has less probability density for the identity and more non-identity displacements. Then, using relation between relations may create a distribution of intervals between note displacements where there are more identity relations than in the distribution of pitch-class displacements, that is a distribution with a smaller entropy.

5.3.5 Additional Observations and Considerations

5.3.5.1 Correlation Between Description Cost and Cross-Entropy

As optimising the transport cost between chords minimises the average pitch class displacement, only a few intervals concentrate most of the probability density function. This raises the idea that there must be some correlation between the global description cost of a sequence (computed with Equation 3.5) and the cross-entropy. Indeed, Figure 5.8 shows a very high correlation between the two values. This result is going along with the idea that there could be an implicit optimisation process governing some aspects of music organisation and it reinforces the approaches in MIR based on the minimum description length principle or related concepts such as entropy or mutual information. Under these hypothesis, using cross-entropy, and then cross-perplexity, to measure the efficiency of a model makes some sense.

Moreover, another very interesting perspective can be drawn from Figure 5.8: whereas the learning phase can be of some use for some prediction or generation tasks, the fact that there is a high correlation between cross-entropy and global transport cost shows that it may be advantageous to describe the structural organisation of a chord sequence using minimal transport model without resorting to an exact estimation of relation probabilities, i.e. no learning phase and (even more important) no training data !

On that basis, the next round of experiments uses the cost function to select the optimal PPP for a given sequence and studies their distribution across our experimental dataset.

5.3.5.2 Distribution of PPPs

As a final experiment in this chapter, we study the behaviour of the $S\&C^X$ model, i.e. the model which selects, for each sequence to describe, the PPP that results in the most economical description.

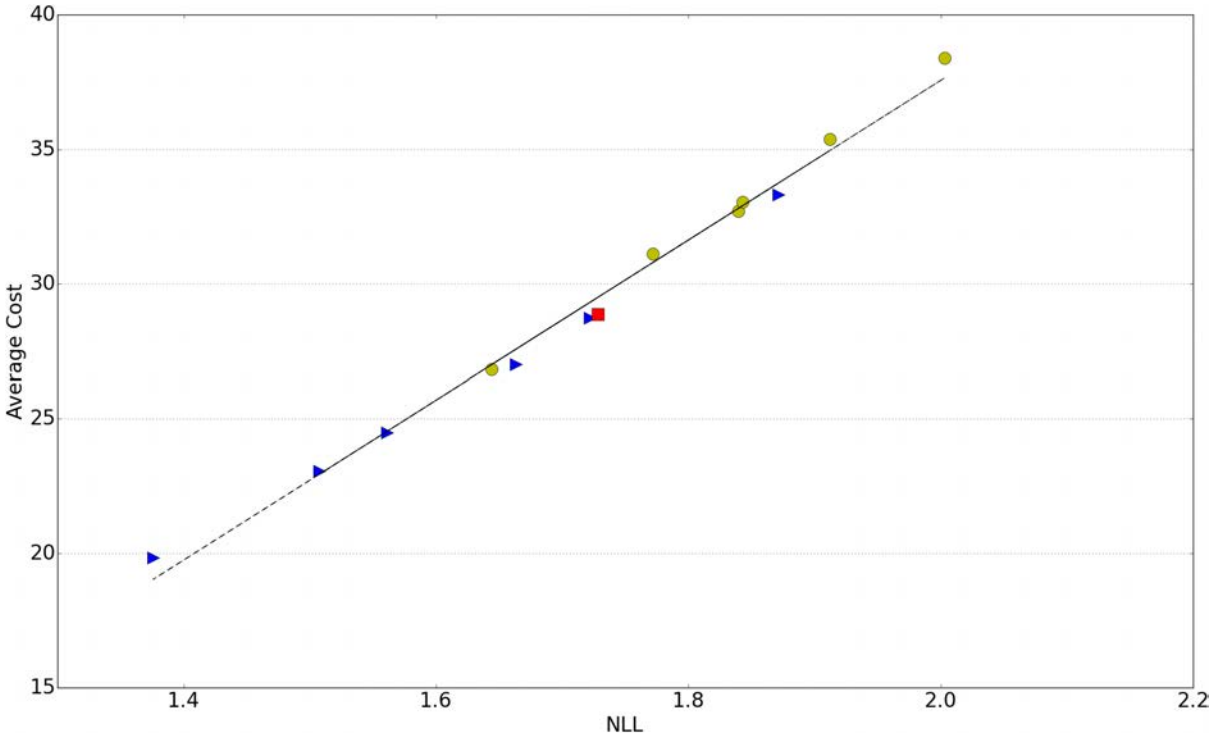


Figure 5.8: Graph of average sum of the optimal transport costs, function of the average cross-entropy on 727 sequences corpus for the Seq (red square), $S\&C_{0...6}$ (blue triangles) and $Sys_{0...6}$ (yellow circles) models. Correlation: 0.996, p-value: $9.76e^{-13}$.

The simplicity of a description can be evaluated by several criteria. In our experiments on model comparisons, the determination of the optimal PPP was obtained by simply picking the PPP with the best cross-perplexity score for the sequence. However, as shown in the previous section, the cross-entropy is highly related to the description cost of a sequence computed with 3.5), in terms of optimal transport cost or rotation distance on a circle. Therefore, another way of evaluating the complexity of a description can be to compute the total relation cost (by using for example Equation 3.5).

The advantage of such a method is that it can be used without any learning step. Moreover, it may be easier to see if two models lead to the same description cost on a sequence. Due to the probability estimation step, it is indeed very rare that two PPPs obtain an identical cross-perplexity value for the same sequence. It results in the fact that one of two PPPs that would give the same score (considering here the description cost), will get discarded as best PPP because of a little difference in term of cross-perplexity. Using cross-perplexity as a criterion for the choice of the PPP may provide better prediction ability to the model; however, it may influence the possible interpretations of the results in terms of structure, because of slight variations in the probability estimation process.

For these reasons, to explore the relative prevalence of the various PPPs, we used the cost criterion. Therefore, for each chord sequence X, it is possible to compute the overall relation cost for each PPP, so as to find which is/are the one(s) yielding to the best score.

Over the 727 sequences, 2 out of 3 have a unique optimal PPP. To build the histogram

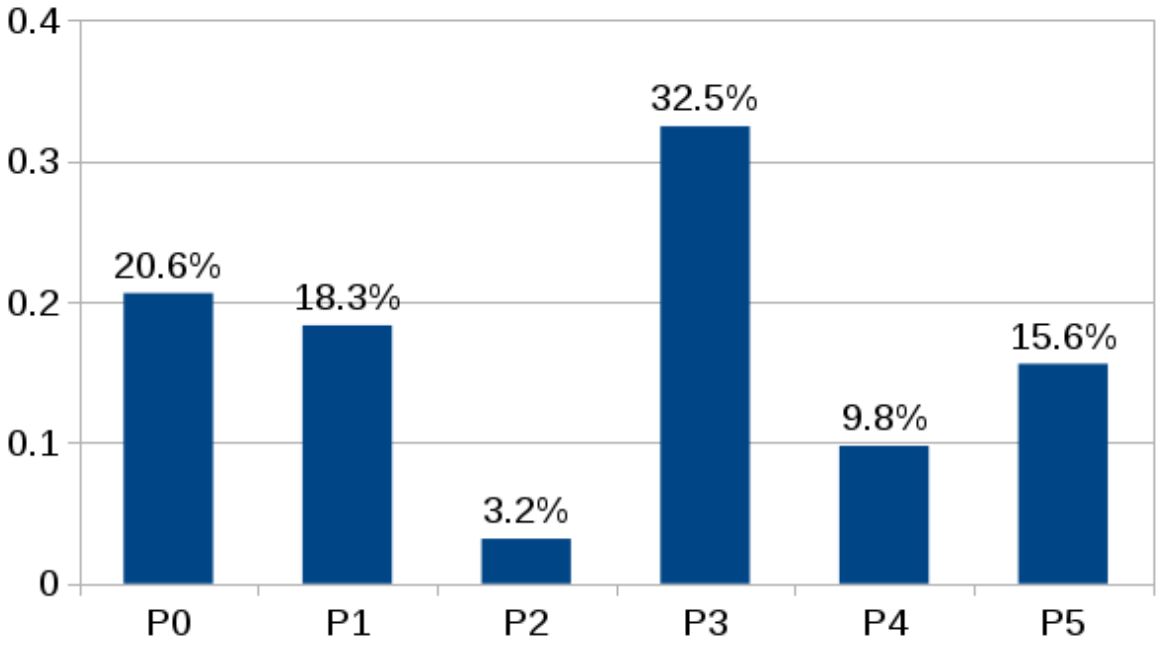


Figure 5.9: Histogram of best PPPs across the test data.

depicted on Figure 5.9, we considered every unique solution as counting for 1. For the other cases (ex-aequos), when the number of optimal PPP, k, is superior to 1, each optimal solution was counted as $1/k$. The relation formalism used to compute the cost of a PPP is the rotation distance over the triad circle of thirds.

On Figure 5.9, permutation PPP_3 (see Fig. 5.5 right) appears as the prevailing one ($\approx$ 33%) and this may be related to the fact that such a PPP is better at describing sequences that correspond to a rather common "antecedent-consequent" form in music (especially, in pop music): $ABAC$. In fact, due to the configuration of each lower-scale sub-system of the PPP_3, if the two halves of the section have the same beginning, then, in the first two sub-systems, $[0, 1, 8, 9]$ and $[2, 3, 10, 11]$, the first and third elements will be identical (or strongly similar) and there will be no contrast as the second and last element would also be the same. These two sub-systems can then be described in a simple manner (as there is no contrast). For the other sub-systems there may be more contrasts due to the difference between the antecedent and the consequent. The upper scale S&C, $[0, 2, 4, 6]$, aims at describing the structure of the first half of the section (that is the antecedent, AB) at a mid-scale level.

On the other hand, the least frequent PPP (PPP_2) displays a frequency of occurrence below 5%. Somewhere in between, the four other permutations see their frequencies ranging within 10% to 20%).

This chapter has focused on the description of the structure of chord sequences, and the various behaviours that the implementations of the PGLR model may exhibit when used for such descriptions. The next chapter investigates on the behaviour of these models for two other musical dimensions: rhythm and melody.

Chapter 6

Rhythm and Melody

In this chapter we extend the application of the PGLR model to other musical dimensions, namely rhythm and melody. In this case, unlike chords, basic elements are not vertical anymore, but horizontal. In fact, in a rhythmic or melodic motif, notes are coming one after another and modifying the time organisation would alter the melody. This is due to the temporal aspects that are at the foundation of melody.

Therefore, one of the main issues faced when working on rhythm and melody instead of chords is that relation formalisms such as optimal transport are not easy to adapt. In this chapter, after reviewing briefly the state of the art in transformations and similarity measures for rhythm and melody, we propose and study a few relation formalisms that satisfy adequate properties for being used with the PGLR model.

As it focuses on rhythm and melody, this chapter also introduces and addresses a key problem: how to handle the phase-shift between the segment boundaries of the melodic section and the accompaniment section that occurs in the case of anacruses or late starts of the melody.

This chapter finally provides the results of several series of experiments that have been conducted to investigate the effectiveness of a multi-scale description of the structure of melodic and rhythmic sections, and the importance of the relation formalisms used to this aim.

6.1 Context and Practical Considerations

As now well-developped in the previous chapters of this thesis, one of the most important aspect of music is that it contains redundancies, and this is the reason why musical objects are perceived as structured. This aspect is especially noticeable by the repetition of rhythmical patterns or melodic motifs which are probably the most salient "agents" of redundancy.

As a consequence, a lot of studies in MIR focus on the detection of these repeated patterns, aiming to find the themes and their repetitions inside a piece of music. But as repeated patterns tend never to be strictly identical repetitions, some methods have been

introduced to measure similarities between melodic patterns. These measures sometimes result in a description that can be understood as the transformation which has to be applied to the first pattern to obtain the second pattern. Once the transformation has been described, it is possible to use this transformation to compute a measure of the distance between the two patterns. If the distance is small, one may consider the second pattern as a repetition of the first one. Beyond a certain threshold, they are considered as distinct.

In this work, the aim is to use some relation formalism which can be applied along with the S&C model, in the sense that it can both describe relations between two melodic (or rhythmic) elements and be used to create some virtual melodic elements by combining transformations (i.e. modelling melodic expectation on the basis of analogical induction).

Therefore, one may be interested in the melodic or rhythmic transformations from the state of the art methods, which are used to compute similarity measures between melodic patterns. To apply them to the S&C description of a melodic section, it is necessary to see if these transformations methods, usually used to compute the distance measures between two melodic patterns, have the property defined in Section 3.3.2 so that they can be safely combined.

One of the most common transformation is the one used to compute the edition distance [Mongeau and Sankoff, 1990, Giraud et al., 2013]. Its most basic implementation consists to give the minimal number of insertions and deletions that need to be applied on the first pattern to obtain the second one. Moreover, additional operations may be used to refine the transformation, such as substitution, fragmentation or consolidation. Note that this problem also occurs in different relation systems that use a more complex insertion/deletion formalism such as the one described by Forth et al. [Forth et al., 2008].

After a brief overview of this type of approaches, it becomes rapidly clear that such transformations cannot be used with the S&C contrast. In fact, the main problem of such transformations is that they may not be applied to every possible pattern. For example, if we consider a transformation which consists in deleting the first note of a pattern, it cannot be applied to a pattern which has no first note! And as our study here focuses on relations between small patterns (a few beats), some of them may contain only silence, for instance... The same issue holds for other comparable operations: how to insert a note where there is already a note? How to consolidate n notes where there are less than n notes? Such a problem may be solved using some rules, but the aim here is to limit as much as possible the creation of such rules, because then the formalism may become too much *ad hoc* and would have a high description cost. Indeed, a more complex space of relations will result in a more complicated explanation of the section's structure.

A second family of transformations that catches our interest are those used for instance by Sioros [Sioros et al., 2018], Ycart [Ycart et al., 2016] or Forth [Forth, 2012]. Such transformations are mostly related to a path in a limited space. For example, Ycart considers rhythms as trees and transformation between rhythms as transformation of these trees. This provides interesting possibilities for comparing rhythms. However, as for editing transformations, there are many cases when combining some transformations requires special rules. For instance, when the combination of two relations would range beyond the limits of the rhythm space, as can be the case when fusing two branches of a

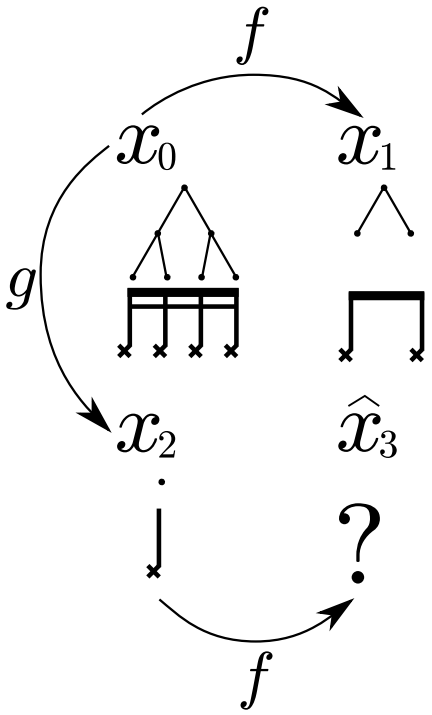


Figure 6.1: Rhythm transformation facing problems for combination.

tree consisting only in one leaf which is the root. If these transformations are very good at describing the relation between two rhythm patterns, they were not designed to be combined, nor to be applied on potentially any rhythmic element.

As an illustration, Figure 6.1 shows a case where f is a fusion operation which applies on two branches at the lower level of the tree and $g = f \circ f$. In that case, $f \circ g$ would be equivalent to applying f to the quarter note... But, considering the corresponding tree, there is nothing to fuse!

Finally, there are other transformations that focus on the onset information of the rhythm, in order to describe the transformation between two rhythmic elements using some translation vectors. But once again, these transformations cannot be easily used with the S&C model because, when applied to a variety of patterns, some notes may begin or end outside the time boundaries of the element.

All the rhythm relations that we have just reviewed show a same drawback: they cannot be easily combined, due to the limits imposed by the subdivision of time or the pattern duration. The main problem is that time flows in one direction, and that reversing it or assuming its circularity would be hard to justify and would result in relations that would be somehow arbitrary. Therefore, we had to search for some alternate formalism in order to express relations between melodic patterns which would (as much as possible) satisfy the properties required for their proper use with the S&C model.

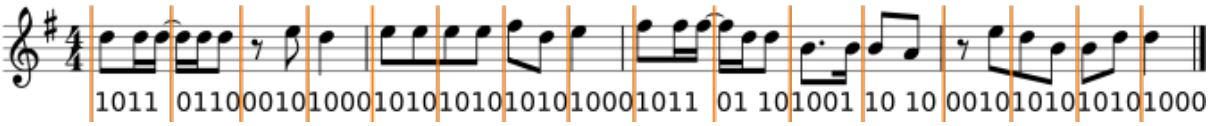


Figure 6.2: Binary encoding of rhythmic information of a melodic section.

6.2 Relation Formalisms

6.2.1 Rhythmic Relations

The relation formalism that is to be used in the PGLR model is intimately related to the representation used to encode rhythmic elements. There are mostly two representations that are common in MIR.

The first one considers rhythmic sections as a set of notes encoded by their onset and duration. The interest of such a representation is that it directly encodes the onset of the note which makes them implicitly ordered in time. Moreover, as the duration is also encoded, it can be used in the relations to compare and model such factors of variability. However, the problem with such a representation is that notes that begin in one element and finish in another one require to be handled, which brings in quite some complication in the modelling scheme.

The second common representation, which we use in our work, simplifies the situation by discarding the duration information and encoding the onset information on a discrete time-scale. Rhythmic information consists in a list of bits, where each bit corresponds to an onset time: if a note in a section starts at an onset-time, the corresponding bit is set to 1, and if there is no note in the musical segment at this onset time, the bit is set to 0. The bits are updated regularly over time with a sampling step (ideally) equal to the smallest common divisor of the time of all durations of the rhythmic pattern.

