Tonal stability modeling from audio chroma features

Agustín Martorell Domínguez

MASTER THESIS UPF / 2009

Master in Sound and Music Computing

Master thesis supervisor:

Emilia Gómez Gutiérrez

Department of Information and Communication Technologies

Universitat Pompeu Fabra, Barcelona

UNIVERSITAT POMPEU FABRA

Acknowledgements

This work is the result of a yearlong interaction of people and circumstances, being impossible any attempt of acknowledge their mutual influences, and for sure there will be important missings here. Xavier Serra, for making possible this opportunity through Sound and Music Computing Master and for moving the wonderful project called Music Technology Group. Emilia Gómez, for her support and supervision, endless discussions and patience, technical and methodological assistance, revisions and friendship. Hendrik Purwins, for his inspiring mathematical discussions on music. Perfecto Herrera, for show me the way through perception and cognition research. SMC teachers Sergi Jordà, Günter Geiger, Enric Guaus, Gustavo Deco, and short-term collaborators Ariadna, Jordi, Pau, Ferran, Enric, Josep María, Merlijn, Maarten, Jordi and Óscar. All SMC classmates and PhD students, particularly to Tan, Anandhi, Rich and Sašo, and the people at Room 312. Special thanks to Mathieu's nonstop brainstoming and his interactive approach to life. All the busy but always helpful MTG people, department administration and library staff. Fred Lerdahl and Carol Krumhansl, for their true inspiration and helpful data sharing. All my family in the distance. All my friends, with particular thanks to Jorge, Laura and Òscar, for their nonconventional support. My musicology people, much particularly Ascensión, Juanjo, Verónica, Patri, Nano, Verónica, Samuel, Olga and Enrique, for their support, advice and musical moments. My music and musicology teachers Juan Carlos, Yvann, María, Conchita and Giuseppe. UPF choir and all the people who made possible Carmina Burana. And finally, my endless gratitude to Emilio Ros, as my teacher, mentor and friend.

Abstract

This work explores the possibility of modeling tonal stability from music signals, in an attempt to approximate in a different way an ubiquitous, yet controversial, topic in music research: that of the perception of tension and relaxation over time.

We rely upon some of the most agreed hypothesis about the hierarchical properties of the tonal system, to define a scenario in which such features could be modeled and exploited from audio signals. From the methodological point of view, we reinterpret current paradigms from music theory and cognitive psychology, in order to define a coherent and manageable concept of tonal stability, around the idea of perceptual degree of closeness between tonal events and their tonal contexts.

Our proposed algorithm extracts chroma information from music signals and builds a representation in which several hierarchies of pitch events, related with traditional categories of pitches, chords and keys, are simultaneously accessible. The different levels on this *keyscape* establish the required hierarchical relationships between tonal events and -within- their tonal contexts. This information is discussed in terms of stability through time and pitch space and, after its mapping into a geometric model of perceptual interkey distances, several metrics of tonal stability are proposed.

Through three case studies, we discuss the applicability of the algorithm for cadence finding, tonal tension peaks modeling and stability-based structural analysis.

Index

List of figures

CHAPTER 1.- INTRODUCTION

1.1 Motivation

Regarding Music Information Retrieval (hereafter MIR) research from audio signals, we identify -unconventionally- two main branches[1]. The first one deals with the problem of modeling musical features from huge collections of music, mostly targeting commercial exploitation, generally around the idea of music recommendation. Some of such systems include timbre/instrument classification, music similarity, genre/style detection or mood characterization, among others. For that, full automation, computationally efficient algorithms and usable evaluation strategies are a requisite. One inherent characteristic of many of these approaches to music analysis is the statistical treatment of the music parameters -and the music itself- in terms of their classification into considerably wide stereotypical categories, which barely account for the attributes of single music works, whose individuality -what makes them unique- is intentionally removed by the methodology used. Typical cases could be music classification in terms of classical/non classical, vocal/instrumental or fast/slow, so don't pretending to say much about music. For applications targeting high-level description, like those related with human emotions, a plethora of machine learning techniques have been applied for trying to bridge the so-called *semantic gap*, and define taxonomies in terms of sensations or mood, like happy, boring or relaxed. Apart from few stereotypical features, most of them strongly arguable -like major/minor modes accounting for happy/sad-, some of these methods don't consider any *a priori* rationale behind those categorizations, since it's expected that taxonomies will *emerge* from the method. In some cases, this outcome is supported and evaluated upon a combination of dozens -or hundreds- of descriptors, often involving low-level features from audio signals -e. g. attack time or spectral centroid- and quite subjective high-level human annotations -e. g. angry or sexy- obtained from a variety of sources, spanning friends ratings to massive anonymous taggings from internet -including fakes-. This complicates the *explanation* of why these parameters may result on the classification, and actually, the interpretation itself is not even considered in many applications, as far as they provide the desired performance in evaluation, which pursues different goals.

The second approach is closer to the traditional view of the philosophical and scientific question of *how music works*. Research here involves perception and cognition, cognitive psychology, musicological concepts, brain sciences, physiology or acoustics. These approaches, closely linked with music theories, aim for understanding the relations between music and humans in perceptual and cognitive terms, considering but not targeting the social aspects of musical experience. These studies usually focus on few descriptors -often just one, like musical pitch-, reduced music material -often synthesized and musically simple- and controlled human annotations, in order to get the best understanding of the process and the scope of the descriptors according to current paradigms. Some applications requiring these refinements go beyond the description of music as a product -related with users/consumers, but objectively separated from them

[1] This is just an additional personal observation aside research community's problem classifications and trends, not normative and compatible with the rest, which motivates this work. We don't pretend in any case either that these arguments are the unique characteristics or they are specific to the mentioned disciplines or methodologies, nor to make value judgments. Maybe, better than *branches*, we should say *attitudes* regarding goals. In my personal interpretation, based on analysis of evaluation methodologies, first approaches tend to model the *What*, while second ones focus on understand the *Why*.

as *listeners*-, and aim for integrating the *experience* itself, in the sense of share some degree of dialogue/interaction between music and humans, but always under the mentioned strong limitations. This includes music analysis and composition tools, interactive performance assistance systems, sensory reinforcement feedback -e. g. music visualization-, educational software and setups for experimental research, among others. Additionally, these descriptors constitute the working material -as hypotheses- for computational approaches to music perception and cognition, in the sense of trying to provide solutions to *inverse problems.* That is, the goal is to derive models of the mechanisms underlying the experience of music from empirical observations, for which multidisciplinary cross-validation and critical discussion are methodological requisites. Although the benefits of such kind of descriptors are evident for analysis of massive music collections as well, methodologies and evaluation strategies are necessarily different. Fortunately, both attitudes are not competing but complementary, and what motivates us is their potential of combination.

During years, MIR applications have been focused on transcriptionist approaches, that is, to extract the *score* from audio signals, but this has proved to be quite complex when dealing with music recordings, providing highly fallible -hence scarcely usable- information. Moreover, most of these attempts have tried to extract isolated features, such as pitch, chords or rhythm, and these plain descriptors -without a proper interpretation- barely account for any kind of musical *meaning*, in the sense of describing experiences somehow coherent for listeners. Thus the growing interest in consider more *holistic* approaches targeting to overpass the semantic gap, for what we believe it's needed to deal with one essential aspect of music experience: its evolution over time. And here we stress the evolution *of the experience*, not of the musical parameters in traditional sense, for instance, not by describing the temporal variations of loudness, but how these variations might be *meaningful* to listeners.

Much of the musical listening experience is about expectation and its fulfillment or disruption. These processes have been hypothesized as relevant for the arousal of emotions and related with the concepts of tension and relaxation. *Musical tension* has been a common topic for composers and theorists, and its relationship with the tonal system -as we know it today- has been strongly supported from a number of disciplines, including comprehensive music theory and cognitive psychology. Most approaches to *tonal tension* are founded upon the idea of *tonal stability*, so, if we were able to model and represent this concept of stability over time, and extract it from music signals, we would have an interesting descriptor related with how music is experienced by listeners. Our main motivation here is to work at higher level than just traditional musical features, but below too high emotional states which often rely on vague and strongly variable adjectives to be described, thus providing a potential -and somehow *neutral*, not too specific, not too subjective- bridge between both.

Such descriptor could assist applications intended for massive music collections, generally around the topic of music recommendation, like music similarity or mood characterization. Additionally, its exploitation within interactive systems -involving user actions, visualization and music generation- might be useful for music education, analysis and compositional purposes, and even provide advanced interfaces for psychological research or music therapy.

1.2 Goals

Our aim is to research, design, develop and evaluate a usable descriptor of tonal stability from music signals, for which the proposed outcome is:

- Critical literature review, digesting relevant state-of-the-art and selecting promising perspectives to address the problem.
- Definition of a system able to capture tonal stability related information from music signals.
- Implementation of the algorithm and general discussion on capabilities and drawbacks, in terms of the information we look for, the particular nature of model's building blocks, and according to current research paradigms.
- Evaluation of the proposed descriptor, by extending the discussion over three specific applications (case studies).

1.3 Structure of this document

The organization of the rest of this document follows. In Chapter 2, an approximation to the state-of-the-art is developed in sections, respectively: a) musical tension and tonal stability; b) tonal description and extraction from audio signals; c) hierarchical and multiresolution approaches. Chapter 3 develops our proposal to tonal stability modeling from audio chroma features, being structured as: a) introductory rationale, derived from the previous chapter's conclusions; b) implementation and general discussion, structured in sections devoted to each main building block. Introduced by some notes about evaluation issues, Chapter 4 discusses the potential applicability of the descriptor, by means of three case studies, namely, cadence finding, tonal tension peaks modeling and stability-based structural analysis, and finishes with some considerations about robustness. Chapter 5, finally, summarize the conclusions, review some of the open issues and suggests future research proposals and applications.

CHAPTER 2.- STATE OF THE ART

In this literature review, we pursue two main goals. The first one, corresponding with the academic nature of this dissertation, is to provide a proper scientific background to the topic under study. Bringing together the variety of paradigms, methodologies, solutions and interpretations involved in this multidisciplinary field requires a tight selection, necessarily far from comprehensive. Additionally, since our proposed model is inspired and evolved from a combination of such perspectives, we also pretend for this chapter to provide a structure which can help the reader to follow the rationale behind the decisions taken.

For that, we organize it in three main sections, converging into a framework in which our proposal may take shape. Thus, we start from the general paradigms under which music stability and tension are described, particularly around the concepts of tonality and tonal system. Then, we present current approaches to tonality description, focused on our main concern of dealing with audio domain and empirical evidence. In a third section we analyze hierarchical and temporal multiresolution solutions, and their potential to provide access to the information we look for. The chapter concludes digesting a selection of topics and techniques conforming our rationale for the proposed model.

2.1 Musical tension and tonal stability

Krumhansl's survey about tonality [17] presents several perspectives from which tonality and tonal system are currently been approached. Musicological, perceptual and cognitive, acoustical, cultural, computational, music theoretical/mathematical and brain sciences findings are posed. In our interpretation of this study, it's remarkable the subtle but critical distinction between the concepts of *tonality* and *tonal system* -not strictly stated in these terms in the paper- tending the former towards human experience and being the latter biased towards theoretical constructions. These differences are not just terminological, but conceptual as well, and seem to be the main friction point within multidisciplinary approaches. Notwithstanding this, there is a general agreement about the relevance that tonality has for experiencing music[2], for instance, having influence on musical memory and learning, providing references which allow for the arousal of *meaningful* musical structures, and affecting emotional states through governing patterns of perceived tension and relaxation.

Regarding all these impressive potential of tonality, theories diverge about the underlying mechanisms, but most of them rely in some interpretation of the idea of *stability*, although not always explicitly using this word[3]. Since to define stability requires a reference from which establish comparisons and elements to be evaluated against it, this concept can be generally understood as the relationships between events

[2] While most studies focus on Western music tradition, wider interpretations of tonality allow almost any kind of music based on systematic pitch organization and octave equivalence, including modal systems, different temperaments, microtonality and even *atonal* systems.

[3] The term *stability* is explicitly referenced at some point in all of the theories discussed here, and -almost- always appears at the very foundations of the tonal system. However, the concept gets embedded by the different paradigms, methodologies and terminologies. Since the present work aims for model this general underlying idea, my purpose here is to digest those approaches from current literature which can be easily interpreted in terms of stability.

and -within- their contexts[4]. This way, contexts provides the references, and events' stability will depend on their *closeness* or *fitting*. Most theories on tonality describe all of them -events, contexts and how they relate each other- from basic musical elements -pitches- and their inherent properties, conforming the foundations of the tonal system. The core concept here, around the idea of *tonic orientation*, is conventionally referred as *pitch hierarchies*.

This has been defined from music theory, experimental psychology, psychoacoustics and connectionist approaches, most of them with some sort of reference to hierarchical relationships. A notable variety arise on the degree of focus and terminology. This way, Schenkerian analysis [6] operates in terms of *Urlinen*, *middle-ground* and *fore-ground*; Lerdahl [21] reinterpret Schenker's concept of *prolongation* and models his *pitch space* according to cognitive-related *pitch hierarchies*; Werts [41] uses his concepts of primary, secondary and tertiary *scale references* to account for essential events, harmonic and non-harmonic *projections*; Parncutt [30] model the perceptual *salience* from the global acoustic stimuli; and Tillmann *et al.* [37] consider the top-down *back reverberation* from key to chord units in their MUSACT model. In Krumhansl's work [15], the relationships between *pitch events* -whether pitches or chords- and their *tonal contexts* are ubiquitous, being actually at the core of her *probe-tone* ratings experimental methodology.

From the structural point of view, stability and *closure* [3] have been related with concepts like *musical coherence*, although this normally refers to higher level concerns -including music style or aesthetics-, and involves many other musical parameters -not just tonality- in order to be properly discussed. Few models from music theory have tried to cover the interaction between different musical parameters to define coherence. One of such tools is Schenkerian analysis, firstly systematized in a pedagogical way by Forte and Gilbert [6], which has remained as a mainstream analytical tool for musicologists since its publication. This theory, despite its known biased aesthetical drawbacks, paved the way to many researchers with its idea of *prolongation*, by which the main pillars of a musical composition scaffolds the details of the *surface* events through a network of hierarchical relationships[5]. Due to the musical corpus for which this tool was devised, Schenkerian analysis became one of the first interpretative theories about tonal coherence as we know the expression today.

Lerdahl and Jackendoff, inspired upon the Chomskyan principles of *generative grammar*, reinterpreted this idea of prolongation in their *Generative Theory of Tonal Music* (hereafter GTTM) [22], which has attracted the attention of computational

[4] As in the previous footnote, my intention to describe stability requires a homogeneous terminology coherent with the different paradigms, but not necessarily using their same terms. Since my proposed model relies in the contextual relationships of tonal events, this review covers those arguments that can be interpreted in such way.

[5] In hierarchical approaches, musical *surface*, as opposing to *deep structure*, is normally understood as those events *less relevant* for musical structure, usually under the denominations of ornaments, passing notes, and the like. Due to their subsidiary consideration, surface events are removed in musical reductions, under the assumption that listeners don't rely much on them to follow -*understand*- the music discourse. Nowadays, traditional *rules* for defining surface have been much criticized, particularly from advocates of expectation modeling [12, 27], as well as the term *prolongation* is also controversial. In this document I will use the term *surface* in a different way, closer to Krumhansl's view, as those events occurring during listening near the present time. In my usage of the term, any event presents a surface level -every event has its present and it's listened- but, depending on its structural importance in tonal hierarchy and musical discourse, it will drill down deeper in the foundations and gain higher status as referential item [15]. My usage of the term *surface* is similar to Lerdahl's *"surface"* (quoted).

musicologists due to its algorithmic, yet flexible, rule-based approach to tonal analysis. Most decisions about the events structure used for the analysis within GTTM are partially covered by defining a set of *well-formedness rules*, accounting for many of the conventions used throughout history of music composition. However, the theory is far for being prescriptive, and explicitly addresses the need of human interpretation to decide about the proper application of the rules, aligned this way with musicological analysis practice. Since it operates in symbolic domain, there have been several attempts to implement it, like the system described in [10]. These authors propose an Automatic Time-span Tree Analyzer (ATTA) from MusicXML encoding of monophonic scores, covering grouping and metrical structure analysis and time-span reduction. Since they aim for total automation, one of the main concerns for Hamanaka and collaborators is how to deal with GTTM's inherent ambiguity respecting the application of preference rules -they apply 17 out of the 26 described by the theory-. The system forces disambiguation by parameterization of the rules, renouncing to the power of the theory's flexibility, however some sort of *stylistic* -analytically- tuning is allowed this way.

Lerdahl review and extend GTTM into his *Tonal Pitch Space* (hereafter TPS) theory [21] by including a considerable outcome from recent psychology research. Apart from this interest on fitting empirical evidence, TPS aims for the integration of rhythmic, melodic and harmonic elements, in a way that all of them interact with each others to build coherent units of musical sense. TPS is not intended for just analysis, motivating interesting applications through its modularity and algorithmic approach to some of its components. As we interpret the theory, and the way it's applied by the author, its true potential relies on the introduction of flexibility and subjectivity as part of the model's foundations, thus providing a framework which might be fruitful in research about cognition and expressivity.

In [21] Lerdahl develops further his model of tonal tension and attraction [20], based upon the prolongational component of GTTM and the tracking of the musical events through pitch space paths at several hierarchical levels. The rationale of the model relies on the metaphor mapping from instability/stability into tension/relaxation, in the specific sense of that created by melodic and harmonic motion over time. Two main components[6] are hypothesized for that: a pitch space distance model, accounting for the relative instability between any two events, and the governing tonic orientation of the piece, providing stability references through a global hierarchical structure. The computational approach to the model requires two main steps, in order to capture properly the mentioned elements.

First, computation of the *sequential tension* between any consecutive pairs of events, for which the simplest version is to calculate the distance between them in pitch space. Tension state for a given chord is established by its distance from the immediate previous one, as a linear combination of the contributions of the three sub-spaces involved, namely, basic, chordal and regional. The chord proximity across regions, applicable when both regions are not far away, is formalized as the *chord distance rule*:

$$\delta(x \rightarrow y) = i + j + k \qquad (2.1)$$

[6] Lerdahl's complete tension model also considers the influence of surface dissonance and voice-leading attractions. However, we only mention in some detail the elements directly related with our proposal. For a detailed discussion over a proper selection of examples, see [21].

