TONAL DESCRIPTION OF MUSIC AUDIO SIGNALS

A DISSERTATION SUBMITTED TO THE DEPARTMENT OF TECHNOLOGY OF THE UNIVERSITAT POMPEU FABRA FOR THE PROGRAM IN COMPUTER SCIENCE AND DIGITAL COMMUNICATION IN PARTIAL FULFILMENT OF THE REQUIREMENTS FOR THE DEGREE OF

–

DOCTOR PER LA UNIVERSITAT POMPEU FABRA

–

WITH THE MENTION OF EUROPEAN DOCTOR

Emilia Gómez Gutiérrez

2006

DOCTORAL DISSERTATION DIRECTION

Dr. Xavier Serra
Department of Technology
Universitat Pompeu Fabra, Barcelona

This research was performed at the Music Technology Group of the Universitat Pompeu Fabra in Barcelona, Spain. Primary support was provided by the Spanish Ministry of Science and Technology through a FPI grant (*Formación del Personal Investigador*, TIC project 2000-1094-C02 TABASCO), and by the EU projects FP6-507142 SIMAC http://www.semanticaudio.org and IST-1999-20194 CUIDADO. This research was partially funded by the European Commission through a Marie Curie fellowship for doctoral students at the Music Acoustics Group, Department of Speech, Music and Hearing, KTH in Stockholm, Sweden.

Abstract

This dissertation is about tonality. More precisely, it is concerned with the problems that appear when computer programs try to automatically extract tonal descriptors from musical audio signals.

This doctoral dissertation proposes and evaluates a computational approach for the automatic description of tonal aspects of music from the analysis of polyphonic audio signals.

In this context, we define a tonal description in different abstraction levels, differentiating between low-level signal descriptors (e.g. tuning frequency or pitch class distribution) and high-level textual labels (e.g. chords or keys). These high-level labels require a musical analysis and the use of tonality cognition models. We also establish different temporal scales for description, defining some instantaneous features as being attached to a certain time instant, and other global descriptors as related to a wider segment (e.g. a section of a song).

Along this PhD thesis, we have proposed a number of algorithms to directly process digital audio recordings from acoustical instruments, in order to extract tonal descriptors. These algorithms focus on the computation of pitch class distributions descriptors, the estimation of the key of a piece, the visualization of the evolution of its tonal center or the measurement of the similarity between two different musical pieces. Those algorithms have been validated and evaluated in a quantitative way. First, we have evaluated low-level descriptors, such as pitch class distribution features and estimation of the tuning frequency (with respect to 440 Hz), and their independence with respect to timbre, dynamics and other external factors to tonal characteristics. Second, we have evaluated the method for key finding, obtaining an accuracy around 80%. This evaluation has been made for a music collection of 1400 pieces with different characteristics. We have studied the influence of different aspects such as the employed tonal model, the advantage of using a cognition-inspired model vs machine learning methods, the location of the tonality within a musical piece, and the influence of the musical genre on the definition of a tonal center. Third, we have proposed the extracted features as a tonal representation of an audio signal, useful to measure similarity between two pieces and to establish the structure of a musical play. For this, we have evaluated the use of tonal descriptors to identify versions of the same song, obtaining an improvement of 55% over the baseline.

From a more general standpoint, this dissertation substantially contributes to the field of computational tonal description: a) It provides a multidisciplinary review of tonal induction systems including signal processing methods and models for tonality induction; b) It defines a set of requirements for low-level tonal

features; c) It provides a quantitative evaluation of the proposed methods with respect to similar ones for audio key finding. This quantitative evaluation is divided in different stages, analyzing the influence of each one; d) It supports the idea that some application contexts do not need a accurate symbolic transcription, thus bridging the gap between audio and symbolic-oriented methods without the need of a perfect transcription; e) It extents current literature dealing with classical music to other musical genres; f) It shows the usefulness of tonal descriptors for music similarity; g) It provides an optimized method which is used in a real system for music visualization and retrieval, working with over a million of musical pieces.

Resumen

En los últimos años, se ha multiplicado enormemente la cantidad de material sonoro que puede estar a nuestra disposición a través de redes informáticas o de sistemas de almacenamiento. Se pueden guardar hasta 15.000 canciones en un solo dispositivo portátil, y esta cantidad aumenta significativamente si consideramos un soporte de almacenamiento de mayor capacidad (como por ejemplo un disco duro) o los recursos que ofrece Internet. Esta proliferación de colecciones digitales de música hace necesario el desarrollo de tecnologías que sean capaces de proporcionar al usuario herramientas de interacción fácil y significativa con dichas colecciones musicales.

Mucha investigación desarrollada a lo largo de los últimos años está relacionada con este tema, y la cantidad de trabajos publicados refleja el tremendo crecimiento de dichas colecciones y la necesidad de realizar búsquedas dentro de su contenido de forma eficiente y efectiva. Dentro de este contexto, la descripción automática de señales de audio se ha impuesto como una de las áreas de mayor interés, ya que permite proporcionar automáticamente una descripción de las características más relevantes de una pieza, como son el ritmo, la armonía y melodía (relacionada con aspectos tonales), o su instrumentación (aspecto tímbrico). La tonalidad es uno de los aspectos musicales que han sido menos tratados en la literatura existente, y es precisamente el objeto del presente trabajo.

La presente tesis doctoral trata sobre tonalidad. Más concretamente, está relacionada con los problemas que surgen cuando se intenta extraer automáticamente una descripción tonal mediante el análisis de una grabación de audio digital utilizando programas informáticos. Esta tesis doctoral propone y evalúa un enfoque computacional para la descripción automática de aspectos tonales de la música a partir del análisis de señales de audio polifónicas.

En este contexto, se define una descripción tonal en distintos niveles de abstracción, diferenciando entre descriptores de bajo nivel que se extraen de la señal de audio (por ejemplo la frecuencia de afinación y la distribución de notas) y etiquetas textuales de alto nivel que requieren un análisis musical y aplicar un modelo de tonalidad (por ejemplo la caracterización de acordes o la tonalidad global de una pieza). También se definen diversas escalas temporales para la descripción, distinguiendo entre algunas características definidas de forma instantánea y otras relacionadas con un segmento de mayor duración, como podría ser un acorde o una sección de una obra musical.

En el trabajo realizado durante esta tesis doctoral, se proponen métodos para procesar directamente grabaciones digitalizadas de señales musicales de instrumentos acústicos con el fin de extraer descriptores tonales. Estos métodos se centran en calcular descriptores de distribución de notas, en estimar la tonalidad de una pieza, en visualizar la evolución del centro tonal o en medir la similitud tonal entre dos piezas diferentes. Dichos métodos validan y evaluan de una forma cuantitativa. En primer lugar, se realiza una eva-luación de los descriptores de bajo nivel, incluyendo la extracción de distribución de notas, la estimación de la frecuencia de afinación (respecto al La 440 Hz) y la independencia de dichos descriptores respecto al timbre (instrumento que toca la pieza), dinámica y otros factores externos a las características tonales. En segundo lugar, se evalúa su validez para estimar la tonalidad de una pieza musical, obteniendo una tasa de aciertos cercana al 80%. Dicha evaluación se ha realizado en una colección musical de unas 1400 piezas de diferentes características y estilos. Se estudia la influencia de ciertos aspectos como el modelo tonal utilizado, la ventaja de usar un modelo basado en cognición musical respecto a métodos de aprendizaje automático, la localización de la tonalidad dentro de una pieza musical y la influencia del estilo musical en la definición de un centro tonal. En tercer lugar, se propone este tipo de descripción como representación de una señal de audio polifónica, y se demuestra que es útil para medir similitud entre dos piezas y para establecer la estructura de una obra musical. Para ello, se ha evaluado la validez de los descriptores tonales para reconocer versiones de una misma obra musical, obteniendo una tasa de aciertos del 55%.

Desde un punto de vista más general, esta tesis contribuye sustancialmente al campo de la descripción tonal mediante métodos computacionales: a) Proporciona una revisión multidisciplinar de los sistemas de estimación de la tonalidad, incluyendo métodos de procesado de señal y modelos cognitivos; b) Define una serie de requerimientos que deben cumplir los descriptores tonales de bajo nivel; c) Proporciona una evaluación cuantitativa de los métodos propuestos respecto al estado del arte actual en estimación de tonalidad. Esta evaluación cuantitativa está dividida en diferentes módulos, analizando la influencia de cada uno de ellos; d) Respalda la idea de que para ciertas aplicaciones no es necesario obtener una transcripción perfecta de la partitura a partir de una grabación, y que se pueden utilizar métodos que trabajan con partituras sin la necesidad de realizar una transcripción automática; e) Extiende la literatura existente que trabaja con música clásica a otros géneros musicales, analizando la problemática asociada a dichos géneros; f) Demuestra la utilidad de los descriptores tonales para comparar piezas musicales; g) Proporciona un algoritmo optimizado que se utiliza en un sistema real para visualización, búsqueda y recomendación musical, que trabaja con más de un millón de piezas musicales.

Resum

En els últims anys, s'ha multiplicat enormement la quantitat de material sonor que pot estar a la nostra disposició a través de xarxes informàtiques o de sistemes d'emmagatzematge. Tan sols en un sol dispositiu portàtil es poden guardar fins a 15.000 cançons, i aquesta quantitat augmenta significativament si considerem un espai d'emmagatzematge major (per exemple un disc dur) o els recursos oferts per Internet. Aquesta proliferació de col·leccions digitals de música fan necessari el desenvolupament de tecnologies capaces de proporcionar a l'usuari eines d'interacció fàcil i significativa amb aquestes col·leccions de música.

Molta recerca desenvolupada al llarg dels últims anys esta relacionada amb aquest tema, i la quantitat de treballs publicats reflexa el creixement enorme d'aquestes col·leccions musicals i la necessitat de realitzar cerques dins d'aquest contingut de forma eficient i efectiva. Dins d'aquest context, la descripció automàtica de senyals d'àudio s'ha imposat com una de les àrees de major interès, ja que permet proporcionar automàticament una descripció de les característiques més rellevants d'una peça, com son el ritme, l'harmonia i la melodia (relacionada amb els aspectes tonals), o la seva instrumentació (aspecte tímbric). La tonalitat és un dels aspectes musicals que han estat menys tractats dins la literatura existent, i és precisament l'objecte del present treball.

Aquesta tesi doctoral tracta sobre tonalitat. Més concretament, aquesta tesi esta relacionada amb els pro-ble-mes que sorgeixen quan intentem extreure automàticament una descripció tonal mitjançant l'anàlisi d'una gravació d'àudio digital utilitzant programes informàtics. Aquesta tesi doctoral proposa i avalua un enfocament computacional per a la descripció automàtica dels aspectes tonals de la música a partir de l'anàlisi de senyals d'àudio polifòniques.

Dins d'aquest context, es defineix una descripció tonal en diferent nivells d'abstracció, diferenciant entre descriptors de baix nivell que es calculen directament a partir del senyal d'àudio (per exemple la freqüència d'afinació i la distribució de notes) i etiquetes textuals d'alt nivell que requereixen un anàlisi musical i aplicar un model de tonalitat (per exemple la caracterització d'acords o tonalitat d'una peça). També es defineixen diverses escales temporals per a la descripció, distingint entre algunes característiques definides de forma instantània i d'altres relacionades amb un segment de major durada, com podria ser un acord o una secció d'una obra musical.

Dins del treball desenvolupat durant aquesta tesi doctoral, es proposen mètodes per a processar directament gravacions digitalitzades de senyals musicals d'instruments acústics amb la finalitat d'extreure descriptors tonals. Aquests mètodes es centren en el càlcul de descriptors de distribucions de notes, en l'estimació de tonalitat d'una peça, en la visualització de l'evolució del centre tonal o en la mesura de la similitud tonal entre dues peces diferents. Aquest mètodes es validen i avaluen d'una forma quantitativa. Primer, es realitza una avaluació dels descriptors de baix nivell, incloent l'extracció de distribucions de notes, l'estimació de la freqüència d'afinació (respecte al La 440 Hz) i la independència d'aquests descriptors respecte al timbre (instrument que toca la peça), dinàmica i altres factors externs a les característiques tonals. En segon lloc, s'avalua la seva validesa per a estimar la tonalitat d'una peça musical, obtenint un percentatge d'encerts proper al 80%. Aques-ta avaluació s'ha realitzat en una col·lecció d'unes 1400 peces de diferents característiques i estils. S'ha estudiat la influencia de certs aspectes com el model tonal utilitzat, l'avantatge de fer servir un model basat en cognició musical respecte a mètodes d'aprenentatge automàtic, la localització de la tonalitat dins d'una peça musical i la influencia de l'estil musical per a la definició d'un centre tonal. En tercer lloc, es proposa aquest tipus de descripció com a representació d'un senyal d'àudio polifònic, que és útil per a mesurar similitud entre dues peces i per a establir l'estructura d'una obra musical. Per a això, s'ha avaluat la validesa de descriptors tonals per a trobar versions d'una mateixa obra musical, obtenint una tasa d'encerts del 55%.

Dins d'un punt de vista més general, aquesta tesi contribueix substancialment al camp de la descripció tonal mitjançant mètodes computacionals: a) Proporciona una revisió multidisciplinària dels sistemes d'estimació de la tonalitat, incloent mètodes de processament del senyal i models cognitius.; b) Defineix una sèrie de requeriments que han de complir els descriptors tonals de baix nivell; c) Proporciona una avaluació quantitativa dels mètodes proposats respecte a l'estat de l'art actual en estimació de tonalitat. Aquesta avaluació quantitativa esta dividida en diferents mòduls, analitzant l'influencia de cadascún d'ells; d) Respalda la idea de que per a certes aplicacions no és necessari una transcripció perfecta de la partitura a partir d'una gravació, i que es poden fer servir mètodes que treballen amb partitures sense la necessitat de realitzar una transcripció automàtica; e) Estén la literatura existent que treballa amb música clàssica a altres gèneres musicals, analitzant la problemàtica associada a aquest gèneres; f) Demostra la utilitat dels descriptors tonals per a comparar peces musicals; g) Proporciona un algoritme optimitzat que es fa servir dins un sistema real per a visualització, cerca i recomanació musical, que treballa amb més d'un milió de obres musicals.

Acknowledgments

There are many people that have made this dissertation possible, and I would need hundreds of pages to express my gratitude to all of them.

I have had the opportunity to work at the Music Technology Work of the Pompeu Fabra University in Barcelona, where I have been able to interact to such a brilliant and stimulating group of people. First I would like to thank Xavier Serra, my supervisor, for giving me the opportunity to joint the MTG and supporting my research along these years. There are two people who I consider as main contributors of this thesis. One of them is Perfecto Herrera, who has been my main guidance along this work. I would like to thank him for reading all my writings, giving ideas and support and devoting much time to me during these years. He also helped me to apply machine learning techniques to the problem of key estimation. The second one is Jordi Bonada, who has provided much help and ideas for improving the signal processing methods and for implementing the algorithms, and who helped to develop the tonality visualization tool.

I have been very lucky to count on such an inspiring team of researchers next to me, who have provided plenty of ideas, interesting discussions and fruitful results: Fabien Gouyon, Beesuan Ong, Sebastian Streich, Enric Guaus, Pedro Cano, Esteban Maestre, Oscar Celma, David García, Xavier Amatriain and Flavio Lazzaretto.

Thanks also to all the remaining MTG people for their support and fun, specially to Oscar Mayor, Alex Loscos, Jordi Janer, Lars Fabig, Joana Clotet, Cristina Garrido, Salvador Gurrera, Carlos Atance, Ramon Loureiro, Maarten de Boer, Pau Arumí, Koppi, Jose Pedro García, Bram de Jong, Nicolas Wack, Gunnar Holmberg, Rafael Ramirez, Martin Kaltenbrunner, Marcos Alonso, Guenter Geiger and Sergi Jordà.

Teaching hours are not always compatible to conferences, and I would like to thank Enric Guaus for shifting my classes when I had to give a presentation.

It was also a pleasure to count on such nice external collaborators for different projects and publications, as Anssi Klapuri, Benoit Meudic, Maarten Gratchen, Leandro T. C. Gomes, Chris Harte, Juan Pablo Bello, Elias Pampalk, Simon Dixon, Gerard Widmer, Geoffroy Peeters, Rui Pedro Paiva, Dan Ellis and Graham Poliner.

One of the best experiences along these years was the time that I spent in KTH, Stockholm, with the Music Acoustics Group. I would like to express my gratitude to the people that made this possible: my supervisor Roberto Bresin, Sten Ternström, Johan Sundberg, Erwin Schoonderwaldt, Sofia Dahl, Anders

Askenfelt, Mikael Bohman, Anders Friberg, Kjetil Falkenberg, as well as other visiting students: Anke Grell, Laura Lehto, Matias Rath, Stefania Serafin and Diana Young.

Many people influenced my research along these years, and some of them provided nice feedback and advises. I would like to thank Wei Chai, Alain de Cheveigné, Elaine Chew, Roger Dannenberg, Takuya Fujishima, Evelyne Heylen, Ozgur Izmirli, Carol Krumhansl, Marc Leman, Micheline Lessaffre, Nicola Orio, Hendrik Purwins, David Temperley and Chen Yang.

Writing is not one of the things I can do best. Thanks to Perfecto Herrera, Xavier Amatriain, Anssi Klapuri (with a special thanks for his valuable review and suggestions to improve this document), Christina Anagnostopoulou, Jorge Chávez and the anonymous reviewers for helping me to get this dissertation better.

I would also like to thank my colleagues and teachers from the DEA ATIAM at IRCAM, specially Thibaut Ehrette, Fabien Gouyon, Jeremy Marozeau, Benoit Meudic, Olivier Lartillot, Vincent Verfaille, Olivier Warusfel, Xavier Rodet, Jean-Claude Risset, Gérard Assayag, Daniel Arfib, Stephen McAdams, Alain de Cheveigné and Geoffroy Peeters.

Last but not least, this work would have never been possible without the encouragement of my husband Jordi, my parents Salvador and Rosario, my sister Rosa, my friends (specially Rocío, a companion in the adventure of getting a PhD), my choir (La Fuga, Lluïsos de Gràcia), the wonderful atmosphere of Barcelona and the air from Sevilla, Vilardell and Zahara de los Atunes.

To Jordi

Contents

Chapter 1

Introduction

This dissertation deals with tonal description of music audio signals. This work proposes a system intended to automatically extract tonal information from polyphonic audio recordings from one or several instruments. We analyze different problems that arise when developing computational models that extract this musical information from audio, such as the extraction of relevant features related to the played notes or the correct use of tonal models for key finding. The goal of this chapter is to present the context in which this thesis has been developed, including the motivation for this work, the research context, some practical aspects related to automatic tonal description and finally a summary of the work carried out and how it is organized along this document.

1.1 Motivation

In the last few years, a great amount of audio material has been made accessible to the home user through networks and mass storage. For instance, using only a portable device, one can store up to 15.000 songs. Nowadays, playing devices only allow to organize and search music pieces by means of editorial information such as the name of the artist or the album. The proliferation of digital music collections necessitates the development of technology that allows an interaction with such collections in an easier and more meaningful way. Interesting searching strategies would be related to musical characteristics of the piece (e.g. played instrument, slow or fast tempo, major or minor key), or to highly semantic descriptors such as the induced mood (e.g. happy vs sad). The main goal of content retrieval and transformation systems is then to allow the retrieval and transformation of audio according to its musical content, so as to provide meaningful criteria to interact with music files.

Much research over the last years have been devoted to music content retrieval. The annual ISMIR conference[1] is the first established forum for those involved in works on accessing digital musical materials. There has been an increasing number of contributions to this conference (the number of published articles

[1] http://www.ismir.net

has evolved from 35 in 2000 to 120 in 2005), and the attendance has increased from 88 people in 2000 to 194 in 2004[2]. At the same time, important contributions to this field have appeared on related events (e.g. the European Conference on Information Retrieval[3], Audio Engineering Society Conventions and Conferences[4] or the Digital Audio Effects Conferences[5], to name a few) and relevant journals (e.g. Journal of New Music Research[6], Computer Music Journal[7], EURASIP Journal on Applied Signal Processing[8], IEEE Transactions on Speech and Audio Processing[9] or INFORMS Journal on Computing[10]). This increasing amount of literature related to music content processing reflects the tremendous growth of music-related data available and the consequent need to provide solutions to search this content efficiently and effectively. As explained in the ISMIR web[10], this vast area challenges those who need to organize and structure musical data, provide tools to search and retrieve, and use these tools efficiently. There are many disciplines involved in this issue, such as signal processing, musicology, statistics and information retrieval.

1.2 Content description of audio

Content description of audio has become very relevant in the context of Music Information Retrieval, because it provides the user meaningful descriptors from audio signals, which can be automatically extracted. The word *content* is defined in a general way as *the ideas that are contained in a piece of writing, a speech or a film*, according to the Cambridge Advanced Learner's Dictionary online[11]. This concept applied to a piece of music can be seen as the implicit information that is related to this piece and that is represented in the piece itself. The concept of *content analysis* is defined as the *analysis of the manifest and latent content of a body of communicated material (as a book or film) through a classification, tabulation, and evaluation of its key symbols and themes in order to ascertain its meaning and probable effect*, according to the Merrian-Webster's collegiate dictionary[12]. That means, the extraction of content information from a piece of music by examining its audio recording.

Any piece of information related to a piece of music, which carries meaningful information to someone can be technically denoted as meta-data. Following this idea, the MPEG-7 standard defines a content descriptor as *a distinctive characteristic of the data which signifies something to somebody*, as appears in Manjuhath et al. (2002). This open view on the nature of music content descriptors has a drawback: as these descriptors represent many different aspects of a musical piece, all meta-data may not be understood by every

[2] http://www.ismir.net/CFH-2007.html
[3] http://irsg.bcs.org
[4] http://www.aes.org
[5] http://www.dafx.de
[6] http://www.tandf.co.uk/journals/titles/09298215.asp
[7] http://mitpress2.mit.edu/e-journals/Computer-Music-Journal
[8] http://www.hindawi.com/journals/asp
[9] http://www.ieee.org/organizations/society/sp
[10] http://joc.pubs.informs.org
[11] http://www.dictionary.cambridge.org
[12] http://www.m-w.com

user. This is part of the *user-modelling* problem, whose lack of precision participates to the so-called semantic gap. According to Smeulders et al. (2000), the semantic gap is defined as *the lack of coincidence between the information that one can extract from the (sensory) data and the interpretation that the same data has for a user in a given situation*. This gap has been signaled by several authors as one of the recurrent open issues in systems dealing with audiovisual content. It is therefore important to consider meta-data together with their functional value and address the question of what means content to each user, and in which application context. The need for descriptive information about what is musically significant addresses a large spectrum of characteristics, from acoustic to musicological and cultural.