Such a binary representation of the rhythm information is shown on Figure 6.2. The interest of such a rhythmic information encoding scheme is that it is easy to split a pattern in sub-patterns by slicing the bit sequence into smaller elements. However, this representation has the disadvantage that it does not encode the duration of the notes. The negative impact of this simplification can be moderated by observations such as those made by Conklin et al. [Conklin, 2013] – where the duration feature does not seem to have a lot of effect on music classification – and Nakamura et al. [Nakamura et al., 2017] – showing that for melodic parts, most of the notes end where the next note begins.

As we did for chords, the rhythmic part of the melodic section is considered as a set of 16 basic elements corresponding to a strong beat. In the case of rhythmic patterns however they are elements of $R = \{0; 1\}^4$. That is, when a melodic section consists in 4 bars of 4 beats, the shortest element is the sixteenth note, whereas when the section is made of 8 bars of 4 beats, the shortest note is the eighth note. Such rhythmic elements can be considered as points in the rhythmic domain defined by Forth et al. [Forth et al., 2008] except that here the length of an element is set to the duration of a strong beat whereas in Forth's work, a point corresponds to a metrical cycle [London, 2012], which itself often corresponds with a bar-length rhythm.

Such a representation may introduce some inaccuracies in the encoding of the rhythm when, for example, a triplet appears in the melody. However, the method that will be presented can also be used with a space larger than R.

Given a section of 16 basic elements from R, the next aim is to define a structured space where we can describe relations between two basic elements, with the constraint that, any relation applied to any elements lives also in R. As mentioned earlier, it is also desirable that each element has only one antecedent for each relation.

There are plenty of spaces that satisfy such properties. However, in relation to the MDL principle, a proper family of such spaces must be simple to describe, in the sense that similar elements must be related through simple transformations. And as we would like to guarantee some musicological relevance of the approach, it is also desirable that the relation encoding can be interpreted in accordance with understandable musical concepts.

Given that times flow in only one direction, it is hard to handle relations that imply displacement of onsets, as some onsets may fall outside of the time limit imposed by one element. Therefore, it appears difficult to define a relation which would result from the combination of the relation between the activated onset of two elements.

However, as there are only 16 elements in R, it is possible to arrange and order them according to some of their properties and then express the relation between two elements as a displacement in that space. In that case, the relation between two elements is global and may not be directly related to relations between their respective sub-elements.

Considering the spaces that were used to describe chord relations, and especially triads, the simplest structure that would ensure the properties required for a convenient use with the S&C model is a circle.

Under this constraint, we end up having to lay out 16 elements on a circle, which opens a number of possibilities equal to the number of permutations of these 16 elements, that is $15!/2 \approx 6.5 * 10^{11}$ (after removing symmetries and rotations of a same circle).

This number of circles is very large and considering all these circles as possible spaces would require an optimisation process over all the circles to find the optimal one... and in what sense?

To reduce further this number of possibilities, we can introduce additional constraints, inspired from the principles of optimal transport, that is considering only circles where elements that are close on the circle also have some proximity in terms of some other metrics or properties. Indeed, on the triadic circle, neighbouring triads have some proximity in terms of optimal transport, and it intuitively makes sense to consider preferentially rhythm circles where neighbours have several activated or inactivated onsets in common.

In order to implement this idea, the resulting approach has been to choose circles where elements that are neighbours only differ by one bit (therefore they shares 3 common bits). As the graph in 4 dimensions, where each of the 16 elements is linked to every other elements differing by one bit, is a 4-cube, the circles that we have just defined correspond to circular Hamiltonian paths traversing this graph.

Such paths exist and a well-known one is the Gray code— resulting from the Gray encoding algorithm [Frank, 1953] (see Figure 6.3). But this is not the unique Hamiltonian

Input: r; the current rhythm, *circ*; the list containing the previous elements in the circle being built

Output: L, a list of Hamiltonian paths in the 4-cube

```
Function HamilCircles(r,circ):
    if |circ| = 15 & circ[0] ∈ Neighbours(r) then
        circ.Append(r);
        return [circ];
    end
    if |circ| = 15 then
        return [];
    else
        L = [];
        for r_n ∈ Neighbours(r) do
            if r_n ∉ circ then
                newcirc = circ;
                newcirc.Append(r);
                L.Extend(HamilCircles(r_n,newcirc))
            end
        end
        return L;
    end
end
```

Algorithm 2: Recursive algorithm such that HamilCircles(r_0, []) (with r_0 being any rhythm) creates all possible Hamiltonian paths on a 4-cube.

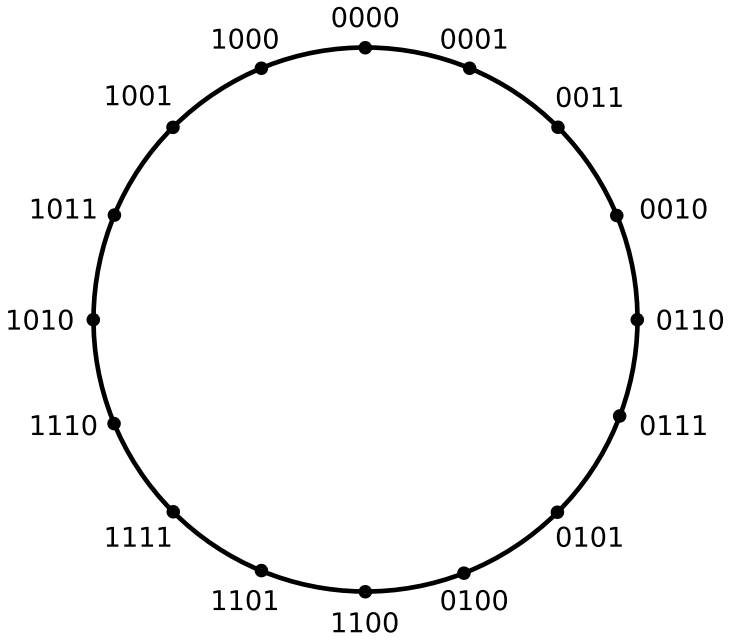


Figure 6.3: A circular organisation of rhythmic patterns generated by the Gray encoding algorithm.

path in the 4-cube, and to generate all the possible circles, we used a recursive algorithm (see algorithm 2) and then removed symmetrical or rotational equivalent circles. As a result, we found 1344 distinct Hamiltonian paths.

A total of 1344 circles is still quite a lot, and appears to be too profuse a number of spaces to describe rhythmical patterns in a meaningful way. It would be surprising that each of them provide a relevant interpretation of the relations between rhythmic patterns, with respect to the circle structure. Therefore we decided to make yet another step to decrease this collection of circles to a more tractable number.

A first approach could be to find out if there are pairs of circles that are equivalent in the sense of the S&C model. Two circles would be equivalent in that sense, if they were to follow the equivalence relation defined by:

$$\mathcal{C}_1 \equiv \mathcal{C}_2 \text{ iff } \forall (a,b,c) \in R, R_{\mathcal{C}_1}(a,b)(c) = R_{\mathcal{C}_2}(a,b)(c) \tag{6.1}$$

Putting property 6.1 into words, two circles are equivalent if, for all triplet of elements, the two circles create the exact same virtual element. In such a case, the two circles result in the same cross-perplexity score for every rhythmic pattern, as the two distributions of rotation probabilities are only permutations of the other one.

For example two circles can be related by a constant multiplier, K. i.e. the elements of $\mathcal{C}_2$ can be obtained using the element of $\mathcal{C}_1$ following the equation:

$$\mathcal{C}_2[i] = \mathcal{C}_1[i * K \pmod{16}] \tag{6.2}$$

In such a case, it is easy to prove that the two circles are equivalent, in the sense of Equation 6.1. Therefore, this equivalence property may be of interest to reduce the

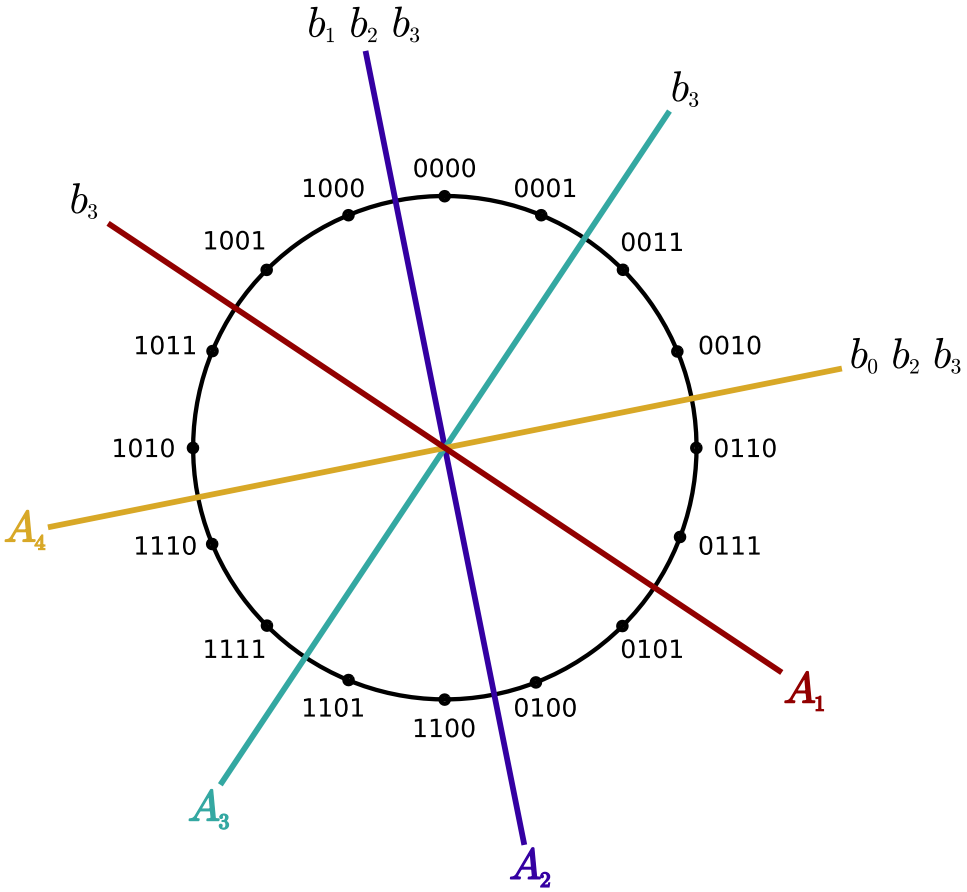


Figure 6.4: Bit-symmetry axes for rhythm circles.

number of circles. However, none of the 1344 circles are equivalent to any of the others...

Another interesting subset of circles are those who exhibit internal symmetries. The idea of considering geometric constraints to reduce the number of possible configurations of a geometric structure was already invoked (in [Louboutin and Bimbot, 2017a]) to reduce the set of 36 permutations for which each lower-scale system is a square in the tesseract to the 6 PPP where the upper-scale system is also a square in the tesseract. Similarly, the symmetry property appears as a logical solution to reduce the number of circles. Indeed, the more internal symmetries a circle possesses, the more "regular" it is.

As a matter of fact, for some of the 1344 circles, it is possible to find axes that split the circle in two halves such that, for a particular bit $n \in [\![0;3]\!]$, each element on the circle has the same n-th bit than its symmetric counterpart w.r.t. the axis. For example, on Figure 6.4, axis A_1 cuts the circle in such a way that each opposite elements have the same fourth bit (b_3). But there may be others axes (say A_2 and A_4) that create some symmetry for other bits, such as represented on Figure 6.4.

Based on this property, it is possible to count, for each circle, the number of axes bit-symmetries and the number of bits that are symmetric for each of them. Following the principle that the more regular the structure, the simpler the explanation, the circles that are ultimately selected are the one that have the most symmetries. The histogram of the distribution of circles as a function of their number of symmetries is shown on

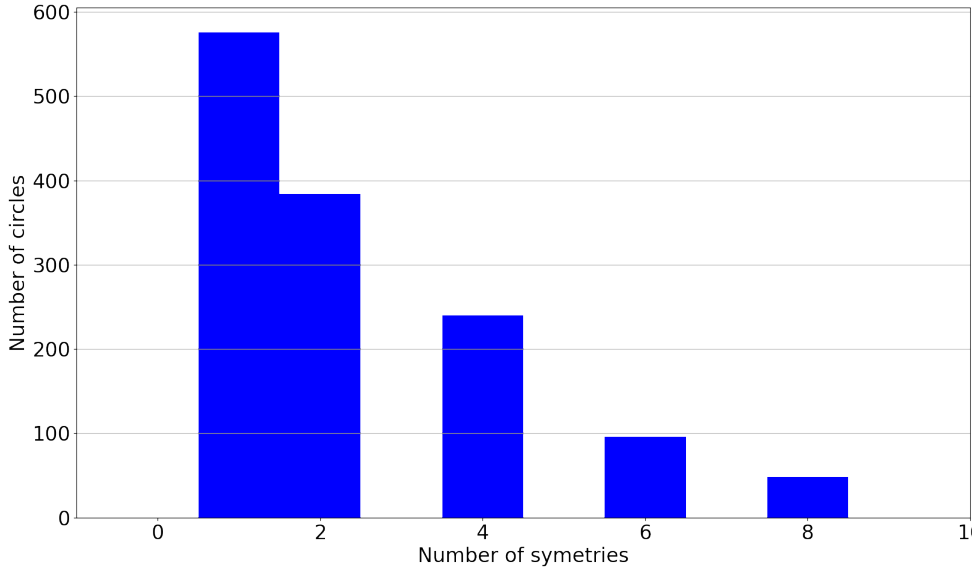


Figure 6.5: Number of circles that have n bit-symmetries (total is 1344)

Figure 6.5.

There are precisely 48 circles that have 8 symmetries (which is the maximum). All of these circles have four axes of bit-symmetry, with two of them creating one bit-symmetry and two others three bit-symmetries. And for each of these circles, there is one bit that has four axes of symmetries, one that has two, and two that have only one. Therefore, the symmetry property can not be used any further to shortlist this set of circle on the basis of the simplicity of their structure. This is not a problem: this inventory of 48 circles provides a very good subset, remembering that we started with 1344... They are in a sufficiently small number to be studied individually while still offering an interesting range of possibilities in modelling rhythmic patterns.

And by construction, we can use the PGLR model to describe a rhythmic section using the same process as for chord sequences with triad circles.

6.2.2 Melody Relations

Once the relations between rhythmic elements have been formalised and internal stability is guaranteed, the formalisation of relations between melodic patterns still faces a few difficulties:

- As two melodic patterns may differ in rhythm, there can be a different number of notes, and then, it is hard to describe the relation between two patterns just with note displacements.
- The need to preserve the order between notes prevents us from using optimal transport as we did for chords. In fact, optimal transport compares all possible permutations, and then chooses the one with the minimum transport. But here, permuting notes does not have any relevance, as neglecting the order of the notes would definitely impact the relevance of the comparison between musical motifs.

- Some patterns may contain only silences. While the formalism of rhythmic relations can describe the transformation of active rhythmic patterns to silence (or silence to active rhythmic patterns), optimal transport was designed to describe relations between chords that have at least one note and the principle used to handle the void chord cannot be easily generalised to handle void rhythms and/or melodies[1].

To overcome these obstacles, an approach could be to complete the representation of the melodic elements using rules that are similar to the one used to complete chord sequence. This completion strategy would be applied to the pitch part of the melodic elements to associate a pitch to unpitched onsets (bit at 0). By having such information encoded, the relation between two melodic elements can be described as the combination of a rhythmic rotation and the encoding of pitch to pitch displacements. Therefore, it is necessary to design a completion method such that each onset (1 and 0) has a corresponding pitch.

To complete a melodic sequence, we used the two following rules (in this order) to define a substitute pitch for unpitched onsets (i.e. an onset whose bit is 0):

1. *Continuation*: when it is possible, a pitch on a 0 onset, is set to the last pitch played (on a 1 onset) in the song. This first "rule" is the same as for chords: it is based on note continuation, which means that when a note is being played, the listener perceives this note until the next note is played. However, in the case of chords, we could trace-back the last chord inside the sequence, whereas for melody it can happen that the previous note is way before the beginning of the segment boundaries of the section.

2. *Anticipation*: when there is no note before the unpitched onset, the pitch is set to the first note being played *after*. This rule is more arbitrary than the previous one, as it is not based on any well-established cognitive perception of pitch sequences. However, it can be justified by the MDL principle. In fact, by doing such a completion, the first note played, if not in the primer, is directly related to the initial pitch which is the simplest relation (the cost of encoding is the same as if it was the pitch on the first onset, if existing). In case the first note played is in the primer, the principle of assuming that the imaginary last note played is the same as the one played is the “simplest” possible explanation too, as melodic movements impliy small displacements in the pitch space.

Note that the solution chosen here is not the only possible one. Another way, similar to what was done with chords, could have been to just remove the 0000 rhythmic pattern and consider each silent pattern as a simple beat pattern with the last note played. There are indeed multiple other solutions, but we decided to use the one described above as it does not change the actual rhythmic information which makes the reconstruction of the sequence easier. It was also the simplest method to implement, given the rest of the work we did on rhythm and chords.

[1]For experiment on chords, void chords N were replaced by the last preceding chord in the sequence. However, here, it is not relevant to fill void rhythms by a repetition of the previous rhythm, as there is nothing comparable as chord persistence or key continuation for rhythm or melody.