Where: *i* stands for the number of shiftings we need to perform over the chromatic circle of fifths to change the diatonic collections supporting both chords (change of region); *j* stands for the number of required shiftings within diatonic circle of fifths to match the root of both chords (chordal distance within region); and *k* stands for the number of distinct pitch-classes on the basic space. Since local keys -regions- are involved, sequential tension requires to perform an appropriate analysis of the piece, which comprises a local interpretation of tonicizations, in the sense that any pair of chords should be self-explained harmonically, as if they were the unique elements to be analyzed, although included in the general context of the prolongational structure. Some notational aspects arise here, for instance, by interpreting the sequence A^0*-Bb* within an *Eb* context as *Eb:vii*^{0}*/V -> I/V* instead as the most reasonable *Eb:vii*^{0}*/V -> V/I*, which only allows such lecture when *Bb* is interpreted as dominant of *Eb* under larger hearing. So, a typical secondary dominant relaxing into dominant is seen in this sequential approach as a dominant falling into a local tonic.

Second, and more important, *hierarchical tension* takes sequential tension -distance-estimations from the prolongational perspective. Put simply, what is computed is the inheritance of distances along the musical discourse, following the ebb-and-flow of tension and relaxation of events down the prolongational tree, which is built according to the hierarchies established by the governing tonic. That means, if we move from a stable event -close to the governing tonic- to a more unstable, the latter increases its tension value by the distance between them[7]. The same way, if we move from one event into a more stable one, the tension of the latter is decreased by their distance. This rule applies for events at the same reductional level and is valid for all levels of the model. Some awkward notation is introduced to manage hierarchical tension, like measuring distances between events in *time reversal*, or introducing tensional contribution between an event and itself in cases of pivot chords -which require to be described in two different contexts-. This can influenced by their interpretation at distinct prolongational levels, as well as by differences between prospective and retrospective hearing.

The main concern for addressing an automated implementation is that this model relies on a proper representation of the hierarchical tree, in which time-span reductions and branching decisions might have an important interpretative contribution from the analyst, depending on the complexity of the music and the focus of the analysis. This becomes particularly problematic for its application to music signals, for that many implementation issues arise, like dealing with note segmentation, expressive timing and voicing transcription. Up to our knowledge, computations of the model have only been addressed manually by human analysts over scores [21, 20, 23, 5]. A number of specific components of TPS have been implemented in interactive pedagogical applications, which allow to navigate through some of the model's functionalities [42], but not intended as tools for analysis.

This overall hierarchical organization of events by listeners is not supported by some researchers, who advocate for more psychoacoustical influence. Bigand and Parncutt's study on perception of long chord sequences [2] favors the relevance of the local harmony for perception. According to their experimental results, in which they evaluated Parncutt's *sensory psychoacoustical* model against GTTM, listeners attend

[7] TPS's hierarchical tension relies upon a complex prolongational structure to define the relative stability of events, involving metric hierarchies, phrasing interpretation and more refined details, which are only accessible from a precise symbolic representation. Most of these elements are beyond state-of-the-art's modeling capacity from audio, so our simplistic description here is based just on the terms we will be able to manage in our model.

primarily to events happening between cadence and cadence, having higher hierarchical information less relevance. This was criticized by Lerdahl and Krumhansl [23], who tested the same musical material -Chopin's *Prelude* in E major- with TPS's tension model and continuous-tension ratings, challenging the short-sliding window hypothesis in favor of a longer-term hierarchical approach[8]. In any case, regardless hierarchy was or not related with the perceptual-cognitive activity on listeners, GTTM was able to model best the cadences *because* it operates in a hierarchical way. Epistemological considerations away[9], and just as a practical conclusion, we derive that nowadays hierarchical models fit better the experimental results.

From the empirical point of view, Krumhansl and collaborators' systematic approach to tonality is founded upon a principle of *perceptual tonal stability*[10]. Their probe tone methodology require that subjects rate *how well* a given tonal event -typically pitches or chords- match a previously established tonal context in purely perceptual terms -so, there are no musical concepts or categories involved in the process-. These contexts can be induced in different ways, typically by preceding the probe-stimuli by scales, tonic chords or chord sequences defining strong cadences. From these experiments, they derive hierarchical properties of pitch events as function of their contexts, being the later described as the concept of musical *key*. This way, the stability of a pitch event, under the influence of a given key, can be quantified and ordered according to the hierarchy established by the key. The relative stability of pitch events in a given key is represented as a 12-dimensional vector, named *key profile* by the authors, and it's described for *tonal* and *harmonic hierarchies*[11].

A step further was done by defining geometrical models related with perceptual distances between keys [15] (chapters 2 and 7 and referred articles). Krumhansl and Kessler suggested quantitative metrics of interkey distances, based on the assumption that if two keys are close each other, they will induce similar patterns of stability on the tones -or chords-. Given this, they propose and quantify a similarity measure -based on correlation- between all profiles for major and minor keys, as *related* with perceptual interkey distance[12]. Then, they derive geometric representations of the distances, by

[8] We support arguments in [23] against the methodology used by Bigand and Parncutt in [2]. Their particular use of a stop-and-rate technique to evaluate their *cadential model* gives little chance to subjects for perceiving any sense of hierarchy beyond the relationships between two consecutive chords.

[9] This controversial question is rather speculative in all related disciplines, in the sense of having been posed as hypothetical underlying mechanism of cognition, but not having been successfully supported. For literature reviews, see [21, 37].

[10] The terms in which this stability is described come from two general assumptions in psychology research about reference points. First, that elements can be rated in terms of *goodness* respecting a given category, which provides a quantitative *hierarchical ordering* of the elements. Second, that this ordering influences measures of perceptual or cognitive processing. Krumhansl's probe tone methodology is based upon these two principles.

[11] To summarize, and assuming the hazards of simplify too much, in [15] *tonal hierarchies* are established by measuring the perception of single musical tones -pitch classes- when they are listened under the influence of a given tonal context. *Harmonic hierarchies* refer to similar considerations regarding the perception of chords instead of tones. In the case of Western music tradition, *tonal key profiles* are described mainly for major and minor keys and equally-tempered pitch-classes, and *harmonic key profiles* for major and minor keys and triad chords rooted on the same pitch-classes. The later can be described in a variety of ways, though, depending on the type of chords and the focus of the study -authors used just diatonic triads, as well as all possible major and minor chords-.

[12] They warn a number of times about the limitations of the assumptions taken: “[...] the patterns arise from the tonal hierarchies and not direct perceptual judgments of key distances”. [15], p. 40.

applying nonmetric multidimensional scaling. This produced optimal four-dimensional spaces, in which the circle of fifths and the relative and parallel major-minor relationships were clearly distributed. Additionally, these spaces allow for direct Euclidian measurements between any two keys. They are shown in the next figure.

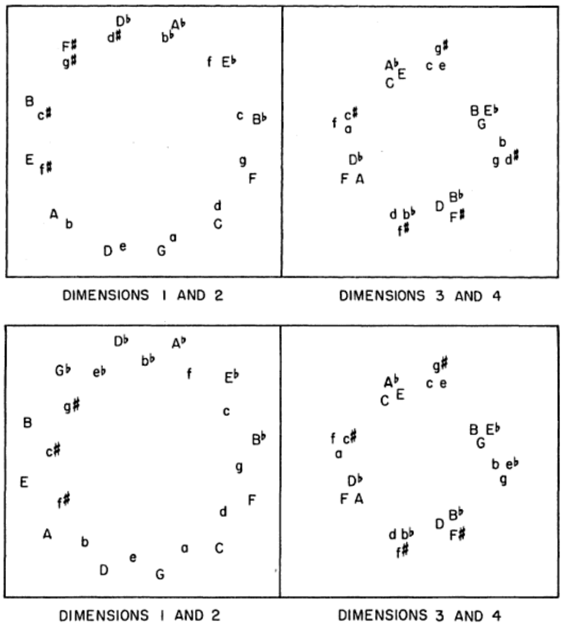


Figure 2.1. Four-dimensional models of interkey distances. Interkey distances derived from correlations between probe tone profiles for all pairs of major and minor keys, as obtained by Krumhansl & Kessler. Top: Derived from tonal-hierarchy ratings. Bottom: Derived from harmonic-hierarchy ratings. These four-dimensional solutions, in which we can observe the circle of fifths and the relative and parallel relationships, allows us to approximate similarity between any two keys by using Euclidian distance. Top: copyright © 1982 by the American Psychological Association. Reproduced by permission of the publisher. Bottom: reproduced from [15], p. 186. Copyright © 1990 by Oxford University Press. Pending for permission.

For visualization convenience, these geometries were also mapped into two-dimensional spaces, shown in the next figure. X-Y axes represent angular information of the underlying toroidal structures over which the tone centers are distributed, so top-bottom and left-right borders represent geometrically the same points. Krumhansl warns about the limitations of this reduction, which does not represent the perceptual distance proposed by their model. To get a proper approximation requires going back to the four-dimensional solution or, obviously, to the direct correlations between profiles.

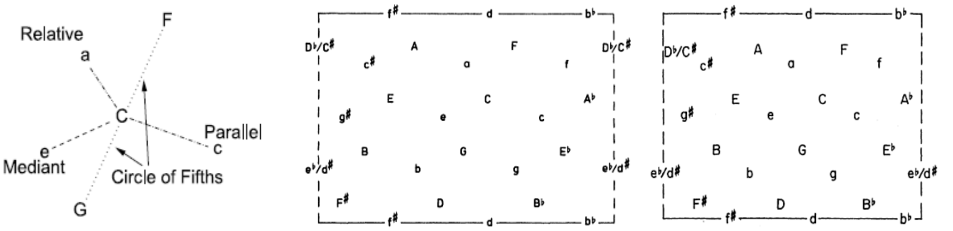


Figure 2.2. Two-dimensionality of interkey distances. Left: spatial distribution around the tonic *C*. Center: distributions of all major and minor tonal centers in a two-dimensional setting (unfolded torus), derived from tonal-hierarchy ratings. Right: distributions from harmonic-hierarchy ratings. X-Y axes represents circular dimensions of angle. NE-SW *diagonals* show the chromatic double-nested circle of fifths (major and their minor relatives). Center: copyright © 1982 by the American Psychological Association. Reproduced by permission of the publisher. Rigth: reproduced from [15], p. 187. Copyright © 1990 by Oxford University Press. Pending for permission.

Music theory and cognitive psychology join their efforts in [23], by evaluating TPS's tonal tension model with empirical results. By trying to define how listeners might are being *analyzing* the music, Lerdahl and Krumhansl approach a solution of the so-called *inverse problems* in computational disciplines. To do that, they propose different prolongational branchings and contributions of dissonance and attractions so as to fit better the tension curves rated by subjects, which were obtained from stop-and-rate and nonstop listening tasks. Some good fittings between predictions and ratings were obtained by this method, particularly at the most stable and unstable events -tensional valleys and peaks-, suggesting interesting hypotheses to be worked out further.

In their study about real-time musical responses, Toiviainen and Krumhansl [39] develop the original probe-tone methodology into their *concurrent probe-tone* setup, which allows for measuring fitting sensation of probe-tones within a non-stop listening task. Their goal were to show how perception of key evolves as complex music unfolds, including relative key strengths, extending previous Krumhansl's experiments about perception of modulation [15] to considerably more complex musical stimulus. By fitting the modeled tonal induction with real-time tension measurements, they roughly support the conventional idea of higher tension arising from tonal regions far away the tonic.

Despite all these efforts, most of these studies don't cover other important contributions to tension, such as dynamics, timbre, thematic/motivic material or a general auditory scene analysis, and thus their results' interpretation are unavoidably speculative. In her PhD dissertation about musical tension, Farbood [5] approaches this problem by considering the parametric contributions of different musical elements, covering harmony, register, melodic expectation, dynamics, onset frequency, tempo and rhythmic regularity. One fresh contribution in this study of interaction between features, is described in an experiment dealing with structural contradiction within a high-level *meaningful* context -perception of tension over time-. This experiment touches a critical point in music cognition. Much work has treated the issue of expectation and its fulfillment or disruption [25, 12, 27], but very few ones have dealt with contradiction between different evolving *discourses*, although it has been a major concern for composers of all times and it's a main source of debate -and enjoyment- among musicologists. As it's introduced by Farbood, who considered harmony, onset frequency, pitch height and loudness, the contribution to tension of single parameters strongly depends on the supporting or opposing combination of the others, following complex behaviors. This point was suggested also by Krumhansl [16], who identified tension dependence of melodic contour, note density, dynamics and tonality.

More realistic listening scenarios increase further the number of variables playing roles in tension induction. For instance, cross-modality between hearing and vision has been experimentally studied in Vines *et al.* [40], who describe evidence of interaction of both senses during tasks about perception of musical phrasing and tension. When information about performer's movement was present at listening, subjects showed significant changes in their sense of phrasing and emotional expectations, which suggests that embodied responses may affect tension, even when those movements are just been seen -not performed- by listeners.

2.2 Tonal description and extraction from audio signals

In this section we review some tonal information representation techniques, with emphasis in those allowing for extract it from audio signals. For that, it's important to enhance one inherent aspect of tonality, extremely relevant for its description but quite often *forgotten* in MIR literature: its ambiguity. From an analytical perspective, the ambiguity of tonality has been treated in depth by Werts [41], to cite just one example, who details with a great amount and variety of musical examples the *ebb and flow* of the induction of multiple tonal centers, especially as part of modulating processes, as one main resource exploited by composers. Wert's *scale references* are probably one of the most fortunate approaches for defining musical key. Scale references are induced -more or less authoritatively- in listeners by a complex interaction of mechanisms, and at the end, which reference(s) is(are) being considered might not be necessarily dominated by one in particular. This can be observed ubiquitously throughout musicological analysis literature, as different interpretations of the same music works in tonal terms. Apart from musicological sources, the topic of tonal ambiguity is covered in [21, 15, 21, 5, 12, 3, 36, 23], among many others.

Krumhansl's experiments about subject's *sense of key* -much particularly on those about modulation- provide great account of evidence about tonal ambiguity [15]. However, as she treats the topic in her celebrated *key-finding algorithm* chapter, her bias to approach a single ground truth overshadows the richness of her quantitative outcome about how all the keys are being induced simultaneously in listeners to different degrees over time. We refer here to how statistical facts about subjects -or algorithms- estimating *correctly* the key of a music excerpt, are expressed in terms of the human -or system's- capacity to derive a single answer -the *good one*- chosen from a set of predefined theoretical categories, despite the analytical fact that this music fragment may be actually ambiguous about its key under the same categorical framework[13]. Somewhat ironically, she explicitly considers the need of knowing the keys to interpret the inherent ambiguity of chords, through their function within the system of interrelated tonalities, but not the inverse relationship.

Here we identify a generalized drawback in many evaluation criteria on the reviewed MIR literature about chord and, much especially, key estimation from audio. As we interpret it, the often cited *semantic gap* in Sound and Music Computing (hereafter SMC) [35] has part of its origin in this treatment of the ground truths, which may have been built with inherent statistical properties or from rough conventions, but they are being used as absolute reference independently of the application. This can be observed even in frameworks aiming for provide evaluation references for research community, like MIREX's Audio Chord Detection task[14]. As it's supported by Lesaffre *et al*. [24], regarding application-dependent manual annotation methodologies in MIR, a

[13] We don't interpret this as a generalization of Krumhansl's reasoning or methodology, since she's usually careful about tonal ambiguity, but we want to warn on her choice of the terms at some points. We refer to her bias -probably unconscious- towards speaking about *correct key* when referring to Bach, and towards *intended key* when dealing with Shostakovich or Chopin. Several studies have used these results -mostly Bach's *Well-Tempered Clavier*- as a sort of baseline for evaluate key-finding algorithms, without a single word about the terms under which this ground truth might be valid, by taking for granted the equivalence between *musical key* and *key of a musical work*. The later, obviously, depends on factors such as historical or aesthetical concerns, and its relation with the music's tonal content itself varies extraordinarily, as it's also discussed by Krumhansl. We don't find any musicological reason for supporting that Bach's *keys* are *more correct* than Chopin's or Shostakovich's in their own terms.

proper ground truth should be built according to the application, much particularly for psychology related modeling, regardless the key on the score or standard -*expert*- chord annotations, which usually serve for other purposes.

Evaluation issues aside, Krumhansl & Schmuckler key-finding algorithm [15] provides usable results at reasonable low computational cost, and it has served as the foundation of a number of implementations for audio signals, most of them focused on chord recognition. Some main extended techniques include a variant of Fujishima's Pitch-Class Profiles (hereafter, PCP) [7]. The general algorithm is based on mapping the harmonic content of the signal into a quantized vector representing the 12 pitch-classes of the Western equally-tempered octave. This *chromagram* is then averaged over some timespan and correlated with ring-shifted chord or key profiles -normally covering all major and minor keys-, which can be obtained from psychology experiments [15, 1], machine learning techniques [32] or built as chroma-templates from theoretical decisions. The highest correlation is taken as the estimated chord or key. Gómez [8] describes a variety of the most commonly used preprocessing methods and decision heuristics. The nature of this kind of descriptors made them only capable of constrained and fallible tonal information estimations in terms of conventional categories, covering major, minor, augmented and diminished chords, making them barely usable for traditional analysis of harmony beyond this level of description. On the other hand, the method provides a rich information of weighted multiple *tonal implication*, in similar terms as those at the foundations of [15]. However, this information is not being much exploited -up to our knowledge-, except for visualization purposes and few specific case studies, and we see here a promising research scenario, particularly regarding perceptual issues.

Three main implementation problems remain for these kind of systems. The first two of them are due to its statistical nature. PCPs are computed by averaging spectral information from audio, so the robustness of the estimation depends on the timbral qualities of the signal. Moreover, as it has been posed before, mere statistical treatment of pitch classes cannot account for the conventional concepts of musical key. Additionally, since harmony in music evolves on a variable timespan basis, fix-sized sliding-window strategies have proven to be a major drawback, and without a properly segmented signal, the quality of the estimations drop significantly. Common cases are root or fifth underestimations over ground truth triads.