1.3 Levels and facets for description

If we study the content related to a piece of music, we can identify different levels of abstraction, as well as many description facets. Any of these levels of abstraction or description facets might be useful for some users (for instance, to a naive listener or a musicologist). Regarding the abstraction levels, we usually distinguish between *low*, *mid* and *high-level* descriptors, as it is explained in Herrera (2006). The term *low-level* is usually employed to denote features that are closely related to the audio signal, which are computed in a direct or derived way. Most of these descriptors do not have much sense for the majority of the users, but they are easily used by computational systems. *Mid-level* features are considered to require an induction operation which allows, after an analysis of the data, to perform some generalizations. Statistics and machine learning are the main disciplines involved in this generalization process, and these mid-level descriptors may have some meaning for certain users. Finally, the step from low or mid-level descriptors to *high-level* descriptors requires to *bridge the semantic gap*, which was mentioned above. These descriptors have a relevant meaning to users, and then they require a modelling process of the user behavior. Many disciplines are involved in this issue, such as music cognition or music psychology.

It is also important to distinguish between description and transcription. Many approaches in the literature have tried to transcribe a piece of music, i.e. extract a score representation from the audio signal. We will show along this dissertation that transcription is not a mandatory step for music description. We can see audio description as different from transcription, as a broader term denoting any significant music description.

In addition to abstraction levels, we also differentiate different facets of music description. The main description facets correspond to the main axes of sound, including melody, harmony, rhythm, timbre, spatial location and dynamics. Combining automatically computed descriptors in those different description facets allows the user to navigate through music collections in a flexible, efficient and truly personalized way.

Audio content description technologies make it possible that this navigation is provided without the need of any manually annotated meta-data. One example of a system for music retrieval based on automatic audio description is MusicSurfer[13], where the results of this PhD dissertation are integrated (see Cano et al. (2005b,a)).

[13] http://musicsurfer.iua.upf.edu

1.4 Tonal description of audio

Tonality is one of the main aspects of Western music. Given its relevance, it deserves a proper description. It is then necessary to develop methods that can extract information related to the tonal content of a piece of music.

As we will see in Chapter 2, tonality is a concept linked to melodic and harmonic aspects, which is mainly related to music analysis. According to Cohen (1977), university music majors are able to sing, though not name, the scale of the key in which a piece is written correctly in 75% of the times after hearing the first measure. Although it could seem that the concept of key only have a sense for users having a musical training, much research has shown the relevance of tonality for naive users. For instance, some studies have shown that listeners acquire knowledge about western tonal regularities through mere exposure to musical pieces in everyday life (we refer to the work by Tillmann et al. (2003) or Tillmann and Bigand (2006)). At the same time, the key (mainly the mode) has been shown to induce a certain mood to the listener (see for instance Juslin and Sloboda (2001)). According to Auhagen and Vos (2000), experimental studies for tonality induction do not agree on how to differentiate between *experienced* and *inexperienced* subjects. Some approaches consider the degree of musical knowledge measured as the number of years of musical studies, or the ability to play or estimate the tonality, and also on the music enculturation in general. More research is necessary to study the difference between different musical training and cultural backgrounds.

1.4.1 Multidisciplinarity

According to Vos (2000), the multidisciplinarity of music research, and hence of tonality induction approaches, is one of the reasons why there is not a convergence between the different theories arising in each field. According to Vos, *because nobody is equally specialized in music theory, music history, psychoacoustics, music psychology, and so on, most theoretical approaches of tonality induction are biased towards one of the various scientific disciplines*. A review of the different approaches for tonality induction is presented in Chapter 2. We will also see that most of the proposed methods for tonality induction are focused on the analysis of score representations of music. When dealing with acoustic inputs, some other disciplines become relevant, such as signal processing or music acoustics. When implementing a computational model, it is also required some knowledge on the computer science field. Figure 1.1 shows a schema of the different disciplines involved in this PhD thesis in different degrees.

This dissertation focuses on computational aspects of the problem of tonality induction from audio signal, analyzing the main issues that appear when a computer is asked to describe the tonal content of a piece of music in audio format.

1.4.2 Practical aspects in automatic tonal description

When developing computational models for automatic tonal analysis, it is necessary to design evaluation strategies, in order to check the performance of the proposed methods. Along the literature, these evaluation

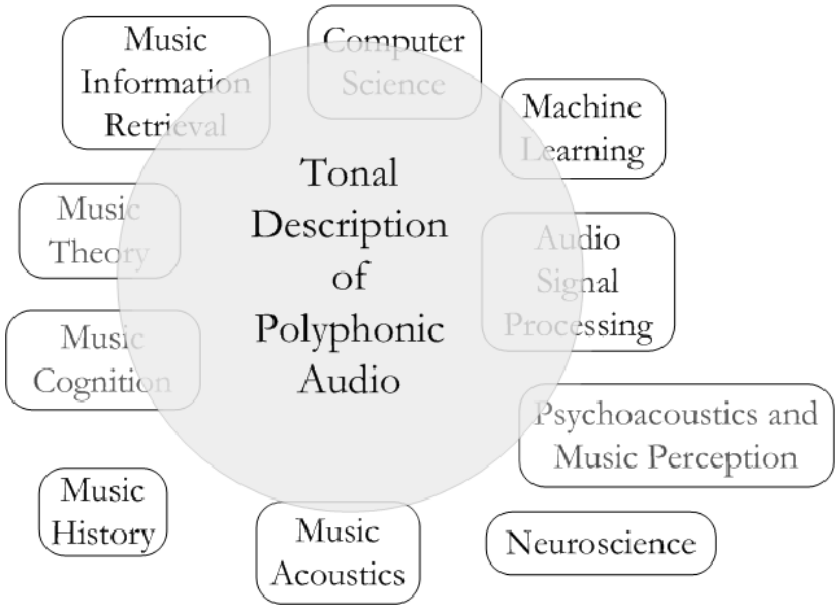


Figure 1.1: Disciplines involved in the problem of describing tonality of polyphonic audio recordings.

strategies have focused on a restricted set of pieces of mainly classical western music, as we will study in Section 2.9. The main reason for this is the availability of musical scores and their corresponding tonal analysis, manually performed by music experts. By comparing manual with automatic analysis, we could have an idea of the utility of the algorithms and models.

This manual analysis, intended to obtain ground truth descriptors, becomes very hard when dealing with a high amount of musical pieces, and when the score is not available, as it is the case in most popular music. Evaluation strategies should then provide quantitative evaluation measures of algorithms for an annotated collection. This collection should be as much representative as possible of the real situation and the variability of the music material.

Another factor that becomes relevant when analyzing huge amounts of data is the computational cost of the algorithms, as the time required for automatic analysis should be much lower than the time required for manual analysis, and acceptable within a real situation.

Along this dissertation, we propose solutions to these issues, including quantitative evaluation methodology and the implementation of the proposed methods within a working system dealing with huge amounts of data.

1.5 Application contexts

Some of the application contexts in which tonal description becomes relevant are listed here:

1. Find music titles that help to convey concepts to naive readers: similar harmonic contour, location of

versions of the same musical pieces (Cano et al. (2005b,a)).

2. Musical analysis of pieces of popular music in audio format, given that most of this type of music do not have an associated score.

3. DJ harmonic mixing, automatically detect the key of songs and adjust it without changing the tempo, in order to mix songs with close tonalities (e.g. Harmonic Mixing [14] or Mixmeister[15]).

This dissertation focuses on the first two application contexts, and it has been carried out in the context of the EU-IST-FP6 project SIMAC, presented in Herrera et al. (2005a). The SIMAC project addresses the study and development of innovative components for a music information retrieval system. The key feature is the usage and exploitation of semantic descriptors of musical content that are automatically extracted from music audio signals. These descriptors are generated in two ways: as derivations and combinations of lower-level descriptors and as generalizations induced from manually annotated databases by the application of machine learning techniques. The project aims also towards the empowering (i.e. adding value, improving effectiveness) of music consumption behaviors, especially of those that are guided by the concept of music similarity.

The description scheme proposed in SIMAC is based on different musical dimensions: rhythm, harmony, timbre and instrumentation, long-term structure, intensity and complexity. The music similarity measures are based on two different sources: audio-based similarity (based on automatically computed descriptors) and information from the web (web-based similarity). Web-based similarity addresses cultural information which can be taken into account, for instance, about artists.

Three software prototypes integrating state-of-the-art automatic audio description and music similarity technology have been developed within the SIMAC project:

- The **Music Annotator** is a tool for the annotation and generation of music meta-data at different levels of abstraction. It is composed of an annotation client that deals with micro-annotations (i.e. automatically extracted information related to a certain song: onsets, chords, beats, etc as explained in Amatriain et al. (2005)) and a collaborative annotation subsystem which manages large-scale manual annotation tasks that can be shared among different research centers (see Herrera et al. (2005b) for a detailed explanation).

- The **Music Organizer and Explorer** demonstrates the visualization and navigation among digital music collections. Two-dimensional maps are used to map songs according to semantic descriptors, and different similarity distance metrics are tried in order to find similar music to a given song.

- Finally, the **Music Recommender** provides recommendations of music titles that are legally downloadable from the WWW. This system, named *Foafing the music* [16], relies on user preferences and listening

[14] http://www.harmonic-mixing.com
[15] Mixmeister DJ Mixing software http://www.mixmeister.com
[16] http://foafing-the-music.iua.upf.edu

habits, which are computed from content analysis of the user's music collection. The system also exploits musical information crawled from the Internet and properly structured and minted to generate musical knowledge. We refer to Celma et al. (2005) for a further description on the recommendation system.

1.6 Goals

We present here an overview of the goals of this PhD dissertation, which are related to the hypothesis that we want to verify:

1. Review current efforts in audio description and tonality induction. This multidisciplinary study comprises signal processing methods, as well as music cognition and music theory approaches. We also study how the current literature related to score can be applied to audio.
2. Justify the role of tonal description for music content description.
3. Study how to establish a relationship between research developed for score analysis and signal processing technologies in order to perform a mid-level content description of audio.
4. Prove that it is possible to automatically extract a tonal description from audio signals, without the need of an exact transcription. Overcome the unavailability of multipitch extraction.
5. Justify that this description is valid to index musical collections and perform search by similarity.
6. Provide a quantitative evaluation of the proposed approaches with a varied music collection of popular music.

1.7 Summary of the PhD work

We believe that this dissertation brings significant contributions to the current state of the art in tonal description of polyphonic audio. As it is seen in Chapter 2, tonal description has only been the focus of literature over the last few years, in contrast with research on tonal induction from score representations. In this sense, the publications generated by this dissertation, which are summarized in Appendix D, belong to this recent literature which conforms the current state of the art in tonal description from audio. In addition to this general remark, we summarize here the work carried out along this PhD.

This dissertation first provides a study of different approaches for tonal induction, including score-based systems and signal processing methods. Based on this study, we have proposed and evaluated a state of the art approach to perform an automatic analysis of the tonal content of a piece of music in audio format.

A preliminary version of our approach was ranked third in the context of the International Conference of Music Information Retrieval and the Annual Music Information Retrieval Evaluation eXchange in 2005 [17]. In

[17] http://www.music-ir.org/mirex2005/index.php/

this context, different algorithms for audio key finding were compared, representing the state of the art at this moment. The system performed with an accuracy of 86.5% of correct key estimation over a data set of 1252 audio files of classical music synthesized from MIDI. The winner's accuracy was just 3% higher, and the best accuracy for symbolic key finding was just 4.8% higher. Based on these results, the method was improved and then evaluated in a quantitative and modular way. First, we evaluated the approach for the extraction of low-level tonal features from audio, including the computation of the tuning frequency, the comparison of different approaches for computing pitch class distribution features, the analysis of the influence of analysis parameters and the correspondence of pitch class profiles with note distributions. Then, the system for audio key finding was tested with a data set of 1450 annotated real recordings from different musical genres. In this context, the accuracy for global key estimation is around 77%. We have analyze different aspects important for audio key finding: the influence of the use of different tonal profiles, a comparison of cognitive versus machine learning modelling strategies, the the use of different segment duration and the performance for different musical genres.

After analyzing the validity of the approach of global key finding, the system has been extended in order to study the evolution of the tonal center in different temporal scopes, providing different views to describe the tonal content of a piece of music in audio format.

Beyond the analysis of a single piece, we think that this dissertation contributes to prove that tonal description provides a powerful tool for a musical meaningful indexing of music collections, and a way to measure similarity between musical pieces. We show that only using the proposed tonal descriptors and a very simple similarity measure, we can get an accuracy of 55% (recall level is equal to 30.8% and F measure of 0.393) when trying to identify different versions of the same piece.

Finally, this work has been integrated in a system for music organization and recommendation based on the analysis of audio signals. The integration of tonal description into the *MusicSurfer* [18] system has allowed to test the proposed automatic tonal description in a real situation, dealing with over a million of pieces, performing 20 times faster than real-time, and being found useful to search and recommend music.

1.8 Organization of the thesis

This dissertation is structured in different blocks. This structure is showed in Figure 1.2.

In this chapter, we have introduced the context of this PhD work and its associated problems. Chapter 2 presents an extensive and multidisciplinary background of studies and computational approaches for the automatic description of tonal aspects of music from audio recordings. We conclude this chapter having got an idea of the current challenges in the field and the contributions of this PhD work. We should note that most of the approaches for tonality description from audio recordings have been developed during the same period that this PhD thesis and their related publications, which are presented in Appendix D.

In Chapters 3 and 4, we propose and evaluate a computational method for the automatic tonal description

[18] http://musicsurfer.iua.upf.edu

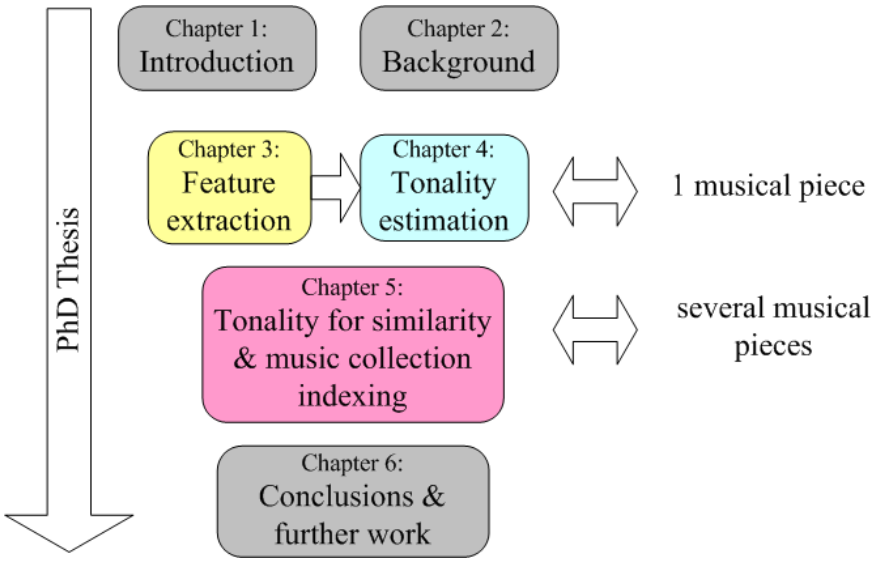


Figure 1.2: Organization and structure of the PhD thesis.

of a given piece of music in audio format. Once single pieces are characterized, Chapter 5 presents some methods for measuring the tonal similarity between two different pieces of music and then faces the problem of characterizing, using tonal descriptors, many pieces of a music collection in order to gain some understanding of this collection. Finally, our conclusions, summary of the main contributions and some future perspectives are presented in Chapter 6.

Chapter 2

Scientific background

2.1 Introduction

The goal of this chapter is to present the research context of this dissertation and place it in the current state of the art. We describe here how tonality has been studied in the literature, which are the most relevant tonal descriptors that have been considered and which are the computational approaches proposed to automatically describe tonality.

First of all, Section 2.2 introduces the most relevant concepts associated to tonality. We see in Section 2.2.4 that the term *tonality* takes different meanings in the literature. This fact and the multidisciplinarity of the different approaches for tonality induction explain the lack of convergence between different studies (as it has been already mentioned in Chapter 1).

Much of the literature related to tonality induction consists in experimental studies with human listeners. Many researchers from different disciplines have been trying to understand how we perceive tonality. Most of the studies concentrate on score representations of musical pieces, in order to simplify the material used for stimuli, as mentioned in Vos (2000). As as result of these studies, some tonal models have been defined, which are then used for find the key of pieces in score representations (mostly coded in MIDI). Most of the findings are coherent with western music theory. We review this literature in Section 2.3. Section 2.4 reviews the research efforts intended to locate modulation within pieces. We finally observe in Section 2.9 that most the approaches for key finding are evaluated in a restricted music collection of classical music.

In contrast to the huge amount of literature related to tonality induction from score representations, it is quite rare to find literature dealing with audio. Only the approach by Leman (2000) and few recent works including the one developed during this PhD thesis try to describe tonality from audio signals. These approaches follow the schema shown in Figure 2.1. The methods for the different blocks in the figure (feature computation, tonal model adaptation and comparison) are reviewed in Section 2.6. We should note that most of them have been developed during the same period as this PhD thesis and their related publications, which

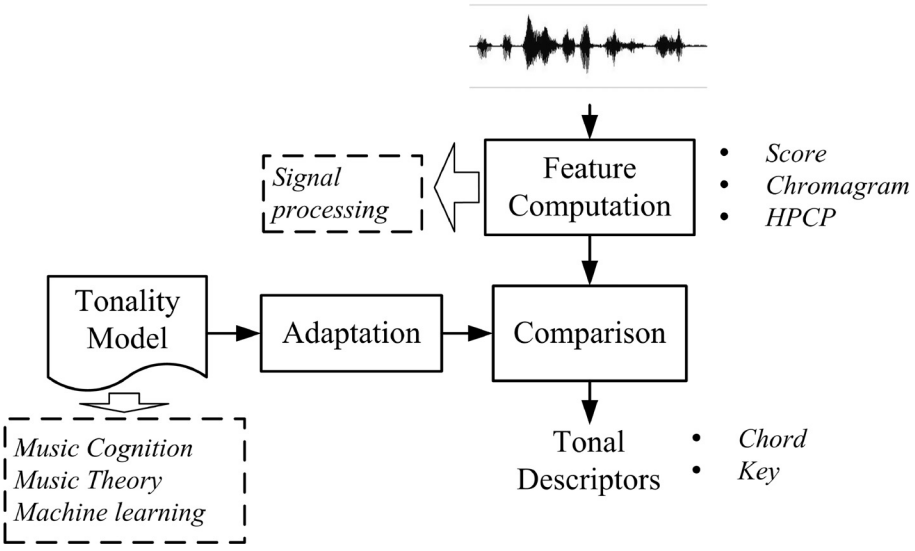


Figure 2.1: Block diagram of systems for tonality description from audio.

are presented in Appendix D. We first present methods for automatic pitch estimation and melodic description, in order to get a score transcription from audio. We see in Section 2.6.2 that the exact transcription is not critical for our purposes. A comparison of the computational models for extracting pitch class distribution features from audio is made in Section 2.6.3, indicating the requirements for this type of descriptors and how the different methods fulfil then. The tonal models are adapted in different ways to this type of audio representation, as reviewed in Section 2.7. This PhD thesis proposes the use of these types of features to perform tonal description. We finally analyze the methods and databases used for evaluation in Section 2.9.

After this overview of the related background, we summarize in Section 2.10 the contributions of this work to the current state of the art: first, it provides a system for tonal description and tonal similarity from audio recordings of popular music. Second, it performs a massive evaluation of the proposed approach in annotated music collections.

2.2 Relevant concepts

2.2.1 Frequency and pitch

The most basic signal feature related to tonality is *frequency* and its corresponding subjective attribute *pitch*. Pitch is one of the main basic dimension of sound, together with loudness, duration, timbre and spatial location.

A simple waveform is represented by a sinusoidal function, as shown in equation 2.1:

$$x(t) = a \cdot \sin(2 \cdot \pi \cdot f \cdot t + \phi) \tag{2.1}$$

where a represent the maximum amplitude, t represents the time (in seconds), f the frequency and ϕ the initial phase.

The frequency f of this simple waveform is defined as the number of times that a cycle is repeated per second (Sadie et al. (2005)). It is usually measured in cycles per second, or Hertz (Hz). For instance, a sinusoidal wave with a frequency equal to A 440 Hz performs 440 cycles per second. The inverse of the frequency f is called the period T, which is measured in seconds and indicates the temporal duration of one oscillation of the sinusoidal signal (equation 2.2).

$$T = \frac{1}{f} \tag{2.2}$$

Most of the pitched sounds are complex waveforms consisting of several components (called *partials* or *harmonics*). The frequency of each component is multiple of the lowest frequency f_0, called the *fundamental frequency*:

$$x(t) = \sum_{n=1}^{N} a_n \cdot \sin(2 \cdot \pi \cdot n \cdot f_0 \cdot t + \phi_n) \tag{2.3}$$

The perceptual counterpart of frequency (e.g. the physical measurement of vibrations per second) is pitch, which is a subjective quality often described as highness or lowness.

Although the pitch of complex tones is usually related to the pitch of the fundamental frequency, it can be influenced by other factors such as for instance timbre. Some studies have shown that one can perceive the pitch of a complex tone even though the frequency component corresponding to the pitch may not be present (denoted as *missing fundamental*) (Schmuckler (2004), pp. 274). Is it out of the scope of this PhD thesis to review the literature dealing with pitch perception. We refer to Schmuckler (2004) and de Cheveigné (2005) for a comprehensive and up to date review on the issue.

In western music, the pitch scale is logarithmic, i.e. adding a certain interval corresponds to multiplying a fundamental frequency by a given factor. Then, an interval is defined by a ratio between two fundamental frequencies f_1 and f_2. For an equal-tempered scale, a semitone is defined by a frequency ratio of:

$$\frac{f_2}{f_1} = 2^{\frac{1}{12}} \tag{2.4}$$

An interval of n semitones is defined by a frequency ratio of:

$$\frac{f_2}{f_1} = 2^{\frac{n}{12}} \tag{2.5}$$

that is, the interval in semitones n between two fundamental frequencies f_1 and f_2 is defined by

$$n = 12 \cdot \log_2(\frac{f_2}{f_1}) \tag{2.6}$$

The first harmonic frequencies of a tone form the approximate intervals with the fundamental frequency represented in Table 2.1 and Figure 2.2.

Harmonic	**Frequency**	**Approximate interval with f_0**	**Pitch class**
1	f_0	unison	A
2	$2 \cdot f_0$	octave	A
3	$3 \cdot f_0$	octave + 5th	E
4	$4 \cdot f_0$	2 octaves	A
5	$5 \cdot f_0$	2 octaves + major 3rd	C#
6	$6 \cdot f_0$	2 octaves + 5th	E
7	$7 \cdot f_0$	2 octaves + 7th	G
8	$8 \cdot f_0$	3 octaves	A

Table 2.1: Intervals between the first 8 harmonics of a complex tone and its fundamental frequency f_0. Example for the harmonics of A.