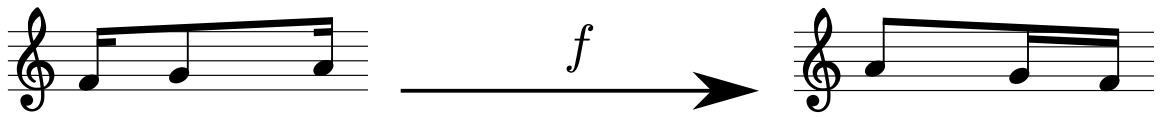


Figure 6.6: Description of the relation between two melodic patterns. Here, the rhythmic transformation (4, i.e. the second part, or fifth argument of the f relation) is obtained using the Gray circle clockwise.

Under this approach, melodic elements of a section are considered as belonging to $\mathcal{M} = [\![0; 11]\!]^4 \times R$, i.e. the melodic space is the combination of a rhythmic space and (the Cartesian product of) four pitch spaces. The relation between two melodic elements is described as the combination of four pitch-to-pitch displacements ($\in [\![0; 11]\!]^4$) and a rotation on a rhythm circle ($\in [\![-7; 8]\!]$) as seen on Figure 6.6.

The melodic space has the properties defined in 3.3.2, which ensures that the formalism can be used with an S&C description. For the rhythmic part, we can use the same arguments as the one used for the triad circles; for the pitch class part, as every onset is associated with a pitch-class, the circular structure of the pitch-class space ensures the properties. The generalisation of the properties to the whole melodic space is straightforward as the two parts (rhythm and pitch) are independent.

6.3 Time Alignment Optimisation

One of the peculiarities of melodic motifs in music is that their boundaries often happen not to be aligned with metrical instants (beats, bars) nor with harmonic changes (governed by chords and other accompaniments). In other words, the surface of melodic sections may present a phase shift with the metrical segment boundaries. This creates of course an additional difficulty when modelling melodic objects and sections. This section addresses this issue.

The first problem occurs in the case of anacruses, i.e when melody begins before the first bar of the sequence. If we synchronise blindly melodic elements on the first beat of the section, it may create a rather complex description of the melodic structure as the element used as the primer may start on the second or third beat of the melodic surface (depending on the length of the anacrusis), and it also potentially contains the beginning of the next motif...

The second problem is dual to the previous one: it happens when the melody of the previous section ends on the first beat of the next section. In that case, picking the melodic content occurring on the first beat of the latter section would result in taking the end of the melody of the former section instead of the real beginning of the melody of the current section. This would affect the efficiency and the interpretability of the resulting analysis.

In fact, a good structure model for melody must account for these various phase shifts of the surface content, and compensate for them in order to synchronise and match motifs

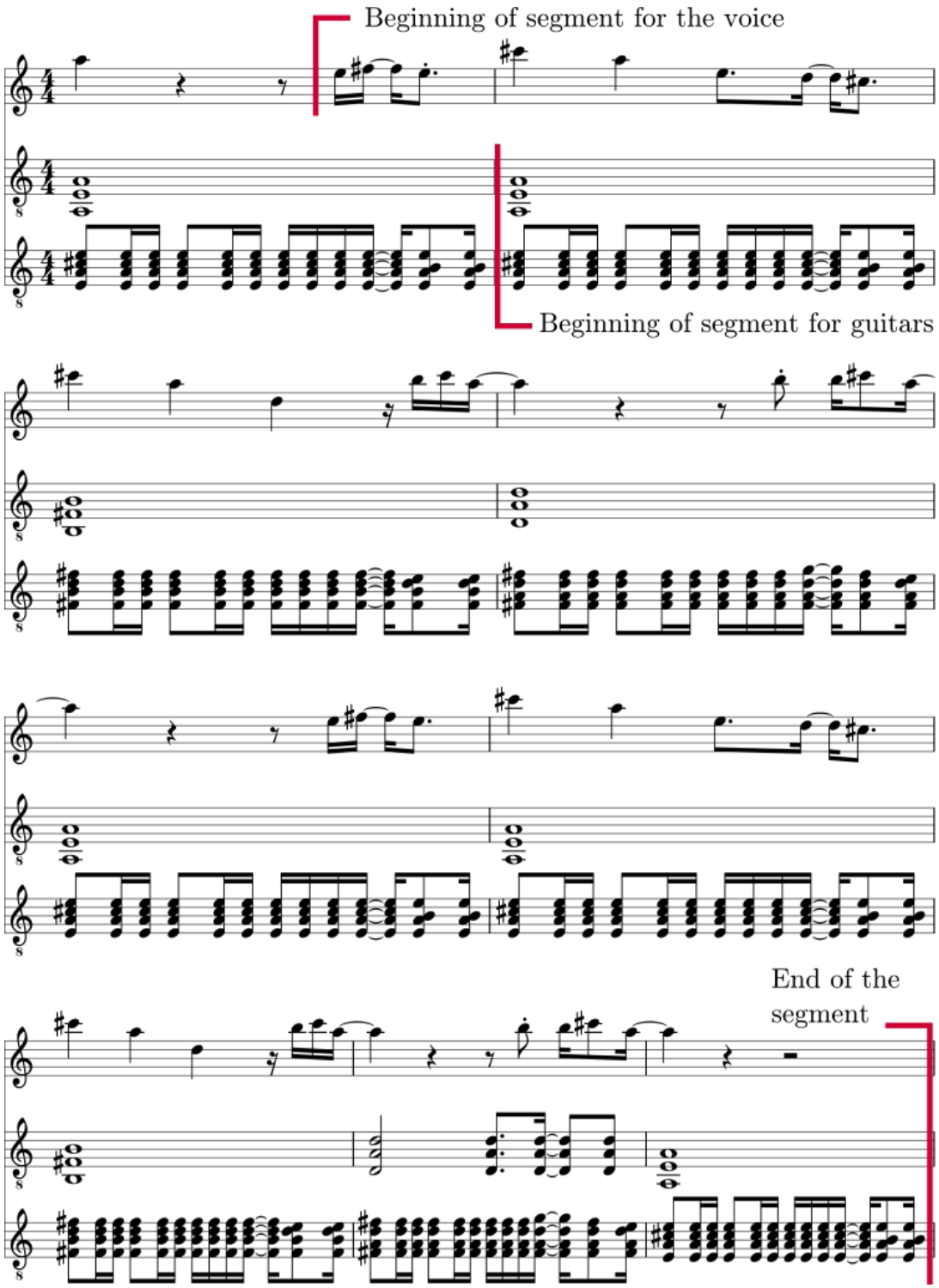


Figure 6.7: Example of a section, taken from the song "What's Up" by *4 Non Blondes*, where the melody does not fit inside the metrical boundaries of the segment for the accompaniment.

before inferring relations between them. Figure 6.7 shows an example where phase-shifts can be observed on a musical section.

In the most general case, each melodic element could be considered as having its own time-shift with respect to the beat, and each of these time-shifts would need to be estimated... However, this would have two drawbacks: (i) a very high complexity for considering all possibilities and (ii) the consequence that it would require to disqualify the difference of phase as a possible structuring relation between two elements in the section and/or for a contrast.

The solution chosen here therefore considers only one possible global phase-shift for the whole melody. In other words, the boundaries of the segment are globally shifted before extracting the melody of the song, then the resulting melody is split in 16 elements. This approach can be related to one of the rules (the 3rd one) used by Temperley in his segmentation algorithm Grouper [Temperley, 2004]. This rule associates a penalty to segments for which metrical position of the first note onset is not the same as the metrical position of the beginning of the other segments.

Under such an hypothesis, the solution becomes much more simple: given an interval of possible phase-shifts, the phase estimation algorithm computes a cost for the section description with each possible phase and then chooses the one that corresponds to the lowest cost.

Put into equation, let ρ be a phase-shift and $X^\rho = (x_i^\rho)_{0 \leq i < 16}$ be the sequence of elements obtained by splitting the "shifted" melody into 16 melodic elements[2]. Assuming we can compute a complexity score, $C_\mathcal{F}(r)$, for the description of a relation, r, using a formalism, $\mathcal{F}$, then, given a description model, M, we can compute a description cost, $C_M(X)$, for the whole sequence:

$$C_M(X^\rho) = \sum_{i=1}^{15} C_\mathcal{F}(r(\Phi_M(x_i^\rho), x_i^\rho)) \tag{6.3}$$

Therefore, given a melodic section, it is possible to compute, for each phase shift $\rho \in \mathcal{P}$, a complexity score. Following the MDL principle, the optimisation process then chooses the phase-shift leading to the minimal complexity score, that is:

$$\rho^* = \arg\min_{\rho \in \mathcal{P}} C_M(X^\rho) \tag{6.4}$$

Except for the additional minimisation over the phase-shift, the criterion defined in Equation 6.3 is identical to that of Equation 3.5 which gives the description cost of a sequence and a relation formalism (which was used to chose the best PPP in the last chapter using a chord relation description cost). Here, the cost of the relations between rhythmic or melodic elements are defined in a similar way as it was done for chord relations:

- the cost with the rhythm formalism, $C_{\mathcal{F}_R}(r)$, can be defined as the absolute value of the rotation angle on the circle:

$$C_{\mathcal{F}_R}(r) = |r| \tag{6.5}$$

[2]It is not really the melody that is shifted but the segment boundaries of the section, but it can also be seen as if we would have shifted the melody in the other direction.

- the complexity cost of a melodic relation, $r = (r_r, r_p)$ associated with the formalism described in Section 6.2.2, $C_{\mathcal{F}_M}(r)$, is defined as the sum of the rhythm relation cost, $C_{\mathcal{F}_R}(r_r)$ (absolute value of the rotation on the circle), and the cost of the pitch part of the relation, $C_{\mathcal{F}_P}(r_p)$ (sum of all the pitch displacements in semi-tones):

$$C_{\mathcal{F}_M}(r) = C_{\mathcal{F}_P}(r_p) + C_{\mathcal{F}_R}(r_r) \tag{6.6}$$

$$= \sum_{i=0}^{3} |r_{p_i}| + |r_r| \tag{6.7}$$

Note that, in the first case, r just encodes a rotation on a rhythm circle while in the second case, $r = (r_p, r_r)$ is the combination of four pitch-class displacements, r_p, and a rotation on a rhythm circle, r_r.

It is important to note that a phase may have a value which is not a integer number of beats. In such a case, every basic melodic element may vary with the change of phase because the sampling of the melody is made after the determination of the optimal phase-shift. Therefore, a phase-shift may create more similarities between basic melodic elements both by, changing the melodic elements and enhancing their similarities with each other, but also by creating a sequence where elements are better aligned and therefore relate more closely to their antecedents in the systemic model.

6.4 Results

After having exposed the principles on which are based the modelling of rhythmic and melodic relations in this work, and explained how the proposed approach can be handled by extending the one previously used with chords, it is time to evaluate and compare the performances of the implementations of the PGLR model on these musical dimensions.

Similarly to what we have done with chord sequences description, we conduct a study of the behaviour of the models on different rhythm circles, in order to see the effect of the choice of a specific relation formalism on the prediction task performances. As in the previous series of experiments, all results are obtained using 2-fold cross-validation strategy on the corpus of the 791 sections of 16 elements extracted from the melodies of RWC POP.

6.4.1 Rhythm Modelling

Figure 6.8 represents the distributions of the cross-perplexities over the 1344 rhythm circles for each model. That is, for each type of model, a box represents the distribution (over the 1344 rhythm circles), of the cross-perplexity of the model averaged over the 791 test sections.

Table 6.1 shows the maximum difference of average cross-perplexity between two models. That is, for two models $\mathcal{M}_1$ and $\mathcal{M}_2$, the value computed is:

$$d^*(\mathcal{M}_1, \mathcal{M}_2) = \max_{0 \leq i < 1344} \left(\tilde{B}(\mathcal{M}_2|C_i) - \tilde{B}(\mathcal{M}_1|C_i) \right) \tag{6.8}$$

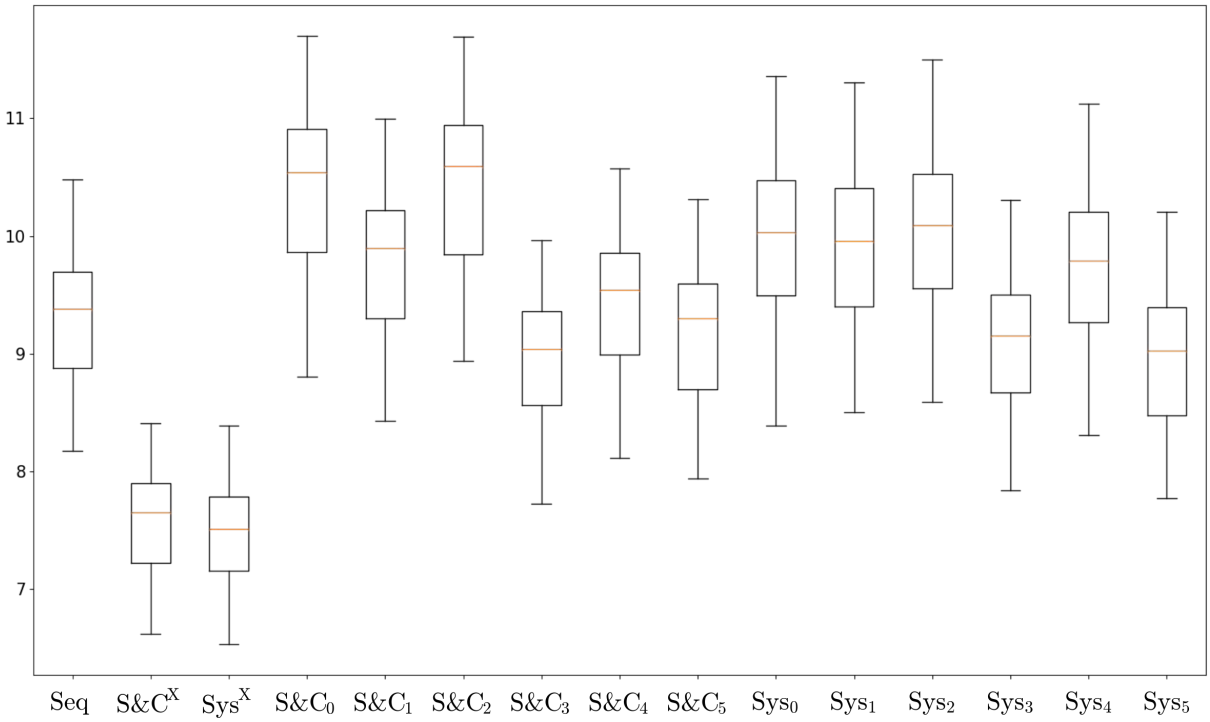


Figure 6.8: Schematic distributions of the average cross-perplexities over the 1344 rhythm circles for each model.

where C_i is the i-th rhythm circle used to describe the relation between rhythmic elements.

If the value is negative, it means that the second model, $\mathcal{M}_2$, is always better than the first one, $\mathcal{M}_1$, and that the difference between the two cross-perplexities is at least equal to the absolute value of the cross-perplexity difference. For example, considering the result on first column and fourth row, the sequential model, Seq, is always better than the static S&C model with no permutation, $S\&C_0$, with a difference in cross-perplexity of at least 0.35. The values in Table 6.1 are used to quantify the difference that can be seen on Figure 6.8. If a negative value appears in the table, it means that, even if the two models have boxes that seem to be at the same level, for all rhythm circle, the cross-perplexity value of the second model is always lower than the one of the first model.

Table 6.2 complements the results shown on Table 6.1 by giving the number of rhythm circles for which the first model performs better than the second model. For example, there are only 3 circles for which the average cross-perplexity of $S\&C_1$ is better than the one of Seq, and looking at the other table (Table 6.1), we can see that, in the best case when $S\&C_1$ "outperforms" Seq, the gain in cross-perplexity value of $S\&C_1$ is only of 0.06 which is only a very small difference.

6.4.1.1 Benefit of the Multi-Scale Organisation

A first observation that can be made by observing Figure 6.8 is that the combined multi-scale models ($S\&C^X$, Sys^X) perform better than the sequential one (Seq). By considering

		Model 2														
		Seq	$S\&C^X$	Sys^X	$S\&C_0$	$S\&C_1$	$S\&C_2$	$S\&C_3$	$S\&C_4$	$S\&C_5$	Sys_0	Sys_1	Sys_2	Sys_3	Sys_4	Sys_5
Model 1	Seq	0.00	−1.21	−1.30	1.75	0.87	1.68	0.11	0.69	0.38	1.31	1.14	1.35	0.31	1.02	0.09
	$S\&C^X$	2.26	0.00	0.09	3.34	2.60	3.35	1.61	2.19	1.93	3.05	2.95	3.13	1.94	2.76	1.83
	Sys^X	2.41	0.26	0.00	3.38	2.67	3.46	1.74	2.24	1.99	3.01	2.92	3.19	1.94	2.77	1.87
	$S\&C_0$	−0.35	−2.18	−2.23	0.00	−0.24	0.48	−0.99	−0.46	−0.75	−0.02	−0.12	0.14	−0.87	−0.17	−0.88
	$S\&C_1$	0.06	−1.78	−1.82	0.96	0.00	1.05	−0.53	0.02	−0.28	0.66	0.45	0.76	−0.37	0.40	−0.45
	$S\&C_2$	−0.37	−2.17	−2.24	0.38	−0.14	0.00	−0.90	−0.62	−0.82	−0.04	−0.08	−0.11	−0.83	−0.40	−1.08
	$S\&C_3$	0.81	−1.10	−1.15	1.83	1.10	1.94	0.00	0.76	0.52	1.52	1.39	1.63	0.39	1.27	0.33
	$S\&C_4$	0.47	−1.41	−1.51	1.32	0.66	1.31	−0.04	0.00	0.04	0.98	0.86	1.00	−0.04	0.60	−0.25
	$S\&C_5$	0.66	−1.26	−1.31	1.54	0.83	1.55	0.05	0.44	0.00	1.33	1.12	1.30	0.22	0.95	0.04
	Sys_0	−0.02	−1.68	−1.73	0.84	0.26	1.02	−0.47	−0.10	−0.32	0.00	0.23	0.36	−0.42	0.10	−0.54
	Sys_1	−0.07	−1.74	−1.81	0.87	0.21	0.94	−0.54	−0.11	−0.36	0.38	0.00	0.42	−0.47	0.14	−0.64
	Sys_2	−0.06	−1.82	−1.90	0.80	0.16	0.71	−0.54	−0.25	−0.42	0.25	0.20	0.00	−0.49	−0.12	−0.73
	Sys_3	0.76	−1.19	−1.28	1.57	0.97	1.79	0.08	0.60	0.35	1.23	1.15	1.43	0.00	1.01	0.18
	Sys_4	0.19	−1.56	−1.66	1.04	0.45	1.06	−0.27	0.00	−0.16	0.50	0.46	0.47	−0.23	0.00	−0.48
	Sys_5	0.84	−0.97	−1.10	1.87	1.13	1.79	0.40	0.74	0.45	1.39	1.28	1.40	0.39	1.03	0.00

Table 6.1: Maximum difference of average cross-perplexity on a rhythm circle between two models. The difference is defined by Equation 6.8.