While this might be problematic for transcriptionist approaches, the usability of these estimations in terms of distance within tonal space might prove not being so critical in many applications. Actually, by substituting at few points the real chords by the estimated, while the surface perception may be disturbed more or less significantly, the overall tonal discourse don't change so much, in terms of longer timespans tonal implication -such as local or global key-. In fact, attending to the nature of the descriptor, these chords are misestimated precisely due to their pitch content similarity in statistical terms. Moreover, compositional practice often avoids hierarchically important notes, such as roots or fifths, and those chords are interpreted -and perceived-

[14] http://www.music-ir.org/mirex/2009/index.php/Audio_Chord_Detection. We don't criticise any particular annotation, like the extensively used Harte's chord labeling for Beatles' songs, but the very idea of non-discussed standard ground truth and evaluation.

functionally without problems in many contexts[15]. In other occasions, this practice seeks to promote intentional ambiguity.

To deal with the sliding window problem, some automated segmentation algorithms have been proposed. Harte *et al.* [11] presents a descriptor to detect harmonic change, based on the idea of *tonal centroid* tracking. Inspired by Chew's *Center Of Effect*, used in her *Spiral Array* [4] to represent key estimations, Harte's centroid is mapped from pitch-class chroma vectors, computed using Constant-Q spectral analysis, into a representation of the hyper-toroidal structure resulting from the combination of the theoretical circles of fifths, minor thirds and major thirds. This six-dimensional centroid is then tracked to detect important harmonic changes by defining a Harmonic Change Detection Function (HCDF), implemented as standard peak detection over Euclidian distance between frames. One interesting point of the geometry proposed is its potential to capture particularly ambiguous events, like augmented and full-diminished chords, by their discrimination along the circles of thirds.

Another proposal for tonal centroids is described by Gómez [8], based on the correlation strength of Harmonic Pitch-Class Profiles (hereafter, HPCP) with key or chord profiles for the 24 standard tonal centers. Her implementation, however, loses its quantitative qualities, since it's optimized by trial and error for chord tracking visual animation by enhancing the relative importance of the strongest tonal center. Moreover, Gómez's tonal space representation is based on Krumhansl & Kessler's two-dimensional reduction of the geometry derived from tonal hierarchy ratings [15]. As it was introduced before, this space is built by using angular information as axes dimensions, resembling so the unwrapped toroidal geometry resulted from the four-dimensional solution. As a consequence, while this two-dimensional space is interesting to visualize the evolution of tonal implication, whether as a centroid or by all 24 key strengths, it does not represent properly the *perceptual* distances between tonal centers, as Krumhansl warns. As it's discussed in [15], four dimensions are required to get a reasonable low-dimensional approximation to interkey distance, which in this case can be computed as an Euclidian metric.

The main problems of Gómez's centroids are their extra summarization -over that performed by the chromagram-, their instability when tonal induction is very ambiguous -e.g. that from symmetric chords-, as well as the difficulty of defining the toroidal space configuration according to the intended application. Here we observe a critical point in terms of interpretation of the musical *meaning* -and thus, usage- of this kind of centroids, based on the geometry of the spaces they evolve within. Tonal centers in toroidal geometries only occupy discrete points at geometry's surface, but this does not hold for summarized information, which is located *outside* the *meaningful space*. So, while this sort of tonal *gravity centers* are informative about relative -weighted- distances in their respective spaces, they can only be associated to categories such as chords or keys in few specific cases[16]. Actually, the main benefit of these centroids is

[15] For instance, *V^7* chords, which don't require neither root nor fifth to impose their function as dominant, which is based upon the leading character of their third and seventh degrees -or upon the *need of resolving* the tritone relation, in a more psychoacoustical approach-.

[16] It's important to clarify one fundamental difference between Gómez's centroids and Harte's or Chew's in terms of representation. Since they map directly pitch-classes, Chew's and Harte's centroids are actually the *theoretical* -just limited by the audio concerns- way of describe chords and keys within their proposed spaces. On the other hand, Gómez's centroids are a summary of the tonal implication after the correlation with tonal profiles has been performed. Harte's and Chew's meaningful spaces are built from pitch classes, while Gómez's -Krumhansl & Kessler's- geometry is based on tonal centers, so they represent

precisely to account for the relative importance of all tonal centers so, by their nature, they are rarely located just over one specific category, although in many cases stay close to them -actually, this behavior can be easily tuned according to our needs-.

Temperley [36] introduces a richer description of tonal content by distinguishing among *tonal implication*, *tonal ambiguity* and *tonalness*, being all of them dependent on the relative strength between all 24 tonal centers. While this is not new in literature [15, 39, 32], the contribution here is to enhance the relevance of describing music in these terms, instead of derive single answers about tonal centers. To deal with ambiguous estimations, Purwins [32] proposes to consider the two -or more- highest correlations and an ambiguity descriptor -the ratio between these strengths- which could serve as weighting parameter for tonal center tracking algorithms.

A geometric approach to tonal space is generally agreed among different disciplines, having been some of them used for applications from audio. Apart from the mentioned empirical evidence from psychology research [15], toroidal mappings have been hypothesized and supported from music and mathematical theories [21, 41, 32] and machine learning techniques [32, 38], although divergence arise regarding dimensions and configurations. Janata *et al*. [13], present some evidence suggesting that the hypothetical -quite speculative- cortical topography of tonal structures in humans differs from cortex to cortex and even from time to time in the same subject, pointing to the lack of stability of any proposed geometry. Additionally, these spaces are only operative for tonal Western music, since toroidal geometries arise from the use of relative and parallel keys [12]. As seems to be a general agreement, further research is needed here targeting specific applications. TPS [21] includes a survey discussion regarding quantitative geometric distances within toroidal and non-toroidal spaces, covering most of the aforementioned disciplines, as well as suggesting how different geometries might account even for stylistic conventions and theoretical changes over history of music. Tillmann *et al*. [37], although not using toroidal mappings, provide a methodological discussion about the interpretation of self-organized techniques, in terms of tonal hierarchical organization, induction and expectation modeling, as well as a survey review of much work on connectionist approaches.

Other tonal systems -polytonal, modal, *atonal*, microtonal, non-equally tempered octaves, non-Western traditions- are sometimes mentioned, exemplified and worked out by SMC community but, away of musicological or compositional focused research and few notable exceptions [21, 15, 28], they receive very little attention, particularly from MIR.

2.3 Hierarchical and multiresolution approaches

Taking into account the proposal in [2] about short-term sliding window approach, Janata [14] describes a data-driven method for tonal representation, based on Self-Organized Maps (hereafter SOM), simulating the effects of tonal induction at two distinct timespans, according to the general agreement from different disciplines about a duration of 2-6 seconds in human sensory memory. In this experiment, several scales, modulating melodies and chord sequences are used as input, and it's shown how the response of the SOM to short-term window integration accounted for individual note events, while the longer timespan captured tonal region shiftings.

An additional interesting point is introduced in Janata's discussion. Observing the dynamics of the activation peaks in the SOM, he noted how the motility of the tonal

chords and keys -even *idealized* ones- differently.

induction was related with the time-constant used as sliding window -the larger the window, the slower the motion-. Given this, he suggest the idea that, if a subject were able to match its perception of harmonic change with trajectories on a SOM simulated for different time resolutions, we could be able to estimate the timescales *operating* in listener's cognitive system. Traditional methodologies in much research about perception usually define the tasks prescriptively, that is, users only react and rate within a tightly controlled setup. Due to the especial subjectivity of the tasks about tonality, it has been suggested [14, 23, 34, 24] instead to allow the users to optimize the stimuli, or to *find their way* on a map of possibilities, according to their perception. The question about the temporal boundaries for perception of tonality has been posed for many studies, and this representation problem is consequently a core issue for tonal extraction from audio.

In his study of tonal center tracking Purwins [32] describes a three-tier multiresolution approach, by considering onset, beat and bar level timescales for analyzing Chopin's *Prelude* in C minor from audio signal. He applies automated segmentation and propose a method for visualize the different resolutions, in which the strongest tonal center estimations are codified with Scriabin's colors and aligned according to the timescale hierarchy. He introduces an interesting mapping of this information into a SOM representation, in which the height of the elevations on the map accounts for the accumulated strength of the tone centers -ambiguity is included by considering the four highest ones- over time, resulting on a *static* topography of the tonal center hierarchies of the piece over which we can trace the tonal discourse. He also proposes an *error function* to measure the quality of the estimations, to be used for evaluation against hand-labeled ground truth. Due to the subjectivity of the interpretation of such rich harmonic language, in which the boundaries between tonicizations and modulations are blurred by analyst's -or listener's, or performer's- intentions, we find Purwins' metric interesting for performance similarity and expressivity modeling applications.

The main problem of these two or three-tier solutions is that they don't guarantee the access to the temporal resolutions required for usable descriptions of the tonal characteristics -e. g. in terms of chords and keys- since this varies extraordinarily in music, even in a single piece. Sapp's proposal [33], devised as an evaluation tool for comparing key-profiles, allow many temporal levels of description at the same time, accessing to tonal information related with notes, chords, brief tonicizations, strong tonal centers and global key. The idea is also attractive for its simplicity. Just by using different sliding-window sizes for averaging pitch information, ranging fractions of seconds to the whole piece, his *keyscapes* show a rich tonal map of the piece. A similar method has been proposed for audio domain by Gómez [8], by averaging HPCP information over different timespans and correlating these pitch-class summaries with key profiles, so as to estimate the best candidate at each resolution over time. A proper graphical representation of such information is pretty informative about the tonal discourse of the piece.

The main drawback of the method is the robustness of the descriptors to keep usability when integrated over different timespans and, much especially, the evolution of the descriptor's *meaning* over distinct resolutions, from the theoretical, perceptual and cognitive points of view, as has been pointed by Segnini's multiresolution approach to timbre summarization [34]. The sliding window policy used at each resolution is much responsible of the definition of the keyscape, and thus, of the usability of the information provided. In the case of Sapp's and Gómez's solutions, the linearity of the temporal scale for computing different resolutions, gives out a fairly usable description of the tonal hierarchies involved in the music piece. To solve this partially, Sapp proposes to plot logarithmically the keyscape's vertical axis, which gives a more intuitive visual information and a much better ratio between the *size* of different tonal categories -pitch, chord, key-. However, his non-overlapping sliding window policy results on a poor definition of relevant tonal boundaries over time. In general, this representation does not solve appropriately the general problem of temporal alignment of the different tonal levels -actually, Sapp uses just one point to represent the tonal estimation whatever the resolution, resulting on a triangular keyscape-. Gómez proposes instead to *spread* the tonal estimation along all its duration, obtaining a more intuitive rectangular plot, in a way that a *vertical* lecture of the keyscape is possible for each frame.

Leman's model of short-term memory [18] also proposes to use different timescales, by considering a two-tier structure in terms of one level for local events -pitches and chords- and one for the context. Leman describes an additional use for temporal multiresolution [19], by means of comparing the tonal induction computed for two distinct timescales. The result is a sort of *tension* measure between the surface events -short time integration- and the tonal center references -large window-, which is closely related with the stability and instability of the tonal discourse. Leman does not develop further this idea -exemplified with pitch mappings of echoic memories with different decay times- but his suggestion touches a critical point. As we have been reviewed so far, temporal multiresolution is starting to be considered interesting for researchers about tonality, but Leman claims the relevance of the *relationship* between those levels, more than the study of the levels themselves. Janata [13] suggests in this respect a more general problem -in his case referred to spatial distribution of memory storage in cortex-, in which the *interplay* of short-term and long-term memories might result on dynamic topographies, point supported as well from psychological evidence in [15] in terms of geometrical models. The need of consider events within their contexts is also one of the open questions in [23]. Regarding how neural networks are able to capture basic properties of pitch space from minimum theoretical assumptions -octave equivalence and chromatic distribution of pitch classes-, Lerdahl and Krumhansl suggest the possible use of temporal multiresolution to describe simultaneously several hierarchical levels, pointing to the problem of finding a proper encoding -representation- for that.

We find the contributions in this last paragraph particularly inspiring, since the idea of relate surface events within wider tonal contexts, a core concept for all the reviewed models about tonal stability or tonal tension, becomes accessible from audio technologies.

2.4 Conclusions

Given the previous theoretical and technological concerns, we can now digest the main ideas, converging in our rationale for the proposed model.

It's assumed by several disciplines a general framework in which music stability is defined: that of the relationship -*closeness*- between sound events and -within- their contexts. Considering a general hierarchical perspective aligned with cognitive approaches to tonality, this is related with surface tonal events -such as notes of a melody or chords- and their tonal references -such as tonicizations, strong tonal centers or global key-.

As we have seen regarding available tonal descriptors from audio, their statistical nature allows us to estimate tonal implication at different temporal resolutions. This way, the average-and-correlate technique from chroma features is usable to some degree for capturing information about chords and different levels of key areas, ranging from short tonicizations, local keys, strong keys and global key -if any- of the piece. Although this description requires a proper interpretation to be compared with such concepts -in terms of *meaning* and *accuracy*-, it can be computed homogeneously, efficiently and simplifies the representation.

We have seen the usability of multiresolution keyscapes as a proper mapping in which summarize in a single plot all this hierarchical information. Additionally, we have described how tonal centroids within pitch space geometries allows for low-dimensional representation of the tonal implication at any timespan considered. Despite the problem of their interpretation in standard music categories, these centroids might provide a usable material for estimate perceptually relevant distances in tonal pitch space and facilitate intuitive visualization.

We have discussed the potential of analyze the relationships between different levels in the hierarchy. Tonal stability-related concepts are founded precisely in this principle, and tonal tension has been hypothesized from this perspective as well. The interest in this multilevel interplay has started to attract the attention of researchers from different disciplines, and the possibilities of modeling it from audio have been posed.

CHAPTER 3.- TONAL STABILITY DESCRIPTION FROM AUDIO

Considering all these issues, we propose a descriptor accounting for tonal stability from audio music signals, which has been developed as an extension of the tonal analysis and visualization tool described in [9].

3.1 Rationale

The overall idea is to consider the relationships between tonal events and their tonal contexts, and provide a metric of stability based on their mutual fitting -closeness-. This is ruled by the hierarchy established by the tonal discourse of the piece, which is governed by tonic orientation principles. For that, we process chroma features from the music signal to build a multiresolution tonal map of the piece, which will conform the required hierarchical framework. This information is translated into perceptually relevant distances through a geometric model of pitch space. A proper interpretation of these distances will result on our proposal for tonal stability description.

This chapter is organized as follows. First, as a guideline, we present a brief overview of the model and a block diagram. Then, the different processes are described in detail, discussing in parallel their capabilities and drawbacks across three main paradigms: the nature of the descriptor itself, their interpretation according to music theories and their relationship with perceptual evidence. For the main concepts, this argumentation is worked out through specific examples computed with our algorithm. Our subdivision will follow the model's build up in order: signal processing to tonal implication, temporal multiresolution (keyscape), geometry of pitch space and model of interkey distances, mapping of multiresolution implication into pitch space (centroidscape) and multiresolution model of distances (distancescape). The section concludes with a summary of main identified capabilities and problems of the proposed model.

3.2 Implementation and general discussion

As a brief introductory summary, the system first processes the audio signal to extract the descriptor we are interested in (HPCP *chromagram*). Then, by using a temporal multiresolution technique, chroma information is averaged over different sliding windows and correlated with tonal profiles. With this, we obtain estimations of tonal information (*tonal implication*) accounting for potential chords, local and global keys. This multilevel information (*keyscape*) is then mapped into a toroidal pitch space geometry, by summarizing the previous tonal implication into a low-dimensional representation (this group of *tonal centroids* conforms our idea of *centroidscape*). Under these conditions, we apply a distance model operating over this space, so as to quantify the perceptual matching -closeness- between different hierarchies of tonal events. This is what we call *distancescape*, from which we derive our metrics related with tonal stability.

The general block diagram follows.

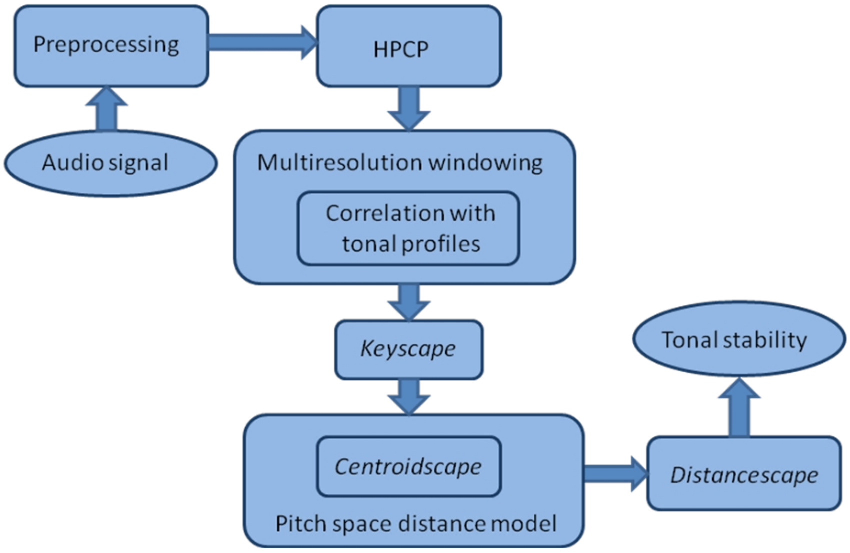

Preprocessing
HPCP
Audio signal
Multiresolution windowing
Correlation with tonal profiles
Keyscape
Tonal stability
Centroidscape
Pitch space distance model
Distancescape

Figure 3.1. Block diagram of tonal stability descriptor

Let's proceed in detail block by block.

3.2.1 From audio signal to tonal implication

The method start by computing high definition chroma information from the audio signal, for which we use the algorithm described in Gómez [8]. Preprocessing includes signal conditioning (stereo to mono, subsampling, normalization), spectral energy computation, spectral whitening, tuning estimation, adding artificial harmonic templates according to the timbral properties of the sound signal, and mapping into a high definition (120, so 10 cents) pitch-class collection. The output of this process is a framewise signal named Harmonic Pitch Class Profiles (HPCP) by the author. For each considered averaging window -see next-, *tonal implication*[17] is computed by correlation of averaged HPCP with ring-shifted tonal profiles to cover all 24 major and minor keys. Several profiles have been proposed in literature, like Krumhansl & Schmuckler's [15] or Aarden's [1], among others. Due to the nature of our implementation, we decide to use ad-hoc chord profiles, since they resulted to provide cleaner representations when the multiresolution method -see next- was introduced. From these correlations, we can just take the strongest one as the best candidate for tonal estimation, as it's usually done for chord and key estimation algorithms, or use a weighted summarization of several or all of them as a sort of tonal centroid. In this study, we will consider both possibilities, as it's discussed later. A graphical representation of this information is illustrated in the next figure.