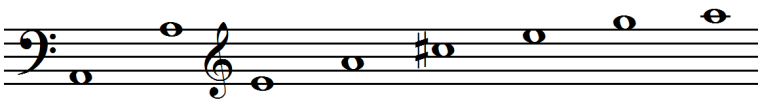


Figure 2.2: Harmonic series from A2.

In western music notation and equal-tempered scale, fundamental frequencies are quantized to pitch values using a resolution of one semitone. The A 440 Hz is considered as the standard reference frequency, although we cannot assume that orchestras and bands will always be tuned to this pitch. We will study this problem in Section 2.6.3 and Chapter 3.

Within a musical context, pitch is represented by a Bi-dimensional model sometimes called the *pitch helix* (see Shepard (1982)) represented in Figure 2.3. This representation is motivated by the observation that some people, particularly trained musicians, perceive tones in a cyclic manner, where the cycle repeats every octave (12 semitones). A pitch is then represented by two descriptors: *height*, moving vertically in octaves, and *chroma*, or *pitch class* determining the rotation position within the helix.

Octave equivalence has been studied in the literature (Deutsch (1999)), and is commonly accepted. Nevertheless, Deutsch (1999) analyzes several musical paradoxes, concluding that these two descriptors (*pitch class* and *height*) are not orthogonal dimensions. This conclusion violates in some way the principle of perceptual proximity under transposition, as explained in Deutsch (1999) pp. 385.

The musical intuitiveness of the chroma makes it an ideal feature representation for note/chord events in musical signals. A temporal sequence of pitch classes results in a time-frequency representation of the signal.

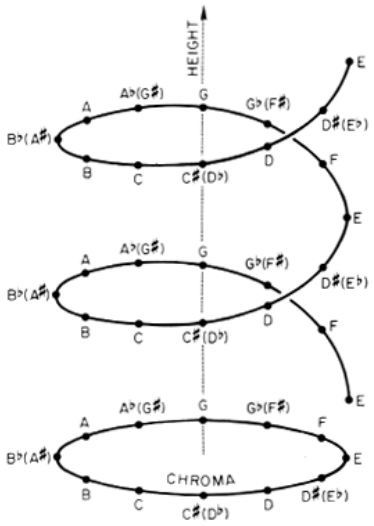


Figure 2.3: Pitch helix representation from Shepard (1982).

2.2.2 Melody

The concept of melody is usually associated to a sequence of pitch notes. This definition can be found for example in Solomon (1996): *"a combination of a pitch series and a rhythm having a clearly defined shape"*, and on Grove Music (Sadie et al. (2005)): *"pitched sounds arranged in musical time in accordance with given cultural conventions and constraints"*. Goto also considers the melody as a sequence of pitches, the most predominant detected pitches at middle and high frequency regions, in opposition to bass line, that can be found at low frequency bands (Goto (1999, 2000)).

2.2.3 Harmony

Harmony is a term that denotes the simultaneous combination of notes, called *chords*, and over time, chord *progressions*. The term is used to describe notes and chords, and also to denote a system of structural principles governing their combination. In the latter sense, harmony has its own body of theoretical literature (Sadie et al. (2005)). In this work we will consider only the aspects of the harmonic content related to the combination of notes into chords, and its relation to the tonality of the piece.

As we have seen, *melody* and *harmony* both refer to the combination of pitches (either *sequential* or *simultaneous*). In this sense, it is quite difficult to separate melody from harmony, as they influence each other, as it is pointed out by Krumhansl (2004).

2.2.4 Tonality

According to Vos (2000), one of the reasons for the lack of convergence between different approaches for tonality induction is the fuzziness of the concept of *tonality*. According to Vos (2000), the term *tonality* was first introduced by Castil-Blaze in 1821. Nowadays, it is primarily used to denote a system of relationships between a series of pitches (forming melodies and harmonies) having a *tonic*, or central pitch class, as its most important (or stable) element (e.g. see its definition in Sadie et al. (2005)). In its broadest possible sense, it refers to the arrangements of pitch phenomena. We describe here how tonality is defined in music theory. Many studies also investigate how is the listener's perception of a tonal center in a piece of music.

The majority of empirical research has been devoted to western music, as shown in Krumhansl (2004). In western tonal music, we define *key* as a system of relationships between a series of pitches having a *tonic*, or central pitch class, as its most important element. Besides the tonic, one of the most important pitch classes of a key is the *dominant* degree, defined as the fifth degree of the scale. Another important degree is the *subdominant* degree, which is the fourth degree of the scale and lies below the tonic as much as the dominant lies above it, that is, a 5th.

There are two basic key *modes*: major and minor. Each of them has different musical characteristics regarding the position of tones and semitones within their respective scales. A *scale* is composed of a sequence of notes; each two notes form a certain interval (see Figure 2.4). Major and minor natural scales are represented in Figure 2.4.

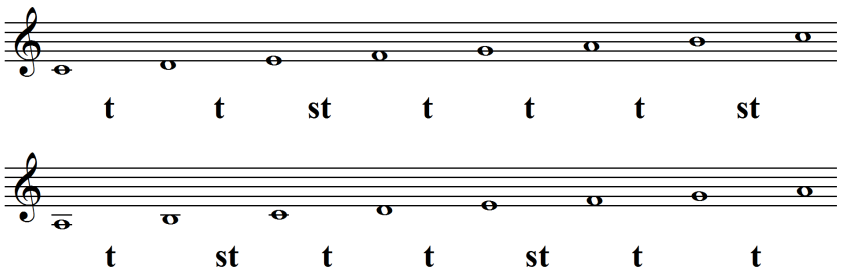


Figure 2.4: C Major and A Minor (natural) Scales. Intervals between consecutive notes are shown. *t*: Tone. *st*: Semitone.

The *natural* minor scale is written with no accidentals. A characteristic of the minor mode in the common practice period of Western music is the use of the seventh tone a half semitone below the tonic, so that the dominant chord is always a major triad. As a consequence, the seventh degree of the scale must be raised with an accidental (G# in Figure 2.4). This leads to the *harmonic* minor scale. The interval between the sixth and the seventh degree is then an augmented second. Some composers have also raised the sixth degree in order to get a smoother ascending melody (F# in Figure 2.4), called *melodic* minor scale.

We can see in Figures 2.2 and 2.4 that the first harmonics of a complex tone belong to the major key

defined by the pitch class of the fundamental frequency, and all of them except the 7th harmonic belong to its tonic triad (C, E, G in C Major, Figure 2.4).

When each tonic manages both a major and a minor mode, there exist a total of 24 keys, given that we choose one of the possible minor scales, if we consider an equal-tempered scale and enharmonic equivalence (i.e. we do not distinguish between notes that sound the same but are spelled differently, such as D♯ and E♭). This corresponds to two different keys for each of the 12 semitones within the chromatic octave. These 24 keys can be arranged in a circle of fifths, as shown in Figure 2.5. Each pair of major and minor modes has the same collection of pitch classes and key signature (i.e. the group of sharp or flat signs representing the key), while the collections of neighboring, 5th-related pairs differ by a one sharp or flat.

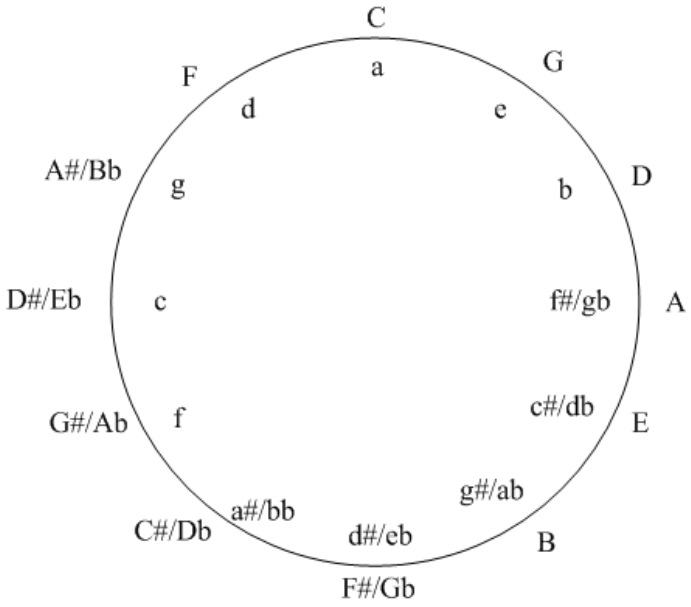


Figure 2.5: Circle of fifths representing major and minor (small caps) tonalities.

There are also some relationships between tonalities: the *parallel* major and minor (e.g., C major and C minor) share the same tonic but have different diatonic collections, while the *relative* major and minor (e.g., C major and A minor) share the same diatonic collection but have different tonics. A key is not limited to the pitch classes within its particular diatonic collection. In certain circumstances, the music can use pitch classes outside its tonic major or minor scale without weakening its sense of orientation towards the tonic. If the orientation towards the tonic is very strong, the music is considered to be very *tonal*, and in the opposite sense there is the concept of *atonality* (Sadie et al. (2005)).

2.3 Tonality induction

There have been many efforts in the literature to model the human cognition of tonality, mainly in the fields of cognitive science and music psychology. Most of them focus their studies on western music (e.g. Longuet-Higgins and Steedman (1971); Temperley (1999); Krumhansl (2000); Chew (2000)), although there have been some efforts to analyze other tonal systems (see for instance Krumhansl (2000)). We focus here on the tonality models and their relation to tonality induction.

As mentioned in Section 2.2.1, Shepard designed a model which spaced all twelve pitches equally over one full turn of a spiral (see Figure 2.3) (Shepard (1982)). This model emphasizes the close relationship between pitches related by octave intervals. Further extensions to incorporate perfect fifth interval relations resulted in double helix structures that still did not explained the major third.

One of the first studies dealing with automatic tonal analysis is the one by Winograd (1968), which proposes a method for the automatic harmonic analysis of a musical piece, using ideas derived from linguistics. Riemann (19th century music theorist), stated that the tonality derives from establishing of significant tonal relationships through chord functions. This theory agrees that the most relevant intervals are the perfect fifth and the major/minor third, which are present in the major and minor triads. Riemman represented these relations in a harmonic network: the *Tonnetz*. The Tonnetz, shown in Figure 2.6, represents the set of pitch classes, where criss crosses horizontals of perfect 5ths with diagonals of major and minor thirds. Lewin and Cohn defined a transformational theory within the *Tonnetz* (Lewin (1987); Cohn (1997)).

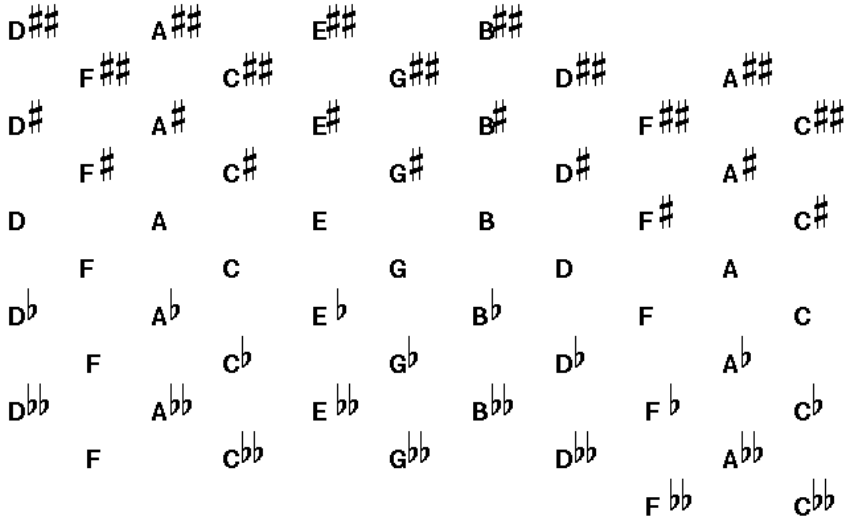


Figure 2.6: Tonnetz or harmonic network (from Sapp (2001)), representing the set of pitch classes, where crosses horizontals of perfect 5ths with diagonals of major and minor thirds.

Longuet-Higgins and Steedman (1971) noted that pitches in a given key are located in a compact neighborhood on the tonnetz or harmonic network. Following this approach, Longuet-Higgins and Steedman (1971) proposed a key finding method based on a Shape Matching Algorithm (SMA). This key estimation algorithm analyzes the different tones of the score in order to find whether they are contained or not in the diatonic major and minor scales. It eliminates the keys having a tone in the sequence which is non diatonic on this key. If all keys are eliminated or there is more than one key candidate, they use a tonic-dominant rule to estimate the key. According to Temperley (1999), we could see this model as a very simple key-profile model, each key having a "flat" key profile where all the pitch classes in the corresponding diatonic scale have a value of one, and all chromatic pitch classes have a value of zero. The input vector is also flat: a pitch class has a value of one if it is present anywhere in the passage, zero if it is not. Choosing the correct key is a matter of correlating the input vector (composed by the set of pitch classes of the score) with the key profile vectors. According to Temperley, we might call it a *flat input/flat key profile* model. The flat key profiles for major and minor (harmonic) modes are represented in Figure 2.7.

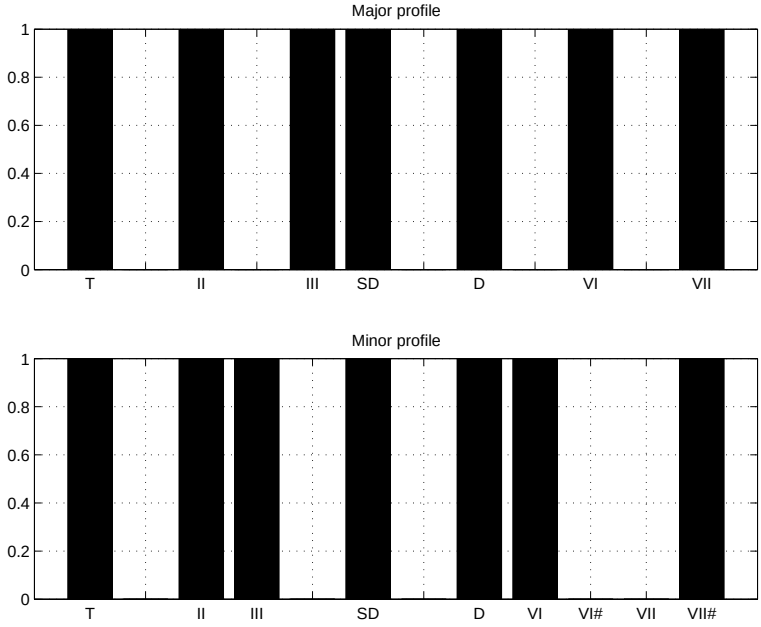


Figure 2.7: Major and Minor (harmonic scale) flat profiles.

Based on the work by Longuet-Higgins, Chew (2000) proposed a spiral representation of pitches by "rolling up" the harmonic network. This representation is illustrated in Figure 2.8. In contrast with the spiral representation of pitch classes defined by Shepard (and shown in Figure 2.3), a vertical step corresponds to a major third, and a rotation within the helix corresponds to a perfect fifth. In this model, each key is represented by a given point, or *center of effect*, obtained by a combination of the tonic, dominant and subdominant triads of the key. Each triad is a composite result of its component pitches. Each chord is represented by a point in

the inside of a triangle formed by joining its component pitches. In the same way, each key is represented by a point in the inside of a triangle formed by joining its component main chords. Based on this model, Chew proposed a key-finding method, called the CEG method, that used pitch and duration information to generate a center of effect which was then compared to the different keys in the spiral array.

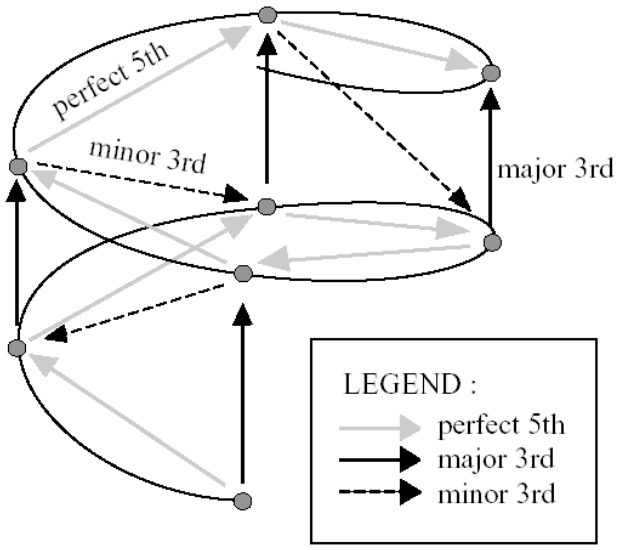


Figure 2.8: Interval representation in the spiral array (from Chew (2000)).

Krumhansl and collaborators (Krumhansl (1990)) defined a basic space by multidimensional scaling, in order to represent pitch proximity (Figure 2.9) and chord proximity (Figure 2.10). They also defined a basic space to represent major and minor keys. This space was obtained using experimental data, and it is shown in Figure 2.11. The horizontal axis corresponds to the angle of each key on the circle in the first two dimensions; the vertical axis corresponds to the angle of each key on the circle in the last two dimensions. Because angle is a circular dimension, top and bottom edges are to be considered the same, and left and right edges are to be considered the same (Krumhansl (1990)), forming the surface of a torus. The horizontal axis of this figure corresponds to major thirds and diagonals to the circle of fifths.

This space was obtained by analyzing, using correlation and multidimensional scaling, a set of tonal profiles, representing the tonal hierarchies of the 24 major and minor keys. They used this map to trace how the sense of key develops and changes over time.

Tonal profiles for major and minor keys are shown in Figure 2.12. Each of them contains 12 values, which are the ratings of the degree to which each of the 12 chromatic scale tones fit a particular key. These profiles were obtained by analyzing human judgements with regard to the relationship between pitch classes and keys (Krumhansl (2004), pp. 78-81). In this study, an unfinished C major scale (without the final tonic) was played in either its ascending or descending form, in order to establish the C major key. After establishing

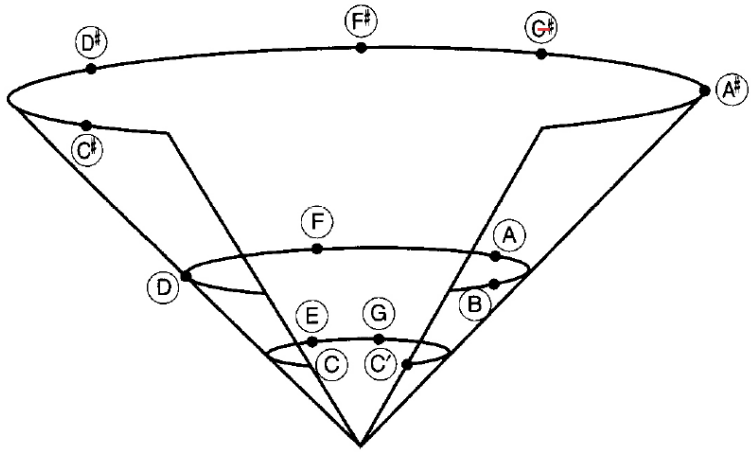


Figure 2.9: Krumhansl's spatial representation derived from empirical data: pitch class proximity. As the radius of the circle grows, the proximity to the considered pitch class (C) decreases. The closest pitch classes, located over the smallest ring, belong to the C major tonic triad. The second ring includes the pitch classes from the diatonic major scale, and the final ring includes also non-diatonic degrees (from Lerdahl (2001), pp. 46 a)).

this context, listeners were presented with one of the 12 chromatic scale tones in the next octave (called the probe tones), and they rated how well each tone completed the scale. Krumhansl and Kessler later extended this method to a variety of ways to establish the key, including chord cadences and both major and minor scales. We can verify in Figure 2.12 that these profiles agree with some knowledge from music theory: the tonic is most predominant, followed by the fifth, third, the remaining scale tones and finally the non-scale tones.

As an application of this key model, Krumhansl and Schmuckler proposed a key estimation algorithm from MIDI representations (Krumhansl (1990)). This approach estimates the key from a set of note duration values, measuring how long each of the 12 pitch classes of an octave (C, C♯, etc.) have been played in a melodic line. In order to estimate the key of the melodic line, the vector of note durations is correlated to the set of key profiles or probe-tone profiles. According to Temperley (1999), this approach is a *weighted input/weighted key profile* model, in contrast with the *flat input/flat key profile* approach from Longuet-Higgins. According to Krumhansl (1990), this method outperformed the one from Longuet-Higgins and Steedman (1971) when applied to Bach's *Well-Tempered Clavier*. One feature of this algorithm is that it allows for the possibility that a number of keys might be quite strongly suggested simultaneously, or that no key is suggested (Krumhansl (2004)).

Temperley (1999) proposed some modifications to the Krumhansl & Schmuckler algorithm: the first one consists on increasing the weight of the seventh scale degree in major and the raised seventh in minor mode. He modified Krumhansl and Schmuckler's profiles, as shown in Figures 2.13, 2.14 and 2.15. A second

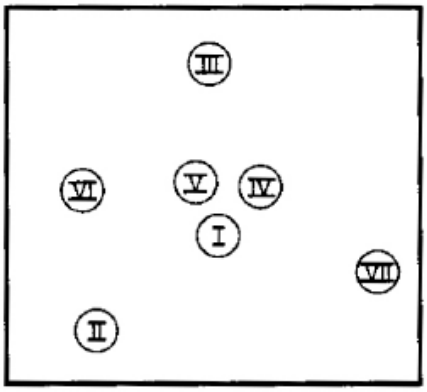


Figure 2.10: Krumhansl's spatial representation of chord proximity derived from empirical data, where tonic (I), dominant (V) and subdominant (IV) chords are close to each other (from Lerdahl (2001), pp. 46 b)).

improvement proposed by Temperley to Krumhansl and Schmuckler's method was to ignore note durations (as Longuet-Higgins approach) by using a *flat input/weighted key profile* approach, and to use a matching formula instead of correlation. When tracking key, a penalty was imposed for changing key from one input segment to the next.

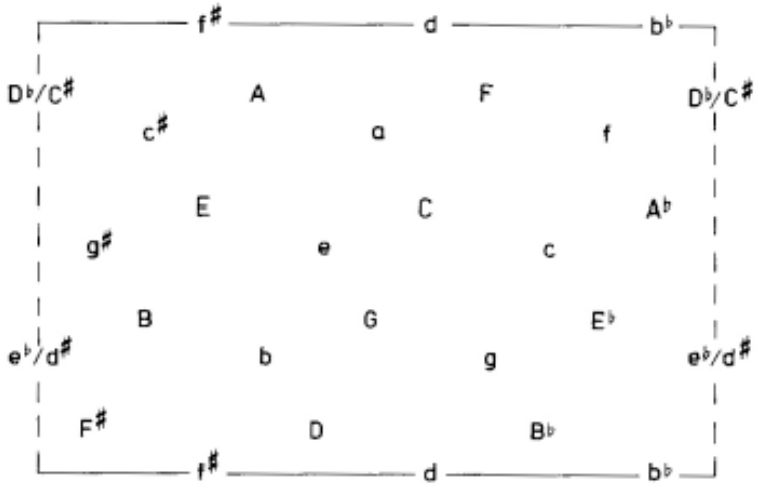


Figure 2.11: Rectangular representation of the multidimensional scaling solution (Krumhansl and Kessler 1982), in which the 24 major and minor (small caps) keys are located in the surface of a torus. The horizontal axis corresponds to major thirds and diagonals to the circle of fifths. The figure is from Krumhansl (1990), pp. 46.