		Model 2														
		Seq	$S\&C^X$	Sys^X	$S\&C_0$	$S\&C_1$	$S\&C_2$	$S\&C_3$	$S\&C_4$	$S\&C_5$	Sys_0	Sys_1	Sys_2	Sys_3	Sys_4	Sys_5
Model 1	Seq	0	0	0	1344	1341	1344	27	1016	301	1344	1344	1344	185	1294	17
	$S\&C^X$	1344	0	63	1344	1344	1344	1344	1344	1344	1344	1344	1344	1344	1344	1344
	Sys^X	1344	1281	0	1344	1344	1344	1344	1344	1344	1344	1344	1344	1344	1344	1344
	$S\&C_0$	0	0	0	0	0	727	0	0	0	0	0	15	0	0	0
	$S\&C_1$	3	0	0	1344	0	1344	0	2	0	1144	1097	1279	0	547	0
	$S\&C_2$	0	0	0	617	0	0	0	0	0	0	0	0	0	0	0
	$S\&C_3$	1317	0	0	1344	1344	1344	0	1344	1333	1344	1344	1344	1296	1344	629
	$S\&C_4$	328	0	0	1344	1342	1344	0	0	3	1344	1344	1344	0	1344	0
	$S\&C_5$	1043	0	0	1344	1344	1344	11	1341	0	1344	1344	1344	346	1344	2
	Sys_0	0	0	0	1344	200	1344	0	0	0	0	333	1007	0	28	0
	Sys_1	0	0	0	1344	247	1344	0	0	0	1011	0	1248	0	32	0
	Sys_2	0	0	0	1329	65	1344	0	0	0	337	96	0	0	0	0
	Sys_3	1159	0	0	1344	1344	1344	48	1344	998	1344	1344	1344	0	1344	91
	Sys_4	50	0	0	1344	797	1344	0	0	0	1316	1312	1344	0	0	0
	Sys_5	1327	0	0	1344	1344	1344	715	1344	1342	1344	1344	1344	1253	1344	0

Table 6.2: Number of circles for which Model 1 performs better than Model 2.

Table 6.2, it appears that there are no circle for which the sequential model outperforms these two models. Moreover, by considering Table 6.1, it appears that $S\&C^X$ always outperforms Seq with a difference of cross-perplexity of at least 1.21 (and even 1.30 for Sys^X).

Therefore, the observation that multi-scale models outperform the sequential model in the chords prediction task seems to extend to rhythm prediction, at least in the framework we consider for rhythm modelling. Note that the multi-scale models ($S\&C^X$ and Sys^X) keep outperforming the sequential model, even after adding the cost of the combination model ($\log_2 6/16$, see Equation 4.11) to the negative log-likelihood before computing the cross-perplexity, even though, after such an operation, the minimal difference of cross-perplexities between $S\&C^X$ and Seq is reduced to 0.34. This confirms the hypothesis that the multi-scale point of view greatly impacts in a positive way the modelling performance of musical sections in terms of prediction.

However, considering each PPP model separately, some of them appear to be less effective than the sequential one, especially $S\&C_0$ and $S\&C_2$ which seem to be the models that have the worst performances (hence the advantage of optimising the PPP for each

section). Considering Table 6.2, there is no circle for which these two models perform better than the sequential model. $S\&C_1$ performs a bit better but is still almost always worst than the sequential model. The remaining three permutations display a better average performance. While $S\&C_4$ performs almost at the same level than the sequential model, $S\&C_3$ and $S\&C_5$ outperform it for a great majority of rhythm circles (1317 out of 1344). In fact, the PPP corresponding to the best performance is the same as for chords prediction, namely $S\&C_3$ which outperforms $S\&C_5$ in most of the cases and wins over all other PPP for all circles. And in the rare cases where the cross-perplexity of $S\&C_5$ is lower than the one of $S\&C_3$, the maximum difference is 0.05, which is, here again, rather small.

Globally, these results confirm the sense that non-sequential dependencies defined on a PGLR structure greatly facilitates the prediction of information in music, here in the case of the rhythm dimension.

Of course, our representation of rhythm is only some approximation of the real rhythm information, as it reduces 2-beat rhythmic cells to four bits (i.e. each bit corresponds to a 8th note). This is a choice that has been made to simplify the description of relations between rhythmic cells. It can be compared to the reduction of chords (of four or more notes) to major and minor triads, in order to simplify the relations from optimal transport to rotations on the triadic circle.

Therefore, it may be kept in mind that part of the results reported here are determined by the type of rhythmic model used in the experiments, which may influence the relative performance of the various approaches. But still, the advantages that can be expected from the static S&C model are patent.

6.4.1.2 Relative Importance of the Virtual Element

Another observation that can be made from Figure 6.8, Table 6.1 and Table 6.2 is that the use of a virtual rhythmic element instead of the primer in the multi-scale description is not as effective as it was for chords sequence prediction. In fact, Sys^X and $S\&C^X$ display almost the same level of performance, with even a slight advantage of Sys^X over $S\&C^X$ on a majority of circles. Sys_0 and Sys_2 also outperform their corresponding model with virtual element ($S\&C_0$ and $S\&C_2$). Moreover, the single permutation multi-scale model that reaches the minimal average cross-perplexity is Sys_5 which is almost always better than its equivalent, $S\&C_5$.

However, it is worth noting that the best $S\&C_i$ permutation model, $S\&C_3$ (which was also the best model for the description of chord sequences) outperforms its equivalent, Sys_3, for a great majority of the rhythm circles. Its performance is also very close to that of Sys_5, as for almost half of the circles, $S\&C_3$ has a lower cross-perplexity and the worst case difference is substantial (Sys_5 better than $S\&C_3$ by at most 0.40 and $S\&C_3$ better than Sys_5 by at most 0.33).

Therefore, for rhythm section description, depending on the section or the circle that is used, the virtual element may have some advantage over the primer as an antecedent for the contrastive element, but this advantage may disappear when considering a large number of sections for a given circle. A reason for this may be that, for rhythms, the

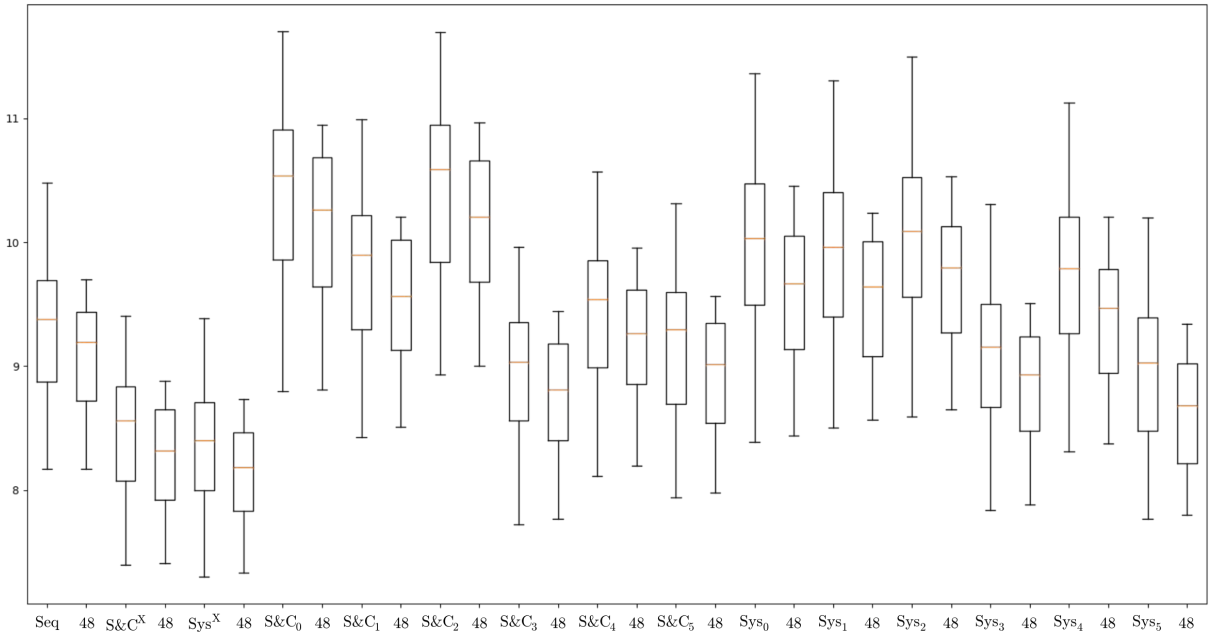


Figure 6.9: Distributions of the cross-perplexities over the 1344 rhythm circles and only 48 circles, alternating for each model.

contrast is more pronounced than for chords. As it is not as often equal to identity, as it was for chords, it may be harder to predict.

6.4.1.3 Importance of the Rhythm Circle

Another observation that can be made by considering Figure 6.8 is that the choice of the adequate relation formalism between rhythmic elements is also very important. In fact for every model, there is a definite difference of average cross-perplexity depending on the rhythm circle chosen. The difference between the maximal and the minimal cross-perplexities varies from one model to another, within about 1.5 points of cross-perplexity, which is a rather strong difference.

Until now, we compared the results of the different models for each of the 1344 rhythm circles. However, to evaluate the relevance of the bit-symmetry property as a way to reduce this number to 48, we compared the results obtained on the 48 circles with those achieved with the entire set of 1344 circles. By doing this, it was possible to evaluate if the 48 circles may be considered as a judicious selection that can reduce the computational complexity without altering significantly the performances of the model.

Figure 6.9 shows the corresponding results by depicting, for each model, (i) the performance obtained on the 1344 circles and (ii) the results on the 48 circles only. The figure shows clearly that "shortlisting" the subset of 48 circles clearly lowers the upper bound of the performance interval (i.e. by getting rid of circles that do not perform well), while it hardly impacts the lower bound (i.e. discarding good circles). In fact if, for some models, the circle that performed the best is not in the subset of 48 circles, the result using the best circle in that subset is never very far from optimal. Therefore, the simplification

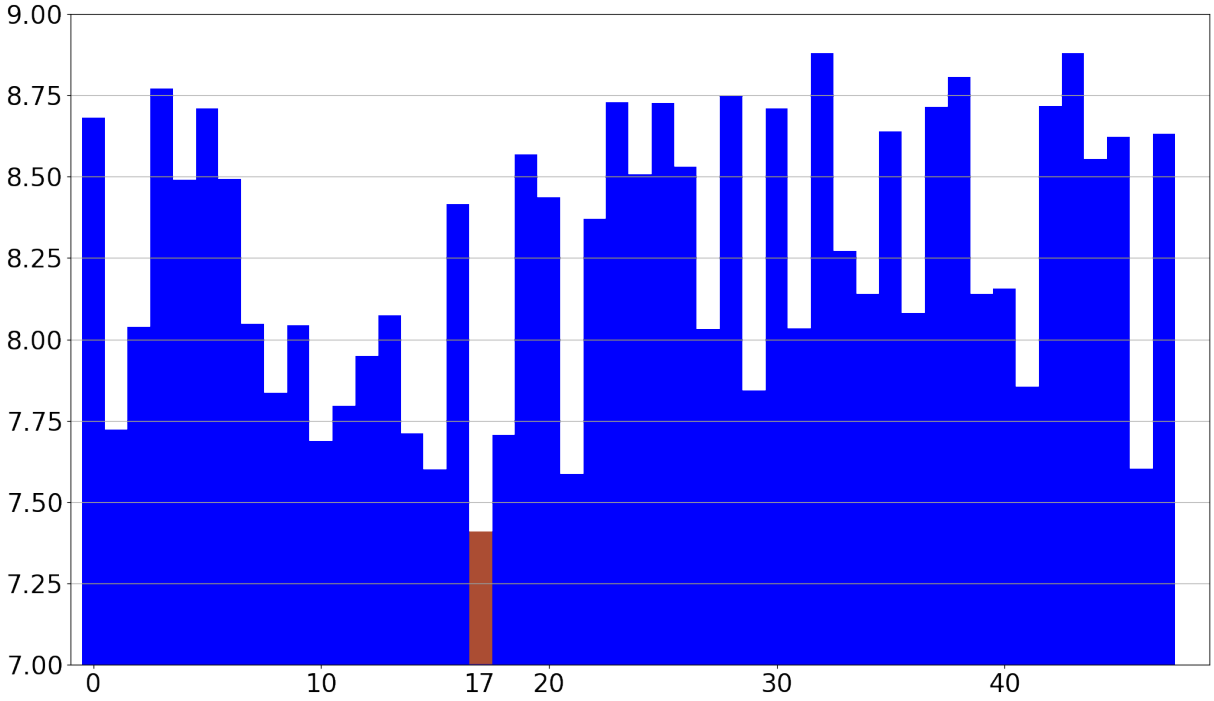


Figure 6.10: Cross-perplexity values obtained by the $S\&C^X$ model over the 48 circles.

seems to be a good compromise and globally all the average scores improve visibly when moving from 1344 circles down to 48.

It is also important to notice that the behaviour of a rhythm circle is almost the same for every possible model. In fact, considering Table 6.2, there are many cases when a model surpasses another one for a great majority of circles. It implies that the ranking of the models, for a given circle, is almost always the same. Therefore, the choice of the good relation formalism is a very important step for the description of rhythmic sections.

As a consequence, we investigated the effect of the choice of circle, over the performances of the $S\&C^X$ model, by focusing on the 48 circles that have 8 bit-symmetries and observe the distribution of cross-perplexities across them. Figure 6.10 shows the average cross-perplexity value obtained with the $S\&C^X$ model associated with each 48 circles. The lowest (i.e. the best) cross-perplexity value (7.41) is obtained with circle $n°17$, which is represented with its associated bit-symmetry axes on Figure 6.11. The difference in cross-perplexity value between this circle and the other ones ranges between 0.17 (minimum) and 1.47 (maximum).

It is interesting to study the structure of the circle to understand what are its main features and why these features may be particularly appropriate to describe relations between rhythmic patterns. Among the observations that can be made, the most striking one is probably that, for this circle, the bit that has the most symmetries is the first bit, i.e. the bit that encodes the strong beat of a basic rhythmic element. Therefore, the reason of the performances associated with this circle may be that it favours regularities between rhythmic elements, with a primary importance given to the strong beat.

Moreover, considering elements that have their first bit active (i.e. an onset set to 1

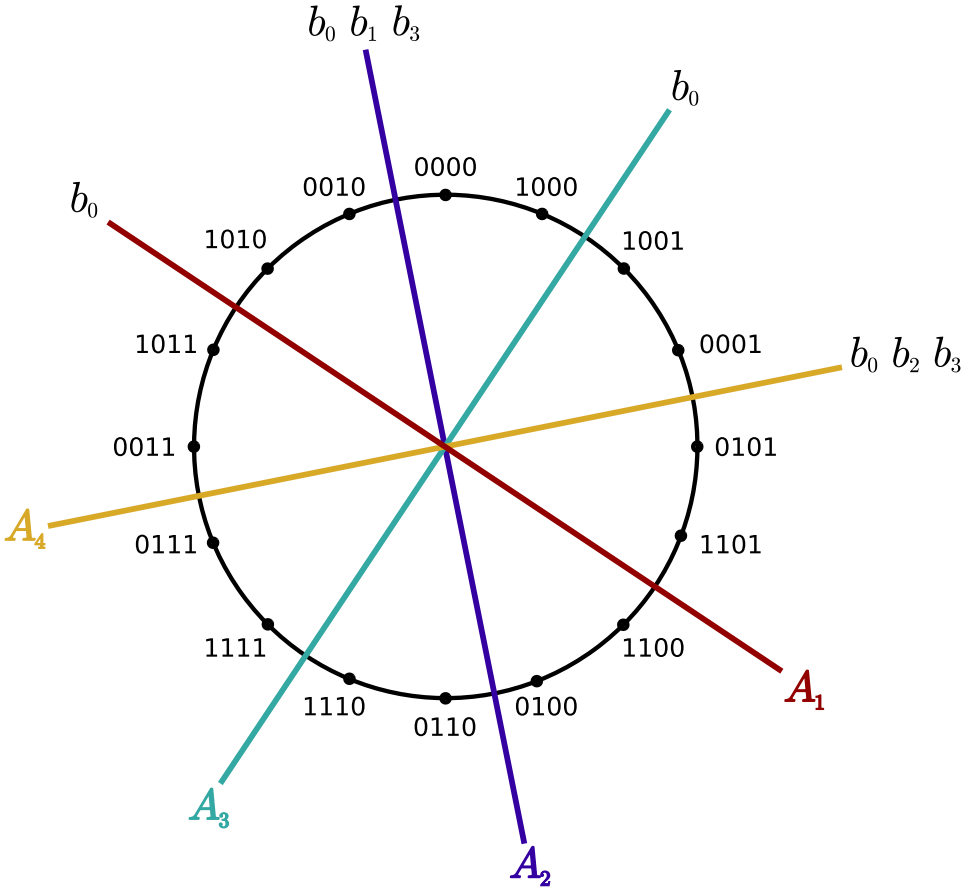


Figure 6.11: The rhythm circle that provides the best cross-perplexity value with the $S\&C^X$ model over the 48 circles.

on the first strong beat), each of them has an immediate neighbour which differs only on the *last* bit (i.e. presumably the weakest beat of the group). Therefore, these differences by one step on the circle link elements that are usually considered as close to each other musicologically. It can be assimilated to the major triad and its relative minor triad that are neighbours in the triad circle of thirds and considered as very close harmonically.