[17] In this study, we will use the term *tonal implication* as the 24-dimensional vector representing the correlation strengths of the frame -whatever the averaging window considered- with all major and minor tonal profiles.

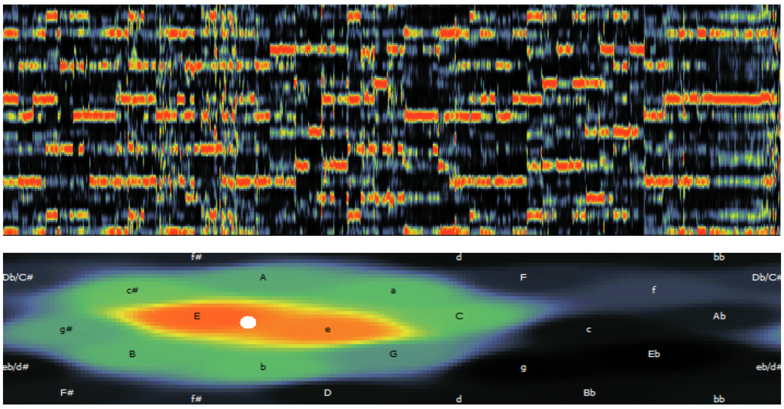


Figure 3.2. Tonal implication computation from audio signals. Top: high resolution chromagram (HPCP) over time (x-axis), with pitch-classes spread along vertical dimension. Bottom: tonal implication for a 1 sec. window corresponding to an *E* chord, showing the correlation strengths -coded in colors- of the averaged chromagram with all 24 tonal profiles. Here it's represented in a two-dimensional space, showing all tonal centers distributed in a convenient way -see later-. In the center of the plot, a white dot shows a centroid summarizing the tonal implication, according to the relative weights of the different tonal centers.

3.2.2 Tonal implication in temporal multiresolution: *keyscape*

Once the HPCP are available and the reference tonal profiles have been chosen for correlation, we build a proper representation of tonal hierarchies in a keyscape. To do that, we modify the description levels and segmentation method proposed by Sapp [33], and implemented as well by Gómez [8], in order to obtain a more usable representation in which we could perform our intended measurements.

Our algorithm, evolved from Gómez's, propose to consider only timescales in *quasi*-logarithmic[18] proportion, instead of compute them linearly and representing them logarithmically, as does Sapp. The rationale for this is that by using linear timescales, we get highly redundant information near the surface level, and relatively low definition at middle and upper level, losing interesting information about key changes boundaries. However, we aim for capturing the context with as higher definition as possible, as well as to do it in a computationally efficient way. By using logarithmic proportions, we cover in few steps all the required levels, from the whole piece up to reaching the shortest surface events. A typical piece during around four minutes reaches the realm of fractions of second in just 8-10 levels. Additionally, a logarithmic scale allows for approximately the same resolution at all levels, and hence chords and strong local keys are represented in a balanced way -i.e. with approximately the same number of levels-, which is a desired feature for our stability description -see later-. Moreover, the use of logarithmic scaling allows for high computational optimization of the chroma information averaging, by reusing the shorter-time estimations to build the immediate higher levels. The choice of base-2 in our logarithmic approach is rather arbitrary[19].

[18] Different timescales follow a $1+2^n$ (frames) relationship, except for the shortest, which has unitary size. The reason is to provide windows with an odd number of frames, so as all windows can be centered at the *present* one. Frame size used is the smallest time resolution considered for tonal estimation, and it serves as well as hop size for the overall sliding window policy.

[19] Our aim is to capture some information at several hierarchical levels of tonal discourse, and higher bases would reduce the number of levels in keyscape considerably. An additional -very weak-

The second significant change of the keyscape algorithm is to substitute the plain segmentation of the signal at every timespan resolution -by which, e.g., for a bipartition of the overall duration, only two estimations are computed, accounting for the first and second halves-, with a homogeneous sliding window strategy, using the same hopsize for all the resolutions. The short hopsize policy is interesting since it provides us a good temporal resolution at high levels on the keyscape, being this point the main drawback of Sapp's and Gómez's solutions. This is essential for capturing tonal center shiftings with some temporal precision[20] and, more important, it provides a uniform criterion for vertical alignment, which we will use later to relate events with their contexts. However, as a by-product of all the methods, residual misestimations at middle-high levels are unavoidable for longer modulating passages or fast alternation of tonal centers, since this information cannot be summarized as a (reasonable) single key of the segment. Some examples of this problem are shown later. The sliding window policy is represented next.

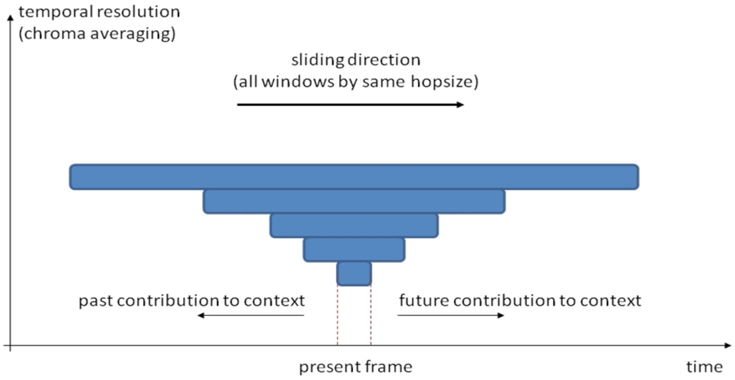

temporal resolution
(chroma averaging)
sliding direction
(all windows by same hopsize)
past contribution to context
future contribution to context
present frame
time

Figure 3.3. Temporal multiresolution sliding window policy. All averaging windows are centered at the present frame, and evolve over time synchronously under the same hopsize basis.

As it's shown in the figure, all the keyscape is defined in the same hopsize terms. For each hop-based frame, chroma information average is computed for a window centered in that frame, spanning the corresponding time resolution -including past and future frames in a balanced proportion-. Tonal context, so, is captured -defined- according to this decision, for which we will discuss some implications later. Let's look at the information provided by these keyscapes.

justification would be the *similarity* with rhythmic reductions in hierarchical models working in symbolic, which operate by non-overlapping segmentation in each level. Even here, this only holds for pure binary metric structures.

[20] Of course, we are conscious about the complex phenomenon of key shifting, far from being straightforwardly described as plain temporal limits. Our aim in this sense is just to clean as much as possible the keyscape's blurring around those *boundaries* -see one example next-.

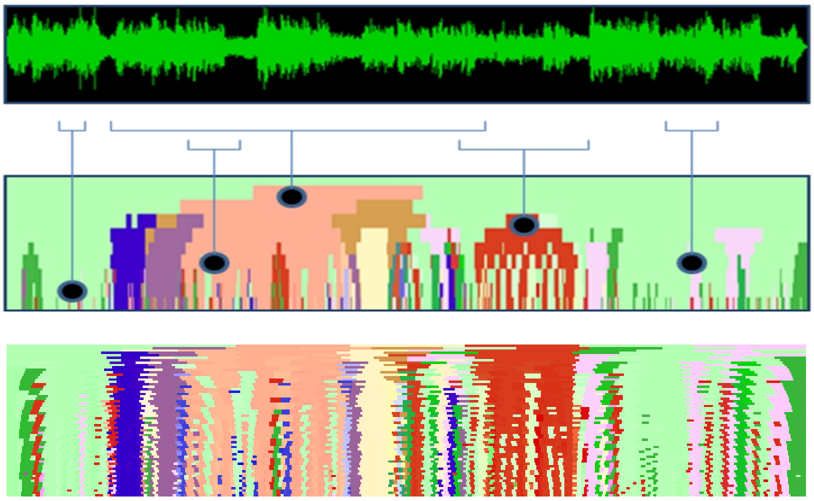


Figure 3.4. Keyscape for Mozart's K.626, *"Dies irae"*. Top: music signal. Center: keyscape, with some example points illustrating the temporal scope summarized at different heights. Bottom: Gómez's keyscape solution (non-overlapping sliding windows and linear scaling for temporal resolutions), for comparison on boundary definition and vertical proportions of the different tonal hierarchies. Color legend: light green = *Dm*, global key of the piece; pink = *Am*; red = *A*.

The figure -center, compared with Gómez's solution on bottom- shows a keyscape computed for a Mozart's *Requiem's "Dies irae"* commercial recording. Time frames are represented in the abscissa, while height is related with the temporal resolution used for averaging: the higher the point in the keyscape, the larger the window used. Colors correspond here to the strongest key estimated by our average-and-correlate method. Bottom layers show information related with short-timed events, like single notes or chords, while higher levels corresponds with key estimations at different degrees. Thus, mid-way levels are likely to represent tonicizations and local keys, and highest points are related to global key[21].

Under the overall context of the music piece -its global key *Dm*, in light green-, different tonal centers are induced during some period. According to the analysis of the piece, the first passage in *Dm* is restated in *Am* (pink) after a transition. Looking into the *Am* section, we can observe a number of lower-level events, which in this case account for chords. Since we pursue to define the relationships between chords and their higher references, we first need to interpret how these contexts are being described by the keyscape. According to our windowing strategy, the context is computed by averaging chroma information, spanning from a number of frames back to the past to the same number of frames forward into the future, centered at the frame we are considering. That means, we average not just the present chord, but the previous and the past ones as well, having as a result an estimation of the local key *best* explaining the involved group of shorter events -*best* according to harmonic content statistics, and under the constraints imposed by the correlation method-. So, by measuring the distance in pitch space -see next- from the surface chord to its local key, we are considering the closeness between just the present and a tonal *summary* including the near past and future events,

[21] We stress again this critical point: all these estimations should always be interpreted in terms of the characteristics and limitations of our algorithm, that is, averaging of chroma information -computed from audio signal's harmonic content mapped into pitch-classes- and correlation with major and minor tonal profiles. Our use of the terms *notes*, *chords*, *tonicizations*, *tonal centers* and *keys* (local or global), unless we argument differently, will always be restricted to this scenario. So, when applied to our algorithm's outcome, these words stand for estimations in these terms.

so, it might be related with how well the surface and the context fit each other. As we interpret the concept of tonal context in TPS or GTTM terms, it is being built generatively[22] from the interaction of near past, present and future events. Under this perspective, a surface event and its context are not independent to each other: context is generated from events, and the later are interpreted under the former, so any sort of *perceptual distance* between both would be best described as a *degree of blend* -see next-. As we move up the keyscape, this concept propagates to possible stronger keys up to the global key of the piece, *resembling* the way GTTM/TPS evolve through different reductions.

3.2.3 Tonal space geometries and model of interkey distances

This approach to keyscapes gives us a usable hierarchical tonal map of the piece. However, as we have mentioned, we are interested in describe the relationships between different levels of this hierarchy in perceptual terms. Put briefly, what we look for is a mapping of keyscape's information into a space in which we can perform perceptually related measurements, for which Krumhansl and Kessler's interkey geometries look like proper candidates.

As was introduced previously, four-dimensional solutions approximate directly the perceptual closeness as Euclidian distances, while the two-dimensional representations are useful for visualization. Of course, a direct computation of the similarity between two keys, whether from tonal-hierarchy or harmonic-hierarchy profiles, is given by the profile correlation values, without the need of reducing the dimensionality of the problem. However, we have two reasons for using these reductions instead. First, direct correlation can only be used if we consider our tonal implication just as the strongest key, losing the potential of summarization mechanisms -to deal with tonal ambiguity, for instance-. Additionally, visualization is always attractive as sensory reinforcement for many applications.

Regarding quantitative metrics, we propose to use Krumhansl's *tonal-hierarchy* four-dimensional toroidal geometry. This decision may seem somewhat arbitrary since other possibilities are available, like using the harmonic-hierarchy instead, and it's based simply on the availability of Krumhansl's numeric data for the four-dimensional version of this model[23] [15]. Actually, our main concern for the applicability of geometrical models to our algorithm, is that our method don't make any distinction between

[22] A terminological warning. Our *generative* process builds tonal information within the realm of our multiresolution short-hopsized sliding-window policy, which operates over statistical summarization (average-and-correlate) of harmonic content mapped into pitch classes, while GTTM/TPS build their hierarchical structures over pitchwise events interpreted categorically according to rhythmic and metric reductions. It might be relevant to mention an additional difference between both processes. In our algorithm, *all* the information is taken into account all the time for all the levels, while Lerdahl's *eliminate* those events considered lees relevant at each reduction. The same way, our method differs from the probe-tone approach for defining contexts, since here the context *includes* the event which is compared to. Additionally, our context is *contemporary* with the surface events, instead of being sequentially defined one after the other.

[23] Four-dimensional spatial coordinates for all tonal centers are available in [15], p. 42. We expect to use harmonic-hierarchy in the future for comparison, as well as other possibilities from those described in literature. We could argue that harmonic-hierarchy space is more appropriate model in our case, since these interkey distances were computed from fitting sensation of chords -and not tones-. However, we have to remember that this geometry also represents interkey distances, not distances between chords and keys-. Both spaces are derived by equally indirect methods, and so we don't find any reason for prefer one in particular.

different tonal levels: all of them are computed by the same means, and so, they represent information in the same terms[24], thus we would be in similar situation by choosing other option.

Considering that our aim is to define the vertical relations of keyscape's layers, let's interpret this constraint in musical and perceptual terms. According to Krumhansl's experiments, the *sense of key* seem not depend much on the way the context was established, whether as a scale, a tonic chord or a strong cadence[25]. This way, we propose to adopt the same metric to measure the distance between different key implication possibilities, including the effect produced by single chords, short tonicizations and strong keys in relation with their upper contexts. Put simply, any level of our keyscape is assumed to be interpreted as a *key* within the realm of its locality -thus the name *keyscape*-, so the distance between a chord and its context(s) might be estimated in terms of interkey distance. This will have surely many implications in quantification, but we consider it a reasonable starting point.

Precisely, our multiresolution method is intended to address the problem of the variable temporal resolution needed for capturing chords or keys, point which is especially relevant when dealing with real music performances. To illustrate this point, let's consider a typical cadential chord progression in a piece from the *common-practice-period*, which is performed by applying a strong ritardando. Many studies have consider the location of ritardandi in music performance as related to tonal hierarchies of the musical discourse, when they are applied at stable closing points -phrase endings and the like-. But interpreted from our current problem, these ritardandi are also the *responsible* of the hierarchical enhancement of the events, by *promoting* the plain chords towards the status of local key via their longer temporal presence. This effect is visible as a product of the statistical treatment of pitch-classes over time, and it can be appreciated in the keyscape, where strong ritardandi are represented as reaching higher in the plot. It's not rare to find performances in which a ritardando slows down the beat in ratios around two to three or even more, which represents two steps up in our keyscape of around ten levels!. In our intuition, this is a resource that can be -and it's probably been- exploited by performers, consciously or not, to put stress according to their own harmonic interpretation of the piece, particularly in short tonicization nuances.

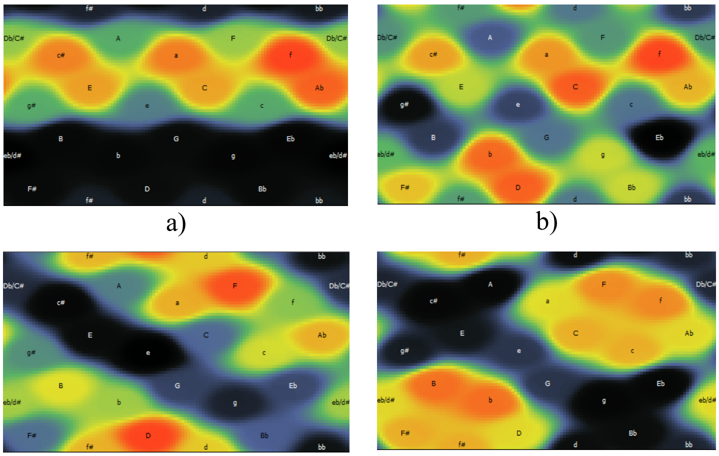
To represent tonal implication in these spaces, different methods can be used. As was introduced before, while some authors advocate for the prevalence of a single key reference in traditional terms, as it's discussed by Werts [41] about Weber's *principle of inertia*, there are empirical evidence of multiple influence of keys [15, 39]. Since these approaches have not been compared systematically -up to our knowledge-, we propose to implement both of them. The first one just considers the position in the geometry corresponding with the strongest tonal center estimation. This fits the traditional *transcriptionist* approach, since we are forcing a single answer for the tonal estimation

[24] Depending on the height in the keyscape, the points in this space might be capturing tonal information related to single notes or chords -at the lowest levels- or different key hierarchies as we move upwards averaging over longer timespans. But in any case a distinction is done to define, for instance, *where* a chord finishes and *where* a key starts vertically. As we will discuss next with an example, the temporal *influence* of concepts such as chords or keys are far from being straightforward, from the perceptual point of view.

[25] Actually, the slight differences led the authors to define their final profiles as a composite of those resulted from tonic chords and strong cadences as contexts.

in terms of standard music theory categories. Additionally, we have developed the idea of *tonal centroid*, evolved from Gómez's solution. As has been said, her centroids are computed by weighting the position according to the correlation strengths with all tonal centers. We have implemented two variants of such centroids, depending on the purpose of the mapping. In the case of their use for visual representation, we apply directly her algorithm over Krumhansl & Kessler's tonal-hierarchy two-dimensional reduction, and an adaptation to the corresponding four-dimensional space has been devised for quantitative account of distances.

To finish this section, just mention two specific problems implicit in these spaces. First, as it's discussed in [15], hierarchical asymmetries are not considered, meaning that the distance between one chord and its local context is measured as equal as if they played the opposite roles, which does not match empirical observations. Additionally, this geometry is intended to capture only perceptual properties of plain major and minor events and contexts, which constraints much of the richness of real music. The way other chord families are mapped in this space might not represent the same perceptual distances, especially when summarized as centroids. To show one extreme case, let's take Temperley's suggestions [36], and have a look to particularly ambiguous events, like symmetric chords[26].



a)

b)

c)

d)

Figure 3.5. Tonal implication for symmetric events. Computed from synthesized piano sounds and represented in the two-dimensional space. a) Augmented chord. b) Whole-tone chord. c) Full-diminished chord. d) Tritone.