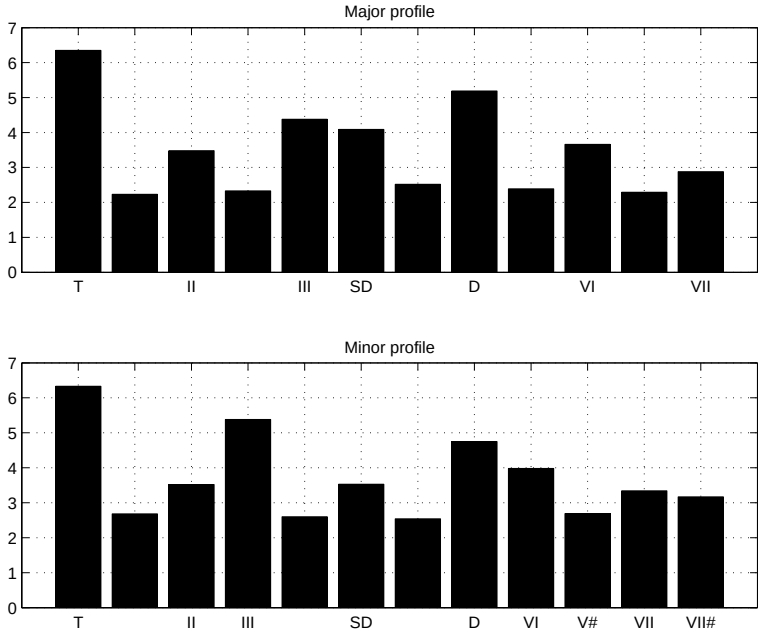


Figure 2.12: Major and Minor profiles as proposed by Krumhansl and Schmuckler (Krumhansl (1990)).

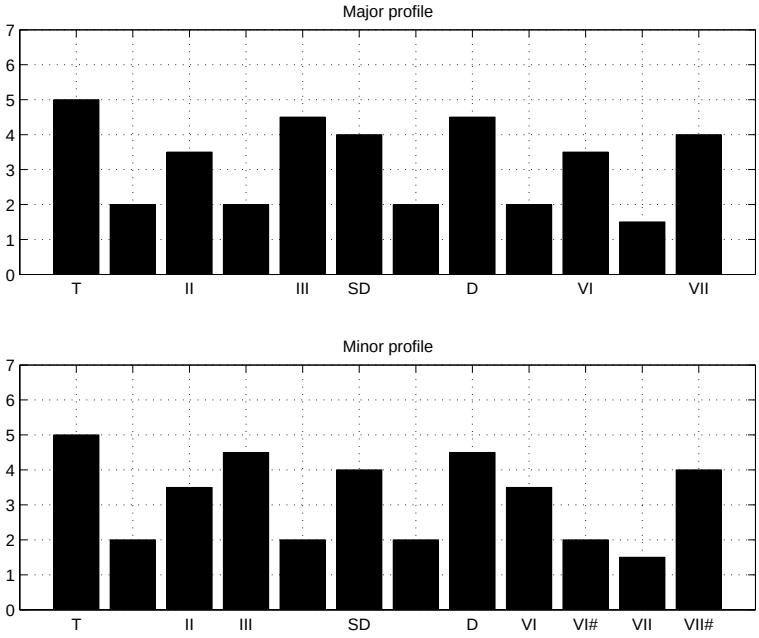


Figure 2.13: Major and Minor profiles as proposed by Temperley (1999).

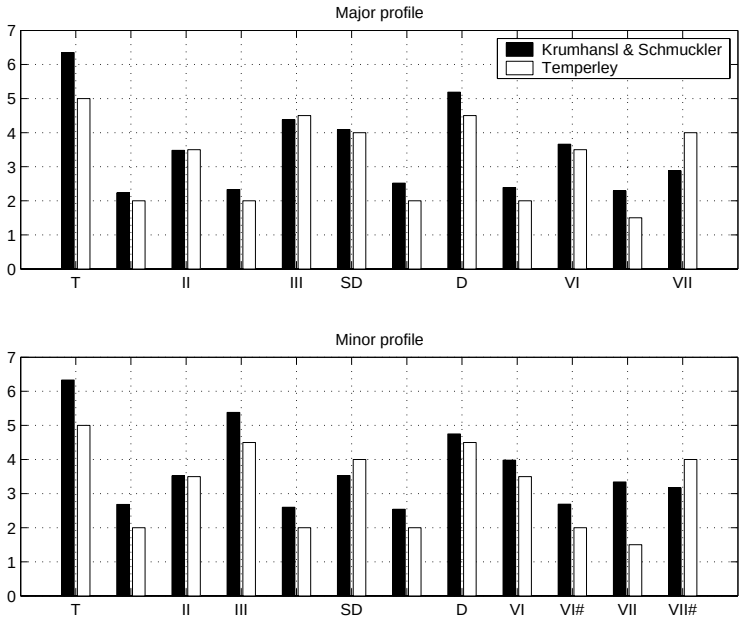


Figure 2.14: Comparison of Major and Minor profiles proposed by Krumhansl and Schmuckler (Krumhansl (1990)) and Temperley (1999).

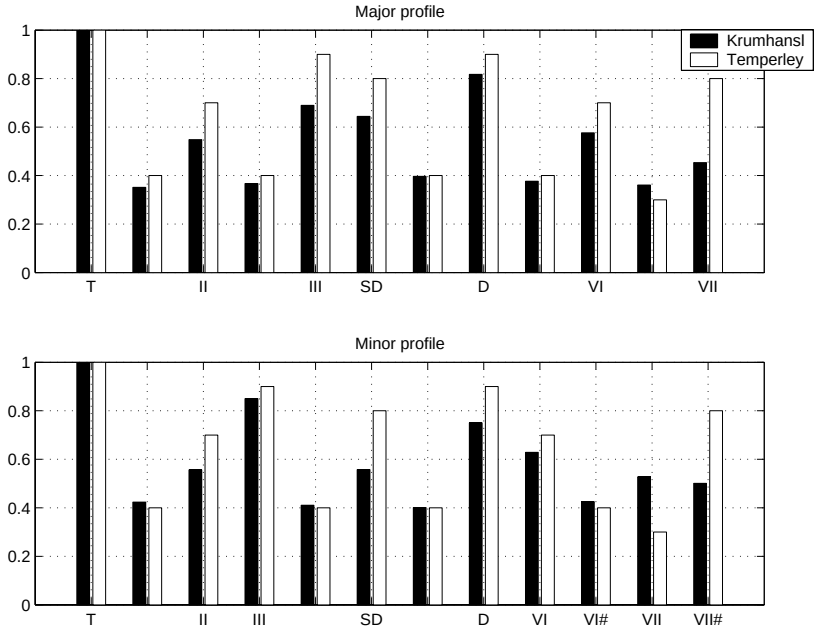


Figure 2.15: Comparison of Major and Minor profiles proposed by Krumhansl and Schmuckler (Krumhansl (1990)) and Temperley (1999), normalized by their maximum value (corresponding to the first degree or tonic (T)).

As an alternative to the profiles derived from human ratings, some studies introduce the idea of learning the profiles from symbolic musical data. Krumhansl reported statistical distributions of pitch classes from classical pieces performed only on the melodic lines Krumhansl (1990) pp. 62-76. These distributions were strongly correlated with the probe tone profiles.

In her PhD thesis, Chai (2005) also obtains a profile by training with 7673 folk music scores. The profile was generated in a similar way than reported in Krumhansl (1990): get the key of each piece, count the number of times that each note appears, average the vectors over all the considered pieces and finally normalize it. Figure 2.16 shows this profile (Chai 2005, personal communication). This approach neglects the relative duration of each of the pitch classes, which is an important factor in the approaches derived from the one by Krumhansl.

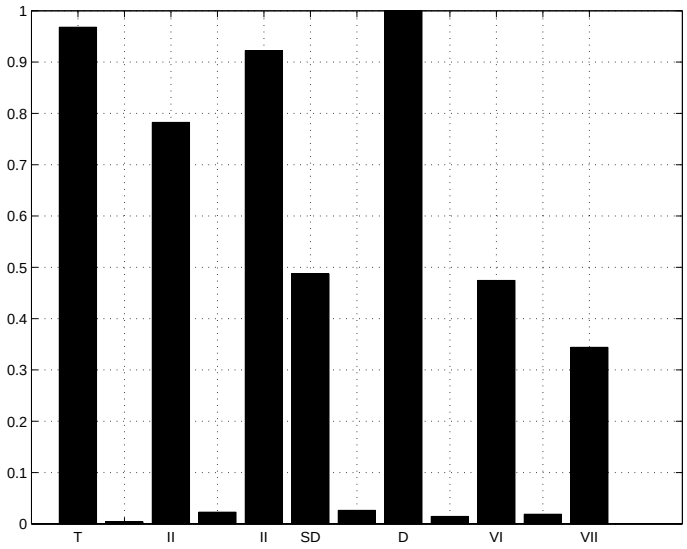


Figure 2.16: Tonal profile proposed by Chai (2005) and obtained by statistical analysis of symbolic data. The profile is normalized by its maximum value (corresponding to the dominant degree).

In her automatic key finding approach, Chai first estimates the key of the piece without considering its mode (e.g. C major equal to A minor) and in a second step the mode is detected by using two profiles for mode which are shown in Figure 2.17. These profiles were empirically obtained, based on the fact that in order to distinguish between a major key and its relative minor key (e.g. C major and A minor), it is necessary to measure the strength of the sixth degree (i.e. the tonic of the minor scale, e.g. A) with respect to the dominant (e.g. G).

Also in Temperley (2005), major and minor key profiles are derived empirically, using a corpus of excerpts taken from the Kostka and Payne music theory textbook (Kostka and Payne (1995)), in which keys are explicitly marked. These profiles are shown in Figure 2.18. Compared to figure 2.13, we can see that the notes belonging to the diatonic sets are emphasized with respect to the others.

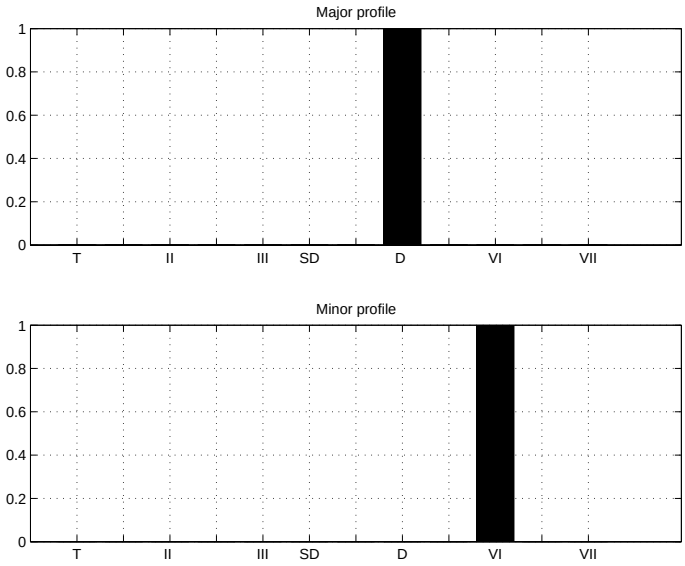


Figure 2.17: Mode profiles proposed in Chai (2005) and obtained empirically.

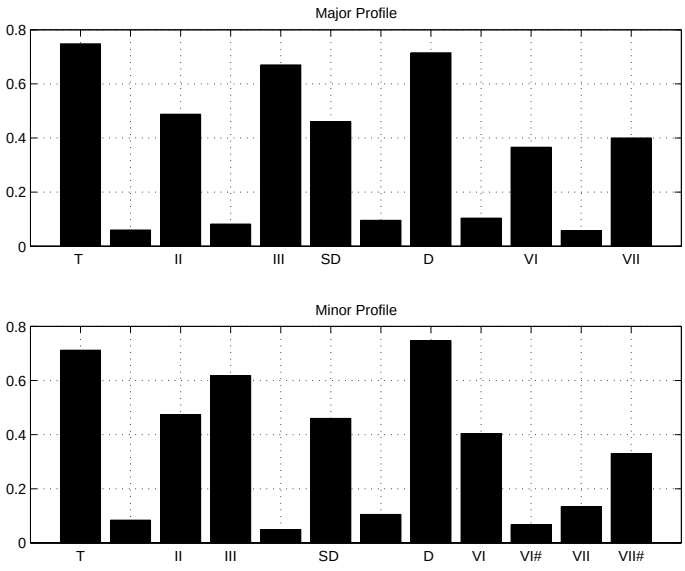


Figure 2.18: Mode profiles proposed in Temperley (2005) and obtained empirically.

Although the approach from Temperley considers modulations or key changes, template-based methods analyzing the sum of the relative durations of pitch classes do not take into account the temporal location of the pitches within the piece. But key attribution may become unambiguous depending on the ordering of tones, as explained in Deutsch (1999). According to Deutsch, *it seems that users make grouping of notes according to pitch proximity and other rules, so that the process of key cognition is very complex, including low-level grouping, knowledge of the pitches in a diatonic collection and of the hierarchies of prominence of tones in different tonalities* (Deutsch (1999) pp. 375). In order to study the influence of the temporal location of the pitches within the piece, Toiviainen and Krumhansl (2003) compared a model based on pitch-class distribution with another one based on the tone-transition distribution (representing sequences of pitch classes), in order to check whether the order of tones provides additional information about tonality. Both models contained dynamic components for characterizing pitch strength and creating pitch memory representations. According to them, *both models produced results closely matching those of the probe-tone data*.

There are some studies made on Krumhansl and Kessler's profiles in order to study whether their findings are the result of the influence of short-term memory into key perception. Huron and Parncutt (1993) shown that a short-term memory model based on Terhardt's pitch model (Terhardt (1979)), plus an echoic memory, explained some of the data reported in Krumhansl and Kessler's study. Following this idea, Leman (2000) describes an auditory model including a short-term memory model that gave an explanation to Krumhansl & Kessler's data. According to Deutsch (1999), pp. 374, *the extent to which ratings were driven by short-term memory, long-term memory or yet other factors remains unresolved.* Also the neural network model by Bharucha (1999) takes into account short-term memory.

As an alternative to template-based methods, Martens et al. (2004) proposed the use of classification trees as a supervised learning technique for key classification.

Finally, as an alternative to the use of acoustic inputs and observation of user responses, brain research introduces the use of brain activity patterns to study tonality induction. Krumhansl (2004) considers the prospects for studying tonality with brain imaging techniques, providing an overview of related literature, which is out of the scope of this dissertation.

There are also some rule-based theories of tonality, the most relevant one being the *Generative Theory of Tonal Music* (GTTM) proposed by Lerdahl and Jackendoff (1983). According to Lerdahl (2001), pp 3., *GTTM attempts to characterize musical structures that are hierarchical and to establish principles by which the listener arrives at them for a work in the Classical tonal idiom. These principles are stated as musical grammar or system of rules*.

The sound signal is considered as a *musical surface* (or sequence of events). This sequence of events acts as input to a rule system (musical grammar) which establishes a structural description, representing the heard structure. The GTTM defines four types of rules, represented in Figure 2.19: grouping structure (related to segmentation into motives, phrases and sections), metrical structure (defining hierarchy between strong and weak beats), time-span reduction (establishing the relative importance of events in the rhythmic units of a piece) and prolongational reduction (hierarchy in terms of perceived patterns of tension and relaxation). These

four hierarchies are integrated as presented in Figure 2.19 (Lerdahl (2001), pp.4). The stability conditions are related to the tonal system, which Lerdhal treated through the concept of *pitch space*.

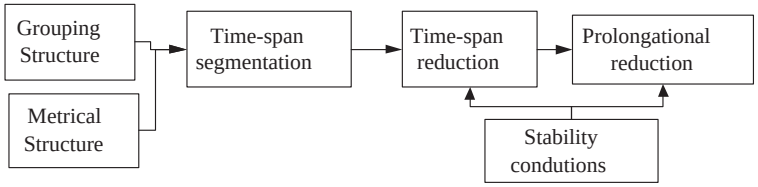


Figure 2.19: A flow chart of GTTM's components (from Lerdahl (2001), pp. 4).

A tonal hierarchy is not related to a sequence of events, but embodies an atemporal hierarchy related to a given tonal system. This tonal hierarchy is acquired by listening.

- The basic space: this space defines different levels for pitch classes, represented in Figure 2.21: octave (*level a*, e.g. C), fifth (*level b*, e.g. C-G), tonic triad (*level c*, e.g. C-E-G), diatonic scale (*level d*, e.g. C-D-E-F-G-A-B) and chromatic scale (*level e*, e.g. C-Db-D-Eb-E-F-F#-G-Ab-A-Bb-B).

- The pitch class level: distance between pitch classes can be measured according to the above mentioned levels, as represented in Figure 2.20.

- The chordal level: chords are represented at level *c*, with level *b* begin the fifth of the chord and level *a* the root (given that we only consider triad chords). In order to compute the proximity of two different chords, this model uses the diatonic circle of fifths and common tones. The chordal space is represented in Figure 2.22.

- The regional level: this level is related to the key, and the region is considered as the diatonic collection. Calculating the regional distance depends on moving the diatonic level (*d*) on the circle of fifths. The regional space, which is represented in Figure 2.23, is created by combining the fifths cycle and the parallel/relative major-minor cycle.

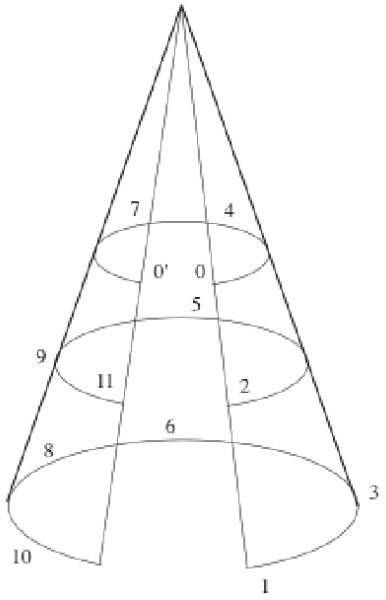


Figure 2.20: The pitch class cone representation by Lerdahl representing pitch class proximity (from 0 to 11 to consider the 12 pitch classes). This representation is analogous and inverted with respect to Figure 2.9. As the radius of the circle grows, the proximity to the considered pitch class decreases. The closest pitch classes, located over the smallest ring, belong to the major tonic triad. The second ring includes the pitch classes from the diatonic major scale, and the final ring includes also non-diatonic degrees (from Lerdahl (2001), pp. 50).

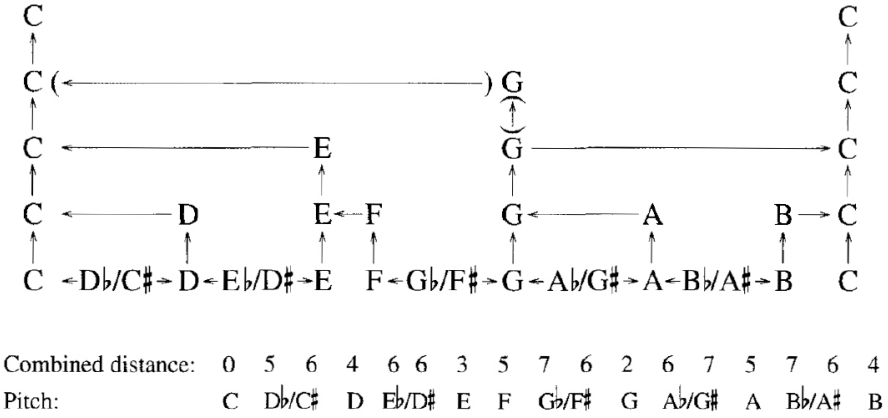


Figure 2.21: Pitch class and pitch proximity in Lerdahl's basic space, representing the stepwise horizontal and vertical pitch paths (from Lerdahl (2001), pp. 49). This arrangement is attributed to Deutsch and Feroe.

vii°	ii	IV	vi	I	iii	V
iii	V	vii°	ii	IV	vi	I
vi	I	iii	V	vii°	ii	IV
ii	IV	vi	I	iii	V	vii°
V	vii°	ii	IV	vi	I	iii
I	iii	V	vii°	ii	IV	vi
IV	VI	I	iii	V	vii°	ii

Figure 2.22: Chordal space, created by combining the diatonic cycle of fifths and common tones between chords. The dashes encloses the "chordal core" (from Lerdahl (2001), pp. 57).

D♯	F♯	f♯	A	a	C	c
G♯	B	b	D	d	F	f
C♯	E	e	G	g	B♭	b♭
F♯	A	a	C	c	E♭	e♭
b	D	d	F	f	A♭	a♭
e	G	g	B♭	b♭	D♭	d♭
a	C	c	E♭	e♭	G♭	g♭

Figure 2.23: Regional space representing the different keys, created by combining the fifths cycle and the parallel/relative major-minor cycle. Small caps represent minor keys (from Lerdahl (2001), pp. 65).

According to Bharucha, *rule-based theories of music, such as the one of Lerdhal and Jackendoff, can be construed as formalizations of constraints on neural processing of music* (Bharucha (1999) pp. 436). That means that we can see either neural networks as implementations of these grammars or grammars as formal descriptions of neural networks. We give now a brief overview of tonal models based on neural networks based on the work by Deutsch and Bharucha. These models are based on considering some aspects of cognition as the result of neural association of patterns (Bharucha (1999), pp. 413).

Deutsch proposed in 1969 a block diagram that would account for some low-level pitch relationships, based on establishing analogies with the visual system (considering aspects such as orientation and angle size) (Deutsch (1999) pp. 352). This diagram, shown in Figure 2.24, consists on two parallel channels, and the information is abstracted in two stages for each of them.

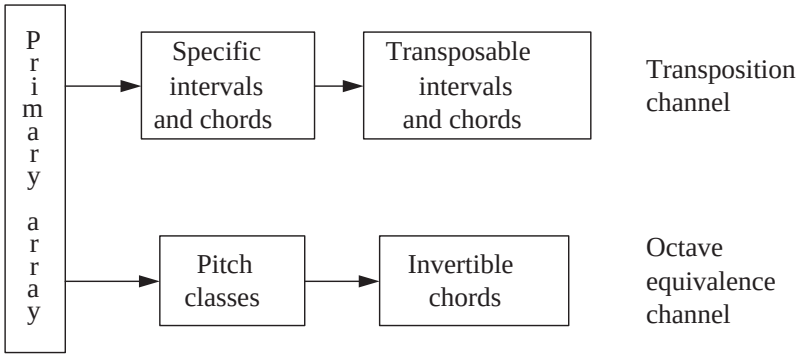


Figure 2.24: Deutsch's model for the abstraction of pitch relationships using two channels: one related to transposition and the other one to octave equivalence (adapted from Deutsch (1999), pp. 352).