Finally, by considering the $S\&C^X$ model and the circle leading to the best performance (namely $n°17$), it is possible to represent the distribution of the PPP providing the best cross-perplexity over the set of sequences in the corpus. In fact, Figure 6.12 is the equivalent for rhythm modelling of what Figure 5.9 was for chord sequence modelling, as both depicts the distribution of the optimal PPP for the $S\&C^X$ model over the test sections (but for different musical dimensions).

The main observation that can be made from Figure 6.12 is that, for the rhythm (as it was the case for chords), PPP_3 is the one that performs the best in a majority of cases (38% for rhythm vs 32% for chords). The ranking of the other permutations is slightly different (but not so much, when comparing the two histograms). PPP_5 comes second for rhythm, with 22% (versus 16% for chords), then PPP_1 and PPP_4 with 12% each (vs. 21% and 10% for chords, respectively), then PPP_1 with 11% comes just after (18% for chords) and finally PPP_2 with 5% (chords: 3%).

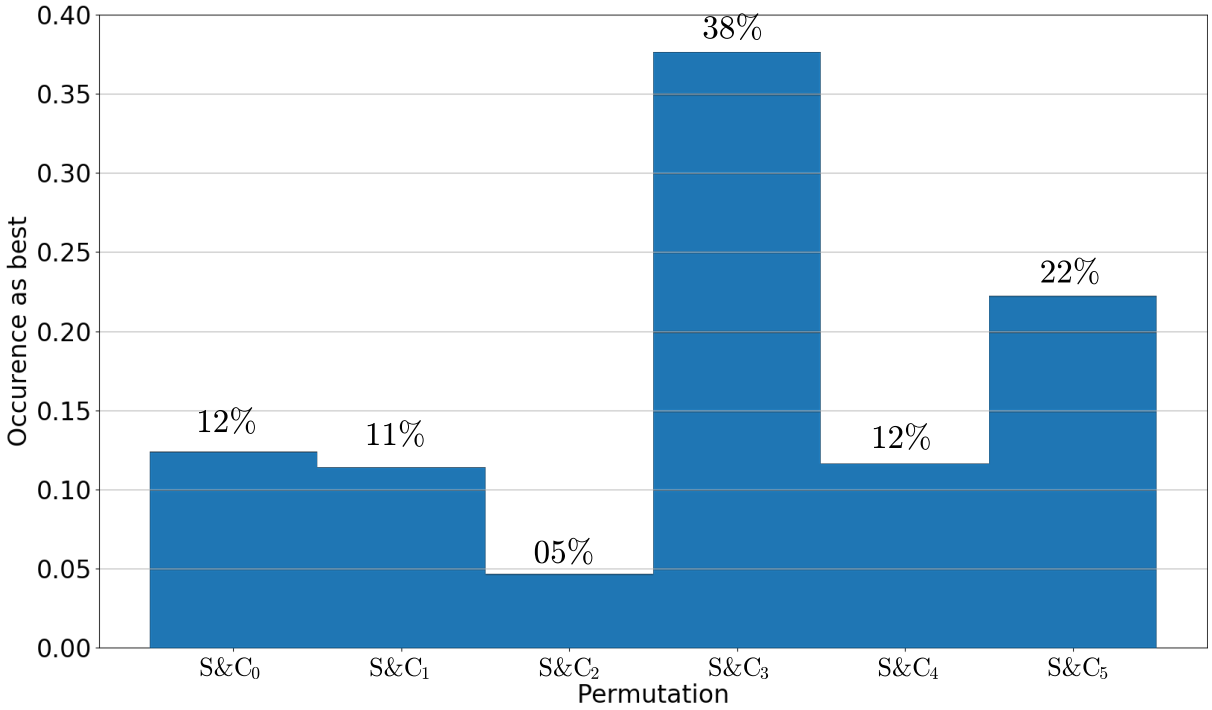


Figure 6.12: Distributions of the PPP ranking best with the optimal circle for $S\&C^X$ over the 791 sections.

The relative similarity of the two distributions illustrates one of the interest of the PGLR model as a common framework for multi-dimensional musical structure description. It makes it possible to compare different models and different musical dimensions within a generic scheme. Here, the fact the results are similar (yet not completely correlated) over the different dimensions of musical information (rhythm vs. chords), shows that there may exist similar trends shared by totally distinct musical dimensions, beyond their different nature at the surface level.

6.4.1.4 Impact of Phase Shift

As developed earlier, the fact that some sections may contain anacruses or late departures is bound to have a clear impact on the results. Therefore, to investigate this aspect, we computed, for each section, each model (Seq, Sys^X, $S\&C^X$ Sys_i, $S\&C_i$) and each circle (out of the 1344 possibilities), the optimal phase-shift, i.e. the phase-shift corresponding to the description with the lowest cost using the method described in Section 6.3. Then, the performance on the prediction task was measured with the phase-compensated shifted sections.

The results are presented on Figure 6.13.

Rather clearly (but not so surprisingly), for every model, the use of the phase optimisation drastically improves the prediction performance. For instance, the drop of the mean average cross-perplexity over the 1344 circles yields a difference of 2.20 cross-perplexity points for Sys_5 and 3.12 for $S\&C^X$. This strong trend clearly supports the hypothesis

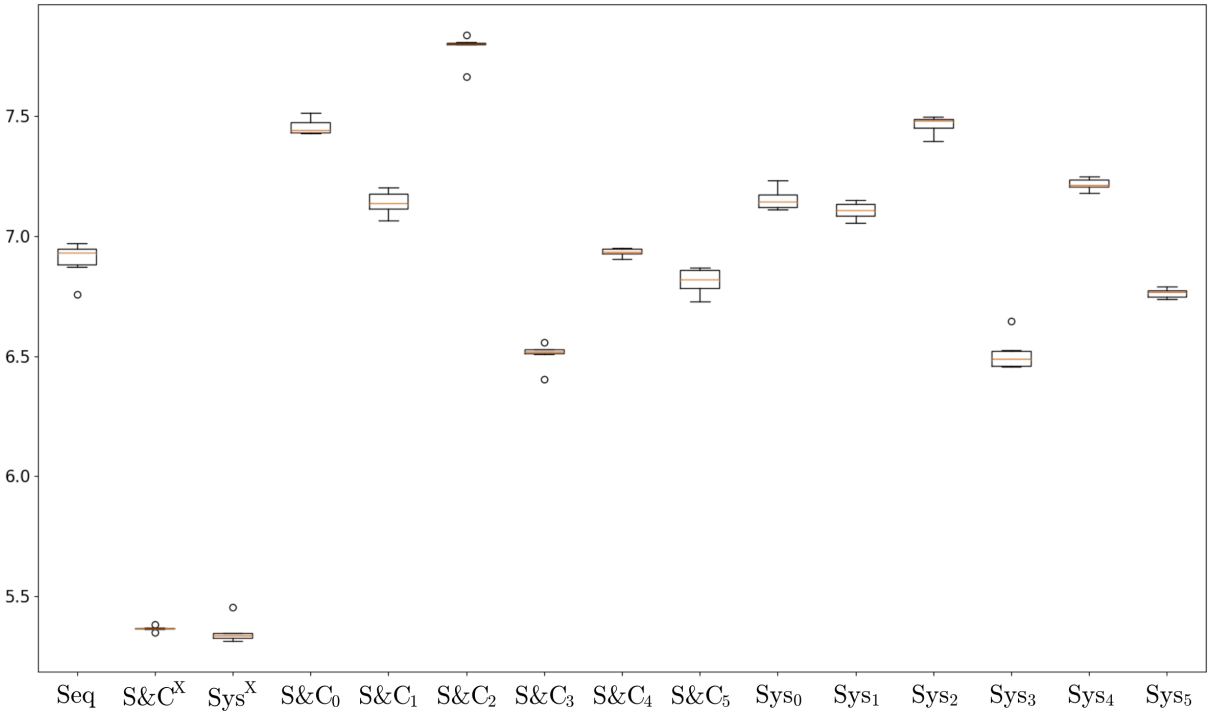


Figure 6.13: Distributions of the average cross-perplexities over the 1344 rythmic circles. The cross-perplexities are computed on the phase-compensated sequences.

that phase-shift compensation is of major importance to model rhythmic sequences. The incorporation of phase optimisation in the criterion greatly improves the prediction of rhythm by creating more coherent basic elements that have a higher similarity with one another.

Another very interesting result is that the difference across circles is much less crucial, as changing the circle does not affect very much the average cross-perplexity. This aspect is particularly visible for the $S\&C^X$ model for which the choice of circle seems to have almost no more effect on the average cross-perplexity.

Yet, as observed on Figure 6.13 the respective position and ranking of static models is almost the same as without optimisation, with only a slight difference for Sys_5 which lost its first position to Sys_3.

6.4.2 Melody modelling

We briefly recall that, by "melody modelling", we mean here, the joint modelling of rhythmic and pitch information in sequences of notes, as exposed above in Section 6.2.2.

Given the additional complexity resulting from this joint modelling, the computation time for the experiments on melodic sections could have become prohibitive. Therefore, given the results observed in our previous experiments, we have limited our experiments in melody modelling to the use of 48 rhythm circles (those that have the 8 bit-symmetries).

Moreover, on a data set like ours (791 sections), the space of melodic relations in its

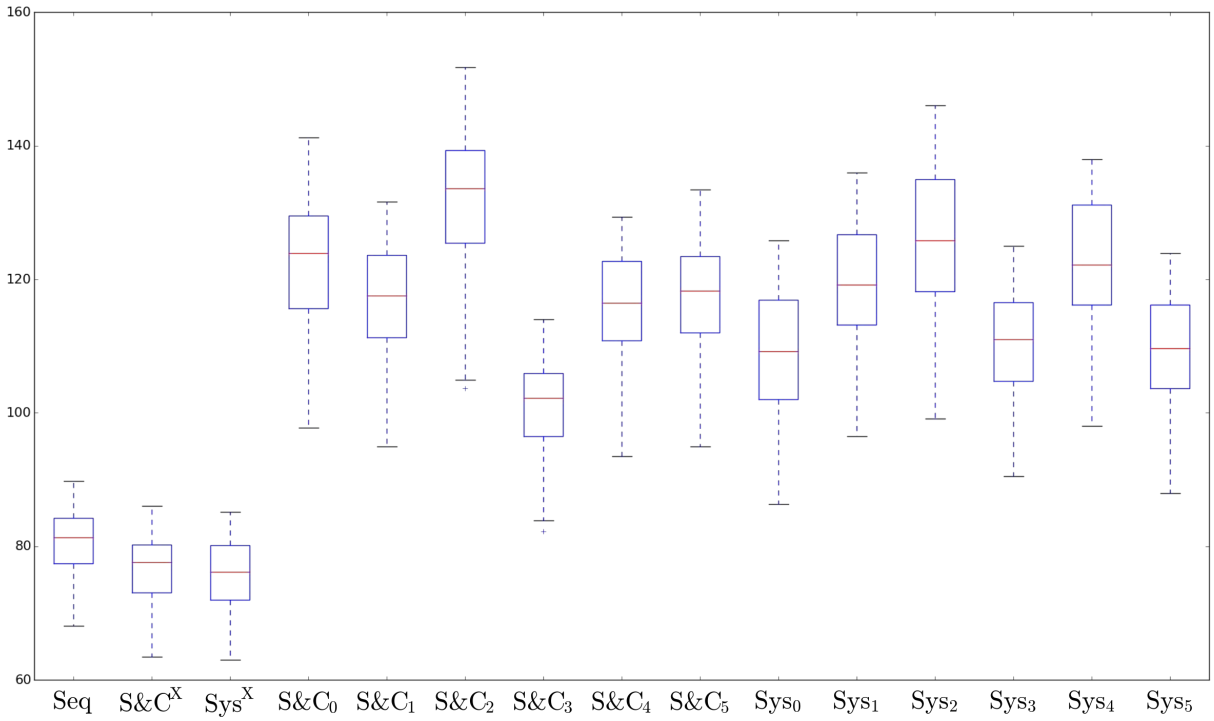


Figure 6.14: Distributions of the average cross-perplexities over the 48 rhythm circles for each model obtained with melodic data.

most general form would be too large to reliably estimate a probability for each relation. Simplifications of the relation space have been made accordingly.

In fact, if we consider a general melodic relation as defined in Section 6.2.2, it would belong to the space $\mathcal{M} = [\![0; 11]\!]^4 * R$ where R is a discrete space of size 16. Therefore, the total number of possible relations would be $12^4 * 16 = 331776...$ Compared to 791 sections in the corpus and only 15 relations per section, i.e. $791 \times 15 = 11865$ observed relations, any reliable probability estimation would be out of reach.

As a consequence, we assume that, given a melodic relation, note displacements are independent from one another, and from the position of the note in the rhythmic pattern. That is, given a melodic relation $((d_0, d_1, d_2, d_3), r)$, the probability distribution estimated for each pitch class displacement d_i is the same. This results in the following approximation:

$$\log_2 P(((d_0, d_1, d_2, d_3), r)) \approx \frac{1}{4} \sum_{i=0}^{3} \log_2 P((d_i, r)) \tag{6.9}$$

under which, the relation space is only of size $12 * 16 = 192$ (which clearly makes any estimation more reliable with 11865 data!) However, this simplification has also some drawbacks which will be discussed later.

Figure 6.14 represents the distributions of the melodic cross-perplexities over the 48 rhythm circles for each model. For each of them, a box represents the distribution (over the 48 rhythmic circles) of the average cross-perplexity obtained on the 791 sections.

A first observation that can be made by observing Figure 6.14 is that the multi-scale models (Sys^X, $S\&C^X$) now performs only slightly better than the sequential one (Seq). And when considering the encoding cost of the multi-scale models ($\log_2 6/16$ added to the negative log likelihood), their performances is not significantly different from that of the Seq model. Sequential and multi-scale models yield approximately the same compression ability.

Moreover, this figure shows that the sequential model clearly outperforms each of the single permutation models. But the respective ranking of the single permutation multi-scale models is similar to the one obtained when considering rhythm only. Indeed, as the rhythm is a strong basis of the melody, it seems logical that the results are similar.

The fact that the multi-scale models do not show a clear advantage over the sequential model may be owed to the representation used to create the basic melodic elements and the method used to estimate the probability of melodic relation. Indeed, the way a basic element is represented is highly dependent on a sequential process. That is, pitch information encoded for a rhythmic bit set to 0 is totally determined from the observation of the last preceding pitch on a rhythmic bit set to 1. This may create an unrelevant similarity relation in the estimation process.

As a consequence, a basic element with no rhythmic bit set to 1 (in a section containing mostly silences), may have pitch class values that are similar (or even identical) to the pitch classes of the direct predecessor of the element in the section, while long range relation may create more gaps between pitch class values. In such a case, sequential relations may become simpler to describe than long range relations. This can explain why the Seq model has similar results to $S\&C^X$ and Sys^X.

Therefore, it appears unnecessary to encode and then predict such information that can be automatically reconstructed on the basis of the pitch information alone, from the knowledge of onsets set to 1 and using the same completion rules as when encoding the initial sections. In fact, once the rhythm is predicted (or generated), the only pitch information that needs to be encoded is the pitch classes that are on bits set to 1.

To measure how the sequential completion affects the results, we computed a modified NLL score, to evaluate the capacity of a model to predict the relevant information, that is pitch class displacements associated to active onsets only. Therefore, for each sequence $X = (((p_0, p_1, p_2, p_3), r)_i)_{0 \leq i < n}$ having $m \leq 4n$ activated onsets (excluding the primer) and denoted as $p^*_0 \dots p^*_{m-1}$, we compute a cross-entropy like score, $\widehat{H^A_M}$, which is restricted to pitch displacements on activated onsets. This score is therefore defined by:

$$\widehat{H^A_M}(X) = -\frac{1}{m} \sum_{j=0}^{m-1} \log_2 P_M((p^*_j, r_i) | \Phi_M(p^*_j, r_i)) \tag{6.10}$$

where, $\Phi_M(p^*_j, r))$ is the pair (p^k, r^k) with:

- $x_i = (P_i, r_i)$ such that p^*_j is the pitch in P_i associated with the j-th activated onset of the section
- $\Phi_M(x_i) = x_k = (P_k, r_k)$

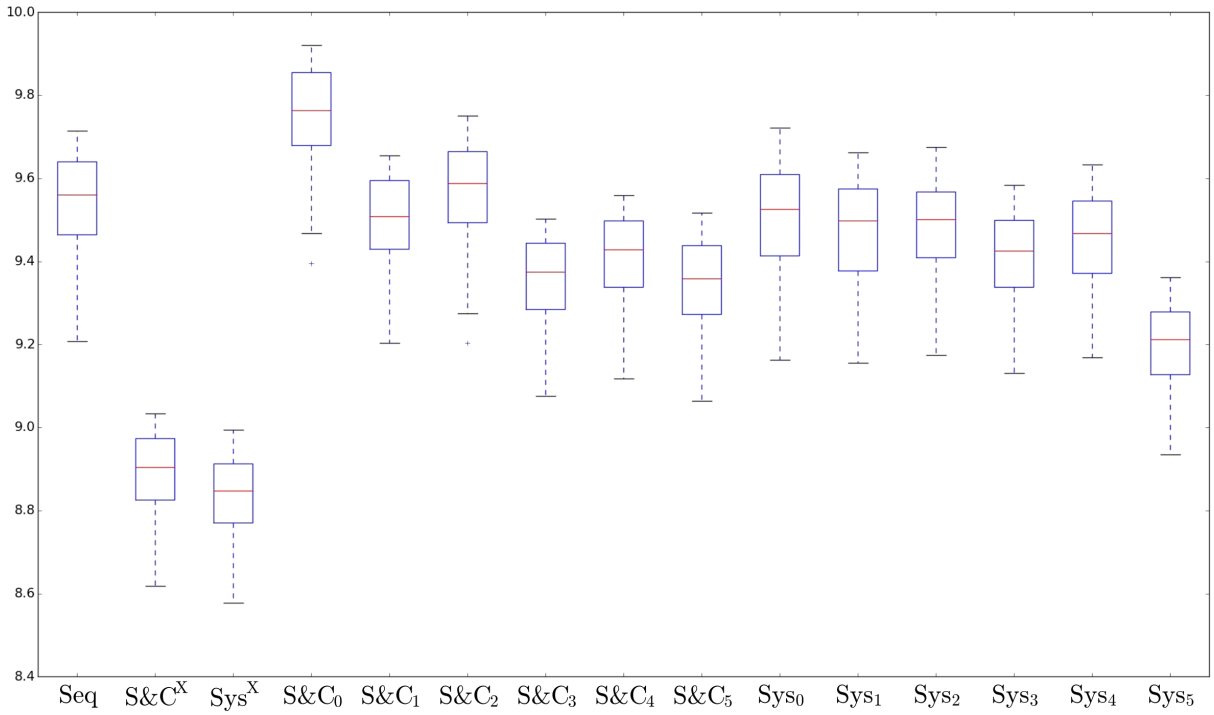


Figure 6.15: Distributions of the average extended cross-perplexity score over the 48 rhythm circles for each model obtained with melodic data.