As we can observe in the two-dimensional representation, symmetric events present tonal implications distributed regularly all over the pitch space -which is actually based on the asymmetries of major and minor keys-, making useless any attempt of summarize them by centroids. These ambiguities are captured in the four-dimensional geometry by locating the centroids centered at the toroid's dimensions 1 and 2 -which actually doesn't exist in the two-dimensional reduction-. Despite all these four families of chords are encoded differently over dimensions 3 and 4, the distances between them are unlikely to characterize any perceptual reality. Some possible solution to this would require empirical research about distance between enriched families of chords -or contexts- and the derivation of one or several spaces that can represent properly such relationships.

[26] This is not limited to chords. Similar representations arise from computing symmetric pitch-collections over longer timespans, capturing information about symmetric tonal contexts, as it happens for instance with Debussy's hexatonic passages or pieces using Messiaen's *modes of limited transpositions*.

3.2.4 Tonal stability over time and pitch space: *centroidscape*

Once we have a usable representation of tonal implication in a proper geometric space of interkey distances, the next natural step is to project all our multiresolution information (keyscape) into it. For that, we propose to use the centroids we just have described, computed for all resolutions, to conform our concept of *centroidscape*[27]. We hypothesize two main characteristics making our centroidscapes relevant for the stability description, if we attend to their temporal and spatial evolution as music events unfold.

a) Centroids group together according to the proximity of the tonal estimations at different temporal levels -i.e. the vertical lecture of the keyscape-, being this related with the tonal stability of the frame. The interpretation of this assertion follows. If all centroids are close together in a single cluster, the system is in its most stable state. Musically, this situation means that the surface chord is the same as its key context, and this relation is kept up to the global key of the piece, thus, the system is on a proper referential state, from which all other possibilities could be measured. On the other hand, if centroids form different clusters away from each others, this can be interpreted musically as chords away of their local contexts or as passages modulated to other regions. The more centroids a cluster has, the more prevalence of the nearest tonal center over time, which means a more stable segment. This can be measured by observing the height of the different areas -colors- on the keyscape: the *higher* a given tonal center spans, the more *influence* it has over its surroundings[28]. This is illustrated next.

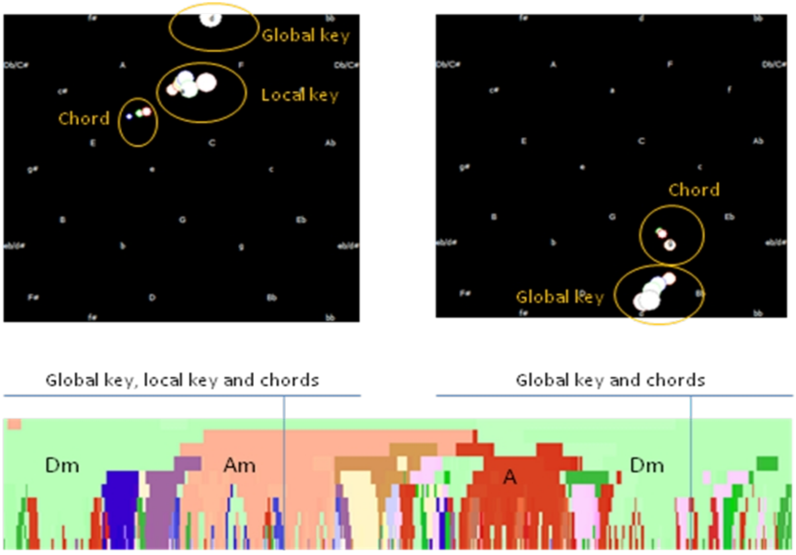


Figure 3.6. Centroid distributions over pitch space. Computed for two different frames of Mozart's *"Dies irae"*. Top left: centroidscape represented over the 2D space of interkey distances (just the relevant portion of the space is shown), for the first frame (left one on the bottom figure). Centroids' size are proportional to height on the keyscape -the larger, the higher-. Top right: centroidscape for the second frame. Bottom: keyscape, with frames marked as blue vertical lines. Color legend: light green = *Dm*, global key of the piece; pink = *Am*; red = *A*. Note that centroidscape's axes represent angular information

[27] We reuse the suffix *-scape* here, since this representation is a mapping of the keyscape. As a difference with keyscapes, however, centroidscapes are not practical for visualization of all the piece as a whole, but for framewise observation of the pitch hierarchies dynamics over time and space. In any case, the suffix is also logical due to the spatial nature of the centroids in pitch space.

[28] Or interpreted from the generative perspective, more stable surroundings contribute to clearer tonal center definition.

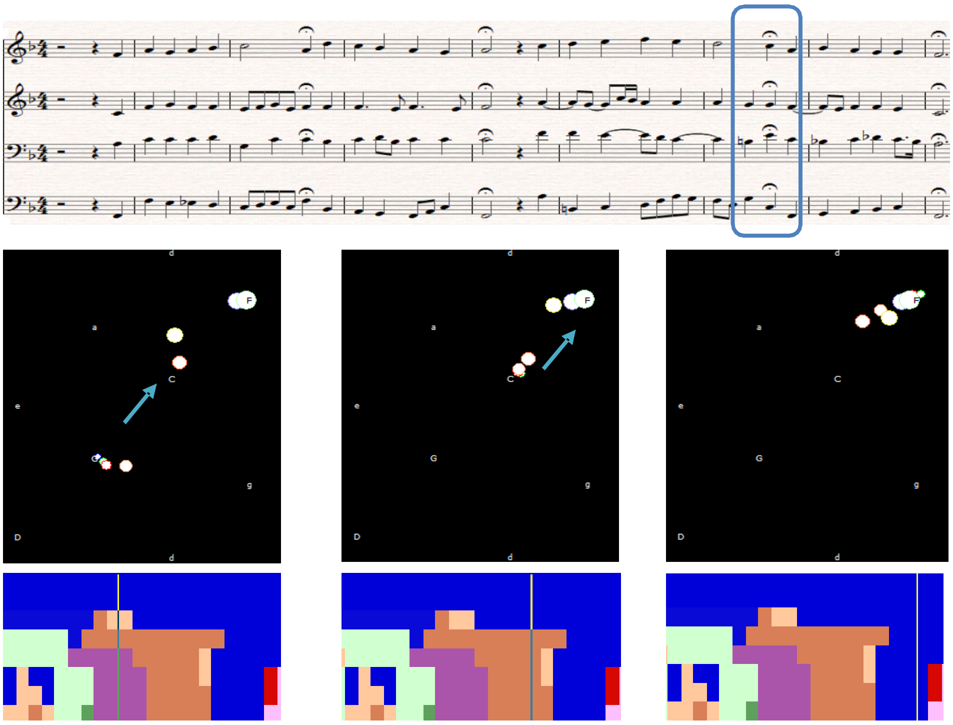
of a toroidal structure, so, top and bottom borders are geometrically the same. Thus the apparent different position of the centroids accounting for the global key -*Dm*-.

In the figure we observe two frames of the centroidscape -weighted centroids, not just strongest key- computed from Mozart's *"Dies irae"*. The first one, taken from the section in *Am*, shows clearly three different regions, accounting for the global key -just represented by one centroid-, local key -pretty much present by means of several centroids- and current chord. For the second frame, in a more stable section of the music, we can only appreciate the influence of the global key -now strongly established, as many centroids stay there- and the present chord.

b) While centroids grouping account for frame's tonal stability according to the location and size of these clusters, if we attend to the temporal evolution of these groups we can actually track the change of stability from one point to another within tonal space, being this directly related with modulating processes. By using centroidscapes, this way, we propose a usable method of observing not just the mere change of tonal center, but the dynamics of these processes as well, by means of *transfer of stability* over time and space. Many music works don't allow us to say unambiguously the exact points where the discourse *has modulated*. Tonal contexts and modulation are complex dynamic processes which evolve in rich behaviors, and thus there is no musicological nor psychological consensus about that[29]. For that reason, we find promising to use descriptions in terms of stability dynamics to assist traditional analysis techniques, particularly in those cases in which talking in terms of conventional categories and zero-time boundaries might obscure the interpretation and diverge from listening experience.

This transfer of stability is observable at any two consecutive *layers* of the keyscape, providing us with a rich tonal dynamics information. One simple -but relevant- example is illustrated in the next sequence of figures.

[29] There exist clear cases and conventions for labeling when a piece of music has modulated to a different region, like the *perfect cadence rule* -if after a modulating process there is a clear perfect cadence, the new tonal center is considered as being established-. This one and other rules, however, are usable only in certain subset of tonal music writing, being mostly useless in many works from 19th Century onwards, in which tonal centers may be passed through -and perceived as it- without real establishment of the tonic chord -neither by cadential nor by any other means-. See, for instance, [21], chapter *Paths in pitch space*.



Applied dominant (*G-C-F*)

Dominant (*C-F*)

Tonic (*F*)

Figure 3.7. Centroidscape for cadential passage in Bach's chorale "Christus, der ist mein Leben". Rendered from MIDI with standard organ patch, covers the final cadence of third phrase and the first chord of the fourth phrase (see score). The hierarchical relationships are visible in the centroidscape dynamics. Left: three groups showing the hierarchical connection between secondary dominant (see footnote 30), dominant and tonic. Center: two groups, after resolving into the dominant. Right: one group, at tonic arrival. Top: score with the three chords highlighted. Center: centroidscape sequence (just three frames and the relevant section of the pitch space are shown). Bottom: keyscape section covering the passage. Color legend: purple = *G*; brown = *C* (light-brown stands for misestimations of *C* in favor of its relative *a*; however, while the strongest key was estimated as *a*, the summarized centroid remains closer to *C*); dark blue = *F*, key of the piece. Cursor in keyscape points to the involved frames.

A typical hierarchical analysis[30] states that the cadence *G-C* at the fermata shouldn't be interpreted as definitive arrival point, due to the presence of the more stable *F*, whose superior status in the hierarchy is shown in the keyscape as spanning from bottom to top (*F* is actually the global key of the piece). Obviously, not all *G-C-F* chord sequences in music are interpretable in this way. This only applies when *F* is a solid stability reference of the music excerpt in a long timespan basis. In the next figure we can

[30] A complete hierarchical analysis of this piece is discussed in detail in [21], including a full graphical representation, and supports our interpretation of this cadential passage in GTTM/TPS terms. Lerdahl's prolongational tree (see Figure 3.8) left-branches *G* chord from the *C* chord, showing the dominant-to-tonic relaxation. *C* chord is interpreted as a right branching (departure) from *F* at the end of the second phrase, which is strongly prolonged (empty circle) into our *F* chord by applying the interaction principle - see *c* to *b* timespan promotion- at the beginning of the fourth phrase. Our use of the term *secondary (applied) dominant* for *G* chord is intentional, since our level of description for stability does not allow us to speak in terms of musical phrases. Lerdahl's analysis, relying on detailed symbolic information, promotes the interpretation of a *structural counterpoint*, and thus he can properly describe *C* chord as *F:I/V* instead of our purely pitch-hierarchical *F:V/I*. Regardless these refinements, our model captures the same essential properties in terms of hierarchy.

confirm this point in Lerdahl's interpretation of such passage -at the highlighted branches-.

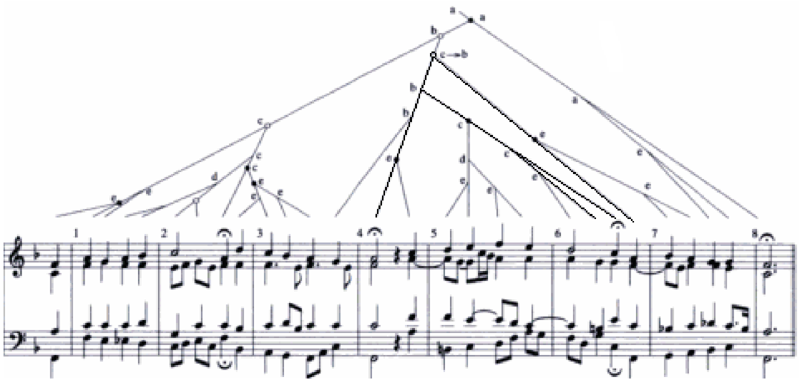


Figure 3.8. Cadential passage in Lerdahl's prolongational analysis of Bach's chorale. The involved branches of the discussed cadential passage (see footnote 30) are highlighted in bold. Adapted from [21], p. 22. Copyright © 2001 by Oxford University Press, pending for permission.

We have found this multicentroid plot over space and time pretty informative about the tonal discourse, including intuitions about tonal stability -see later-, when watched in animation along with the sound.

3.2.5 Metric for tonal stability: *distancescape*

Once we have shown how centroidscapes represent tonal functional behaviors over pitch space, the next step is to use them for proposing a metric of stability through the discussed interkey distance model over which this space is built. For that, we propose different possibilities.

a) *Bottom-up distance*. Measures distances from centroids at the shortest time resolution (tonal surface) to the centroids at the rest of levels, up the hierarchical structure. Depending on which level we look at, we have an estimation of the relative stability between surface events and the different contexts. This way, we can measure the distance from one chord to its immediate tonal reference, or go higher in the hierarchy up to strong or global keys. This metric is thus suggested for capturing local tonic orientation, such as cadences.

b) *Top-down distance*. Measures distances from the largest time resolution (highest tonal reference) to the centroids at the rest of levels, down the hierarchical structure. Depending on the depth we choose, we can reach key nuances at different levels -like strong keys or short tonicizations- and study their relation within wider contexts. Obviously, surface level is accessible from top as well. This metric, so, is proposed to track tonal center evolution.

Both elements of the tonal discourse -cadences and tonal center shifting- are ubiquitous in the referred literature about tonal tension, being also critical parameters in algorithmic approaches, like in TPS. In some perceptual studies, the larger variations in tension sensation have been associated to tonal center shifting in relation with stable segments [16, 39]. Over these main fluctuations due to key shifting, local peaks superimpose, for which Krumhansl suggests that might account for melodic contours or local dynamics. To some extent, this *additive* behavior is also proposed by Lerdahl's pitch space distance model, which is built as a linear combination of the contributions of locations of tonal centers in pitch space and local relations within regions. In our third case study -see Evaluation chapter next- we will discuss this further.

c) *Averaged upwards or downwards distance*. Computed by averaging all curves from one of the previous possibilities. For many applications we might be interested in just one curve, and this metric provides a sort of summary of the overall behavior, in cases of prominent accumulation of curves, as happens in cases of clean and well-balanced keyscapes. The interpretation of such single curve in terms of stability, however, is far from direct and will depend on the specific scenario.

d) *Cumulative upwards or downwards distance neighborwise*. Accumulates distances (bottom-up or top-down), as in the previous version, but in an level-incremental way: from one level to the immediate next one, instead of use just one fixed reference. We suggest this metric as related with the mentioned additive contribution of tonal shifting and local chords in a given region, and it also provides global metric of distances between all levels. In cases of clean keyscapes, for which few clusters of centroids account for each category of chords, local keys (if any) and global key, it represents the overall sum of distances between immediate categories, since the sum of distances within each cluster will be minimum. However, it could lead to strong peaks in cases of residual isolated misestimations at middle levels, since these noisy distances will be reckoned twice.

3.3 Conclusions related to the model

To finish this section, we will summarize the main identified capabilities and drawbacks of our model, all of them mostly derived from the same conceptual assumptions taken.

3.3.1 Capabilities

- It operates directly from audio signals and relies upon a high resolution chromagram as the unique descriptor.
- Tonal implication can capture weighted key induction and ambiguity.
- Temporal multiresolution tonal description uses a homogeneous technique for averaging, leading to a uniform representation of tonal information.
- Different levels of tonal contexts are also *uniformized* by the sliding window policy, which allows a vertical lecture in the same terms between any two layers.
- The distance model is also homogeneous, providing an interpretation of tonal distances in the same terms.
- Keyscapes facilitate the lecture of the tonal hierarchies along the musical discourse over time. They provide *static* visual intuitions about it.
- Centroidscapes map our information into pitch space, establishing the relationships between hierarchies by a metric of distance. They provide *dynamic* visual intuitions about hierarchical tonal discourse and tonal stability, when visualized along music.
- Distancescapes provide a configurable variety of measurements, accessing to global and local tonic orientation information, and their combinations.

3.3.2 Drawbacks

- Tonal implication is fallible respecting conventional categories of tonal description.
- The homogeneity of the method don't allows for parameterization. All tonal categories are described equally as a *key*.
- Tonal context definition considers equally past and future, which is the source of several representational problems in keyscape.
- The rationale for mapping tonal implication into the distance model is rather indirect, since it's applied over already indirect models, complicating any interpretation. Geometric spaces don't allow for asymmetries between an event and its context.
- All our model operates in a *framewise* basis. We don't know about empirical account of listeners' *framewise* perception of *tonal* music to be compared with[31].

[31] Pressnitzer *et al.* [31] have studied the relationship between the perception of timbral tension and psychoacoustic roughness in a *quasi* frame-based fashion, due to the nature of the music examples used in the experiment, consisting on orchestrated chord sequences not inducing sense of tonality -actually, taken from a *non-tonal* piece-. This kind of music, although it can be obviously followed event by event, promotes its listening in terms of timbre, which is more *framewise* related.

CHAPTER 4.- EVALUATION OF THE MODEL

In this section we will discuss some functionalities of our model of tonal stability. As it has been stated a number of times in introduction and literature review chapters, evaluation of systems which aim to model perceptual and cognitive human behavior should always be taken with extreme prudence. Since our model is not an exception, this chapter is organized as follows. First, an introductory section will cover particular aspects of evaluation in our case. Then, three case studies will serve for discussing the model in the context of musically related applications, namely, cadence finding, tonal tension peaks modeling and stability-based structural analysis. The section finishes with some comments about robustness.

4.1 Evaluation issues

One of the main motivations of MIR applications is to achieve full automatic algorithms applied to massive music collections, and our technique is proposed precisely to manage audio signals, ideally from commercial recordings. Notwithstanding this, for the moment, our work aims for contribute to the discussion about tonal stability, by means of proposing and interpreting one model which integrates several of the current paradigms in the field. The problem of tonal stability -or tension- has no consensus even for its definition, however we identify some applications strongly related with the concept, through which we can test indirectly our algorithm. At our stage of research, we find more fruitful a cross-discussion over these applied scenarios, in particular covering the similarities and differences of the involved models, than mere statistical evaluations in terms of *accuracy* over arbitrary databases. One fundamental aim in any science is to find models which can accurately describe and predict the phenomena under study. However, here we have two main arguments discouraging it.