The first channel mediates the perceptual equivalence of chords and intervals under transposition. It consists of two stages:

1. The first stage is formed by first-order units that respond to tones. They project, in groups of two or three, into second-order units representing intervals and chords. Such linkages are only defined for tones which are separated by an octave or less.

2. Second-order units, representing combination of tones, project into third-order units, so that tones in the same relationship project to the same unit. For instance, all the major third project to the same unit, and all the second-order units activated by a major triad project to the same third-order unit.

The second channel mediates the perceptual equivalence of tones having the same chroma or pitch class. The two stages of this channel are the following:

1. First-order units respond to tones. They project to second-order units so that tones related by octave correspond to the same unit. The second-order units respond to a given pitch class, regardless of the octave.

2. Second-order units project in groups of two or three into third-order units, which respond to combinations of pitch classes.

This channel then represents the perceptual similarity of chords related by inversion. According to Deutsch (1999) pp. 353, the type of architecture that this block diagram proposes agrees with some basis of the auditory system.

Bharucha later developed a neural network sharing some aspects with the system described above Bharucha (1999). Systems based on neural network as Bharucha (1999); Leman (1991, 1994) provide key estimation as a result of passive exposure to music.

2.4 Locating modulations

Although the models presented above have been used to trace how the sense of key evolves through time, there has been less research explicitly devoted to locate modulations. The first problem to solve when trying to segment a piece according to its tonality is how to correctly identify regions of stable key centers and regions of modulations. Some approaches apply a sliding analysis window to the piece to generate a set of localized key estimations (as Shmulevich and Yli-Harja (2000) and Chew (2004)). This set of measures gives a good description of the key evolution of the piece, but calls for the setting of a suitable window size, which normally depends on the tempo, musical style and the piece itself. According to Shmulevich and Yli-Harja (2000), there seems not to be a general rule for choosing the length of the sliding window.

In order to smooth out the local oscillations and remove impulses, Shmulevich and Yli-Harja review the use of median-based filters and propose a graph-based smoothing method (Shmulevich and Yli-Harja (2000)).

According to Leman (2000), his model of short-term memory *assumes that context is built up at different time scales, at least one time frame for local events, such as pitches and chords, and one time frame for global events, such as the induced tonal context*. Leman (1994, 1995a) proposes a system for tonality tracking based on the idea of a retroactive process, where a system for tonal center recognition (based on a self-organizing map Kohonen (1984)) is extended by an attractor dynamics model. This tone center attractor dynamics (TCAD) model describes recognition in terms of attractors, stable states and transitions. Leman (1994) refers to tone center *interpretation* instead of *recognition*. The idea is that if the system is presented with a cadence consisting of IV-V-I, then the first chord is considered as a tonic chord and therefore in the key corresponding to the tonic of IV. However, as soon as the systems considers the third chord, the ambiguity is solved. The first chord IV is indeed in the key of the tonic of this chord. A snail-like metaphor is described in Leman (1994) (pp. 192) to model this behavior, which might be influenced by the force of different attractors.

Sapp (2001) introduced the use of multi-timescale visualization for displaying the output of key-finding algorithms from MIDI representations.

Recently, Chai proposed the use of Hidden Markov Models to track the evolution of the key of piano pieces in audio format (Chai (2005); Chai and Vercoe (2005)). More details of this approach are given in

Section 2.7.

2.5 Tonality and popular music

As mentioned above and will be shown in Section 2.9, most of the approaches for tonality induction focus on modelling tonal hierarchies of western music (Vos (2000)). These models seem to be also valid for actual popular music which are derived from western music tradition.

Temperley discusses in Chapter 9 of his book Temperley (2001) how modality and tonicization are defined in rock compositions, analyzing the modal character of much rock music. Given a pitch class collection from a diatonic scale in C major (composed by the notes shown in Figure 2.7), we can find different possible tonal centers corresponding to different modes in addition to the ”major” mode. The most relevant are the following: Ionian mode (having C as a tonic), mixolydian mode (having G as a tonic), dorian mode (having D as the tonic) and aeolian mode (having A as the tonic and corresponding to the minor natural scale). According to Temperley (2001), pp. 260, it is common to shift between different modes along a piece of rock music, but it is rare to find pitches out of the pitches provided by the four common rock modes. Then, it is possible to define a *supermode* including all the pitch classes of the four most common rock modes. A *supermode* flat profile is shown in Figure 2.25. We could think in using this kind profile to estimate the

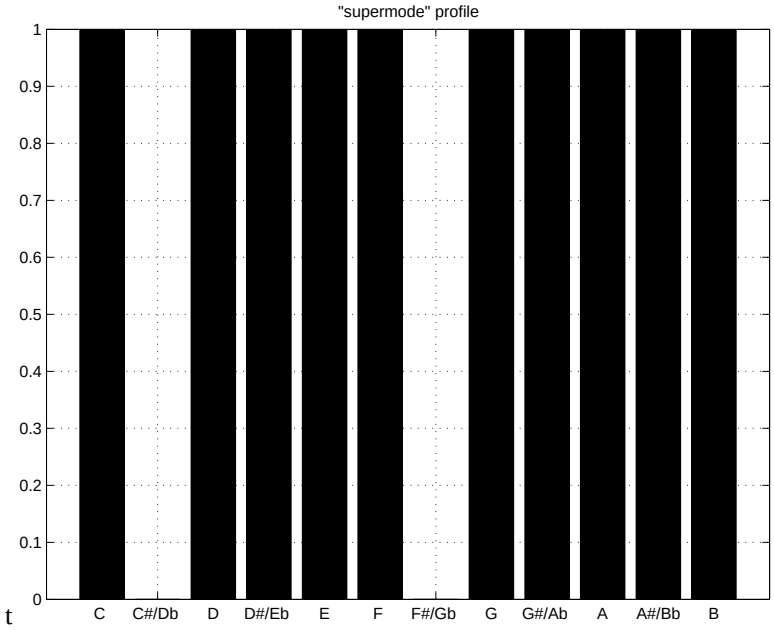


Figure 2.25: A ”supermode” profile, formed from the union of the four common rock modes (adapted from Temperley (2001), pp. 260).

tonality of the piece, but this profile would not work if all the modes are not equally used in a piece (which

happens most of the times) and it would then generate many estimation errors. According to Temperley, *there is more to tonic-finding in rock that simply monitoring the scale collection in use*. He proposes the use of weighted profiles (as shown in Figure 2.13) where the pitch classes corresponding to the tonic triad are emphasized, considering the importance of the tonic triad in tonicization. Then, this resulting profile should be combined with the supermode profile for each tonic center.

2.6 Audio feature computation

In order to describe the tonal aspects of a music piece in audio format, it its necessary to automatically extract information related to the played tones by analyzing the audio signal.

We review here the different approaches for feature extraction from audio, including transcription-oriented methods for the estimation of pitch and pitch class distribution descriptors.

2.6.1 Approaches based on automatic transcription

As most of the literature dealing with tonal induction is oriented to the analysis of scores, a straight solution to apply these methods to audio would be to automatically extract score information from audio signals. This is the goal of automatic transcription systems, and it was the approach adopted within the first stages of this PhD work. We will see below (Section 2.6.2) that we have finally proposed the use of pitch class profiles computed without the need of an exact transcription. For this reason, in this review we focus on the analysis of methods for the extraction of pitch class distribution features from audio (Section 2.6.3). We only summarize here the main approaches for automatic transcription of polyphonic audio. We focus on the estimation of fundamental frequencies, although there are other related tasks (e.g. note segmentation or estimation of the melody line from different voices) that have a strong relevance within an automatic transcription system.

The goal of this section is to give an idea of the main techniques for fundamental frequency estimation for monophonic and polyphonic signals, in order to see which is the current state of the art and the accuracy rates that we can achieve. We refer to Gómez (2001, 2002); Gómez et al. (2003), Klapuri (2004) and de Cheveigné (2006) for a detailed review.

2.6.1.1 Fundamental frequency estimation from monophonic signals

Fundamental frequency is the main low-level feature to be considered when describing tonality. Due to the significance of pitch detection for speech and music analysis, a lot of research has been made on this field. We can also find several surveys and evaluations of pitch detection algorithms for different purposes, such as Hess (1983); Roads (1996); Romero and Cerdá (1997); Kostek (2004); de Cheveigné (2005).

Much literature deals with the analysis of monophonic audio recordings in order to estimate its fundamental frequency. The first solution in the literature was to adapt some of the techniques proposed for speech, as presented in Hess (1983). Later, other methods were specifically designed for dealing with music.

All the fundamental frequency estimation algorithms give us a measure corresponding to a portion of the signal (analysis frame). According to McKinney (cited in Hess (1983)), the fundamental frequency detection process can be subdivided into three main steps that are passed through successively: the pre-processor, the basic extractor, and the post-processor (see Figure 2.26). The basic extractor performs the main task of measurement: it converts the input signal into a series of fundamental frequency estimates. The main task of the pre-processor is to process the input signal in order to facilitate the fundamental frequency extraction (e.g. denoising or normalization). Finally, the post-processor is a block that performs more diverse tasks, such as error detection and correction, or smoothing of an obtained contour.

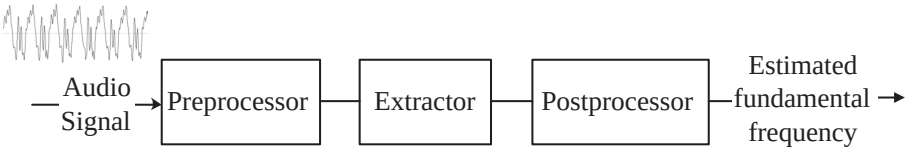


Figure 2.26: Steps of the fundamental frequency detection process.

There are many ways of classifying the different algorithms. One could classify them according to their processing domain. Following this rule, we can discriminate between time-domain algorithms, dealing with the signal in time domain, and frequency-domain algorithms, that use the signal in frequency domain (the spectrum of the signal). This distinction between time-domain and frequency-domain algorithms is not always so clear, as some of the algorithms can be expressed in both (time and frequency) domains, as the Autocorrelation Function (ACF) method. Another way of classifying the different methods, more adapted to the frequency domain, could be to distinguish between spectral place algorithms and spectral interval algorithms, classification proposed in Kostek (2004). The spectral place algorithms, such as the ACF method and the cepstrum analysis, weight spectral components according to their spectral location. Other systems, as those that are based on envelope periodicity or spectrum autocorrelation computation, use the information corresponding to spectral intervals between components. Then, the spectrum can be arbitrarily shifted without affecting the output value. These algorithms work relatively well for sounds that exhibit inharmonicity, because intervals between harmonics remain more stable than the places for the partials.

Time-domain algorithms try to find the periodicity of the input sound signal in the time domain. **Zero-crossing rate** (ZCR) is amongst the first and simplest techniques toward estimating the frequency content of a signal in time domain. This method is very simple and inexpensive, but not very accurate. The value of ZCR has also been found to correlate strongly with spectral centroid, so that it can have more to do with timbre than with pitch.

Time-domain **Autocorrelation** function (ACF) based algorithms have been amongst the most frequently used fundamental frequency estimators (see Medan et al. (1991), Talkin (1995) and de Cheveigné and Kawahara (2002)). The autocorrelation can also be computed in frequency domain, as ACF is the inverse FFT of the

power spectrum. According to Kostek (2004), systems based on ACF can be called spectral place type fundamental frequency estimator. ACF based fundamental frequency detectors have been reported to be relatively noise immune (Romero and Cerdá (1997)) but sensitive to formants and spectral peculiarities of the analyzed sound (Kostek (2004)).

Other methods are based on the analysis of the **envelope periodicity** (EP). They are considered to be spectral interval oriented Kostek (2004), as they analyze the frequency difference between frequency components. The most recent models of human pitch perception calculate envelope periodicity separately at distinct frequency bands and then combine the results across channels (Meddis and Hewitt (1991)). These methods attempt to estimate the perceived pitch, not pure physical periodicity, in acoustic signals of various kinds. The algorithm proposed by Terhardt represents an early and valuable model (Terhardt (1979); Terhardt et al. (1981)). Except for the simplest algorithms, that only look for signal periodicity, "perceived pitch" estimators use some knowledge about the auditory system at their different steps: when pre-processing, extracting or post processing data. Then, they could be considered as pitch estimators. However, as the psychoacoustic knowledge is only applied to improve the periodicity estimation and no complete model of pitch perception is applied, some auditory phenomena are not explained.

Parallel processing approaches are based on several processes working in parallel. One example is the fundamental frequency detector defined by Gold and later modified by Rabiner (Gold and Rabiner (1969); Rabiner and Schafer (1978)), shown in Figure 2.27. This path has been little explored, but is plausible from the human perception point of view, and might be very fruitful. Bregman remarks: *"I believe that there is a great deal of redundancy in the mechanisms responsible for many phenomena in perception"* Bregman (1998). Several different processes analyze the same problem, and when one fails, the other succeeds.

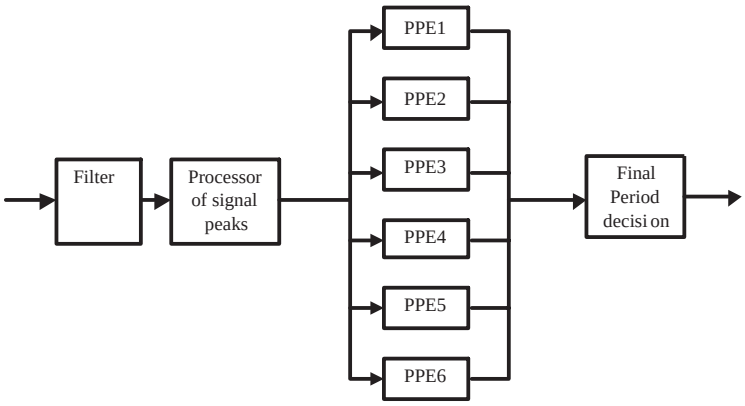


Figure 2.27: Parallel processing approach.

Frequency-domain algorithms use the spectral information of the signal, obtained by a short-time

Fourier Transform or another transformation. The **cepstrum** pitch detection algorithm from Noll (1967) was the first short-term analysis algorithm that proved realizable on a computer. The cepstrum is the inverse Fourier transform of the power spectrum logarithm of the signal. The pulse sequence originating from the periodicity source reappears in the cepstrum as a strong peak at the "quefrency" (lag) T_0, which is readily discovered by the peak-picking logic of the basic extractor. Cepstrum fundamental frequency detection has close model level similarity with ACF systems. Unlike ACF systems, cepstrum pitch estimators have been found to perform poorly in noise and to have good performances with formants and spectral peculiarities of the analyzed sounds (Kostek (2004); Romero and Cerdá (1997)).

Spectrum autocorrelation methods are based on detecting the period of the magnitude spectrum using its autocorrelation function. A nice implementation of this principle can be found in Lahat et al. (1987).

Harmonic matching methods try to extract a period from a set of spectral maxima of the magnitude spectrum of the signal. Once these peaks in the spectrum are identified, they can be compared to the predicted harmonics for each of the possible candidate note frequencies, measuring how do they fit. This approach has been widely used in Piszczalski and Galler (1979), Maher and Beauchamp (1993) or Doval and Rodet (1991). The solution presented in Maher and Beauchamp (1993) is to employ two mismatch error calculations and is illustrated in Figure 2.28. The first one is based on the frequency difference between each partial in the measured sequence and its nearest neighbor in the predicted sequence and the second is based on the mismatch between each harmonic in the predicted sequence and its nearest partial neighbor in the measured sequence.

The **wavelet** transform (WT) is a multi-resolution, multi-scale analysis that has been shown to be very well suited for music processing because of its similarity to how the human ear processes sound. In contrast to the short-time Fourier transform, which uses a single analysis window, WT uses short windows at high frequencies and long windows for low frequencies.Some wavelet based fundamental frequency algorithms have been proposed (e.g. Jehan (1997)) for voice signals. Following with the idea of the constant Q frequency analysis, Klapuri (2000b) proposed an algorithm for periodicity analysis that calculates independent fundamental frequencies estimates at separate frequency bands. Then, these values are combined to yield a global estimate. This solves several problems, providing robustness in the case of badly corrupted signals, where only a fragment of the whole frequency range is good enough to be used.

In Table 2.2 we can find a summary of the listed algorithms. The information about the performances has been extracted from different sources such as Kostek (2004); Romero and Cerdá (1997) or from the authors' related work. Certain algorithms performs well in different situations, but there is not yet a solution for all the instrument sources in all conditions.

Fundamental frequency detection algorithms suppose that a fundamental frequency is present, but this is not always the case. We could have segments where no pitch can be found, as for instance at silences, percussion solos or noise segments. A segmentation process should distinguish between pitched and unpitched periods, so that the fundamental frequency detection will be only performed for pitched parts of the signal. However, most of the techniques used for pitched/unpitched segmentation already use the estimated

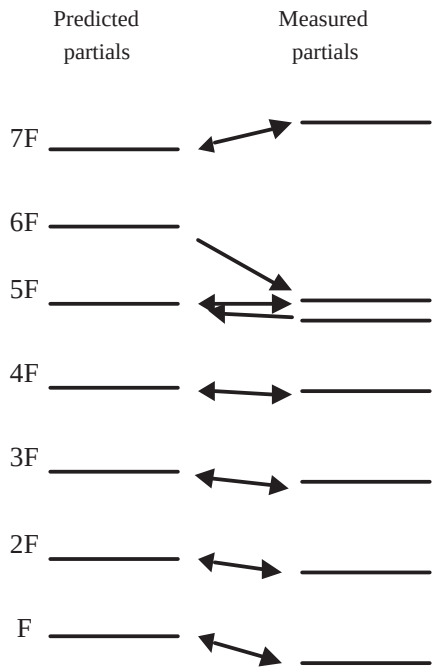


Figure 2.28: Two-Way Mismatch procedure.

fundamental frequency to decide whether this information is valid or corresponds to an unpitched signal, in addition to other features computed from the signal (see for instance Cano (1998)).

Pre-processing methods are intended to suppress noise and to enhance the features that are useful for fundamental frequency estimation. Some of the pre-processing methods used in speech processing are detailed in Hess (1983), including techniques for the isolation of the first partial, moderate low-pass filtering to remove the influence of higher formants, inverse filtering to remove vocal tract influence and estimate the voice source (excitation), comb filtering to enhance harmonic structures, non linear processing in the spectral domain, center clipping (that destroys the formant structure without destroying the periodic structures of the signal), signal distortion by an even nonlinear function, envelope detection, instantaneous-envelop detection, etc. These algorithms can also be used for fundamental frequency detection of music signals. Some pre-processing methods have been defined for musical signals. The method proposed by Klapuri applies the principles of RASTA spectral processing tries to remove both additive and convolutive noise simultaneously (see Klapuri et al. (2001)).

The fundamental frequency contour that is the output of the different algorithms is normally noisy and sometimes badly affected with isolated errors, so different **postprocessing methods** for correcting them have been defined. The most usual way to smooth a function is the convolution of the input signal with the impulse response of a low-pass filter. As presented in Hess (1983), the application of low pass filters removes much of

Method	**Domain**	**Spectral Place/ Interval**	**Simplicity**	**Noise**	**Spectral Peculiarities**
ZCR	Time	SP	Very simple		
ACF	Both	SP	Simple	Relatively immune	Sensitive
EP	Time	SI	Simple		
Rabiner (parallel processing)	Time	SP	Relatively simple		
Cepstrum	Frequency	SP	Simple	Poor Performance	Relatively immune
Spectrum AC	Frequency	SI	Simple		
Harmonic Matching Method	Frequency	both	Quite complex	Relatively immune	Relatively immune
Wavelet based method	Frequency (WT)		Quite complex	Immune	
Bandwise Klapuri	Frequency	both	Quite complex	Relatively immune	Relatively immune

Table 2.2: Summary table of the different methods for fundamental frequency estimation.

the local jitter and noise, but it does not remove local gross measurements errors, and, in addition, it smears the intended discontinuities at the pitched-unpitched transitions. Hence, some kind of non-linear smoothing might be more appropriate. In a paper by Rabiner et al. (1975), median smoothing is proposed as a non-linear method.

Another approach is described in Laroche (1995). The procedure consists of storing several possible values for the fundamental frequency for each analysis frame, assigning them a score that represents the estimation goodness. The goal is then to find a ”track” that, following one of the estimations for each frame, will have the best score. Other post-processing methods are dependent on the algorithm used for sound analysis and fundamental frequency detection. One example is the approach used in Serra (1996). Fundamental frequency is estimated using spectral peak information. Spectral peaks are computed frame by frame using windowing and FFT analysis. In the windowing procedure, window size is updated depending on the fundamental frequency. If an estimation error is produced, then the window size for the following frames will not be correctly chosen. In order to correct this type of errors, a reanalysis is made over a group of frames beginning at the last one and finishing at the first one.

Fundamental frequency history and future frames are also used to choose between candidates with the same Two-Way Mismatch error (see Cano (1998)), having a smooth evolution with neighbor frames. The phase of the peaks can finally be useful to modify the search among fundamental frequencies candidates. In this case, the pitch contour is smoothed according to sound properties, in opposition to median techniques.

Fundamental frequency tracking using other knowledge sources. We can also used other available meta-data for guiding the fundamental frequency detection and tracking process. Content information has been used in the form of internal sound source models (Kashino et al. (1995)). Martin (1996) also used musical rules to transcribe four-voice polyphonic piano pieces. When some assumption about the type of signal is made or when the algorithm is adapted to some of the signal properties, we are also taking advantage of some information that can be considered as context information.

For instance, we could use some pre-processing methods dependent on the played instrument. Goto considers the different pitch ranges of melody vs bass lines and discriminates them using different band-pass filters (Goto (2000, 1999)). Another possibility could be to adapt some of the parameters of an algorithm to the played instrument. Some of these ideas are the basis of some multitimbre pitch detectors (Goto (2000, 1999); Anderson (1997)).

2.6.1.2 Multipitch estimation

It is generally admitted that single-pitch estimation methods are not appropriate as such for multipitch estimation (Kostek (2004)), although some of the algorithms used in monophonic pitch detection can be adapted to simple polyphonic situations. In Anderson (1997), it is described how some of the methods applied to monophonic fundamental frequency estimation can be adapted to polyphony. Also, the TWM procedure can be extended to duet separation, as explained in Maher and Beauchamp (1993), trying to find two fundamental frequencies that best explain the measured spectral peaks.