- p^k has the same index in P_k than p_j^* in P_i

Here, the probability $P_M((p_j^*, r)|\Phi_M(p_j^*, r))$ is estimated as the frequency in the learning corpus of the pair $((p_j^* - p^k) \mod 12, (r - r^k) \mod 16)$. Note that $\Phi_M(p_j^*, r)$ may be related to a non-activated onset, and for this reason it is still important to have an initial representation with pitch information on non-activated onsets.

This new score provides the likelihood that the model predicts the pitch class displacement and the rhythm change (which can be identity) for newly activated pitch classes. As there may be a difference of activated locations across sections, this is not a normalised score for comparing a given model on different sequences. However, this score can be used to compare different models on the same set of sections : as the number of basic elements with onset is only dependent on the section and does not change with the model , the normalisation factor (m) is the same for all models. It therefore provides a score that is comparable across models.

Figure 6.15 represents the distributions of the modified score, $\widehat{H_M^A}$, obtained with the 48 rhythm circles for each model. That is, for each model, a box represents the distribution, over the 48 rhythm circles, of the average score obtained using a rhythm circle on 736 test sections (out from the 791), as the 55 others were containing no activation location (i.e. $m = 0$).

The results corresponding to this new scoring function are depicted on Figure 6.15 and they now show the same tendency than the one observed with chords and rhythm. By considering only activated onsets, the multi-scale models (Sys^X and $S\&C^X$) greatly outperform the sequential model (except for the $S\&C_0$ model). Here, however, the virtual

element does not seem to have a positive effect on the performance, but this observation was not further investigated.

6.4.3 Summary

Results presented in this chapter point towards and strengthen the potential of the PGLR model as an attractive approach for music structure modelling along several dimensions.

It would then be very interesting to conduct further studies based on this framework, as there are several aspect addressed in this work that can be further improved or extended in different directions. Examples of such perspectives are briefly considered in the next and last chapter of this thesis.

Chapter 7

Conclusion and Perspectives

7.1 Contributions of this work

The work developed in this manuscript has been studying the formalisation, the extension and the implementation of the System&Contrast (S&C) model for music structure modelling in a multi-scale and multi-dimensional framework. Even though the present work has not yet reached a fully operational status, it opens new tracks for modelling music structure and offers an attractive paradigm for information processing in music.

One hypothesis behind the S&C model is that, the relations between the elements forming a musical section and the logical implications that can be induced from them, provide a relevant description of the structure of this section.

To consolidate this hypothesis, the S&C framework has been investigated, extended and evaluated in this thesis, along several directions:

- *in a multi-scale framework*, with the PGLR (Polytopic Graph of Latent Relations) as a general scheme for describing the systemic structure of regular musical sections,
- *for several musical dimensions*, i.e. considering chords, rhythm and melody,
- *considering a variety of formalisms* to represent the relations between musical elements, based on optimal transport or circular displacements.

As a common background to this work is the Minimum Description Length principle, which has been called for in multiple occasions for defining cost criteria whose optimisation has provided structural descriptions of musical contents.

In this framework, the Polytopic Graph of Latent Relations (PGLR) model assumes that relations of dependency between the elements of regular sections lie on a corresponding polytope (square, cube, tesseract...), and under such an assumption, different models have been built to describe sections of 16 elements, in particular:

- a sequential model, *Seq*, which relates each element in the section with its direct predecessor (bi-gram);

- six multi-scale tree systemic models, Sys_i, which consist in four lower-scale systems where all elements are described in relation to the primer of the system and one upper-scale system linking each primer of the lower-scale systems. Each model is associated to a Primer Preserving Permutation (PPP) which is a permutation such that the graph formed by the five systems has some particular geometrical properties when projected on the tesseract.

- six static multi-scale S&C models $S\&C_i$, which consist in four lower-scale S&Cs and one upper S&C linking each primer of the lower-scale S&C. The difference with the tree-systemic models comes from the fact that the fourth element in each S&C is defined in relation to a virtual expected element instead of the primer. Each multi-scale static S&C model is also associated with a PPP.

- two models, Sys^X and $S\&C^X$, which consider all the PPPs and choose, for a section X, the PPP that best describes it, by minimising a description cost.

- a dynamic multi-scale S&C model, Dyn^X, which, for each element in a contrastive position, construct a S&C to describe it as a local surprise.

- a relational multi-scale static S&C, $RS\&C$, which consists in describing lower-scale S&Cs as S&Cs of relations instead of S&Cs of elements.

Each of these models is associated with an *antecedent function* formalising the fact that they are all first-order models (in the sense that each element has only one antecedent). In some cases, the antecedent may be *virtual* in the sense that it may not be part of the elements observed in the section.

For all these models, we have considered different transformations to formalise the latent relations between an element and its antecedent, and we showed that, within the PGLR framework, these transformations have to satisfy some properties. On this basis, we have designed a number of spaces that can be used to describe relations between chords, rhythms and melodies.

A number of combinations of spaces and models have been tested on a corpus of pop-music (symbolic data) to evaluate the interest of the multi-scale models. Performance has been measured using the cross-perplexity, i.e. the ability of a model to predict unseen sections after a training step, on other sequences.

These experiments have systematically shown an advantage of the multi-scale models over the sequential approach, with a particularly strong benefit for the modelling of rhythmic and chord sequences. If for rhythm and melody, the introduction of a virtual element for predicting the contrast does not seem to have a strong advantage, its effectiveness on chord sequence prediction seems more convincing.

As a consequence, this study has provided strong arguments to support that the PGLR extension of the S&C model displays interesting properties for music structure description.

7.2 Extensions and perspectives

7.2.1 Improving the model

This thesis has provided and experimented formalisms that can be very useful for music structure modelling. The proposed scheme results from the combination of two aspects of music that are related but may be considered separately. The first one is to model the structure of the dependency relations between the elements of a section, and this resulted in our case in the PGLR. The second aspect is the formalism effectively used to describe the relations (or equivalently the transformations) between the elements.

On this second aspect, some solutions have been designed by considering spaces of relations which form some sort of external algebra (i.e. such that, for any pair of elements, it is possible to compute the relation between the two and apply it to any other element of the section). We have proposed and/or designed such spaces of relations for harmonic, rhythmic and melodic elements, which can be used in this framework.

But there is still a lot to improve along these two aspects. First, the only type of sections considered in this work are "regular" sections of 16 elements, where each group of four elements may form a S&C. Therefore, there is still some work needed to generalise the implementation of the PGLR framework to sections of any size. One promising track is the approach adopted by Guichaoua [Guichaoua, 2017], where sections of various sizes are modelled as deformations of regular sections by removing or adding auxiliary elements, at the expense of some cost which is included in the MDL criteria and integrated in the optimisation process. In this case, the optimal PGLR (which is no more a n-cube) is a by-product of the structure modelling process.

A complementary track would be to include explicitly the possibility of triangular polytopes (and other predetermined structures) for accommodating ternary constructions, or others.

This means that it is necessary to extend the S&C formalism such as illustrated in Figure 7.1 for sections of size, 3, 6 and 5 elements. Moreover, increasing the number of possible S&C structures also increases the combinatorial complexity of the problem, which results in a higher encoding cost of the PGLR model as well as a longer computation time.

As regards now the definition of adequate spaces for describing relations between musical elements, a lot can be done to improve and generalise the work that has been presented here. Not only can one imagine many variants of the formalisms of relations between chords, rhythmic patterns and melodic motifs, but there are way more properties that would be useful in order to describe the multiple musical dimensions that may play a part in the structure of the section. For instance, the instrument playing, the dynamics, the local metric (which would allow to have basic elements of different size), but also properties that may be implicit.

Tonality is one of them and, indeed, once the listener hears the beginning of a section where all elements belong to the same tonality, the most logical outcome would be that the last part of the section be in the same tonality, resulting in the fact that notes and

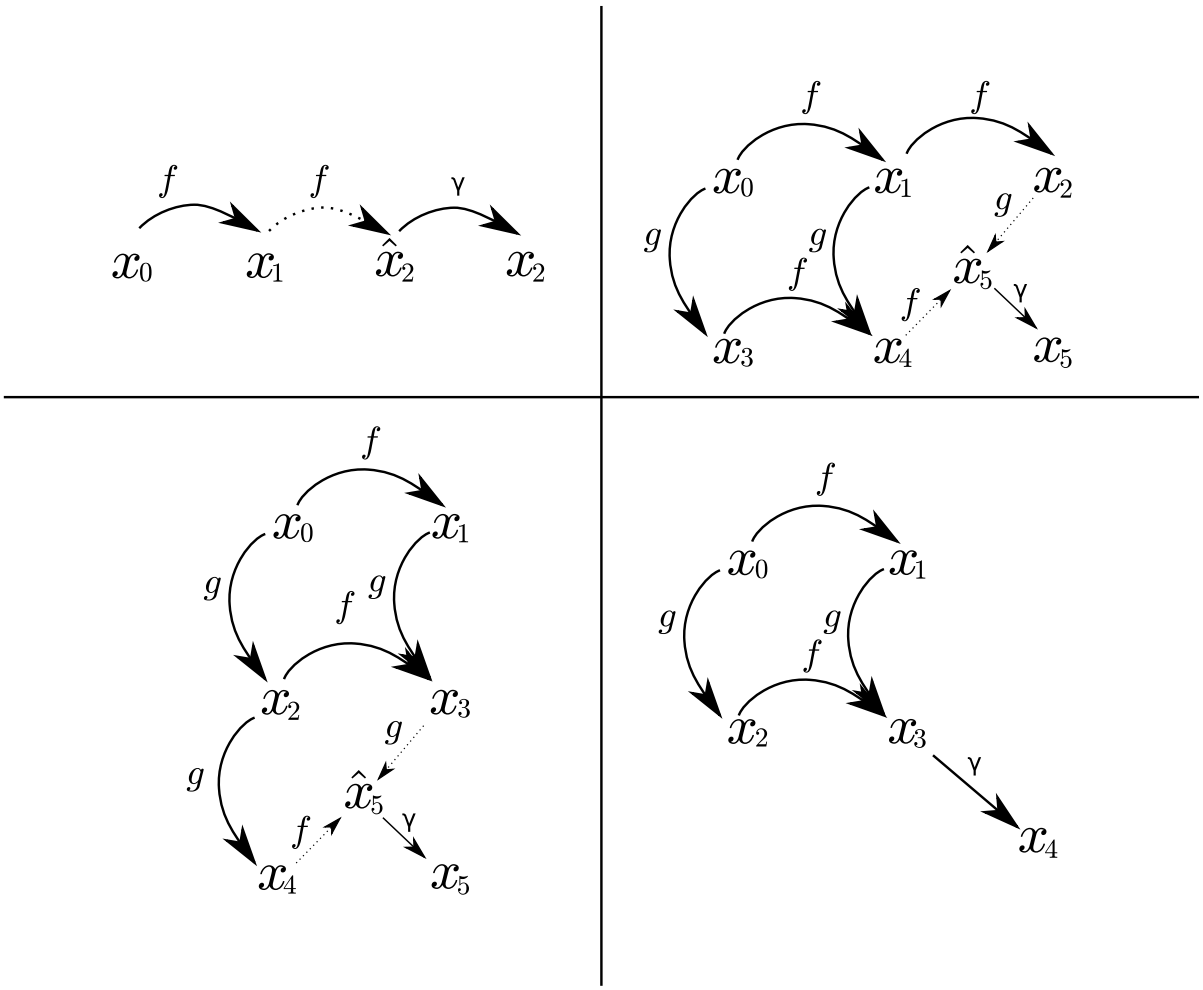


Figure 7.1: A few possible formalisms of S&Cs for non-square sections.

chords that do not belong to this tonality create a strong contrast and mark a radical change in the harmony (which indeed sometimes happens !)

Coming back to the musical dimensions studied in this thesis, there is still a lot to do, as for chords, it could be interesting to develop relation formalisms that incorporate musicological rules such as diatonic displacement instead of simple chromatic displacement. Another promising track for rhythm, would be to find a relation space which has the right properties (such as rhythm circles) but which does not requires to sample the rhythmic patterns at fixed instants (which, in our case, creates problems for triplets, for instance). But, as explained in Section 6.1, there is– for rhythm– no obvious circularity of the properties like the one of the pitch class space. The only solution presented in this work for rhythm relation description results in a compression scheme which is not lossless, i.e. it is not always possible to reconstruct the exact rhythmic section using its PGLR description. There may therefore be a need there, to improve the relation formalism, in order to make it lossless.

Finally, all experiments presented in this study focus on a unique *stream* of data at a time. Even melody, which involves two dimensions in our approach (rhythm and pitch), forms in fact a single stream of data (sequences of note onsets and pitches). Modelling

the multi-stream aspect of music is essential though, to account for the dependencies that exist between, for instance, tonality, chords and melody, which has not been addressed in the present work.

In a musical section, the parallel structures of the melodic stream and of the chord progression may differ at some points but they are still related. Therefore, to improve the method that has been presented in this work, one may focus on the combination of the relations formalisms to handle several dimensions *at the same time*. This could also be very useful for drum structure description where the kick, the snare, the hit-hat and the other instruments may each be represented using a separate rhythmic dimension but may be combined to form a unique description of the musical section.

7.2.2 Music Cognition

Assuming that the S&C model and its PGLR implementations are able to catch some structural information inside a section, it then appears as a very interesting approach to analyse and understand some aspects of music construction. In fact, the present work is strongly related to music cognition and music perception, in the sense that it models musical contents in relation to some expectation, which may (or may not) play a role in the musical experience of listeners.

Such an investigation of the perception of listeners in front of a musical section may be useful to find a clear or formal characterisation of the actual surprise in a musical section. In fact, considering the contrast as the surprise in the section, it can be fully described as the relation between the expected element and the heard one. Therefore, having a perception evaluation of this surprise can help define and refine measures to quantify "surprise". The interest of having a measure able to quantify the surprise in a section is that it can relate musical sections that are totally different in terms of composition principles by their structure and their degree of complexity.

More generally speaking, our work confirms that music can be viewed as the sequential presentation of a multi-scale content, which is structured by the distribution of surprises at specific instants. Extensive perceptual tests (with a proper methodology) would be extremely useful to further establish this conception. However, a major difficulty to anticipate in such a study is the entanglement of two types of surprises in music: the "cognisable" surprise (i.e. the surprise created by the deviation of an element from its purely logical induction), and the "cultural" surprise (namely the observation of an unconventional event in a given context). This duality of the concept of surprise is discussed in more details in [Bimbot et al., 2016], in relation with different types of information, and remains to date an open question.

7.2.3 Music Generation

Finally, while most of the work presented in this thesis is focused on theoretical aspects of the System & Contrast model and experimental validations, we briefly investigated some simple and direct application of the PGLR model. Considering that we have at hand,

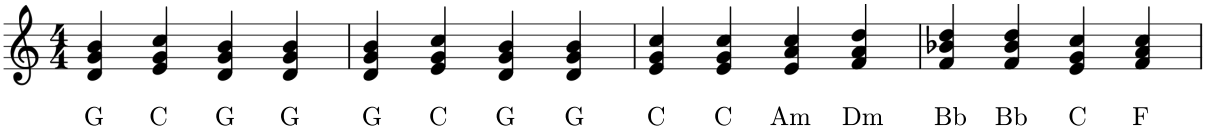


Figure 7.2: Example of a generated sequence by $S\&C_3$ model.

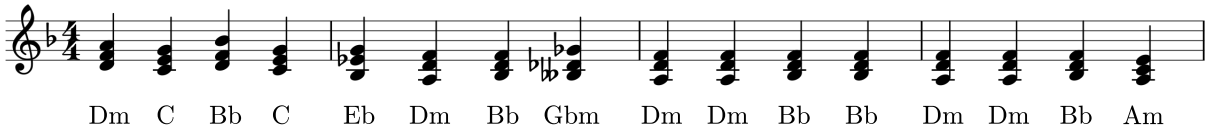


Figure 7.3: A second example of a generated sequence by $S\&C_3$ model.