First, the scarcity of empirical ground truth to compare our model with. For that, several studies focus on very small and specific musical stimuli -often, just few bars of a single piece-, which are normally tested with few highly musically trained subjects in controlled situations [1, 5, 15, 16, 23, 39]. These works provide the required empirical evidence, obtained by expert teams, and they are critically discussed. However, as most of these authors warn, the validity of these finding in terms of statistical significance should be taken with extreme prudence, especially regarding their use as ground truth references for general applications, involving average listeners and more realistic listening scenarios. Obviously, much more care is required to interpret this evidence as solutions of inverse problems in human cognition.

Additionally, other point should be stressed about comparing different models, especially when quantitative metrics are involved. This point has been vividly touched by Meyer, regarding *primary* and *secondary parameters* in research methodologies and the hazardous temptation -and practice- of derive *explanations* from quantitative correlations of curves obtained from them without appropriate supporting hypotheses [26]:

> “[...] quantification must occur in the context of hypotheses that are *explicit* and *specific enough* that the predictive inferences derived from them can be tested -provisionally confirmed or definitively disconfirmed.”

> “[...] "tension" [...] is only loosely defined and [...] the roles of the various musical parameters (especially those that are not pitch-specific) that presumably create it are not carefully isolated and delineated.”[32]

As we have largely stated along previous chapters, our hypotheses don't provide the required *direct* connection between our model's parameters and Lerdahl's or Krumhansl's, whose proposals despite many efforts are also notably supported over mere intuitions. Thus, we can only suggest a qualitative correlation regarding the explicitly shared aspects of all three models: the relative stability of events in terms of hierarchies and distances in pitch space. Although these concepts are present in all the models, they are embedded differently by their paradigms and representations, and we find very unlikely the possibility of isolating them from the whole, for objective comparison.

Taking all this into account we conclude that, for the scope of this dissertation at the current stage of research, the best ground truth and evaluation methodology is the discussion itself in terms of what we know about our model's description capabilities. This is also what motivates us to choose our music materials in the next sections. For the first two applications, we selected the same examples used for the main contributors in the field dealing with similar problems. The large discussion covering such examples from different models is a good context in which insert our perspective. The choice of examples is not casual. For both first case studies, experimental methods have been carried out by expert psychologists to measure human responses to those musical stimuli. Additionally, from all the analytical possibilities available at the theoretical side of the problem, we will use Lerdahl's approach because its hierarchical nature, although diverging in many points, is closer to our concepts.

Up to what extent our algorithm might capture relevant information about perception of tension in the rich variety of real music experience scenarios, is something that necessarily requires further research. At the current stage of this study, we believe it would be dishonest to try a quantitative account of perception of musical tension, and too pretentious to point out how these processes might be operated by human cognition. If this work might contribute somehow in this sense, is just as suggestions which could be exploited in future studies, when developed enough as hypothesis and proper methodologies, including the required quantitative treatment.

4.2 Case study I. Cadence finding

Bach's chorale "Christus, der ist mein Leben" will serve our purposes, due to the nature of this kind of composition, the deep analyses available in literature and the availability of empirical data regarding tension perception. In [23] Lerdahl & Krumhansl modeled the hierarchical tension for this piece and compared it with subjects' tension ratings, obtained by stop-and-rate and continuous-rate experimental setups.

Since this musical style is strongly chordal and cadential, and its harmonic discourse does not deviate much from the stable tonal reference, we will look for those surface events which reach the top of the hierarchy all along the pathway up in the keyscape. For these events, our centroidscape will tend to group all resolutions into one single cluster, located over the global key -as in Figure 3.7c-. In the keyscape, this is

[32] [26], p. 468-469. Stress (cursive) is mine, quotations on *tension* as published. These comments are done in the context of a critical review of the first movement of Mozart's K.282 analyses by Narmour, Lerdahl, Gjerdingen, Bharucha, Krumhansl and Palmer, who apply and cross-cite different models related with tension and stability. Since we will use the same music material and two of such approaches, we find this assertion particularly well suited.

represented as the same estimation spanning from bottom to top. We have several possibilities to capture such conditions with our descriptor. The one proposed here is to use our bottom-up distancescape, for which distances from bottom to all the rest of levels are expected to be minimum at the points we look for. To enhance the discrimination of those events, we propose a cumulative version of the distancescape, by adding all curves, and we use centroids forced into the strongest estimation[33]. So, our most stable points will be located at zero values -all centroids exactly over the global key of the piece-.

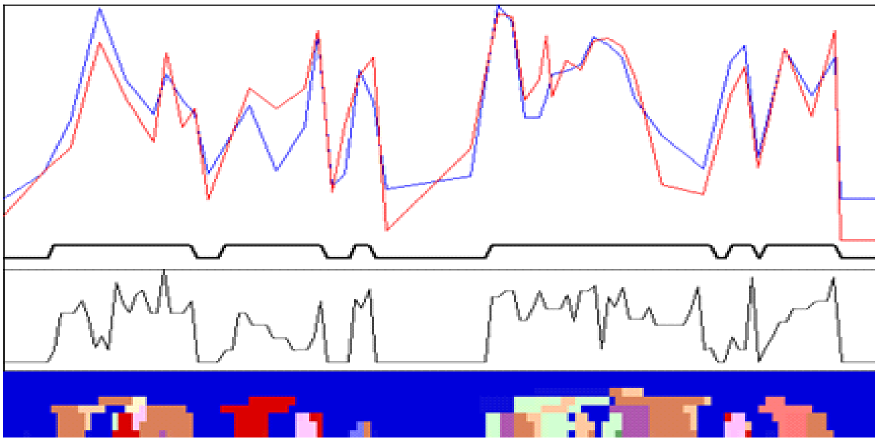


Figure 4.1. Cadence finding in Bach's chorale. Top: Lerdahl & Krumhansl hierarchical tension (including attractional influence) prediction (blue) and subjects' judgments (red). Numerical values from Lerdahl's personal communication, duration information has been added to the raw data -eventwise without rhythm-, in order to be aligned properly with our sound events (here, at onsets). Center: discrimination function for cadence finding (in bold), consisting on locate the zeros of the cumulative *bottom-up distancescape* computed from strongest tonal center estimations (just below). Down: keyscape. Color legend: dark blue = *F*, key of the piece. Audio was rendered from a MIDI version of the score in [21], p. 22, using a standard organ patch.

As we can observe in the figure for this *easy* task[34], tonic arrivals were captured as zero value in the distancescape (center), which corresponds with maximal stability points at strongest local minima in both Lerdahl and Krumhansl's tension predictions (top, in blue) and experimental ratings (top, in red). This is coherent with Lerdahl's prolongational analysis, from which predicted tension is derived, as it's shown in the next figure.

[33] Other configurations, like top-down distancescapes or weighted centroids, provide similar results. Also notice that the *cumulative* version is just proportional to the *averaged* one introduced in the last chapter. In the figure all the distances are shown, showing the final curve, used to discriminate our events, on top.

[34] We qualify it as *easy*, due to the optimal conditions of this example. This includes metronomical timing, organ patch for rendering (to which our chroma descriptor performs the best), and the properties of the tonal discourse of the piece, which promotes the easy location of tonic arrivals -both theoretically and perceptually-. This is observable in the clarity of the keyscape, in which the most stable points span from bottom to top. Dark blue stands for *F* estimations, key of the piece.

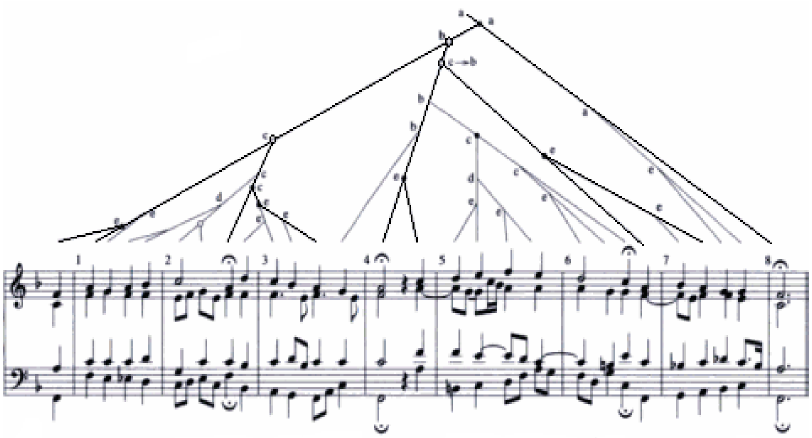


Figure 4.2. Cadence finding in Lerdahl's prolongational analysis of Bach's chorale. It shows (in bold) the involved branches captured by our model, all of them corresponding with prolongations -whether strong or weak- of the global pitch event reference of the piece (last *F* chord). Adapted from [21], p. 22. Copyright © 2001 by Oxford University Press, pending for permission.

All the events captured as zero in our distancescape correspond in Lerdahl's interpretation with prolongations -weak or strong- of the global pitch reference of the chorale (*F*). Just one weak prolongation was not captured, at first beat of third bar. While the surface was defined correctly by our algorithm as *F* (blue in the keyscape), the influence of the previous *Bb* chord (in red) was estimated as stronger when larger timespan was considered, probably because of the contribution of pitches *Bb* and *D* in the next beat (our windowing policy is symmetric). This might be related by the fact that this theoretical prolongation is not visible in the perceptual ratings as a strong relaxation. A similar effect accounts for the slight delay in capturing the cadence at the begining of fourth phrase, due to the prolonged influence of the previous *C* chord. This serves to illustrate one characteristic drawback of the average-and-correlate method when it's applied in multiresolution, which is to produce this sort of residual tonicizations. Since in this case our chroma information was pretty clean, this also serves to show why mere statistical treatment of pitches and durations does not account for the theoretical concepts of key we normally manage, and the rationale behind reported *inaccuracies* of key finding algorithms based on this method.

Overall, this example does not seem too impressive if we attend to the relatively low complexity of the music and the artificial conditions used to generate it. However, we want to point that the algorithm captured branching decisions in terms of tonal stability -actually, just prolongations of the main reference-, which in our opinion is a rather *high-level* information, and this was performed directly from audio just from chroma information. Neither note segmentation nor labeling of any kind was used.

4.3 Case study II. Tonal stability peaks modeling

For this task, the first 8 bars of Mozart's K.282 will be analyzed. In [21] Lerdahl discusses in depth this excerpt to illustrate his tonal tension model, and detailed computation of all steps are provided, covering sequential, hierarchical and attractional tension contributions. Additionally, experimental measurements of tension were described by Krumhansl [16]. This music example is significantly more complex in terms of tonal discourse than the previous Bach's chorale. In particular, the excerpt includes the presentation of the first theme group of a sonata-allegro structure[35] and the transition leading to the second theme group. In order to speak in the most approximated terms, we will compare a version of our top-down distancescape with just the hierarchical contribution to tonal tension from Lerdahl's model[36]. Because of the notable differences between both models, we will first try to introduce in which terms they could be compared.

Lerdahl's hierarchical tension computation requires a prolongational structure to model how events tense and relax into each others, and a model of distances in pitch space to quantify the amount of tension being involved in these *transactions*. As was introduced in last example, our distancescape might be able to capture rough branching decisions in terms of strong stability. Additionally, now we have to deal with a general tonal shifting from the beginning to the end. This could be a proper context in which apply our top-down distancescape. Since it measures the distance between the global key level and the rest of resolutions, it can access to regional shifting, as well as reach the surface to find those strongly stable events. Due to the characteristics of our keyscape, in which chords and keys are well balanced in vertical representation, our distancescape presents an accumulation of curves -i. e. several centroids clustered or several distance curves accounting for the same hierarchy-, and so, we decide to use an average of all these curves -actually, just the sum, since it's proportional and we don't care now for absolute quantification-. Distancescape was smoothed by applying a leaky integrator of 1 sec. half decay, corresponding with one beat of the music[37]. As a last parameter, we use just the strongest estimation for locating our centroids in pitch space, so as to measure distances in the most theoretical way as possible, since it's the same categorical approach in Lerdahl's model. This way we also guarantee minima near zero for very stable events.

Both models are compared in the next figure. The audio signal used was rendered from a MIDI version of the sonata (first 8 bars, full score, not reduced), using a standard piano patch and 1 sec/beat metronomic tempo. Lerdahl's hierarchical tension values from [21] (pp.157-158) were aligned by applying the corresponding temporal axis, in this case at event onsets.

[35] Even when this movement is not an *allegro*, we find more musicological sense in using the term *sonata-allegro structure*, than the traditional expression *sonata form*, which leads to easy misinterpretations.

[36] Lerdahl's attractional tension modeling is far beyond the capabilities of our descriptor, due to the involvement of individual voicing.

[37] In [16], where Lerdahl's tension predictions were compared with experimental ratings, Krumhansl suggested that listeners *integrate* their tension perception over time, producing smoothness and delay. Our integration over one beat is arbitrary, just based on reasonable visual smoothing without too much delay, and no attempt for best correlation has been performed.

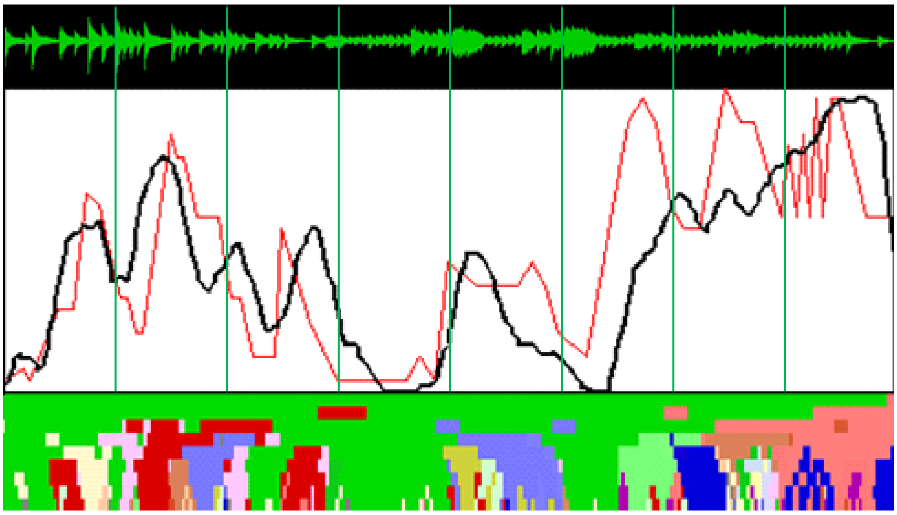


Figure 4.3. Tonal stability for Mozart's K.282 compared with Lerdahl's hierarchical tension. Mozart's K.282, bars 1-8 of the full score (not reduced) rendered from MIDI with standard piano patch. Top: audio signal, showed for alignmet with vertical green lines at bar separations. Center: cumulative top-down distancescape (bold black), computed from strongest tonal centers, compared with Lerdahl's hierarchical tension analysis (red) (adapted from data in [21], pp. 157-158, for temporal alignment of original event-based values, here fixed at event onsets). Down: keyscape. Color legend: green = *Eb*, key of the piece.

By observing the peaks and valleys provided by both algorithms, qualitative correlation appears at least in the first five bars. In our distancescape we can appreciate generally the same three big peaks in the first three bars, a similar decay to minimum in fourth bar and an additional elevation followed by decay to minimum in fifth bar. This looks approximated even in the relative peak's height (both curves were normalized to maximum ranges). The last three bars show more difference, but a general elevation -with no strong relaxation- can still be appreciated in both curves, corresponding with the modulation towards the second theme group.

Apart from the non accurate alignment, partially due to residual high level estimations, we observe a noticeable difference respecting the peaks' shapes in both models. Theoretical tension curve raises faster and decays slower than distancescape's estimation. Lerdahl's model interprets the beginning of some peaks -at the first, second, sixth and seventh bars- as left branchings of the next strongly stable point -the next valley-. So, maximum tensional value is assigned to first events, being this tension progressively released as we reach the main branch, which is interpreted as the expected arrival for all the events in between. Our algorithm, however, considers a balanced windowing strategy for all levels. All averaging windows are centered on the present-frame, defining higher contexts in a balanced way from past to future. This *democratic* decision does not provides us with the flexibility of bias the tension estimation towards more inertial or more expectative stable states, and consequently our peaks are more *symmetric*.

On the other hand, empirical results seem to point to the opposite behavior in listeners perception of tension. In Krumhansl's analysis of subject's ratings of tension for the full sonata movement [16], she suggests that tension peaks tend most of the times to be asymmetric, but increasing gradually and falling rapidly, resembling a more inertial accumulation of tension. Our algorithm could capture these different behaviors by modifying the windowing strategy, as in the figure.

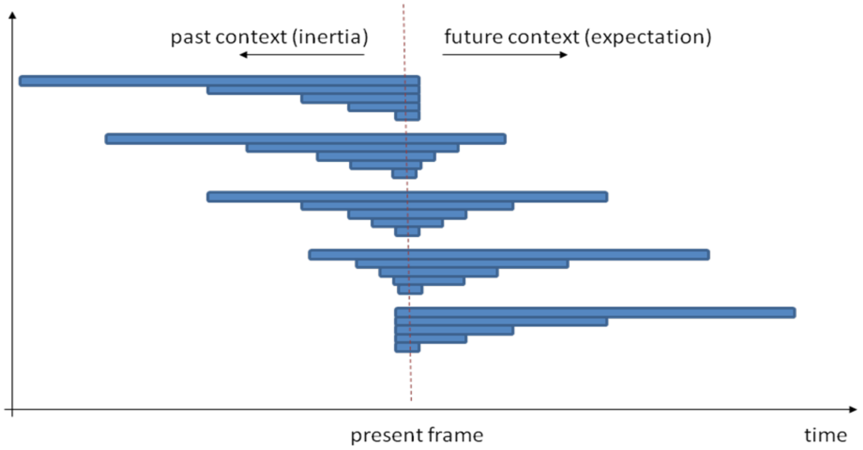


Figure 4.4. Past-future balance in sliding window strategies. Five sliding window strategies for a five-resolution keyscape, allowing for different models of tonal context description. Whatever the chosen strategy, all windows evolve framewise simultaneously. Top window set maximizes inertial modeling, i.e. context only considers present and past-. Central window set corresponds with the equally balanced context discussed here. Bottom window set maximizes expectation, i.e. context is built only from present and future. Between them, different degrees of bias.

This way, we could model our contexts only from the past or from the future, or even balance it gradually according to our analysis needs, to allow our distancescape to measure stability in a more or less inertial or expectative way. Actually, the top one was implemented in our initial stages of research, providing a purely inertial context description, which could be used in real-time applications for which we don't know the future. Despite this model captured properly a past-based context, we lost intuitiveness for interpreting keyscapes, much particularly after modulations -as expected-.