Multipitch estimation is oriented toward auditory scene analysis and sound separation: if an algorithm can find the pitch of a sound and not get confused by other co-occurring sounds, the pitch information can be used to separate the partials of the sound from the mixture. Indeed, the most successful multipitch estimation methods have applied the principles known from human auditory organization.

Kashino et al. (1995) implemented these principles in a Bayesian probability network, where bottom-up signal analysis could be integrated with temporal and musical predictions. A recent example following the same principles is that of Walmsley et al. (1999), who use the Bayesian probabilistic framework in estimating the harmonic model parameters jointly for a certain number of frames. Other statistical approaches include the work Davy and Godsill (2003) and Cemgil (2004). Godsmark and Brown (1999) have developed a model that is able to resolve melodic lines from polyphonic music through the integration of diverse knowledge. The system proposed by Goto (2000, 1999) is more application-oriented, and is able to detect melody and bass lines in real-world polyphonic recordings by making the assumption that these two are placed in different frequency regions. Other methods are listed in Klapuri (2000a), where a system is described following an iterative method with a separation approach. This algorithm operates reasonably accurately for polyphonies at a wide fundamental frequency range and for a variety of sound sources.

The state-of-the-art multipitch estimators operate reasonably accurately for clean signals, the frame-level error rates progressively increasing from two percent in two-voice signals up to about twenty percent error rates in five-voice polyphonies. However, the performance decreases significantly in the presence of

noise, and the number of concurrent voices is often underestimated. Also, reliable multipitch estimation requires significantly longer time frames (around 100 ms) than single-pitch estimation (Klapuri (2000a); Kostek (2004)).

Recently, some systems for *predominant* pitch estimation have been evaluated in the context of the ISMIR conferences 2004 and 2005[1][2]. The goal of the evaluation has been to compare state-of-the-art algorithms for predominant pitch estimation within polyphonic signals. A summary of the evaluation results for the first edition of the contest is presented in Gómez et al. (2006). According to these experiments, the overall accuracy is around 71%. This performance can then decrease when estimating multiple voices within a polyphonic signal. This estimation error then affects the accuracy for a higher-level tonal description computed after this multipitch estimation.

2.6.2 Transcription and tonal description

According to the current state of the art in automatic transcription from audio, it is not still possible to get a reliable score representation for any piece of music in audio format. However, is it really necessary? One of the hypotheses of this dissertation is that it is not necessary at all having the transcription for obtaining a complete and multilevel tonal description, including pitch class distributions, chords and keys, and allowing the computation of tonal similarity and the organization of music collections according to key. This idea has been followed by several authors, as reviewed in Section 2.6.3.

We will show in the following chapters that pitch class distribution features are extracted from audio signals with no loss of accuracy or information due to estimation. The schema of this approach is presented in Figure 2.29. The use of pitch class distribution features overcomes the problem of automatic transcription, by describing tonal aspects without the need of a perfect transcription. On the other hand, it introduces an additional issue on how to apply tonal models, which have been designed to work with score representation, to these new representations.

Finally, the estimated tonality could be used to improve the automatic transcription process. For instance, key or scale information could be used to give a higher expectancy to fundamental frequency values that match the scale. The distinction between transcription and description has been already signaled in Herrera (2006) for different aspects of sound.

2.6.3 Pitch class distribution features

We observe that many efforts have been devoted to the analysis of chord sequences and key in MIDI representations of classical music (see Section 2.3 and Section 2.9), but little work has dealt directly with audio signals or other genres. The use of MIDI-oriented methods would need a previous step of automatic transcription of polyphonic audio, which has been justified to be a very difficult task.

[1] http://ismir2004.ismir.net/melody_ contest/results.html
[2] http://www.music-ir.org/evaluation/mirex-results/audio-melody

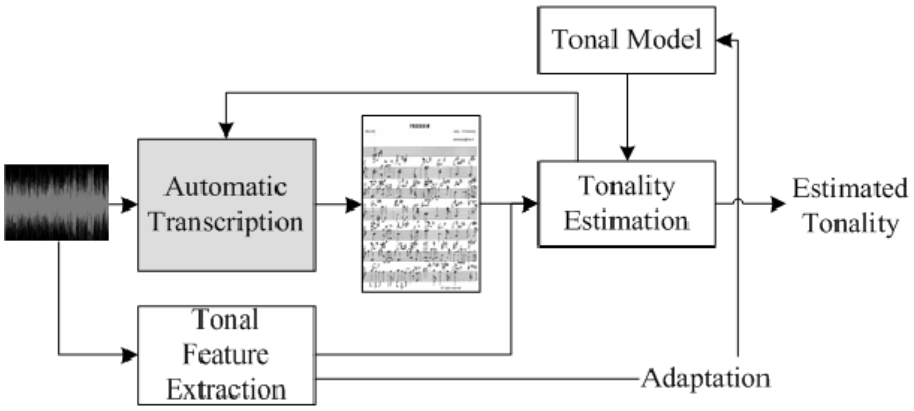


Figure 2.29: Schema of our approach, where there is no need of having the transcription of the audio recording in order to have a tonal description of the piece in audio format.

As mentioned above, the unavailability of automatic transcription makes that many approaches work directly with audio recordings to extract information related to the pitch-class distribution of music. The pitch-class distribution of music is, somehow, directly related to the chords and the tonality of a piece. Chords can be recognized from the pitch-class distribution without precisely detecting which notes are played. Tonality can be also estimated from the pitch-class distribution without a previous chord-estimation procedure. Reliable pitch class distribution descriptors should fulfil the following requirements:

1. Represent the pitch class distribution of both monophonic and polyphonic signals.
2. Consider the presence of harmonic frequencies. We can see in Figures 2.2 and 2.4 that the first harmonics of a complex tone belong to the major key defined by the pitch class of the fundamental frequency, and all of them except the 7th harmonic belong to its tonic triad.
3. Robustness to noise that sound at the same time: ambient noise (e.g. live recordings), percussive sounds, etc.
4. Independence of timbre and played instrument, so that the same piece played with different instruments has the same tonal description.
5. Independence of loudness and dynamics.
6. Independence of tuning, so that the reference frequency can be different from the standard A 440 Hz.

All the approaches for computing the instantaneous evolution of pitch class distribution follow the same schema shown in Figure 2.30. In addition to this schema, it is necessary a further processing of the instantaneous descriptors when describing a larger segment of a piece.

We analyze in the following sections the different approaches (summarized in table 2.3) for each of the steps of the general schema for tonality induction from audio. The nomenclature for this type of features

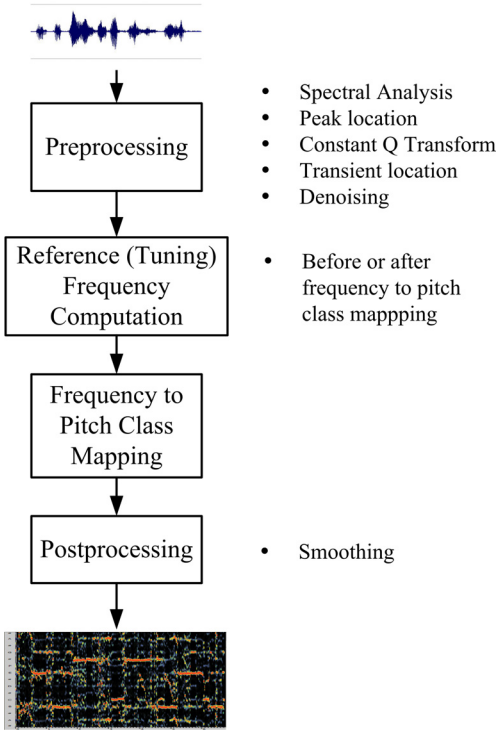


Figure 2.30: General block diagram for methods for pitch class distribution computation from audio.

is quite varied. The first approach for key induction from audio was proposed by Leman (1991, 1995b, 2000); Martens et al. (2004)). The approach presented in Leman (2000) extracts a set of *pitch patterns* in a similar procedure to some methods for multipitch estimation mentioned in Section 2.6.1.2 using a model of the human auditory system. The use of *pitch histograms* appears later in Tzanetakis (2002). Fujishima (1999) proposed a system for chord recognition based on the *pitch-class profile* (from now PCP), a low-level feature. PCP, as formulated by Fujishima, is a twelve dimensional vector representing the intensities of the twelve semitone pitch classes. This chord-recognition system compares this vector with a set of chord-type templates to estimate the played chord. The approach presented in the thesis uses an extension of the PCP, called *Harmonic Pitch Class Profile* or HPCP, which is described in detail in the following chapters. Constant Q profiles have also been used to characterize the tonal content of audio, as the *constant-Q profile* by Purwins et al. (2000) and the *pitch profile* by Zhu et al. (2005). Another recently proposed feature by Pauws (2004), that he calls the *chromagram*, is also used for key estimation.

2.6.3.1 Pre-processing

The main task of the pre-processing step is to prepare the signal for pitch class distribution description and enhance the features that are relevant for this kind of features. Pre-processing should help to fulfil the third

requirement mentioned above, providing robustness to noises: ambient noise, percussive sounds, speech sounds, etc.

All the approaches found in the literature are based on performing a frequency analysis of the audio signal. In the approach by Fujishima (1999, 2000), the input sound is transformed to a Digital Fourier Transform (DFT) spectrum (defining analysis frames of 2048 samples under 5.5 KHz, with a duration of 400 ms). DFT is also used in our approach (see Chapter 3 and Gómez (2006, 2004); Gómez and Herrera (2004)), and also in Pauws (2004); Chuan and Chew (2005); Izmirli (2005), using similar window duration.

It is also common to restrict the analysis to a given frequency region in order to eliminate non audible frequencies and to select the most relevant frequencies for pitch distribution descriptors. Fujishima (1999) selects a frequency region between 63.5 and 2032 Hz. Our approach also uses only a frequency region of the signal between 100 Hz and 500 Hz (see Chapter 3 for further details), Pauws (2004) uses frequencies between 25 and 5000 Hz, Zhu et al. (2005) from 27.5 to 3520 Hz, Chuan and Chew (2005) from 32 to 1975 Hz and Izmirli (2005) from 50 to 2000 Hz.

Instead of selecting a single frequency band, some methods consider a multiband approach. In Leman (2000), an auditory model decomposes the input signal into different frequency bands represented as nerve patterns, also known as the auditory nerve image. Fifteen channels are used with center frequencies ranging from 141 to 5173 Hz (see Martens et al. (2002)).

The constant-Q transform was introduced by Brown (1991), and Brown and Puckette (1992) later proposed an algorithm for its efficient computation. This transform has been employed by Purwins et al. (2000) and later by Zhu et al. (2005) for low-level tonal description. The constant-Q transform can be seen as a bank of filters with geometrically spaced center frequencies f_i. The letter "Q" is related to the constant ratio between the center frequency and the bandwidth of each filter. f_i is defined as:

$$f_i = f_{ref} \cdot 2^{\frac{i}{size}} \tag{2.7}$$

and the constant ratio of center frequency to bandwidth as:

$$Q = \frac{f_i}{\Delta_i} = (2^{\frac{1}{size}} - 1)^{-1}, \;\; i = 0...size - 1 \tag{2.8}$$

$size$ determines the number of filters per octave (12, 24, 36, etc.). This parameter is set to $size = 120$ (0.1 semitone resolution) in Zhu et al. (2005), in order to improve the determination of the tuning frequency. The computation of the Q-transform is achieved by choosing a different window length N_i for each frequency bin. For integer values of Q, the i-th bin of the constant-Q transform is equal to the Q-th bin of the spectrum (DFT) with a window length of:

$$N_i = Q \cdot \frac{f_s}{f_i} \tag{2.9}$$

The procedure for computing the constant-Q transform is then the following one: first choose the smallest frequency f_{ref} and the number of bins per octave ($size$) and consider equations 2.8 and 2.9. Then, the i-th

bin of the constant-Q transform $QTX(i)$ is equal to the Q-th bin of the DFT transform $X_{N_i}(Q)$ considering a window length of N_i, which is defined as:

$$QTX(i) = X_{N_i}(Q) = \frac{1}{N_i} \sum_{n<N_i} x[n] \cdot \omega_{N_i}[n] \cdot e^{-j2\pi n \frac{Q}{N_i}} \tag{2.10}$$

where $x[n]$ is the time-domain sampled audio signal and $\omega_{N_i}[n]$ is the chosen analysis window. This method is not very efficient computationally, as different FFTs should be computed.

In addition to a frequency analysis of the input signal, other pre-processing steps mentioned by Fujishima (1999) include non-linear scaling and silence and attack detection to avoid noisy features; they are not further explained nor evaluated in his work. Transient detection is also used in Gómez (2006) as a preprocessing step.

Another interesting pre-processing in order to reduce noise is the inclusion of a peak selection routine only considering the local maxima of the spectrum (Gómez (2006, 2004), see Chapter 3 for further details). Chuan and Chew (2005) also introduce this procedure. Finally, Pauws (2004) also mentions an enhancement procedure of the spectral components to cancel spurious peaks, although this procedure is not explained either in his paper. We are unaware on any formal evaluation of pre-processing methods and their usefulness for better descriptors.

2.6.3.2 Reference frequency computation

As explained in Section 2.2.1, the A 440 Hz is considered as the standard reference frequency for pitch class definition. According to this, the majority of approaches for pitch class distribution computation use a fixed frequency as a reference Fujishima (1999); Purwins et al. (2000); Pauws (2004); Martens et al. (2002); Chuan and Chew (2005); Izmirli (2005), which is usually set according to the A 440 Hz.

Nevertheless, we cannot assume that orchestras and bands will always be tuned to this pitch. Some pieces may not be tuned to 440 Hz, so that we should estimate the tuning frequency in order to assure robustness to tuning (the sixth requirement of our list). This procedure can be performed either above or after the computation of pitch class distribution, and the final feature vector must be adjusted to this reference frequency.

Fujishima (2000) included a procedure for adjusting the PCP values to the reference frequency after PCP computation. In this procedure, the octave profile, obtained using a frequency resolution of 1 cent, is divided into twelve equal parts of one semitone width, and these twelve parts are summed up to obtain a semitone profile. This semitone profile is ring-shifted, and mean and variance values are computed for each shifted semitone profile. The minimum variance value gives the genuine peak position, and the reference frequency is computed using the mean value for this position and the amount of shift of the semitone profile.

The procedure for tuning frequency determination proposed by Zhu et al. (2005) is performed before computing the pitch class distribution, and is used as the reference frequency for logarithmic frequency to pitch mapping. This method is based on the statistical analysis of the frequency positions of the prominent

peaks of the constant-Q transform, and consists of four different steps:

1. Detection of the local maxima of the constant-Q transform having its energy above a threshold.

2. Grouping of the detected peaks into groups according to their deviations (modulus 12) from the 440 Hz reference frequency.

3. Choose the maximum group as the tuning frequency for this analysis frame, given that its value is prominent enough.

4. Build an histogram of the tuning frequencies of all the frames and choose the maximum point as the reference frequency.

The reference frequency is computed with a resolution of 0.1 semitone (10 cents, compared to 1 cent resolution by Fujishima), and only a short segment (30 seconds) of the piece is needed to estimate the tuning frequency, given the assumption that the reference frequency is usually constant over the piece. The approach proposed in this dissertation uses a similar procedure for tuning frequency determination (see Chapter 3 for further explanations). As for pre-processing methods, we are unaware of any formal evaluation of methods for reference frequency computation.

2.6.3.3 Frequency determination and mapping to pitch class

Once the reference frequency is known and the signal is converted into a spectrogram by means of DFT or constant-Q analysis, there is a procedure for determining the pitch class values from frequency values.

The approaches by Leman (2000) and Tzanetakis (2002) are oriented to multipitch estimation, and a periodicity analysis is applied on the output of the filter-bank using autocorrelation, in order to extract a set of K predominant frequencies f_{pk}, where $k = 1 \ldots K$, which are the ones used for tonal description. Leman (2000) then uses the predominant frequencies and matches them to pitch classes using log mapping with respect to the reference frequency:

$$PCD(n) = \sum_{f_{pk} \ s.t. \ M(f_{pk})=n} 1 \tag{2.11}$$

where PCD represents the pitch class distribution features. We will keep the name PCD during this dissertation, to have a coherent formulation. In this equation $n = 1...12$, f_{pk} represents the predominant frequencies and $M(f_{pk})$ is a function which maps a frequency value to the PCD index, following a logarithm mapping:

$$M(f_{pk}) = round(12 \cdot \log_2(\frac{f_{pk}}{f_{ref}}) \ mod \ 12) \tag{2.12}$$

where f_{ref} is the reference frequency that falls into $PCD(0)$. Although pitch class C is often assigned this bin, we consider here pitch class A, so that f_{ref} would be 440 Hz if the piece is tuned to this value.

Instead of using only predominant pitches, Fujishima (1999) considers all the frequencies of the DFT, where the weight of each frequency to its corresponding pitch class is given by the square of spectral amplitude:

$$PCD(n) = \sum_{i \ s.t. \ M(i)=n} |X_N(i)|^2 \tag{2.13}$$

where $n = 1...12$, $|X_N(i)|$ is the linear magnitude of the spectral bin i, $i = 0...N/2$ where N is the size of the DFT. M(i) is now a table which maps a spectrum bin index to the PCP index, following a logarithm mapping:

$$M(i) = \begin{cases} -1 & \text{if } i = 0 \\ round(12 \cdot \log_2(\frac{\frac{f_s \cdot i}{N}}{f_{ref}}) \ mod \ 12) & \text{if } i = 1, 2, ...N/2 \end{cases} \tag{2.14}$$

where f_s is the sampling rate, f_{ref} is again the reference frequency that falls into $PCP(0)$. The term $\frac{f_s \cdot i}{N}$ represents the frequency of the spectrum bin i.

Other approaches as that of Purwins et al. (2000); Chuan and Chew (2005); Izmirli (2005) do not consider the square $|X_N(i)|^2$ as the weight of each frequency but the magnitude $|X_N(i)|$ corresponding to each frequency value:

$$PCD(n) = \sum_{i \ s.t. \ M(i)=n} |X_N(i)| \tag{2.15}$$

Our approach introduces a weighting scheme using a cosinus function (described in Chapter 3). Pauws (2004) also introduces some weighting according to the frequency position (frequencies contribute less following a decreasing exponential function) and to the perceived loudness (arc-tangent function). The contribution is defined by a maximum likelihood procedure.

Izmirli (2005) proposes the addition of a filtering stage using the spectral flatness measure (SFM). The spectrum is divided into frequency bands of half octave and for each band the SFM is computed. Only those bands having a SFM value higher than 0.6 (i.e. having significant peak information) are added to the PCD vector.

PCD computation methods should consider the presence of harmonics. Our method considers the presence of harmonics when adapting the used tonal model, as well as in Izmirli (2005) (as it will be explained in the next chapters). Pauws (2004) also considers the presence of harmonics of each tone, taking into account a total of 15 harmonics of each pitch class that contribute to the pitch class value. Zhu et al. (2005) also consider this fact by adding a filtering method, called consonance filtering, after the constant-Q transform procedure. In this step, they extract only the partials which are consonant, according to a diatonic scale. This could also be considered as a post-processing step, as the consideration of harmonic frequencies is not made during the PCD computation. When considering the presence of harmonics, we get closer to automatic transcription approaches, as the ones from Leman (2000) and Tzanetakis (2002). As for the other steps, we

are unaware of any systematic evaluation of the effectiveness of different weighting schemes.

2.6.3.4 Interval resolution

One important parameter for pitch class distribution descriptors is the frequency resolution used. One semitone (12 PCD bins, 100 cents) is the resolution that is usually chosen to represent pitch class distribution, as we can see in Pauws (2004); Chuan and Chew (2005); Izmirli (2005).

Increasing this resolution can help improving robustness against tuning and other frequency oscillations. Fujishima defines 12 bins for chord estimation, although 1 cent (1200 values) resolution is employed in the first stages of the algorithm for frequency folding and tuning. This is also the case in Zhu et al. (2005), where the interval resolution for pitch profile generation is one semitone (12 values) for key estimation, although 10 cents resolution (120 values) is set for the determination of the tuning frequency.

In the same way, Purwins et al. (2000) and Gómez (2006, 2004) define 36 values (1 third of semitone) to improve PCD resolution. The resolution is often decreased to 12 values when comparing to a tonal model. To obtain 12 values, all the amplitudes within each semitone are typically summed up (as in Fujishima (2000)).

2.6.3.5 Post-processing methods

As it is usual in fundamental frequency estimation methods, some post-processing methods are used after computing the pitch class distribution.

One of the mentioned requirements for pitch class distribution features is the robustness against variations on dynamics. This is usually obtained by normalization. Gómez (2006, 2004) propose to normalize the PCD vector for each frame by its maximum value. Chuan and Chew (2005) also use a normalization by the maximum value and the sum of all values of the overall segment PCD vector. This normalization can also be implicit in the feature computation procedure or included as a post-processing step.

Leman (2000) proposes to add to each incoming feature vector (called image) a certain amount of the old one, specified by a half-decay time. The half-decay time is the time it takes for an impulse signal to reach half of its original value (Martens et al. (2004)).

Fujishima (2000) proposes a peak enhancement procedure: first, the PCP is ring shifted by an amount of n semitones ($n = 1...11$), and the correlation between the original PCP and the shifted version is computed for each value of n. All these correlated profiles are summed in order to obtain an enhanced PCP. According to Fujishima (2000), this profile has sharper or more prominent peak contours than the PCP. It is said that this procedure may be omitted for simplicity, although no evaluation is performed to prove its utility.

Other ideas mentioned in Fujishima (1999), though not further developed, include smoothing (average) and chord change sensing (monitoring the direction of the PCP vector). Izmirli (2005) also proposes a summation procedure over 5 second segments.

2.6.3.6 Segment descriptors

Pitch class distribution, as an instantaneous feature, has been used for chord estimation, as in the system described in Fujishima (1999). Most of the approaches for global key estimation use the accumulation of instantaneous pitch class distribution into a global vector, as found in Purwins et al. (2000); Martens et al. (2002); Gómez (2006, 2004); Pauws (2004); Zhu et al. (2005) and shown in Table 2.3. Zhu et al. (2005) also introduce a normalization procedure so that the sum of the values equals to one.