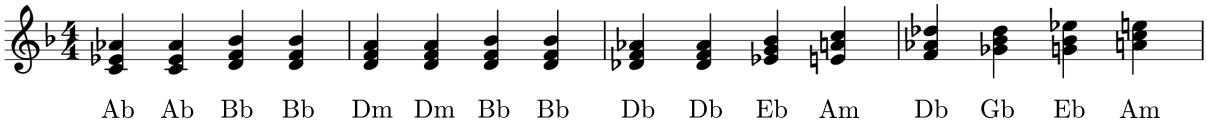


Figure 7.4: A third example of generated sequence by $S\&C_3$ model.

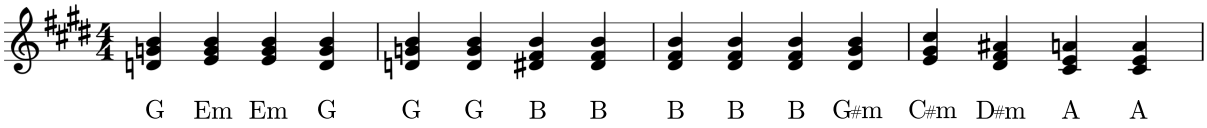


Figure 7.5: Example of a generated sequence by *Seq* model.

from our experiments, some estimations of the probability distributions for the relations between chords, these can be used for music generation, using a multi-scale framework. In fact, given a graph of dependencies, each associated relation can be randomly generated using the distribution that has been computed for that PGLR to describe the structure of the sections of RWC POP.

Figure 7.2, Figure 7.3, and Figure 7.4 show some examples of chord sequences obtained with the prevalent static S&C configuration ($S\&C_3$) using the trained distribution of rotations over the circle of triads. Here, only the model, $S\&C_3$ and the very first chord were chosen, all other chords were obtained by randomly assigning relations on the edges of the PGLR (following the trained distribution). In comparison, Figure 7.5 shows a chord sequence generated by the sequential model.

These are only examples, and the following comments are therefore informal. But a first observation that can be made is that the sequence generated by the *Seq* model and the sequences obtained using the $S\&C_3$ model are very different in term of structure. And despite some tonal peculiarities (in all sequences), the organisation of the multi-scale examples may seem more familiar in their organisation. For the sequence generated by *Seq*, changes of chords seem more erratic and, unlike $S\&C_3$-based sequences, it is quite difficult to subdivide it into consistent lower-scale groups of chords.

In fact, considering the first example (Figure 7.2), one can easily identify a structural pattern *aabc* as the system formed by the last four chords is highly contrastive due to the occurrence of the *Bb* chord and something that could be interpreted as a perfect cadence

in F or a half cadence in Bb (even though this is pure chance, as it is not built in the model). The modulation may be a bit abrupt as all chords beforehand were in C key and Bb is the only chord in the full sequence that is not in C. Still, such a sequence could be considered as a reasonable basis for a composition. Indeed, as the Dm chord belongs both to C, F and Bb keys, it can be viewed (a posteriori) as a passage chord and it makes the musical flow rather fluid. In fact, even though these musicological considerations were not "injected" in the model, we believe that the strong structure of the sequence which has been generated does help supporting these impressions.

The second sequence (Figure 7.3) feels a bit more unusual and appears as a structure which could be described by the pattern $abcc'$ which is not that common in music. However, except for the Gbm chord at the end of the second bar, which sounds really off-tonality, the rest of the chord flow sounds plausible (with some medieval-like ending at the end...) The third example (Figure 7.4) has a form which can be seen as $aba'c$ where a' is a transposition of a with a small contrast on its second half. This is a very conventional pattern. Yet, because of the absence of a tonal model, the chords sequence tends to ramble in different tonalities in an erratic way. In summary, the chords sequences generated by the $S\&C$ show a priori a good structural coherence as they exhibit, by construction, well-identifiable structural patterns at different scales, which is not the case for the sequential model.

However, both models are facing the problem of tonal stability. In fact, given the generation process, some generated chords that occur can very well be out of key. This problem is well illustrated by the sequence on Figure 7.4 where, in the same section, all pitch classes except B are present. This problem stems from the fact that models are first-order models and therefore have no mechanism to constrain tonality. As for chord relation description, it could be interesting to add such musical constraint in order to improve the results in generation and get sections that are therefore more consistent. More generally, an interesting move would be to couple the chord model with a tonality model (which could itself be sequential... or systemic...)

7.2.4 Broadening the scope

In the work presented in this thesis, the musical dimensions— chords, rhythm and melody— have been studied separately. A next step of this study would certainly be to combine these dimensions into a single model. One way to address this problem could be to find ways to inter-operate the PGLR with Conklin et al.'s Multiple Viewpoints model [Conklin and Witten, 1995] or the Information Dynamics of Music (IDyOM) approach designed by Pearce et al. [Pearce and Wiggins, 2012, Sears et al., 2018].

In fact, these approaches consider a stream of event as a stream of object having multiple features called *viewpoints*: chromatic pitch, chromatic pitch interval, onset, duration, inter-onset interval, scale degree, mode, tonic pitch, or even, scale degree thread first in bar. All these viewpoints can then be used to improve the prediction of the next musical object. One may use such descriptions of musical objects to consider multiple dimensions at the same time. For example, one may consider the melodic dimension as a combination of two viewpoints: chromatic pitch class interval and rhythm pattern.

However, to be able to use the multiple viewpoint approach along with the PGLR paradigm, it is necessary to design for each viewpoint a relation formalism which would have suitable properties as defined in Section 3.3.2, which may require some inventiveness[1].

A second characteristic of Information Dynamics in Music as developed by Abdallah et al. [Abdallah and Plumbley, 2009] or Pearce et al. [Pearce and Wiggins, 2012] is that expectation in the music flow is modelled on a sequential and statistical basis (typically n-grams). PGLRs, as developed in this thesis, consider an interesting alternative paradigm, where dependencies between musical units within sections are assumed to be primarily governed by multi-scale relations (which are therefore not necessarily contiguous) and where its inner structure is based on its internal relations of analogy.

So it makes it very tempting to investigate how these two conceptions could be merged in an inter-operable framework, in which statistical information and description complexity could cooperate to a robust modelling of information dynamics, which would jointly integrate statistical *and* complexity criteria to describe the evolution and the organisation of the musical flow. For a lack of time, this has not been addressed within the scope of the present work, but it appears as one of the most exciting perspective that could follow.

7.2.5 Final Words

One fascinating aspect of music is certainly the incredible number of ways to approach it, across so many different facets. Hopefully, this journey throughout the multi-scale and multi-dimensional universe of music structure, will have taken readers of this work into a new landscape which will enrich and enhance their own approach to music.

Talking of my own experience, I now listen to music with much more open-mindedness, and with a better ability to understand pieces that I may not have instinctively liked a few years ago.

Beyond its theoretical contributions and its experimental results, a great achievement of this work would be not only to inspire colleagues and students for future scientific work in music data processing and information retrieval but also to create interest with musicologists and composers, who may find new tools and sources of inspiration from this emerging framework.

[1] As an alternative, the PGLR approach could be used in its systemic version to start with, leaving temporarily out the contrastive part of the model.

Remerciements

Une thèse n'est pas une chose simple à réaliser et il est de nombreuses personnes qui m'ont apporté encouragements et soutient dans cette dure épreuve. Leur présence à mes côtés est sans nul doute ce qui m'a permis de surpasser les obstacles que j'ai pu rencontrer pour finalement soutenir. Il m'apparaît alors à la fois évident et important de les en remercier.

La personne par qui je me dois de commencer est bien évidemment mon directeur de thèse, Frédéric, pour avoir réussi à me motiver à entreprendre un chantier aussi immense, pour m'avoir guidé tout au long de cette épreuve et avoir cru en moi. J'ai énormément appris à ses côtés et ai été ravis de travailler et discuter en sa compagnie. Tout cela n'aurait bien sûr pas pu être possible sans Anna, qui a accepté que je m'accapare le temps précieux de Frédéric et qui a compensé les mauvais jours que je lui faisais passer.

Mais plus qu'un directeur, je tiens particulièrement à remercier toutes les personnes qui ont contribué à développer un environnement de travail saint et plaisant dans lequel je puisse m'épanouir. Cela comprend notamment Stéphanie et Armelle pour l'aide, administrative mais pas que, et tous les beaux voyages qu'elles m'ont organisés. Mais aussi tous les membres des équipes PANAMA et TEA avec lesquels j'ai eu la chance de partager cafés, galettes des rois, raclette, bières et autres joyeusetés culinaires. Plus particulièrement, je tenais à remercier toutes les personnes qui ont eu pour charge de me supporter dans leur bureau : Emmanuel, Jeremy, Alexandre, Nicolas, Maxime, Cassio, Hymalaya, Mohammed, Xavier, Victor, Maël. Je tiens à m'excuser auprès de tous ceux qui m'ont vu zomber dans leur bureau quand je m'ennuyais et les remercie pour leur accueil.

Ma thèse ne s'étant pas uniquement concentrée sur ce laboratoire rennais qu'est l'IRISA, il me faut remercier les personnes avec qui j'ai eu l'occasion de collaborer à commencer par Mathieu Giraud et Matthew Davies ansi que leurs deux équipes respectives.

En outre, je souhaite remercier toutes les personne qui ont travaillé avant et avec moi sur ce sujet qu'est la structure musicale : Emmanuel Deruty, Gabriel Sargent, Corentin Guichaoua, Maxime Lecoq, Valentin Guillot, Nathan Libermann, Tudor... Vos points de vue et les discussions que l'on a pu avoir m'ont énormément apporté.

Et pour conclure sur le plan professionnel, je tiens à remercier tous les membres du jury de ma thèse pour avoir accepté de lire ce manuscrit et de venir à Rennes pour m'écouter présenter mes travaux à une heure bien matinale.

Mais une thèse, ce n'est pas quelque chose qui se cantonne à la vie professionnelle, aussi il me faut remercier toutes les personnes qui m'ont soutenu pendant ces trois ans

et demi. Merci papa, merci maman, et merci les frangins, la frangine, le chat, merci pour votre soutient. Et merci à tout le reste de la grande famille, et tout particulièrement à Loane qui m'a appris à prendre des instants de détente sur diverses plages du monde.

Au delà de la famille, il y a les amis, je n'aime pas du tout les classer, aussi, comme tous ont joué une part importante dans la maîtrise de ma psychologie et ce au travers de longues conversations, vais-je faire une liste tout aussi longue. Un immense merci à Aurel, Ben, Grégoire, amis de longue dates, mais aussi à tous ceux que j'ai pu rencontrer plus récemment : Maria, Maëlle, Estelle, Yann, Nico, Florian, Quentin, Valou et de manière générale à toutes les personnes que j'ai côtoyé dans le monde de la musique, principalement les membres de mes groupes : Nicolle, Old Future, Acid Drop, Funky-T, Improvidenz et Sairen. Merci aussi à toutes les autres personnes que j'ai rencontrées dans la musique qui m'ont apporté des points de vue intéressants sur le sujet ainsi qu'une curiosité motivante.

La liste d'amis est longue car ils sont nombreux à m'avoir encouragé, je finirais par remercier un trio que j'ai eu la chance de rencontrer pendant ma thèse, merci à Cindy, Chris et Anta pour leur soutient et les belles tranches de rire qui m'ont toujours redonné la motivation de finir ma thèse.

Et enfin, merci à ma volonté d'avoir tenu le coup !

Me connaissant, il est fort à parier que des personnes aient été oubliées, et ce malgré mes efforts pour que cela n'arrive pas. Milles excuses à ces personnes qui ne méritent pas moins mes remerciements.

Bibliography

[Abdallah and Plumbley, 2009] Abdallah, S. and Plumbley, M. (2009). Information dynamics: patterns of expectation and surprise in the perception of music. *Connection Science*, 21(2-3):89–117.

[Agres et al., 2018] Agres, K., Abdallah, S., and Pearce, M. (2018). Information-theoretic properties of auditory sequences dynamically influence expectation and memory. *Cognitive science*, 42(1):43–76.

[Amiot, 2013] Amiot, E. (2013). *The Torii of Phases*, pages 1–18. Springer, Mathematics and Computation in Music: 4th International Conference, MCM 2013, Montreal, QC, Canada, June 12-14, 2013. Proceedings, Berlin, Heidelberg.

[Bharucha, 1987] Bharucha, J. J. (1987). Music cognition and perceptual facilitation: A connectionist framework. *Music Perception: An Interdisciplinary Journal*, 5(1):1–30.

[Bigand and Parncutt, 1999] Bigand, E. and Parncutt, R. (1999). Perceiving musical tension in long chord sequences. *Psychological Research*, 62(4):237–254.

[Bimbot et al., 2012a] Bimbot, F., Deruty, E., Sargent, G., and Vincent, E. (2012a). Semiotic structure labeling of music pieces: concepts, methods and annotation conventions. In *Proc. ISMIR*.

[Bimbot et al., 2012b] Bimbot, F., Deruty, E., Sargent, G., and Vincent, E. (2012b). System & Contrast : a Polymorphous Model of the Inner Organization of Structural Segments within Music Pieces (Original Extensive Version). Research Report IRISA PI-1999, IRISA.

[Bimbot et al., 2016] Bimbot, F., Deruty, E., Sargent, G., and Vincent, E. (2016). System & Contrast : A Polymorphous Model of the Inner Organization of Structural Segments within Music Pieces. *Music Perception*, 33:631–661. Former version published in 2012 as Research Report IRISA PI-1999, hal-01188244.

[Bimbot et al., 2010] Bimbot, F., Le Blouch, O., Sargent, G., and Vincent, E. (2010). Decomposition into autonomous and comparable blocks: A structural description of music pieces. In *International Society for Music Information Retrieval Conference (ISMIR)*, Utrecht, Netherlands.

[Bittner et al., 2014] Bittner, R. M., Salamon, J., Tierney, M., Mauch, M., Cannam, C., and Bello, J. P. (2014). Medleydb: A multitrack dataset for annotation-intensive mir research. In *ISMIR*, volume 14, pages 155–160.

[Bod, 2002] Bod, R. (2002). Memory-based models of melodic analysis: Challenging the gestalt principles. *Journal of New Music Research*, 31(1):27–36.

[Bresson et al., 2010] Bresson, J., Agon, C., and Assayag, G. (2010). Openmusic–visual programming environment for music composition, analysis and research. In *ACM MultiMedia (MM'11)*.

[Brown et al., 1992] Brown, P. F., Pietra, V. J. D., Mercer, R. L., Pietra, S. A. D., and Lai, J. C. (1992). An estimate of an upper bound for the entropy of english. *Computational Linguistics*, 18(1):31–40.

[Calvo-Zaragoza et al., 2016] Calvo-Zaragoza, J., Vigliensoni, G., and Fujinaga, I. (2016). Document analysis for music scores via machine learning. In *Proceedings of the 3rd International workshop on Digital Libraries for Musicology*, pages 37–40. ACM.

[Cambouropoulos, 2001] Cambouropoulos, E. (2001). The local boundary detection model (lbdm) and its application in the study of expressive timing. In *ICMC*.

[Cambouropoulos, 2006] Cambouropoulos, E. (2006). Musical parallelism and melodic segmentation:: A computational approach. *Music Perception*, 23(3):249–268.

[Caplin, 1998] Caplin, W. E. (1998). *Classical form: A theory of formal functions for the instrumental music of Haydn, Mozart, and Beethoven*. Oxford University Press.

[Celma, 2010] Celma, O. (2010). Music recommendation. In *Music recommendation and discovery*, pages 43–85. Springer.

[Chaitin, 1966] Chaitin, G. J. (1966). On the length of programs for computing finite binary sequences. *Journal of the ACM (JACM)*, 13(4):547–569.

[Chang et al., 2004] Chang, C., Jiau, H. C., et al. (2004). Representative music fragments extraction by using segmentation techniques. In *Proc. of International Computer Symposium*, pages 1156–1161.

[Cho, 2011] Cho, T. (2011). Manually annotated chord data set of us pop songs and popular music collection of rwc music database. url: https://github. com/tmc323. *Chord-Annotations (cf. page 41)*.

[Clercx, 1935] Clercx, S. (1935). La forme du rondo chez carl philipp emanuel bach. *Revue de Musicologie*, 16(55):148–167.

[Cohn, 2011] Cohn, R. (2011). *Audacious Euphony: Chromatic Harmony and the Triad's Second Nature*. Oxford University Press.

[Conklin, 2003] Conklin, D. (2003). Music generation from statistical models. In *Proceedings of the AISB 2003 Symposium on Artificial Intelligence and Creativity in the Arts and Sciences*, pages 30–35.

[Conklin, 2013] Conklin, D. (2013). Multiple viewpoint systems for music classification. *Journal of New Music Research*, 42(1):19–26.

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

[de Haas et al., 2011] de Haas, B., Magalhaes, J. P., Veltkamp, R. C., and Wiering, F. (2011). Harmtrace: Improving harmonic similarity estimation using functional harmony analysis. In *ISMIR*, pages 67–72.

[de Haas et al., 2009] de Haas, B., Rohrmeier, M., Veltkamp, R. C., and Wiering, F. (2009). Modeling harmonic similarity using a generative grammar of tonal harmony. In *ISMIR*, pages 549–554.

[De Haas et al., 2013] De Haas, W., Magalhães, J., Wiering, F., and C Veltkamp, R. (2013). Automatic functional harmonic analysis. *Computer Music Journal*, 37(4):37–53.

[De Haas et al., 2009] De Haas, W. B., Rohrmeier, M., Veltkamp, R. C., and Wiering, F. (2009). Modeling harmonic similarity using a generative grammar of tonal harmony. *Proc. ISMIR*.

[Déguernel et al., 2017] Déguernel, K., Nika, J., Vincent, E., and Assayag, G. (2017). Generating equivalent chord progressions to enrich guided improvisation: application to rhythm changes. In *SMC 2017-14th Sound and Music Computing Conference*, page 8.

[Deruty et al., 2013] Deruty, E., Bimbot, F., and Van Wymeersch, B. (2013). Methodological and musicological investigation of the System & Contrast model for musical form description. Research Report RR-8510, INRIA. hal-00965914.