We don't pretend from this single example that our model can be comparable to Lerdahl's hierarchical tension for other cases -even for similar ones-. As was said, they only share very general principles -hierarchical tonic orientation and model of distances in pitch space-, but they operate very differently. Apart from the evident contrast in terms of representation and *accuracy* of information, and quantitative issues of the pitch spaces, we interpret their main difference as follows. TPS's model computes tensional states between consecutive events, and builds up the general tension discourse by inheritance, following the hierarchical patterns of the prolongational structure. Our algorithm, on the other hand, models the stability *directly* for each surface event by a vertical lecture of the keyscape, which contains all the contextual information required. It considers past and present events, but *digested* in the contexts. Thus, the impossibility of compare even single components of both models.

Let's contrast now Krumhansl's perceptual measurements of tension for the same music excerpt with other kind of distancescape. In this occasion, empirical measurements are not sustained over a model from which we can compare, but still we have to decide a choice of parameters for our algorithm. In the last chapter, we described a potential distancescape called *cumulative neighborwise*, which accumulates distances between adjacent levels in the keyscape. Since some sort of cumulative contribution of key shifting and local tonal variations was suggested by Krumhansl [16], we propose a similar interpretation with this variant of distancescape. Additionally, we will use a more general approach for our representation of tonal implication by means of weighted centroids, since there is no specific requirement to reduce it to a single tonal center, allowing this way for tonal ambiguity influence. The comparison is shown in the next figure.

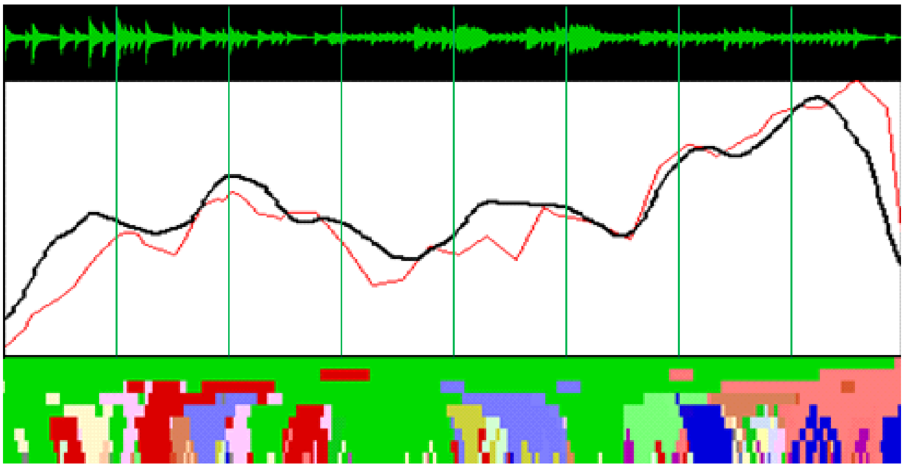


Figure 4.5. Tonal stability for Mozart's K.282 compared with tension ratings. Mozart's K.282, bars 1-8 of the full score (not reduced) rendered from MIDI with standard piano patch. Top: audio signal, showed for alignmet with vertical green lines at bar separations. Center: cumulative neighborwise top-down distancescape, computed from summarized centroids, compared with Krumhansl's tension ratings (data from Krumhansl's personal communication, adapted for temporal alignment -see text-). Bottom: keyscape. Color legend: green = *Eb*, key of the piece.

The perceptual tension ratings curve, from those described in [16] (data from Krumhansl's personal communication), was the same used for comparison with Lerdahl's tension predictions -by averaging across several experiments-. According to the author, ratings for all three experiments agree well enough to perform this average. Due to the nature of the music example used in those experiments, whose original source was not available at the moment of the present work, the data have been adapted in order to be properly aligned with our sound signal. The original music was a synthesized version of a real performance, including dynamics and expressive tempo, and tension ratings from subjects -moving a slider- were registered during nonstop listening every 250 ms. Precise timing for beats and approximated for the rest of events were facilitated by the author -error due to sampling rate-. We select just the tension value closest to the onset estimations for each event. After that, we aligned those tension rates according to our metronomic timing. As was pointed by Krumhansl in her fourth experiment [16], the manipulation of the performed tempo -from expressive to metronomic- had little effect on tension judgments for the excerpt considered here[38]. Since we don't attempt any quantitative approach, we think this approximation can be reasonable and there is no need of refined interpolations.

Visual comparison shows an overall rough similitude between both profiles, in terms of general trends towards peaks and valleys. In this case, we cannot derive a proper interpretation in terms of how our algorithm fits subject's perception of tension, not just because there are no models to compare with, but especially since these ratings were not devised to capture specifically tonal tension, but general tension -moreover, we avoid the term *tension* in our model-. Again, we don't pretend that our algorithm is capable of

[38] The expressivity of this specific audio source was the focus of the analysis by Palmer [29] -same issue of the journal-, where annotations of dynamics, tempo and pedaling where registered directly from the performance on a digital piano. In the first 8 bars of the piece, the most remarkable changes were a light ritardando on the last beat of bar 3, and a strong one -more than twice the duration- in second beat of bar 8.

capture musical tension from just this single example. Our proposal here is just to show their general similarity, which could be explained in part by our conception of tonal stability.

4.4 Case study III. Stability-based structural analysis

Let's move to a more informal case. For this task, we propose to analyze a commercial recording of the famous *Ciaccona* for violin and continuo, long attributed -wrongly, as it's assumed nowadays- to Tomaso Vitali, in this case with organ accompaniment. This piece consists on a large series of variations over a theme, based upon an eight-bar melodic-harmonic schema subdivided in two semiphrases. Harmonic progression, however, is leaded by a descending Phrygian tetrachord *ostinato* falling into the dominant every four bars, and the same metric -four bars- is the timespan of most of the variations[39].

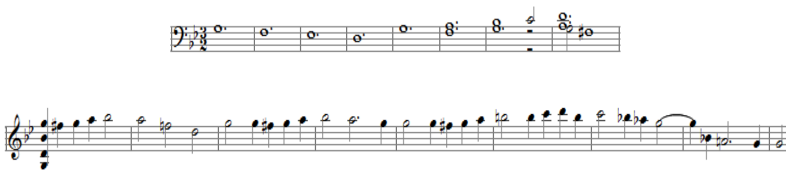


Figure 4.6. Harmonic progression and theme of *Ciaccona* attributed to Vitali. Top: bass descending tetrachord and first elaboration. Bottom: theme.

Apart from the many expressive and virtuosic variations, musical interest is increased by moving over different keys after few cycles of the theme. Let's look at the raw information provided by our distancescapes, computed for the complete piece, which lasts about nine and half minutes.

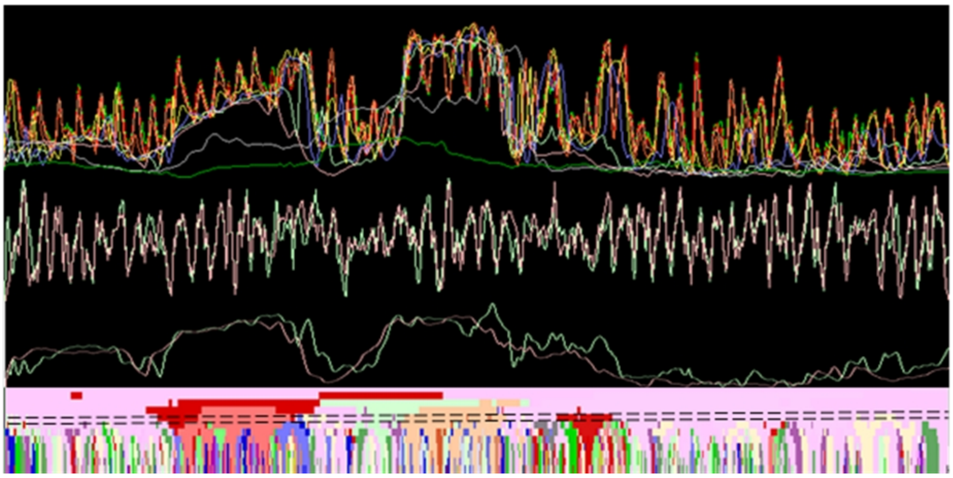


Figure 4.7. Stability based structural analysis of *Ciaccona* attributed to Vitali. Commercial recording performed by violinist Jascha Heifetz. Top: raw *top-down* distancescape, showing all the distances from highest level. Center-top: *bottom-up* distancescape for just two levels at local key timespan resolution, showing local stability oscilations without key shifting influence. Center-bottom: *top-down* distancescape for the same two levels, showing key shifting contribution isolated from local stability variations. Bottom: keyscape. Color legend: light purple = *Gm*, global key of the piece, dark pink = *Bbm*, blue = *Fm*, orange = *Am*. Horizontal black dashed lines show the two resolutions used for local key discrimination. Weighted centroidscapes -not just strongest estimation- and a leaky integrator (1 sec. half-decay) for smoothing were used for all cases.

[39] Only those variations promoting clearly a melodic line span over eight bars, however the four-bar harmonic structure is kept.

In the top-down distancescape we can observe an accumulation of curves forming clear peaks and valleys rather regularly, corresponding with the resolutions at chord level -which in this case spans about four-five levels in keyscape, due to the tempo of the piece-, so representing the distance between those surface chords and the context of the global *Gm*. Every cycle on the harmonic progression -a complete descending Phrygian tetrachord on the bass line- is here represented by one of such peaks, being the most stable point at the beginning of the progression. Additionally, we can appreciate the different regions to which music modulates by a general elevation of the distancescape, over which we can still trace the evolution of the different variations within these keys, by means of peaks and valleys superimposed. We didn't choose here a summarization of all curves, in order to observe the different contributions to stability, as they are visualized in figure's central plots and discussed next.

Local stability -that operating around local tonic orientation- can be extracted by a *bottom-up* distancescape which only considers the proper resolutions (marked in keyscape as horizontal dashed lines), as it can be seen in center-top plot. Another way of isolate this information would be to consider a *center-down* distancescapes, in which all lower levels -chords- would be measured against the mentioned references capturing key shifting. By this selection of curves, we remove the key shifting influence -a sort of *DC level* of the signal-, but still accessing to short-term stability information. In other words, chords are measured only within their local key reference. Or interpreted from our generative perspective, this local key is the consequence of grouping these chords in a mid-term scope, being this resolution the averaging window used at the considered levels in the keyscape.

On the other hand, these same levels measured from top -*top-down* distancescape just for the levels isolating key influence-, provide a representation of key shifting without local variations, as it's shown in center-bottom plot. Since here we don't consider chord levels, local stability -local tonic orientation- is not captured.

This example serves to illustrate one main problem of our model. In our selection of levels to describe key shifting by top-down distancescape, less represented keys -like *Fm* (in blue)- were not captured for requiring shorter averaging windows to be accessed[40]. If we want to capture them by selecting lower layers in the keyscape, however, we cannot avoid to include many *chord's tops*, at similar temporal resolutions, introducing more noise in our attempt of isolate just the local key domain. Actually, this is not just a specific problem of the model, but a consequence of real music's properties. As was discussed before, even time-related music nuances like *accelerandi* or *ritardandi* can contribute to perception of tonality, and in general the levels of reference might vary along the piece. This drawback is well-known among the community working on key and chord extraction from audio, for whom segmentation and temporal resolution are core issues.

Notwithstanding this, our multiresolution technique is at least able to identify those temporal resolutions a piece need to be described by, so we expect future improvements in terms of automation. In this respect, we have tested manually the benefits of increasing the number of temporal resolutions just around the problematic area interfacing chords and keys. Doing so, we obtained better balance between both hierarchical categories, which allowed us to discriminate chords from local keys spanning differently up the keyscape. We believe this interacting area is the most

[40] Since this local tonal reference doesn't last long in time -just few chords at surface-, it does not reach higher on the keyscape.

relevant for tonal analysis, as it's somehow supported -of course, considering the extreme comparison- by Schenkerian practitioners, who achieve the most interesting contributions by focusing in middle-ground understanding. Actually, it's around those resolutions where occur the interplay between tonal categories, and their temporal nature suggests a connection with Janata's comments about the mutual influence of short- and long-term memory [13].

It's also evident from the plots that raw distance curves are quite noisy, particularly during modulating processes[41], but still we can observe how stability related information is encoded by the model. At our present state of research, this model is not intended yet for precise structural analysis of music, but we expect to refine it in near future to do so. What we find really interesting of such kind of representation is not just the possibility of segmenting music, but the potential for accessing *inside* these segments in terms of meaningful sensations -stability-, which is one of the motivations of this study.

4.5 Robustness considerations

Next, just few words about robustness of the proposed model. Robustness can be considered in many different scenarios, so in particular, we will discuss here how the algorithm is expected to perform with different versions of the same piece, which keep untouched the tonal discourse. We are, so, interested in compare keyscapes in terms of tonal content description, and distancescapes in terms of stability. Here, we will consider two kinds of alterations: timbral changes and transposition.

For the first case, we will use the third movement (Presto) of Vivaldi's *Violin concerto* Op.8, nº2 "*Summer*", in radically different versions by Karajan and Joe Satriani commercial recordings. Both versions are strictly respectful with the score, however, Karajan's use of a baroque-sized string chamber orchestra, and Satriani's powerful distorted guitar, bass and drums (including filling passages over whole chord durations), contribute to pretty different timbral qualities of both audio signals.

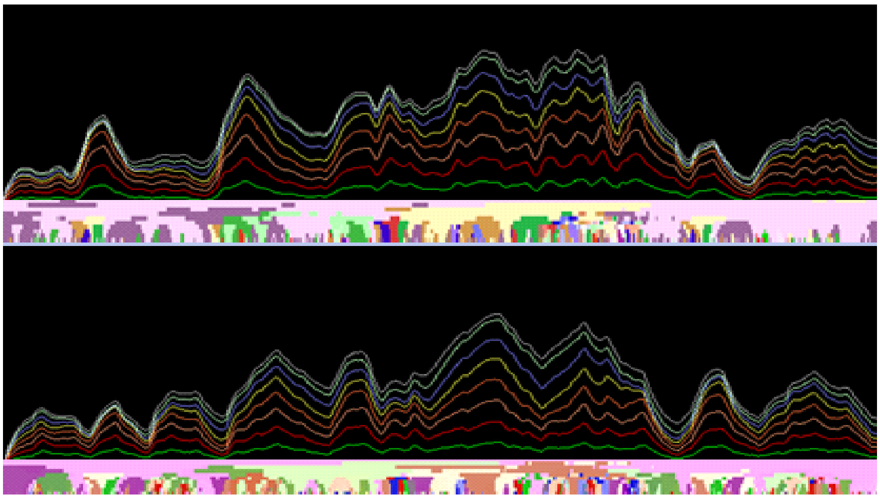


Figure 4.8. Robustness to timbral variety. Keyscapes and top-down cumulative distancescapes computed for the third movement (Presto) of Vivaldi's Op.8, nº2 "*Summer*". Top: Karajan's version. Bottom: Satriani's version.

[41] In this case, there aren't sofisticated modulating processes, and key changes occur abruptly from one variation and the next.

Both versions can be compared in the figure. We can observe many differences on the tonal estimation on the keyscapes -which only show the strongest estimation-. However, the essential behavior of the stable points were well aligned[42]. Despite its highly distorted signal, Satriani's version resulted to provide better definition of the structure of the piece, which contrasted sections change at valleys and some of the peaks in distancescape. Actually, chroma information was cleaner -less noisy- in this version, for which probably the powerful bass lines are much responsible. This is the case of the third peak clearly visible on Satriani's version, but almost flat on Karajan's, accounting with the respective prominence of the bass line in the audio. It seems logical to think that our stability descriptor will be robust in similar terms as the underlying tonal information representation is also robust. That is, it will depend on the capability of the correlation method to derive proper locations of centroids, according to the harmonic content of audio signals as captured by HPCPs (which in our implementation are tuned to be robust to timbre, through spectral whitening among other mechanisms).

Since distancescapes are computed from the relative positions between centroids in tonal space, it's expected that this metric will be robust to transposition. This point is illustrated next.

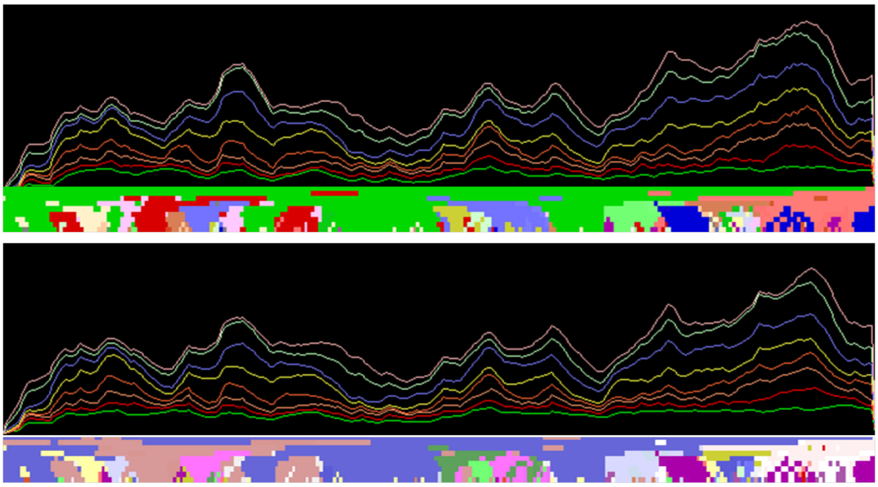


Figure 4.9. Robustness to transposition. Top-down cumulative neighborwise distancescape and keyscapes computed for Mozart's K. 282, bars 1-8. Top: Original key (*Eb*). Bottom: Transposed to *E* (half-tone up). Transposition was performed in MIDI and rendered with the same piano patch as the original version. Color legend: green = *Eb*, global key of the first version, blue = *E*, global key of the transposed version.

In the figure we observe distancescapes computed for two different versions of the same music material -first eight bars of Mozart's K. 282-, being the second one transposed a half-tone up (so, changing the key from *Eb* to *E*). As we can see, all the estimations on the keyscape were changed accordingly to reflect the new situation. This change is notable when mapped into our space, since a half-tone away is quite far in the toroidal geometry. However, the distancescape showed minimum differences, since centroids' absolute location is not relevant here.

[42] Satriani's version is about 12 seconds shorter. Our check by listening of the events at local minima shows quite accurate alignment between both versions.