Method	**Pre-processing**	**Reference Frequency Computation**	**Frequency to Pitch Class Mapping**	**PCD Resolution**	**Post-processing**	**Segment description**	**Application**
Pitch patterns Leman (2000)	Filter bank (141-5173 Hz), periodicity analysis (correlation)	No	Equal weight	Frequency analysis	Echoic memory module	Sum	Key induction
PCP Fujishima (2000)	DFT (63.5-2032 Hz), non-linear scaling, silence and attack detection	PCP shifting procedure	Square of spectral magnitude	$\frac{1}{2}$ tone (1 cent for tuning frequency)	Peak enhancement, smoothing (average), chord change sensing	None	Chord recognition
Constant-Q profiles Purwins et al. (2000)	Constant Q transform	No	Spectral magnitude	$\frac{1}{6}$ tone	None	Sum	Key estimation and tracking of classical music, comparative analysis of composers
Chromagram Pauws (2004)	DFT (25-5000 Hz), spectral peaks enhancement	No	Likelihood, Spectral magnitude, 15 harmonics (decreasing contribution), arc-tangent weighting according to loudness	$\frac{1}{2}$ tone (12 bins)	None	Sum and normalization	Key estimation from piano classical music

Pitch profile Zhu et al. (2005)	Constant Q transform (27.5-3520 Hz)	Analysis of peak deviations	Equal weight	$\frac{1}{2}$ tone (10 cents for tuning frequency)	None	Consonance filtering	Key estimation
Pitch class and strength Chuan and Chew (2005)	DFT (32-1975 Hz), peak estimation	No	Spectral magnitude	$\frac{1}{2}$ tone	Normalization	None	Key recognition from audio generated from MIDI, classical
Chroma summary vector Izmirli (2005)	DFT (50-2000 Hz), filtering based on SFM over half octave bands	No	Spectral magnitude	$\frac{1}{2}$ tone	Summation of 5 seconds windows	Sum	Key estimation from audio, classical
HPCP Gómez (2006, 2004), Chapter 3	DFT (100-5000 Hz), transient detection	Analysis of peak deviations	Square of spectral magnitude and weighting scheme	$\frac{1}{6}$ tone (frequency resolution for tuning frequency)	Normalization	Sum	Key estimation

Table 2.3: Summary table of the different methods for pitch distribution computation.

2.7 Adaptation of tonal models to audio

We have noticed that computational methods for tonality induction are based either on symbolic representations (or scores) or on acoustic signals. Most of the literature analyzes score representation. According to Vos (2000) pp. 406, *if one respects the claim that a model should be ecologically valid, the approach based on the acoustic signal is certainly superior to the first*. Working with acoustic input requires an understanding of the basics of how the human auditory system works. Up to our knowledge, the work by Leman (Leman (1991, 1995b)) is the first example of a tonality induction system from acoustic input. This system is based on an auditory model using a Kohonen map (a self-organizing map created using artificial neural network techniques Kohonen (1984)) with trained tonal centers for the recognition.

The approaches for tonality estimation from audio adapt the findings of score-based models and music theory to audio features, as seen in Figure 2.1. We present in this section how tonal models are adapted to work with audio features.

Most of the approaches for comparing audio features and tonal models are based on template-based distances, as it happens for score-based systems (explained in Section 2.3). Fujishima (1999) uses an approach based on pattern matching with a set of Chord Type Templates ($CTT(p), p = 0 \cdot 11$). $CTT_c(p)$ is set to 1 if the chord type c at root A (corresponding to f_{ref}) has the pitch class p in it. This profile is ring-shifted in order to obtain other roots than A. This procedure corresponds to use a set of flat profiles for chords, as shown in Figure 2.7 and explained in Section 2.3. Different profiles have also been used for key estimation. Krumhansl and Kessler's probe tone profiles are the most commonly used (as for instance in Purwins et al. (2000); Gómez (2006, 2004); Pauws (2004)), as well as Temperley's extensions, used in Izmirli (2005). Chuan and Chew (2005) use the spiral array CGE algorithm from Chew (2000) adapted to work with audio features.

In order to establish a direct correspondence between tonal models used for score and audio features, one fact has to be taken into consideration, which was mentioned in Purwins et al. (2000) pp. 273: in the audio features, all the harmonics of the played tones are present. It is necessary then to either modify the tonal models (for instance, the template profiles) or consider the presence of harmonic frequencies in the feature computation procedure, in a way closer to transcription, in order to eliminate them from the final pitch class distribution. In Gómez (2006), we propose the consideration of harmonic frequencies when adapting the tonal model. This is also the approach followed by Izmirli (2005), where harmonic templates are obtained from monophonic audio samples from piano. In these approaches the timbral characteristics of the input sound become relevant. This type of methods may be dependent on the considered instruments and their spectral envelope, as for instance using piano samples for computing templates in Izmirli (2005). Our approach also introduces an adaptation of the profiles defined for melodic lines to consider polyphonic situations where chords are dominant (see Chapter 4 for further explanations).

Once the profiles are chosen, several distance methods have been used. Two different similarity measures

were proposed in Fujishima (1999): nearest neighbor and weighted sum. **Nearest neighbor** is defined as

$$Score_{nearest,c} = \sum_{p=0}^{11}(T_c(p) - PCP(p))^2 \tag{2.16}$$

$$n = \arg\min_c(Score_{nearest-c}) \tag{2.17}$$

where $T_c(p)$ is a PCP defined from the the original $CTT_c(p)$.

Inner product is computed as follows:

$$Score_{nearest,c} = \sum_{p=0}^{11} W_c(p) \cdot PCP(p) \tag{2.18}$$

where $W_c(p)$ is a weight vector, defined as well from $CTT_c(p)$. There are no further details in Fujishima (1999) on how to compute the vectors $T_c(p)$ and $W_c(p)$. Inner product is the distance metric used in Fujishima (2000); Pauws (2004) and Izmirli (2005).

Purwins et al. (2000) propose the use of the **fuzzy distance**, defined as follows: let y and σ be the mean and standard deviation of some statistical data. The fuzzy distance of some value x to y regarding σ is defined by

$$d_\sigma(x,y) = |x-y| \cdot (1 - \frac{\sigma}{|x-y|+\sigma} e^{-\frac{|x-y|^2}{2\sigma^2}}) \tag{2.19}$$

The fuzzy distance is similar to the Euclidean distance, but the greater the uncertainty, the more relaxed is the metric.

In addition to template-based distances, Martens et al. (2002, 2004) introduced the use of classification trees. In Gómez and Herrera (2004), different machine learning methods are compared to the use of Krumhanl's and Kessler model (see Chapter 4 for further details). Other methods introduce the use of Hidden Markov Models (HMMs) to analyze the evolution of pitch distribution features and track the evolution of chords, (as in Sheh and Ellis (2003) and Chai (2005)) and key (as in Chai (2005)). The disadvantage of this type of methods based on training is that they require a big amount of annotated meta-data. Some musical knowledge is usually incorporated in order to minimize the required quantity of labelled data. Chai (2005), for instance, considers the observation probability distribution of the HMMs, defined as the probability at which a chromagram is generated by a given key or mode, using either a flat diatonic profile (shown in Figure 2.7) or a profile obtained from analyzing MIDI data (see Chai (2005)).

Finally, the approach by Zhu et al. (2005) is based on first determining the scale root and then the mode. The root is determined by finding the cluster of 7 consecutive tones (in the circle of fifths) by accumulating the occurrence values of the tones. The mode is found by evaluating the weight of the tonic and dominant in the scale. As it is based on accumulation, we can consider this approach as a flat profile-oriented method using an inner product metric. A pre-processing step is made by removing the effect of the augmented seventh degree (G# in A minor), so that we consider a minor natural scale. All the reviewed approaches are summarized in

Table 2.4.

Method	Tonal model	Distance measure	Adaptation
Chord estimation from PCP. Fujishima (2000)	Flat	Nearest neighbor and inner product	None
Key estimation from pitch histograms. Martens et al. (2002, 2004)	None	PCA and classification trees	None
Key estimation from constant-Q profiles. Purwins et al. (2000)	Krumhansl's and Kessler (Krumhansl (1990))	Fuzzy distance	Harmonics into profiles
Key estimation from chromagram. Pauws (2004)	Krumhansl's and Kessler (Krumhansl (1990))	Inner product	Harmonics into features
Key estimation from pitch profile. Zhu et al. (2005)	Flat	Inner product	None
Key estimation from pitch class and strength. Chuan and Chew (2005)	Spiral Array (Chew (2000))	Center of Effect Generator	None
Key estimation from chroma. Izmirli (2005)	Temperley (1999)	Inner product	Harmonic into templates
Key and chord tracking from PCP. Chai (2005)	Flat vs Obtained by analysis of MIDI data	HMMs	None
Key estimation from HPCP. Gómez (2006, 2004); Gómez and Herrera (2004)	Krumhansl's and Kessler / Machine learning	Inner product	Harmonics into profiles. adaptation to polyphony

Table 2.4: Summary table of the different methods for tonal description from pitch class distribution feature.

2.8 Application contexts for audio tonal description

Audio features related to pitch class distribution have been tested for different purposes.

Fujishima proposed a system for chord estimation, obtaining 94% accuracy from electric piano and CD recordings, as it appears in Fujishima (1999, 2000). Based on this work, Sheh and Ellis (2003) and later Chai (2005) also introduced hidden Markov models to estimate the chords within an audio recording. Sheh and Ellis obtained a maximum of 26% frame accuracy using an evaluation corpus of 20 Beatles songs, which, according to the authors, was not yet sufficient to provide usable chord transcriptions of unknown audio.

When looking at a larger temporal scale, pitch class distribution is useful to estimate the key of a piece or a given segment, as in the approaches by Martens et al. (2004, 2002); Leman (2000); Purwins et al. (2000); Gómez (2006, 2004); Gómez and Herrera (2004); Pauws (2004); Zhu et al. (2005); Chuan and Chew (2005); Izmirli (2005); Chai (2005). Pauws performed an evaluation with 237 pieces of classical piano sonatas, obtaining an accuracy of 75.1%. Zhu et al. performed an evaluation with 60 pop songs and 12 classical

pieces, obtaining an accuracy of 90% and 83.3% respectively. Chuan and Chew evaluated their method using the first 15 seconds of 18 Mozart symphonies generated from MIDI, obtaining a 90% accuracy. Izmirli evaluated his system in the starting segment of 85 pieces of classical music with a 86% accuracy. We present the results of our evaluation in Chapter 4.

Goto and Muraoka (1999) also introduced the computation of a histogram of frequency components to estimate the beat of drumless audio signals. This histogram is mainly used to detect chord changes. This method did not require chord names to be identified, as the goal was to use these histograms to track beats at different rhythmic levels.

Some researchers use pitch class distribution for structural analysis, e.g. to identify the chorus of a song (see Bartsch and Wakefield (2001) and Chai (2005)), and for audio to MIDI alignment, such as in the system by Hu et al. (2003). Another application introduced by Purwins et al. (2003) is the use of these features for the characterization of different composers.

2.9 Evaluation

Evaluation is an important issue. This section reviews which are the music collections that have been used in the literature to evaluate tonal description systems, as well as evaluation metrics. We are unaware of any formal evaluation of pre-processing methods (robustness to noise, etc), frequency resolution and tuning frequency determination procedures. Only Zhu et al. (2005) make a small evaluation of the algorithm for tuning frequency computation: the result of tuning pitch determination is based on the concentration of the population on the tuning pitch, which is the ratio between the summing of the 3 bins around the tuning pitch against the whole population of the histogram. Some efforts are currently being made inside the Music Information Retrieval (MIR) community, with the goal of establishing a common evaluation database and metric for key estimation[3].

As mentioned above, we review here which music collections have been used to evaluate algorithms for tonal description (mainly key and chord estimation). First, we review MIDI-oriented approaches, and then methods working with audio recordings. Most of the evaluation in this field is made by comparing the performance of automatic description algorithms to manual human annotations by experts.

The music collections that have been used to evaluate tonal description system from score representation (mostly in MIDI format) are described here and summarized according to the author in Table 2.5. Systems working from symbolic representations have been evaluated over a small corpus of data. The Well-Tempered Clavier (from now WTC) from J.S. Bach (preludes and fugues) is the most popular collection found in the literature to evaluate systems for key estimation. One reason is that these preludes and fugues move through the entire set of 24 possible major and minor keys.

Each prelude begins quite clearly in the key of the key signature. That is one of the reasons why the algorithm in Krumhansl (1990) is evaluated for the initial segments of the 48 preludes (see pp. 81-89). In

[3] http://www.music-ir.org/mirexwiki

Author	Collection	Application
Longuet Higgins (1971)	24 fugues subjects, WTC, Bach	Key estimation
Krumhansl (1990)	Initial segments of 48 preludes, WTC, Bach 24 preludes, Shostakovich 24 preludes, Chopin 48 fugue subjects, WTC, Bach 24 fugue subjects, Shostakovich Bach C minor prelude, Book II	Key estimation Key tracking
Temperley (1999)	First half of Courante, Cello Suite in C major, Bach Gavotte, French Suite No. 5 in G major, Bach Excerpt of Mazurka op. 17 no. 4, Chopin 48 fugue subjects, WTC, Bach 46 excerpts, Kotska-Payne theory textbook	Key estimation
Chew (2000)	48 fugue subjects, WTC, Bach Siciliano, Schumann Minuet in G and Marche in D, "A Little Book for Anna Magdalena", Bach Excerpt of Sonata Op. 13 "Pathetique", Beethoven Excerpt of Op. 33, Schubert	Key estimation

Table 2.5: Summary table of the different collections used for the evaluation of tonal description algorithms working with symbolic notation.

addition to Bach's preludes, they consider 24 preludes by Shostakovich and 24 Chopin preludes, which were composed in homage to Bach. In order to define the duration of the initial segments, only the first 4 notes were selected, also counting simultaneous notes (for instance getting the first triad chord and one note of the next chord).

Fugue subjects also define quite clearly the key of the key signature. Longuet-Higgins and Steedman (1971) analyzed 24 fugue subjects to evaluate their approach for key estimation. 48 fugue subjects from Bach's WTC are also analyzed by Krumhansl (1990) (pp. 89-96), Temperley (1999) and Chew (2000). The use of the same evaluation collection allows the direct comparison of approaches. Krumhansl (1990) also considers the analysis of 24 fugue subjects by Shostakovich, composed in homage to Bach.

In addition to this small and more standard collection, only single pieces have been analyzed. Krumhansl (1990) analyzes J.S. Bach's C Minor Prelude, Book II, in order to measure the presence of modulations. Later, Temperley (1999) analyzes the first half of Courante of Bach's Cello Suite in C major (BWM 1009), a Gavotte from Bach's French Suite No. 5 in G major (BWM 816), an excerpt from Chopin's Mazurka op. 17 no. 4 (a piece famous for its tonally unstable nature) and a set of 46 excerpts from the Kotska-Payne theory textbook.

Chew (2000) focuses her evaluation on piano pieces. In addition to the 48 fugue subjects from Bach, she analyzes the following pieces: Siciliano by Schumann; Bach's Minuet in G (in order to extract the estimated key and chords) and Marche in D from "A Little Notebook for Anna Magdalena"; a passage of Beethoven's Sonata Op. 13 "Pathetique", in order to analyze the evolution of chords and finally a passage of Schubert's

Op. 33.

When dealing with audio analysis, there may be some differences in the performance of an algorithm when dealing with the same piece under different recording conditions, played by different performers, by different instruments, etc. We find in the literature a broad variety of music collections in audio format used for evaluating different approaches, which are summarized in Table 2.6.

This table shows that there is no standard collection for evaluation. Some researchers work with audio from synthesized MIDI. In our opinion, the results of these evaluations do not cover the complexity of dealing with real recordings. Other studies only focus on a given set of acoustic instruments (e.g. the piano), which does not cover yet the timbre complexity found in music. The musical genres used are also varied, even though classical music prevails.

There is also no general agreement on the metrics used for evaluation. Researchers usually consider the percentage of correct estimation as a measure of the system accuracy. In Zhu et al. (2005), the piece is divided into segments and an estimation is provided for each of the segments. The global estimation is considered correct if there are more correct than wrong estimations. Some authors also consider, when evaluating methods for key finding, that close tonalities (relative major or minor, parallel or related by fifth) are partially correct.

This analysis shows that there is still much work left to reach a general (in terms of recording conditions, musical instruments and styles), modular (analyzing the influence of the different steps) and quantitative evaluation procedure, which can represent the complexity of music audio material. Up to our knowledge, this dissertation presents the most exhaustive and comprehensive evaluation carried out to date, which is given along Chapter 3 (pitch class distribution features), Chapter 4 (key estimation) and Chapter 5 (tonal descriptors and music similarity).

Author	Collection	Instruments	Application
Leman (1991, 1994)	"Through the Keys", Bartók Excerpts from Sextet no. 2, Brahms Excerpt of Prélude no. 20, Chopin Excerpt of Arabesque no.1, Debussy	Piano String (violin, viola, cello) Piano Piano	Key estimation and tracking
Fujishima (1999)	Sounds from a YAMAHA PSR-520 electronic keyboard 208 chords	Synthesized ocarina, strings and grand piano CD recordings	Chord recognition
Purwins et al. (2000)	C minor prelude, op. 28, no. 20, Chopin	Band	Key tracking
Sheh and Ellis (2003)	20 songs from the Beatles	Rock band	Chord recognition
Martens et al. (2002, 2004)	Passage of Invention no. 1 in C major, Bach Excerpt of "Eternally" by Quadran (dance music) Excerpt of Robert Mile's "Children"	Piano Band Band	Key estimation
Pauws (2004)	237 performances of classical sonatas	Piano	Key estimation
Zhu et al. (2005)	60 pop songs 12 classical pieces Vivaldi's "The four Seasons" (only 2 minutes for scale determination)	Band Orchestra instruments Orchestra instruments	Key estimation
Chuan and Chew (2005)	first 15 seconds of 18 Mozart symphonies	MIDI synthesizer	Key estimation
Izmirli (2005)	first 7.5-10 seconds of 58 classical pieces	Orchestra instruments	Key estimation
Chai (2005)	10 classical pieces	Piano	Key estimation

Table 2.6: Summary table of the different collections used for the evaluation of tonal description algorithms working with audio recordings.

2.10 Summary, current directions and hypothesis

After analyzing the current state of the art in tonal description from audio, we summarize here the main conclusions and open issues, and we summarize the work of this PhD, described in next chapters.

2.10.1 Multi-level tonal description

We have seen that different tonal descriptors are defined in different *temporal scopes*. Some of the descriptors are defined as instantaneous (for instance, fundamental frequency, pitch class distribution or chord) or related to a certain segment or to the whole musical piece (e.g. global pitch class distribution or key). In the same way, features are related to different *abstraction levels*: low-level signal descriptors, musical descriptors (applying musical knowledge or tonal models and requiring an induction process) and semantic descriptors (user-centered descriptors).

This idea has already appeared in the literature, as in Dannenberg (1993); Leman (2002). Dannenberg proposed musical representations at different levels, from the most abstract (where hi considers musical scores) to the most concrete level (the audio signal). He mentioned the need for a multiple-hierarchy scheme to represent music, which should be extensible to support new concepts and structures. According to Leman (2002); Leman et al. (2002), different representational levels for music exist, and they provide cues for content representation of musical audio.

Following this idea, this PhD proposes the use of a multi-level representation of tonal descriptors. Different temporal scopes are used, defining two categories of descriptors:

- Instantaneous: we define features attached to an analysis frame, which is the minimal temporal unit that we consider. We define a set of instantaneous features.
- Global: here we define some features related to a wider audio segment, which for example could be a phrase, a chorus or a whole song.

We also distinguish three **levels of abstraction**:

- Low-level signal descriptors: features related to the audio signal.
- Mid-level descriptors computed from low-level descriptors after a data modelling and inductive inference procedure, and related to the musical content of the signal (e.g., chord, key, etc.).
- These low and mid-level descriptors can be used for meaningful similarity measures related to a high abstraction level, as shown in Chapter 5.

Table 2.7 shows examples of descriptors belonging to each of the defined categories. These representations are available for content retrieval and navigation across digital music collections and are useful in different contexts, as we will present in the next chapters.

Table 2.7: Example of descriptors classification.

Name	Temporal Scale	Level of abstraction
HPCP	Instantaneous	Low
Chord	Instantaneous	Mid
Average HPCP	Global	Low
Key	Global	Mid

2.10.2 Challenges for low-level tonal descriptors

We have reviewed different methods for low-level tonal descriptors computation (pitch and pitch class distribution descriptors). Some of them are oriented towards automatic transcription and some not, although their goal is to represent the pitch content of the piece in audio format.

In Section 2.6.3 we listed the main requirements for this type of descriptors:

1. Be representative of the pitch content of both monophonic and polyphonic signals.
2. Consider the presence of harmonic frequencies.
3. Be robust to noise.
4. Independence of timbre and played instrument.
5. Independence of loudness and dynamics.
6. Independence of tuning.

The current state of the art introduces many procedures for succeeding. We have reviewed a set of pre-processing methods and some tuning frequency computation methods. But we have seen that no formal evaluation has been performed for testing these requirements. Robustness to noise has not been tested, neither independence of timbre, dynamics and tuning. Finally, classical music has been the main focus of the studies, more research is needed to have descriptors robust to percussion found in popular music.

This PhD proposes a set of features representative of the pitch class distribution of music, general for all the musical styles and fulfilling all the mentioned requirements. These features will be developed and evaluated in Chapter 3.

2.10.3 Tonal models and audio

We saw that most of the studies for tonality induction have been developed looking at the score. Tonal models should be adapted to work with audio signal in polyphonic situations in order to increase the performance of the systems. On the other side, models can be obtained by analyzing the low-level descriptors following a machine learning approach. In Chapter 4 we propose some mechanism for this adaptation, evaluating the performance of the proposed approach for chord and key estimation. We compare the use of different tonal

profiles proposed in the literature, and we compare the use of these approaches with some machine learning techniques.

2.10.4 Tonality visualization

We will see how giving a single key is often poor in terms of tonal description. We explore in Chapter 4 different ways of visualizing the tonal content of a piece of music in audio format, including tonality tracking and multi-resolution description.

2.10.5 Tonal similarity

Work on tonal similarity from audio recordings is in its early stages. We propose in Chapter 5 a set of features used to compute similarity and a set of distances. We present some examples of how this similarity measures can be used to identify different versions of the same musical piece.

2.10.6 Bridging the semantic gap

Descriptors related to tonal aspects of music are now limited to key and mode, and therefore there is a need to explore new semantic descriptors measuring, for instance, how *tonal* a piece of music is. There is also motivation to study how tonality is related to other high-level concepts such as *mood*, *artist* or *genre*. This aspect has been only started in Purwins et al. (2003). We will see in Section 5.3 how the proposed tonal features can be used to organize music collections of popular music.

Chapter 3

Feature extraction: from audio to low-level tonal descriptors

3.1 Introduction

The goal of this chapter is to present our approach for the computation of low-level tonal descriptors from polyphonic audio signals. These features represent the pitch class distribution of the analyzed piece. As mentioned in Section 1.3, low-level descriptors are closely related to the audio signal and computed in a direct way. We call these features the *Harmonic Pitch Class Profile* (HPCP), which is based on the Pitch Class Profile (PCP) proposed by Fujishima (1999). We present in this chapter the different steps for computing this profile and compare it with other approaches reviewed in Section 2.6.3. The overall scheme is shown in Figure 3.1: first, there is a preprocessing step, followed by the computation of the HPCP vector and finally some postprocessing is applied to the computed vector.