[Duerksen, 1972] Duerksen, G. L. (1972). Some effects of expectation on evaluation of recorded musical performance. *Journal of Research in Music Education*, 20(2):268–272.

[Durand and Essid, 2016] Durand, S. and Essid, S. (2016). Downbeat detection with conditional random fields and deep learned features. In *ISMIR*, pages 386–392.

[Eerola and Toiviainen, 2004] Eerola, T. and Toiviainen, P. (2004). Midi toolbox: Matlab tools for music research.

[Farbood, 2006] Farbood, M. M. (2006). *A quantitative, parametric model of musical tension*. PhD thesis, Massachusetts Institute of Technology.

[Ferradans et al., 2013] Ferradans, S., Papadakis, N., Rabin, J., Peyré, G., and Aujol, J.-F. (2013). Regularized discrete optimal transport. In *International Conference on Scale Space and Variational Methods in Computer Vision*, pages 428–439. Springer.

[Ferrand et al., 2003] Ferrand, M., Nelson, P., and Wiggins, G. (2003). Memory and melodic density: a model for melody segmentation. In *Proceedings of the XIV Colloquium on Musical Informatics (XIV CIM 2003)*, pages 95–98.

[Flexer and Stevens, 2018] Flexer, A. and Stevens, J. (2018). Mutual proximity graphs for improved reachability in music recommendation. *Journal of new music research*, 47(1):17–28.

[Forth, 2012] Forth, J. (2012). *Cognitively-motivated geometric methods of pattern discovery and models of similarity in music*. PhD thesis, Goldsmiths, University of London.

[Forth et al., 2008] Forth, J., McLean, A., Wiggins, G., et al. (2008). Musical creativity on the conceptual level. In *Fifth International Joint Workshop on Computational Creativity, Ciudad Universitaria, Facultad de Informatica, Madrid, Spain*, pages 17–19.

[Frank, 1953] Frank, G. (1953). Pulse code communication. US Patent 2,632,058.

[Gaultier et al., 2017] Gaultier, C., Kitić, S., Bertin, N., and Gribonval, R. (2017). AUDASCITY: AUdio Denoising by Acaptive Social CoscarsITY. In *2017 25th European Signal Processing Conference (EUSIPCO)*, pages 1265–1269. IEEE.

[Giraud et al., 2013] Giraud, M., Groult, R., and Levé, F. (2013). Subject and countersubject detection for analysis of the well-tempered clavier fugues. In *From Sounds to Music and Emotions*, pages 422–438. Springer.

[Goto, 2006] Goto, M. (2006). Aist annotation for the rwc music database. In *ISMIR*, pages 359–360.

[Goto et al., 2002] Goto, M., Hashiguchi, H., Nishimura, T., and Oka, R. (2002). RWC Music Database: Popular, Classical and Jazz Music Databases. In *ISMIR*, volume 2, pages 287–288.

[Grill and Schlüter, 2015] Grill, T. and Schlüter, J. (2015). Music boundary detection using neural networks on combined features and two-level annotations. In *ISMIR*, pages 531–537.

[Guichaoua, 2017] Guichaoua, C. (2017). *Modèles de compression et critères de complexité pour la description et l'inférence de structure musicale*. PhD thesis, Université Rennes 1.

[Guichaoua and Bimbot, 2018] Guichaoua, C. and Bimbot, F. (2018). Inférence de segmentation structurelle par compression via des relations multi-échelles dans les séquences d'accords. In *Journées d'Informatique Musicale (JIM 2018)*.

[Hepokoski and Darcy, 2006] Hepokoski, J. and Darcy, W. (2006). *Elements of sonata theory: Norms, types, and deformations in the late-eighteenth-century sonata*. Oxford University Press.

[Homburg et al., 2005] Homburg, H., Mierswa, I., Möller, B., Morik, K., and Wurst, M. (2005). A benchmark dataset for audio classification and clustering. In *ISMIR*, volume 2005, pages 528–31.

[Hu et al., 2003] Hu, N., Dannenberg, R. B., and Tzanetakis, G. (2003). Polyphonic audio matching and alignment for music retrieval. In *Applications of Signal Processing to Audio and Acoustics, 2003 IEEE Workshop on.*, pages 185–188. IEEE.

[Huron, 2006] Huron, D. B. (2006). *Sweet anticipation: Music and the psychology of expectation*. MIT press.

[Jelinek et al., 1977] Jelinek, F., Mercer, R. L., Bahl, L. R., and Baker, J. K. (1977). Perplexity—a measure of the difficulty of speech recognition tasks. *The Journal of the Acoustical Society of America*, 62(S1):S63–S63.

[Juhász, 2004] Juhász, Z. (2004). Segmentation of hungarian folk songs using an entropy-based learning system. *Journal of New Music Research*, 33(1):5–15.

[Jusczyk and Krumhansl, 1993] Jusczyk, P. W. and Krumhansl, C. L. (1993). Pitch and rhythmic patterns affecting infants' sensitivity to musical phrase structure. *Journal of experimental psychology: Human perception and performance*, 19(3):627.

[Kantorovitch, 1958] Kantorovitch, L. (1958). On the translocation of masses. *Management Science*, 5(1):1–4.

[Kolmogorov, 1965] Kolmogorov, A. N. (1965). Three approaches to the quantitative definition ofinformation'. *Problems of information transmission*, 1(1):1–7.

[Lattner et al., 2017] Lattner, S., Grachten, M., and Widmer, G. (2017). Learning transformations of musical material using gated autoencoders. In *Proceedings of the 2nd conference on computer simulation of musical creativity, CSMC*, pages 11–13.

[Lerdahl and Jackendoff, 1985] Lerdahl, F. and Jackendoff, R. S. (1985). *A generative theory of tonal music*. MIT press.

[Lévy, 2004] Lévy, F. (2004). *Complexité grammatologique et complexité aperceptive en musique: étude esthétique et scientifique du décalage entre la pensée de l'écriture et la perception cognitive des processus musicaux sous l'angle des théories de l'information et de la complexité*. PhD thesis, Paris, EHESS.

[Lewin, 1998] Lewin, D. (1998). Some ideas about voice-leading between pcsets. *Journal of Music Theory*, 42(1):15–72.

[Li et al., 2010] Li, T. L., Chan, A. B., and Chun, A. (2010). Automatic musical pattern feature extraction using convolutional neural network. In *Proc. Int. Conf. Data Mining and Applications*. sn.

[Lin and Zhang, 2018] Lin, J. and Zhang, B. (2018). A music retrieval method based on hidden markov model. In *Intelligent Transportation, Big Data & Smart City (ICITBS), 2018 International Conference on*, pages 732–735. IEEE.

[London, 2012] London, J. (2012). *Hearing in time: psychological aspects of musical metre*. Oxford University Press, Oxford, UK, 2nd edition.

[Louboutin and Bimbot, 2016] Louboutin, C. and Bimbot, F. (2016). Description of Chord Progressions by Minimal Transport Graphs Using the System & Contrast Model. In *ICMC 2016 - 42nd International Computer Music Conference*, Utrecht, Netherlands.

[Louboutin and Bimbot, 2017a] Louboutin, C. and Bimbot, F. (2017a). Modeling the multiscale structure of chord sequences using polytopic graphs. In *18th International Society for Music Information Retrieval Conference*, Suzhou, China.

[Louboutin and Bimbot, 2017b] Louboutin, C. and Bimbot, F. (2017b). Polytopic Graph of Latent Relations: A Multiscale Structure Model for Music Segments. In Agustín-Aquino, O. A., Lluis-Puebla, E., and Montiel, M., editors, *6th International Conference on Mathematics and Computation in Music (MCM 2017)*, volume 10527 of *Lecture Notes in Computer Science book series*, Mexico City, Mexico. Springer.

[Louboutin and Meredith, 2016] Louboutin, C. and Meredith, D. (2016). Using general-purpose compression algorithms for music analysis. *Journal of New Music Research*.

[Loy, 2017] Loy, D. G. (2017). Music, expectation, and information theory. In *The Musical-Mathematical Mind*, pages 161–169. Springer.

[Maddage et al., 2004] Maddage, N. C., Xu, C., Kankanhalli, M. S., and Shao, X. (2004). Content-based music structure analysis with applications to music semantics understanding. In *Proceedings of the 12th annual ACM international conference on Multimedia*, pages 112–119. ACM.

[Mavromatis, 2009] Mavromatis, P. (2009). Minimum description length modelling of musical structure. *Journal of Mathematics and Music*, 3(3):117–136.

[McKay et al., 2010] McKay, C., Burgoyne, J. A., Hockman, J., Smith, J. B., Vigliensoni, G., and Fujinaga, I. (2010). Evaluating the genre classification performance of lyrical features relative to audio, symbolic and cultural features. In *ISMIR*, pages 213–218.

[Meredith, 2012a] Meredith, D. (2012a). Music analysis and kolmogorov complexity. *Proceedings of the 19th Colloquio d'Informatica Musicale (XIX CIM)*.

[Meredith, 2012b] Meredith, D. (2012b). Music analysis and Kolmogorov complexity. *Proceedings of the 19th Colloquio d'Informatica Musicale (XIX CIM)*.

[Meredith, 2013a] Meredith, D. (2013a). Analysis by compression: Automatic generation of compact geometric encodings of musical objects. In *The Music Encoding Conference (MEC 2013)*.

[Meredith, 2013b] Meredith, D. (2013b). COSIATEC and SIATECCompress: Pattern discovery by geometric compression. In *International Society for Music Information Retrieval Conference*.

[Meredith, 2018] Meredith, D. (2018). Music analysis and data compression. In *Oxford Handbook of Sound and Imagination*. Oxford University Press.

[Meyer, 1956] Meyer, L. B. (1956). *Emotion and meaning in music*. University of chicago Press.

[Meyer, 1957] Meyer, L. B. (1957). Meaning in music and information theory. *The Journal of Aesthetics and Art Criticism*, 15(4):412–424.

[Monelle, 2014] Monelle, R. (2014). *Linguistics and semiotics in music*. Routledge.

[Mongeau and Sankoff, 1990] Mongeau, M. and Sankoff, D. (1990). Comparison of musical sequences. *Computers and the Humanities*, 24(3):161–175.

[Nakamura et al., 2017] Nakamura, E., Yoshii, K., Dixon, S., Nakamura, E., Yoshii, K., and Dixon, S. (2017). Note value recognition for piano transcription using markov random fields. *IEEE/ACM Transactions on Audio, Speech and Language Processing (TASLP)*, 25(9):1846–1858.

[Narmour, 1989] Narmour, E. (1989). The "genetic code" of melody: Cognitive structures generated by the implication-realization model. *Contemporary Music Review*, 4(1):45–63.

[Narmour, 1992] Narmour, E. (1992). *The analysis and cognition of melodic complexity: The implication-realization model*. University of Chicago Press.

[Narmour, 2000] Narmour, E. (2000). Music expectation by cognitive rule-mapping. *Music Perception: An Interdisciplinary Journal*, 17(3):329–398.

[Nika et al., 2016] Nika, J., Chemillier, M., and Assayag, G. (2016). Improtek: introducing scenarios into human-computer music improvisation. *Computers in Entertainment (CIE)*, 14(2):4.

[Noland and Sandler, 2006] Noland, K. C. and Sandler, M. B. (2006). Key estimation using a hidden markov model. In *ISMIR*, pages 121–126.

[Ockelford, 2006] Ockelford, A. (2006). Implication and expectation in music: A zygonic model. *Psychology of Music*, 34(1):81–142.

[Pachet et al., 2017] Pachet, F., Papadopoulos, A., and Roy, P. (2017). Sampling variations of sequences for structured music generation. In *Proceedings of the 18th International Society for Music Information Retrieval Conference (ISMIR'2017), Suzhou, China*, pages 167–173.

[Pearce, 2005] Pearce, M. T. (2005). *The construction and evaluation of statistical models of melodic structure in music perception and composition*. City University London.

[Pearce and Rohrmeier, 2018] Pearce, M. T. and Rohrmeier, M. (2018). Musical syntax ii: empirical perspectives. In *Springer Handbook of Systematic Musicology*, pages 487–505. Springer.

[Pearce and Wiggins, 2012] Pearce, M. T. and Wiggins, G. A. (2012). Auditory expectation: the information dynamics of music perception and cognition. *Topics in cognitive science*, 4(4):625–652.

[Peeters and Bisot, 2014] Peeters, G. and Bisot, V. (2014). Improving music structure segmentation using lag-priors. In *ISMIR*, pages 337–342.

[Peeters and Deruty, 2009] Peeters, G. and Deruty, E. (2009). Is music structure annotation multi-dimensional? a proposal for robust local music annotation. In *Proc. of 3rd Workshop on Learning the Semantics of Audio Signals*, pages 75–90. Citeseer.

[Piston, 1948] Piston, W. (1948). *Harmony*. Norton.

[Raffel and Ellis, 2016] Raffel, C. and Ellis, D. P. (2016). Optimizing dtw-based audio-to-midi alignment and matching. In *Acoustics, Speech and Signal Processing (ICASSP), 2016 IEEE International Conference on*, pages 81–85. IEEE.

[Rohrmeier, 2007] Rohrmeier, M. (2007). A generative grammar approach to diatonic harmonic structure. In *Proceedings of the 4th Sound and Music Computing Conference*, pages 97–100.

[Rohrmeier, 2011] Rohrmeier, M. (2011). Towards a generative syntax of tonal harmony. *Journal of Mathematics and Music*, 5(1):35–53.

[Schmuckler, 1989] Schmuckler, M. A. (1989). Expectation in music: Investigation of melodic and harmonic processes. *Music Perception: An Interdisciplinary Journal*, 7(2):109–149.

[Schoenberg, 1983] Schoenberg, A. (1983). *Theory of harmony*. Univ of California Press.

[Schoenberg et al., 1967] Schoenberg, A., Stein, L., and Strang, G. (1967). *Fundamentals of musical composition*. Faber & Faber London.

[Sears et al., 2018] Sears, D. R., Pearce, M. T., Caplin, W. E., and McAdams, S. (2018). Simulating melodic and harmonic expectations for tonal cadences using probabilistic models. *Journal of New Music Research*, 47(1):29–52.

[Shannon, 2001] Shannon, C. E. (2001). A mathematical theory of communication. *ACM SIGMOBILE mobile computing and communications review*, 5(1):3–55.

[Shao et al., 2004] Shao, X., Xu, C., and Kankanhalli, M. S. (2004). Unsupervised classification of music genre using hidden markov model. In *ICME*, volume 4, pages 2023–2026. Citeseer.

[Sioros et al., 2018] Sioros, G., Davies, M. E., and Guedes, C. (2018). A generative model for the characterization of musical rhythms. *Journal of New Music Research*, 47(2):114–128.

[Sloboda, 1991] Sloboda, J. A. (1991). Music structure and emotional response: Some empirical findings. *Psychology of music*, 19(2):110–120.

[Smith and Chew, 2013] Smith, J. B. and Chew, E. (2013). Using quadratic programming to estimate feature relevance in structural analyses of music. In *Proceedings of the 21st ACM international conference on Multimedia*, pages 113–122. ACM.

[Snyder and Snyder, 2000] Snyder, B. and Snyder, R. (2000). *Music and memory: An introduction*. MIT press.

[Steedman, 1996] Steedman, M. (1996). The blues and the abstract truth: Music and mental models. *Mental models in cognitive science*, pages 305–318.

[Straus, 2003] Straus, J. N. (2003). Uniformity, balance, and smoothness in atonal voice leading. *Music Theory Spectrum*, 25(2):305–352.

[Temperley, 2004] Temperley, D. (2004). *The cognition of basic musical structures*. MIT press.

[Temperley, 2014] Temperley, D. (2014). Probabilistic models of melodic interval. *Music Perception*, 32(1):85–99.

[Tenney and Polansky, 1980] Tenney, J. and Polansky, L. (1980). Temporal gestalt perception in music. *Journal of Music Theory*, 24(2):205–241.

[Tillmann et al., 2014] Tillmann, B., Poulin-Charronnat, B., and Bigand, E. (2014). The role of expectation in music: from the score to emotions and the brain. *Wiley Interdisciplinary Reviews: Cognitive Science*, 5(1):105–113.

[Tymoczko, 2006] Tymoczko, D. (2006). The geometry of musical chords. *Science*, 313(5783):72–74.

[Tymoczko, 2008] Tymoczko, D. (2008). Scale theory, serial theory and voice leading. *Music Analysis*, 27(1):1–49.

[Van den Oord et al., 2013] Van den Oord, A., Dieleman, S., and Schrauwen, B. (2013). Deep content-based music recommendation. In *Advances in neural information processing systems*, pages 2643–2651.

[Velarde et al., 2018] Velarde, G., Cancino Chacón, C., Meredith, D., Weyde, T., and Grachten, M. (2018). Convolution-based classification of audio and symbolic representations of music. *Journal of New Music Research*, pages 1–15.

[Villani, 2003] Villani, C. (2003). *Topics in optimal transportation*. Number 58 in Graduate Studies in Mathematics. American Mathematical Soc.

[Vitányi and Li, 2000] Vitányi, P. M. and Li, M. (2000). Minimum description length induction, bayesianism, and Kolmogorov complexity. *Information Theory, IEEE Transactions on*, 46(2):446–464.

[Weitzmann and Saslaw, 2004] Weitzmann, C. F. and Saslaw, J. K. (2004). Two monographs by carl friedrich weitzmann: Part i:" the augmented triad"(1853). *Theory and Practice*, 29:133–228.

[Wiering et al., 2009] Wiering, F., de Nooijer, J., Volk, A., and Tabachneck-Schijf, H. J. (2009). Cognition-based segmentation for music information retrieval systems. *Journal of New Music Research*, 38(2):139–154.

[Ycart et al., 2016] Ycart, A., Jacquemard, F., Bresson, J., and Staworko, S. (2016). A supervised approach for rhythm transcription based on tree series enumeration. In *International Computer Music Conference (ICMC)*.