CHAPTER 5.- CONCLUSIONS AND FUTURE WORK

To finish this work, we will summarize the main conclusions regarding our model, pointing to some improvements and future work which could be devised from here.

5.1 Contributions of this work

According to our proposed goals, the contributions of the present study include:

- Critical literature review regarding tonal stability modeling from audio. Several paradigms have been abstracted to model a homogeneous idea which could be manageable by different disciplines, in order to allow a proper cross-discussion. This has included music theoretical concepts, empirical evidence from cognitive psychology research and implementation concerns of audio technologies. The concept of tonal stability was modeled around the idea of fitting -closeness- of tonal events and their contexts.
- Design of a tonal stability model from audio in those terms. Our main building blocks were audio chroma features, temporal multiresolution tonal estimation -including ambiguity and summarization- and perceptually related pitch space geometries and distances. Our principal contribution here is the extension of Sapp's keyscapes into our idea of centroidscapes and distancescapes.
- Implementation and discussion of such algorithm. We stressed the discussion of the model and its building blocks, avoiding non-intuitive mathematical treatment and trying a balance across three perspectives -music theory, psychology and technology-. The algorithm has been implemented as an extension of an audio and music analysis tool (SMS Tools), which includes computation and visualization of its main components.
- Qualitative evaluation and discussion of the functionalities and limitations of the model, through three case studies. The potential applications were cadence finding, tonal tension peaks modeling and stability-based structural analysis.

5.2 About the stability descriptor and music

The model generally performed according to our expectations. As it has been stated all along the discussion, the main capabilities and the main drawbacks are due to the same conceptual assumptions and technological concerns. This way, the advantages provided by a homogeneous method for capturing information at each step (efficiency of average-and-correlation method, simplification of the representation, vertical alignment of multi-level tonal contexts, and Euclidian metrics in pitch space) are also the responsible of the main undesired side-effects (tonal misestimations, residual by-products in keyscape, deallocation of centroids in pitch space, and a general difficulty for interpretation of distancescapes).

The model has proved to capture stability related information in some specific scenarios. Strong cadences seem feasible to be modeled in cases of non-modulating music, for which keyscapes are likely to be clean. Roughly, this behavior seems to correspond to some perceptual evidence about tonal tension, and to be consistent with tension paradigms from music theory, particularly around the most stable points -minimal tension-. General modulations are captured around middle levels of the keyscape, but the interfacing area between chords and keys tend to be quite noisy during these processes. It's possible to segregate local and global tonic orientation from clean keyscapes.

The main concern remains in the assumptions taken for define keyscapes. Although we believe that tonal implication description and summarization techniques can be polished in near future, we identify the main problem at temporal level. Despite our sliding window policy improve both the temporal boundary definitions and the vertical alignment over Sapp's and Gómez's solutions, it's clear how a single multiresolution policy does not suffice for describing musicological approaches to harmonic analysis or perceptual intuitions. Our preliminary tests with pure inertial context definitions provide the evidence. They performed well the task of defining only past-based contexts, however, it's obvious that this cannot capture the reality of modulation -whether theoretical or perceptual-, for which some sort of *reset* would be needed after the key change. However, this solution would not comply with our requirement of keeping all the hierarchical information of the piece. The same could be inferred regarding purely expectation-based contexts (this has not been tested yet).

Our study of dozens of keyscapes for varied music, along with listening and musicological analysis, confirm our logical intuitions: real music tonal discourse don't follow homogeneous rules. Music works, as forms of art, are not *constrained* to our statistical treatment -actually, to any statistical treatment-[43], and in our case this affects particularly to duration of tonal events. Global key is not required to be *present* along any specific length, or at any given position of the piece, and the same holds for local keys or chords, which can span very differently over time. This affects to the overall hierarchy, and our assumptions can result in many cases in rather *surprising* estimations, which are especially relevant -for our problem- at middle-high levels in the keyscape.

At structural level of music -musical form- this is the case, for instance, of many baroque works, in which a series of modulations don't guarantee stable temporal presence of the tonic. Classical style, for which is conventionally expected a refined tonal equilibrium, is not an exception. A typical allegro-sonata movement, may vary extraordinarily the duration of the theme groups and development sections and separate from conventions regarding keys, and this holds even for the most *paradigmatic* classical composers, like Haydn. Pop music alternation of chorus and verse is not a guarantee for even establish the global key, depending on the modulations, although it generally produces proper keyscapes.

Regarding key induction mechanisms, local keys can be suddenly established without preparation, they can arise from unstable and long transitions, they can stay long without the presence of the tonic, or they may be defined by long melodic lines stressing just few degrees of the scale, to cite just some possibilities. To capture such variety of techniques requires different windowing policies, in terms of temporal resolutions and vertical alignment. But over all these concerns, tonality is ambiguous at any level of representation. A fragment of music can *be* in a given key, but *more or less over time*. Notes, chords and keys influence each others' interpretation dynamically, and this involves complex and subjective prospective and retrospective listening processes.

[43] Forms of art, as product of human mind, necessarily require different epistemological treatment than natural sciences, since all of them, and much particularly music, tend to escape *by nature* from standardization. Music can be analyzed statistically to derive *rules* that can be used more or less *successfully* for specific applications, but this should not be misinterpreted as underlying *governing laws explaining* music. What we extract from the plain application of such techniques is no more than mere censual -hence fallible- information about music parameters, and no single piece of music is *required* to fit mainstreams -if any-. So, they may provide usable descriptive account of music parameters, but not prescriptive account about how music *should* behave logically or rationally. This point is at the core of human-related scientific methodologies, like cognitive psychology, as discussed in [15].

As personal observations, centroidscapes have resulted pretty informative about tonal discourse and provide intuitions about tonal stability -local and global- in visualizations along listening, particularly by observing centroid clusters evolution: generation, splitting/grouping and movement over time and space. To follow distancescapes along listening has resulted roughly coherent with phrasing sensation in non complex music, but this has only been tested informally with few listeners and selected pieces.

Most of these individual capabilities and problems are interpretable in terms of music theory, psychological models and technology concerns. However, a global evaluation is difficult due to the lack of consensus from different disciplines, whose paradigms embed the idea of stability in quite different ways. The same way, different applications might require distinct choice of the algorithm's parameters, and we would only suggest its use for collections of music of any size after further development over specific music material. More generally, musical tension -for which tonal stability concept might be just one component- is far beyond state-of-the-art.

5.3 Future work and potential applications

From all these unsolved issues, we identify some points in which the system could be improved in near future.

- Tonal implication provides rich information which is not being much exploited, in terms of description of enriched tonal events and contexts, including summarization.
- Study other possibilities for keyscapes, like higher vertical resolution or combination of different sliding window policies -inertial, balanced and expectative-, focused on describe the interfacing area between chords and keys. Combination of overlapping sliding policies with segmentation.
- Evaluation and tuning over specific problems, like kinds of cadences, types of modulating processes or temporal patterns in tonal discourse. Once isolated these scenarios over a proper music dataset and a discussed ground truth, a quantitative *accuracy* of the algorithm can be approached for specific applications.
- Research distance models based on non-summarized tonal implication and enriched families of tonal events, beyond major and minor.
- Combination with other techniques for tonal description and structural analysis of music, like self-similarity matrices based on chroma information.

Among the potential applications of the descriptor, we identify:

- Music analysis and composition.
- Music similarity.
- Mood modeling from audio signals.
- Music visualization.
- Interactive performance systems, for which our model could provide visual sensory reinforcement.
- Interactive interfaces for psychology research.

Bibliography

[1] Aarden, B. J. (2003). *Dynamic Melodic Expectancy*. Ph. D. diss. Ohio State University.

[2] Bigand, E. and Parncutt, R. (1999). "Perceiving musical tension in long chord sequences". *Psychological Research*, 62, pp. 237-254.

[3] Bregman, A. (1990). *Auditory Scene Analysis. The perceptual organization of sound*. Cambridge, MA: MIT Press.

[4] Chew, E. (2000). *Towards a Mathematical Model of Tonality*. Ph. D. diss. MIT.

[5] Farbood, M. M. (2006). *A Quantitative, Parametric Model of Musical Tension*. Ph. D. diss. MIT.

[6] Forte, A. and Gilbert, S. (1982). *Introduction to Schenkerian analysis*. New York: Norton.

[7] Fujishima, T. (1999). "Realtime chord recognition of musical sound: a system using common lisp music". In ICMA, editor, *International Computer Music Conference*, Beijing, pp. 464–467.

[8] Gómez, E. (2006). *Tonal description of music audio signals*. Ph. D. diss. Universitat Pompeu Fabra, Barcelona.

[9] Gómez, E. and Bonada, J. (2005). "Tonality visualization of polyphonic audio". *Proceedings of International Computer Music Conference*, Barcelona.

[10] Hamanaka, M., Hirata, K. and Tojo, S. (2006). "Implementing a 'Generative Theory of Tonal Music'". *Journal of New Music Research*, 33/4, pp. 249-277.

[11] Harte, C., Sandler, M. and Gasser, M. (2006). "Detecting Harmonic Change in Musical Audio". In *Proceedings on the 1st ACM workshop on Audio and Music Computing Multimedia,* Santa Barbara, CA, pp. 21-26.

[12] Huron, D. (2006). *Sweet anticipation: music and the psychology of expectation*. Cambridge, MA: MIT Press.

[13] Janata, P., Birk, J. L., Van Horn, J. D., Leman, M., Tillmann, B. and Bharucha, J. J. (2002). "The cortical topography of tonal structures underlying Western music". *Science*, 298, pp. 2167-2170.

[14] Janata, P. (2007). "Navigating Tonal Space". *Tonal Theory for the Digital Age. Computing in Musicology*, 15, pp. 39-50.

[15] Krumhansl, C. L. (1990). *Cognitive Foundations of Musical Pitch*. New York: Oxford University Press.

[16] Krumhansl, C. L. (1996). "A perceptual analysis of Mozart's Piano Sonata K.282: Segmentation, tension and musical ideas". *Music Perception*, 13/3, pp. 401-432.

[17] Krumhansl, C. L. (2004). "The Cognition of Tonality - as We Know it Today". *Journal of New Music Research*, 33/3, pp. 253-268.

[18] Leman, M. (2000). An auditory model of the role of short term memory in probe-tone ratings. *Music Perception, Special Issue in Tonality Induction*, 17/4, pp. 481–509.

[19] Leman, M. (2003). "Foundations of Musicology as Content Processing Science". *Journal of Music and Meaning*, 1. Online: http://www.musicandmeaning.net.

[20] Lerdahl, F. (1996). "Calculating tonal tension". *Music Perception*, 13/3, pp. 319-363.

[21] Lerdahl, F. (2001). *Tonal Pitch Space*. New York: Oxford University Press.

[22] Lerdahl, F. and Jackendoff, R. (1983). *A Generative Theory of Tonal Music*. Cambridge, MA: MIT Press.

[23] Lerdahl, F. and Krumhansl, C. L. (2007). "Modeling tonal tension". *Music Perception*, 24, pp. 329-366.

[24] Lesaffre, M., Leman, M., Baets, B. D., and Martens, J. P. (2004). "Methodological considerations concerning manual annotation of musical audio in function of algorithm development". In *International Conference on Music Information Retrieval*, Barcelona.

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

[26] Meyer, L. B. (1996). "Commentary". *Music Perception*, 13/3, pp. 455-483.

[27] Narmour, E. (1990). *The analysis and cognition of basic melodic structures: the implication-realization model*. Chicago: University of Chicago Press.

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

[29] Palmer, C. "Anatomy of a Performance". *Music Perception*, 13/3, pp. 433-453.

[30] Parncutt, R. (1988). "Revision of Terhardt's psychoacoustical model of the root(s) of a musical chord". *Music Perception*, 6, pp. 65-94.

[31] Pressnitzer, D., McAdams, S., Winsberg, S. and Fineberg, J. (2000). "Perception of musical tension for nontonal orchestral timbres and its relation to psychoacoustic roughness". *Perception & Psychophysics*, 62/1, pp. 66-80.

[32] Purwins, H. (2005). *Profiles of Pitch-Classes. Circularity of Relative Pitch and Key: Experiments, Models, Computational Music Analysis and Perspectives*. Ph. D. diss. Berlin University of Technology.

[33] Sapp, C. S. (2005). "Visual Hierarchical Key Analysis". Computers in Entertainment, 4/4, pp. 1-19.

[34] Segnini, R. (2006). "Timbrescape: a Musical Timbre and Structure Visualization Method Using Tristimulus Data". *9th International Conference on Music Perception and Cognition*, Bologna.

[35] Serra, X., Bernardini, N., Leman, M., Widmer, G. and De Poli, G. eds. (2009). *Sound and Music Computing Network Roadmap*. Current open version. http://smcnetwork.org/roadmap.

[36] Temperley, D. (2007). "The Tonal Properties of Pitch-Class Sets: Tonal Implication, Tonal Ambiguity and Tonalness". *Tonal Theory for the Digital Age*. *Computing in Musicology*, 15, pp. 24-38.

[37] Tillmann, B., Bharucha, J. J. and Bigand, E. (2000). "Implicit Learning of Tonality: A Self-Organizing Approach". *Psychological Review*, 107/4, pp. 885-913.

[38] Toiviainen, P. (2007). "Visualization of Tonal Content in the Symbolic and Audio Domain". *Tonal Theory for the Digital Age*. *Computing in Musicology*, 15, pp. 187-199.

[39] Toiviainen, P. and Krumhansl, C. L. (2003). “Measuring and modeling real-time responses to music: The dynamics of tonality induction”. *Perception*, 32, pp. 741-766.

[40] Vines, B. W., Krumhansl, C. L., Wanderley, M. M. and Levitin, D. J. (2006). “Cross-modal interactions in the perception of musical performance”. *Cognition: International journal of cognitive science*, 101/1, pp. 80-113.

[41] Werts, D. (1983). *A theory of scale references*. Ph.D. diss. Princeton University.

[42] Williams, J. K. (2007). “An Interactive, Multimedia Environment for Exploring Tonal Pitch Space”. *ICL, International Computer Aided Learning Conference*, Villach (Austria). Available at: http://music.uncg.edu/etps/.

Appendix I - Audio material

- Mozarts's *Requiem* in Dm, K.626 - *Sequenz I "Dies irae"*. Karl Böhm conducting Vienna Philharmonic Orchestra.
- Synthesized piano chords: major, augmented, whole-tone, full-diminished and tritone.
- Bach's chorale "Christus, der ist mein Leben". MIDI transcription of score on [21], p. 8, rendered with standard organ patch. Metronomic tempo of 1 sec/beat, no fermatas.
- Mozart's *Piano sonata* in Eb, K.282, first movement, bars 1-8. MIDI transcription of score (no reduction), metronomic tempo of 1 sec/beat. Rendered with standard piano patch.
- Mozart's *Piano sonata* in Eb, K.282, first movement, bars 1-8. MIDI transcription of score (no reduction), transposed half-tone up (from Eb to E), metronomic tempo of 1 sec/beat. Rendered with standard piano patch.
- *Ciaccona* for violin and continuo, attributed to Tomaso Vitali, performed by Jascha Heifetz with organ accompaniment.
- Vivaldi's Violin concerto op.8, nº2 "Summer", third movement (presto). Herbert von Karajan conducting Berlin Philharmonic Orchestra.
- Vivaldi's Violin concerto op.8, nº2 "Summer", third movement (presto). Joe Satriani.

-

Appendix II - Glossary of frequently used terms

Centroidscape. Multiresolution tonal implication mapped into a pitch space geometric model. Shows a *vertical* lecture of the keyscape for each frame, with their hierarchical relationships distributed across pitch space. 2D version is used for visualization. 4D version is used to measure Euclidian distances between event hierarchies.

Chroma features / information. Descriptor which maps harmonic content of audio signals into pitch-classes, which can be used for tonal estimations. Usually, it covers the chromatic octave in 12 pitch-classes. Our implementation as HPCPs has a resolution of 120 (10 cents) instead.

Distancescape. Multiresolution distance model between hierarchical tonal events. It's computed from centroidscape, by measuring Euclidian distances between centroids. *Top-down* version measures distances from global key down the rest of levels in keyscape. *Bottom-up* version measures distances from surface chords up the rest of levels in keyscape. *Averaging* version summarizes all the curves by arithmetic mean. *Neighborwise* version measures distances from one level to the immediate one, up or down the keyscape, instead of taking a fixed reference.

Keyscape. Temporal multiresolution tonal implication, inspired by Sapp. Averaging window sizes range from fractions of seconds to the whole piece.

GTTM. Generative Theory of Tonal Music. Theory by Lerdahl and Jackendoff

Interkey distance. Model of perceptually relevant distances between keys, based on empirical studies by Krumhansl and associates. The model used here is derived from tonal-hierarchies ratings, and it's applied between any two levels of keyscape.

Pitch hierarchies. Foundation of tonal system. Hierarchical ordering of the pitch-classes of a given musical system. It's related with stability through tonic orientation principles. The higher the hierarchical status of a given pitch, the higher its stability. This hierarchy rules beyond pitch-classes, governing chord and keys relationships as well.

Schenkerian analysis. Analytical methodology developed by Schenker, which allows for hierarchical descriptions of complex music.

Surface events. In my usage of the term, those events occurring during the *listening present*. All events have surface level, but can reach deeper structural functions in musical discourse depending on their stability.

Temporal multiresolution. Technique based on using several sliding window sizes to summarize features from audio signals. Applied to tonal estimation, allows for estimate chords, local keys and global keys simultaneously.

Tonal context. Tonal implication of the near surroundings of a given event. Tonal context of a chord is its local key. Tonal context of a local key, is the global key.

Tonal implication. Estimation of tonal information from a segment of music. It's computed by averaging chroma information and correlating them with tonal profiles. It's a 24-dimensional vector, showing the relative strengths of all major and minor keys.

Tonal stability. The goal of this work. Stability of a surface event depends on its relationship with its tonal context. The better their fitting -the closer they are in pitch space-, the higher the stability. This relationship is measured between all levels of the keyscape. The most stable event -ideally- is the root of a tonic chord in a local key which is the same as the global key of the piece.

Toroidal geometry. Applied to tonality, geometric space in which tonal centers -or pitches- are distributed according to some pattern of *distances*. This distribution varies across different models. Krumhansl proposes two different 4-dimensional structures, devised to represent perceptual closeness between keys.

TPS. Tonal Pitch Space. Theory by Fred Lerdahl.