We study here how the proposed low-level features fulfill the requirements presented in Chapter 2 (Section 2.6.3): first, they should represent the pitch class distribution of both monophonic and polyphonic pieces; second, pitch class distribution features should consider the presence of harmonic frequencies; third, features must be robust to noise (e.g. noise from live recordings or percussion); finally, they must not depend on the played instruments, dynamics and tuning.

Evaluation of pitch class distribution features is a difficult task, mainly for popular music: first, there is no ground truth of pitch class distribution features of polyphonic audio, so that the only data to compare with are some pitch class profiles computed from symbolic representation of classical music, that we will use in Section 3.8.5. Second, it is difficult to find other implementations of these pitch class distribution features in order to compare them to the proposed ones. Thanks to some collaborations with researchers working on the same field (Dan Ellis and Alexandre Sheh (Sheh and Ellis (2003)), Hendrik Purwins (Purwins (2005)) and Chris Harte (Harte and Sandler (2005))), we have had the opportunity to compare our approach

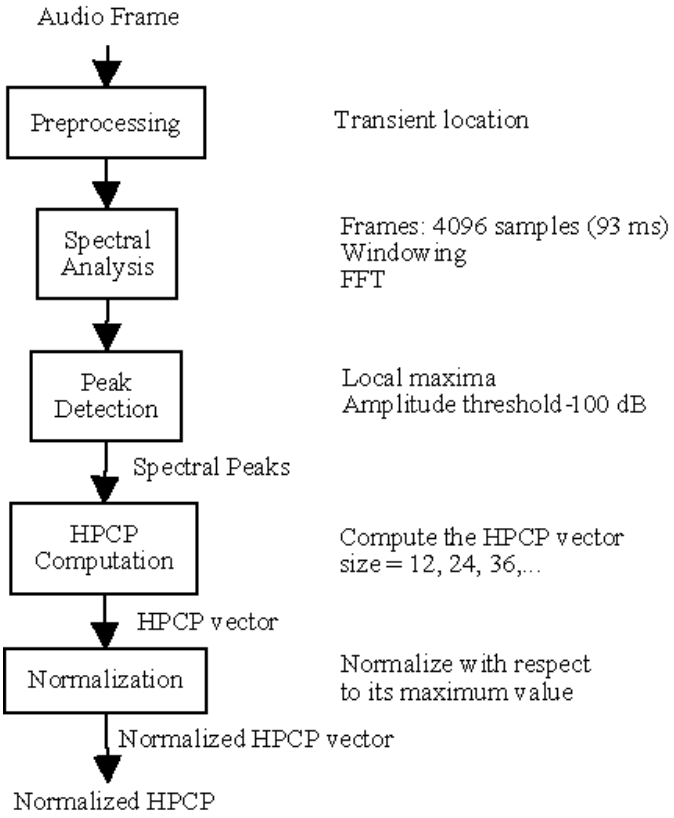


Figure 3.1: Block diagram for HPCP computation.

to similar ones, which have been introduced in Section 3.8.5. As mentioned in Section 3.8, We have tackled the problem of evaluation by presenting some modular evaluation experiments, as well as some examples and case studies. We perform a quantitative evaluation of some of these aspects: the algorithm for reference frequency computation and a quantitative comparison of HPCP vs similar chroma features obtained from audio and score for a reduced music collection.

As a wider objective, this chapter is also intended to justify that it is possible to automatically extract a low-level tonal description from polyphonic audio, without the need of performing a full transcription. These features then overcome the difficulty of multipitch extraction methods. This corresponds to the objective number 4 of this dissertation as mentioned in Section 1.6.

3.2 Pre-processing

The goal of this step is to prepare the signal for the computation of the pitch class distribution vector. Our approach performs a spectral analysis of the signal in order to obtain a representation in the frequency domain, and include some other steps to make the features robust to noise, including a transient location procedure

and frequency selection. We explain these pre-processing steps in this section.

3.2.1 Transient location

As a first step, we include a transient detection algorithm to eliminate regions where the harmonic structure is noisy, so that the areas that are located $50\ ms$ before and after the transients are not analyzed. This pre-processing also decreases the computational cost of the HPCP computation. On the other hand, it does not affect either the resolution or the effectiveness of the tonality descriptors, so that the accuracy for key estimation is similar when discarding the transients.

The transient detection algorithm used in this work is the one proposed by Jordi Bonada in the context of a time-scale audio modification algorithm, which is explained in Bonada (2000). In this method, transients are located by analyzing the evolution of several spectral features: bank filter energies, Mel cepstrum coefficients and their derivatives. The bank of filters used in the algorithm has 42 bands with frequencies between 40 Hz and 20 KHz, following a Mel scale. Several simple rules are applied combining these inputs, in order to find points of maximum increasing slope. Figure 3.2 shows an overall block diagram of the algorithm.

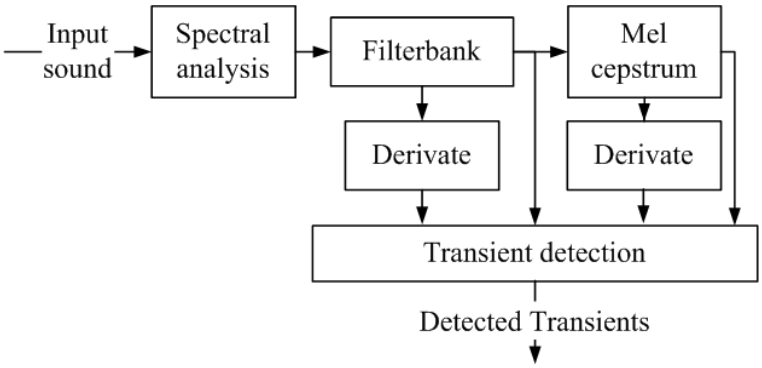


Figure 3.2: Block diagram for transient location (adapted from Bonada (2000)).

3.2.2 Spectral analysis

For the regions which are not located near a transient, the sampled audio signal is analyzed in order to obtain its representation in the frequency domain. The main steps for spectral analysis are Windowing, zero-padding and DFT computation, as illustrated in Figure 3.3.

The sampled input signal $x(n)$ is split into a series of analysis frames of size N_{frame}, $x(n + l \cdot N_{hop})$ where $n = 0, 1, \ldots N_{frame} - 1$, l indicates the number of frame that is analyzed and N_{hop} the time advance for each frame, measured in samples, or hop size.

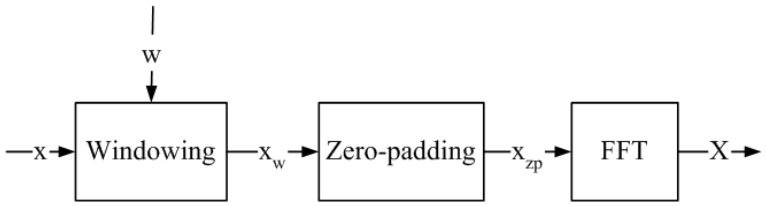


Figure 3.3: Steps for spectral analysis.

3.2.2.1 Windowing

The first step to perform a spectral analysis of a time-domain signal is the windowing of the signal, illustrated in Figure 3.4. Each of the frames $x(n + l \cdot N_{hop})$ is multiplied by a window function $w(n)$ to obtain the windowed signal $x_w(n)$.

$$x_w(n) = x(n + l \cdot N_{hop}) \cdot w(n) \;\; for\; n = 0...N_{frame} - 1 \tag{3.1}$$

All standard windows are real and symmetric and have a frequency spectrum with a sinc-like shape. The choice is mainly determined by two of the spectrum's characteristics: the width of the main lobe and the highest side-lobe level. Most common windows are the rectangular window (main-lobe width of 2 bins and side-lobe level equal to -13 dB), the Hanning window (main-lobe width equal to 4 bins and side-lobe level equal to -23 dB), the Hamming window (main-lobe width of 4 bins and side-lobe level equal to -43 dB), the Blackman-Harris window, the Kaiser window, etc. We use a *Blackman Harris 62 dB* window for analysis. The formula for a L-term Blackman-Harris is defined as:

$$w(n) = \frac{1}{N_{frame}} \sum_{l=0}^{L-1} \alpha_l \cdot cos(\frac{2nl\pi}{N_{frame}}), \quad n = 0, 1, \ldots N_{frame} - 1 \tag{3.2}$$

Blackman-Harris 62 dB has the following parameters: $\alpha_0 = 0.44859$, $\alpha_1 = 0.49364$ and $\alpha_2 = 0.05677$, having side-lobe level equal to -62 dB. Finally, the windowed data is centered on the time origin before DFT computation, in order to get a zero-phase window:

$$x_{wc}(n) = x_w(n + \frac{N_{frame}}{2}) \;\; , \quad n = -\frac{N_{frame}}{2}, \ldots, \frac{N_{frame}}{2} - 1 \tag{3.3}$$

3.2.2.2 Discrete Fourier Transform

After windowing and zero padding, we compute the Discrete Fourier Transform (DFT) to get the frequency spectrum, $X(k)$, as explained in McClelland et al. (1998); Zölzer (2002). $X(k)$ is defined by the following formula:

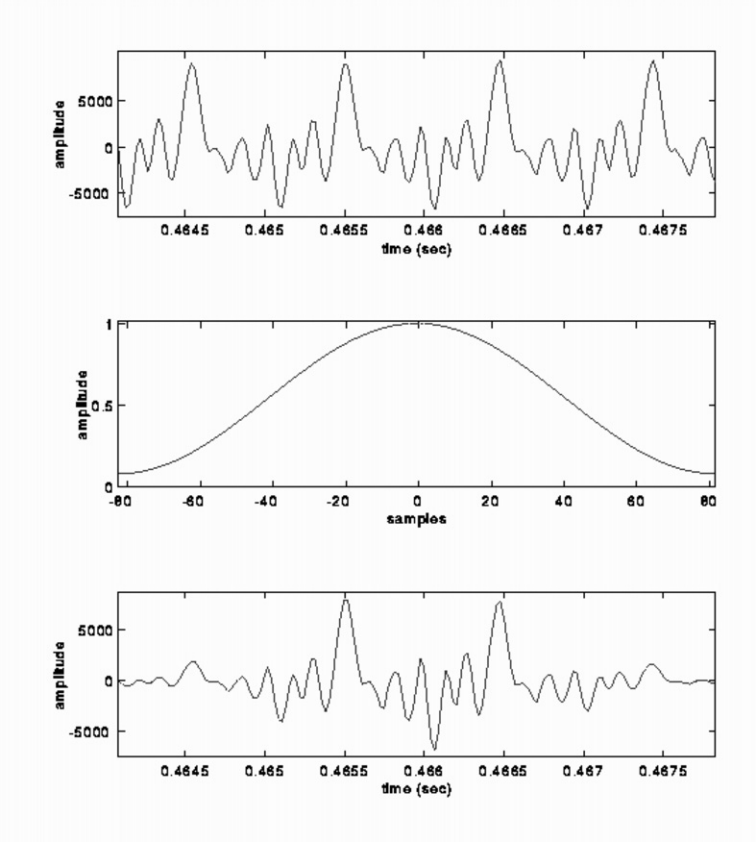


Figure 3.4: Sound frame blocking and windowing. a. Portion of a violin sound to be used in the analysis of the current frame. b. Hamming window. c. Windowed sound. (from Serra (1996)).

$$X(k) = DFT[x(n)] = \sum_{n=-\frac{N_{frame}}{2}}^{\frac{N_{frame}}{2}-1} x_{wc}(n) \cdot e^{-j2\pi nk/N_{frame}} \tag{3.4}$$

where $k = 0, 1, ...N_{frame} - 1$. These N_{frame} samples of the DFT are complex values having a real X_r and imaginary X_i part. We can compute the magnitude $|X(k)|$ and phase $\phi(k)$ of the spectrum following the formulas below:

$$|X(k)| = \sqrt{X_r(k)^2 + X_i(k)^2} \tag{3.5}$$

$$\phi(k) = arctan\frac{X_i(k)}{X_r(k)} \tag{3.6}$$

where $k = 0, 1, \ldots N_{frame} - 1$. The number of points of the DFT $X(k)$ is then equal to N_{frame}, and positive frequencies start from $0\ Hz$ up to $\frac{f_s}{2}\ Hz$. These frequency points are then given by $k \cdot \frac{f_s}{N_{frame}}$, and the frequency resolution is equal to $\frac{f_s}{N_{frame}}$.

3.2.2.3 Zero-padding

In order to increase this frequency resolution for spectrum analysis, we should then increase the frame size N_{frame}. A way to do it without increasing the number of samples analyzed from the input signal $x(n + l \cdot N_{hop})$ is to get N_{frame} samples from $x(n + l \cdot N_{hop})$ and add zero samples until reaching the size N_{FFT} providing the required frequency resolution $\frac{f_s}{N_{FFT}}$. The number of added zeros is then $N_{FFT} - N_{frame}$, and the zero-padding factor is defined as the ratio $\frac{N_{FFT}}{N_{frame}}$. The zero-padded signal is defined as follows:

$$x_{zp}(n) = \begin{cases} 0 & for\ n = -\frac{N_{FFT}}{2} \ldots -\frac{N_{frame}}{2} - 1 \\ x_{wc}(n) & for\ n = -\frac{N_{frame}}{2} \ldots \frac{N_{frame}}{2} - 1 \\ 0 & for\ n = \frac{N_{frame}}{2} \ldots \frac{N_{FFT}}{2} - 1 \end{cases} \tag{3.7}$$

This windowed and zero-padded signal is then the input signal for the computation of the Fourier transform, which, in combination with Equation 3.7, is expressed as follows:

$$X(k) = \sum_{n=-\frac{N_{FFT}}{2}}^{\frac{N_{FFT}}{2}-1} x_{zp}(n) \cdot e^{-j2\pi nk/N_{FFT}} = \sum_{n=-\frac{N_{frame}}{2}}^{\frac{N_{frame}}{2}-1} x_{wc}(n) \cdot e^{-j2\pi nk/N_{FFT}} \tag{3.8}$$

where $k = 0, 1, \ldots N_{FFT} - 1$. The number of points of the DFT $X(f)$ is now equal to N_{FFT} spread over the original sample rate f_s, and the positive frequencies start from $0\ Hz$ up to $\frac{f_s}{2}\ Hz$. These frequency points are then given by $k \cdot \frac{f_s}{N_{FFT}}$, and the frequency resolution is equal to $\frac{f_s}{N_{FFT}}$. The process of zero padding and FFT computation is illustrated in Figure 3.5.

3.2.3 Resolution and spectral analysis parameters

One of the most important parameters for frequency analysis is the analysis frame size N_{frame}. The compromise between having a good temporal resolution (using short windows) or a good frequency resolution (using long windows) is one of the problems of the FFT analysis, as seen in the literature (see for instance Harris (1978) or Amatriain et al. (2002), pp. 379).

When performing a tonal analysis of an audio signal, it is necessary to get a good frequency resolution, so that large frames are defined. We use a frame size of $N_{frame} = 4096$ samples, that is, $T = 93$ ms for a sample rate of 44.1 KHz. This is also the case in all the approaches for tonality estimation including DFT computation, mentioned in Section 2.6.3. Some approaches define even larger frame sizes (e.g. 400 ms in Fujishima (2000)).

Consecutive analysis frames are spaced N_{hop} samples. In our approach, we define a hop size of $N_{hop} = 512$ samples (that is approximately $T_{hop} = 11$ ms).

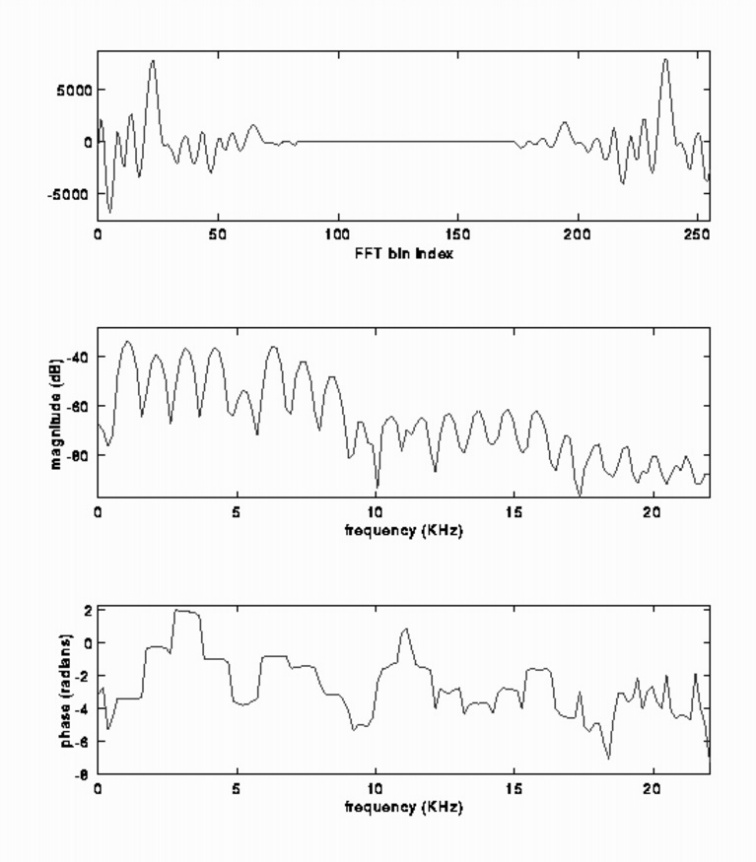


Figure 3.5: Computing the FFT. a. packing of the sound into the FFT buffer for a zero phase spectrum. b. Magnitude spectrum. c. Phase spectrum (from Serra (1996)).

3.2.4 Peak detection

The peak detection algorithm used in our system has been developed in the context of the Sinusoidal Modelling Synthesis framework (SMS), proposed in Serra (1996) and also explained in Amatriain et al. (2002), which has been mainly used to represent monophonic signals. The Fourier theorem states that a frame of N samples can be represented perfectly with N frequency components, and sinusoidal model assumes that a spectrum $X(k)$ can be represented by a smaller number of frequency components, called sinusoids. A sinusoid is a frequency component that is stable both in frequency and amplitude. Each sinusoid represents a partial, having a well-defined representation that is the Fourier transform of the window used for analysis. There are some interactions between frequency components that makes it difficult to estimate the partial by analyzing a single spectrum.

In SMS, we define a ”peak” as a local maximum in the magnitude spectrum, where the only constraints are that its frequency belongs to a certain range and its magnitude is higher than a given threshold. Due to the sampled nature of the spectra, each peak is accurate only to within half a spectral bin. A spectral bin

represents a frequency interval of $\frac{f_s}{N_{FFT}}$ Hz, where f_s is the sampling rate and N_{FFT} is the FFT size. An example of the peaks detected from a spectrum is shown in Figure 3.6.

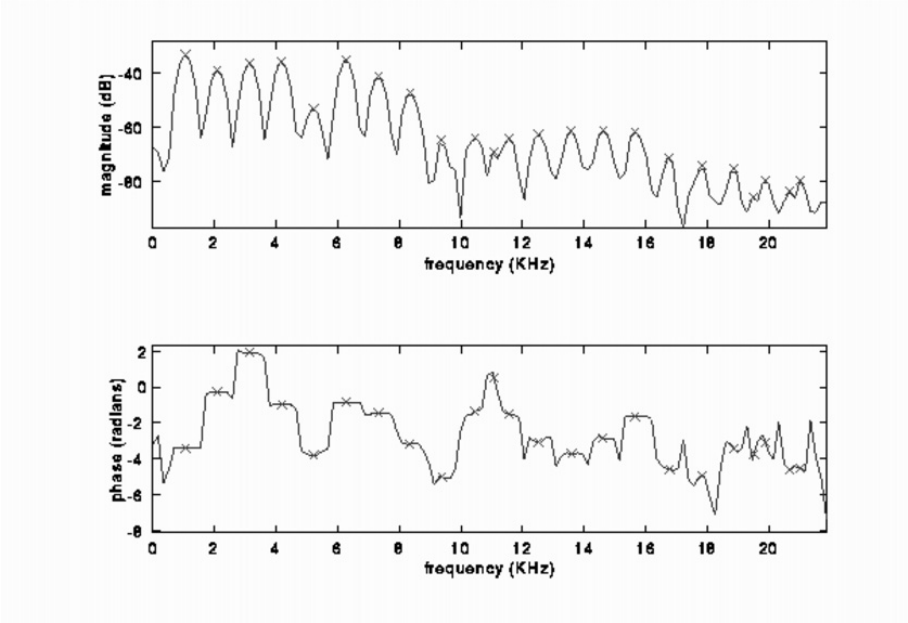


Figure 3.6: Peak detection. a. Peaks in the magnitude spectrum. b. Peaks in the phase spectrum (from Serra (1996)).

As we have seen above, zero-padding increases the frequency resolution and then the accuracy of a simple peak detector. According to Amatriain et al. (2002) pp. 383, if we use a rectangular window, the required zero-padding factor is 1000 if we want to increase the frequency resolution on the level of 0.1% of the width of the window transform main-lobe. In order to decrease this required zero-padding factor, an additional procedure is used, where quadratic spectral interpolation is performed using only the samples which are close to the maximum-magnitude frequency sample. In this procedure, the three closest points to the maximum-magnitude frequency are considered to be in a parabola defined by the following equation:

$$y(x) = a(x-p)^2 + b \tag{3.9}$$

where p is the center of the parabola, a is a measure of its concavity and b is the offset. Let's consider the maximum frequency point k_β being a local maximum:

$$\begin{aligned} y(-1) &= \alpha = 20\log_{10}|X(k_\beta - 1)| \\ y(0) &= \beta = 20\log_{10}|X(k_\beta)| \\ y(1) &= \gamma = 20\log_{10}|X(k_\beta + 1)| \\ &\alpha \leq \beta \geq \gamma \end{aligned} \tag{3.10}$$

If we consider that they define a parabola, we can then obtain the value of the center of the parabola p, defining the interpolated peak location:

$$p = \frac{1}{2} \cdot \frac{\alpha - \gamma}{\alpha - 2\beta + \gamma} \tag{3.11}$$

so that the peak location in frequency bins is equal to

$$k^* = k_\beta + p \tag{3.12}$$

and the peak magnitude is equal to

$$y(p) = 20 \cdot \log_{10} |X(k^*)| = \beta - \frac{1}{4}(\alpha - \gamma)p \tag{3.13}$$

The magnitude threshold used for peak detection is set to -100 dB with respect to the maximum possible magnitude, so that only local maxima higher than this threshold are selected.

3.2.5 Frequency filtering

As many of the approaches for pitch class distribution features determination explained in Section 2.6.3, we perform a selection of the detected spectral peaks according to their frequency. We only consider those spectral peaks i whose frequency belongs to the frequency interval $f_i \epsilon [100, 5000]$ Hz, not considering high frequencies in our analysis. The reason is that the predominant audio objects are more noisy in this region, due to some percussion and instrumental noise (blowing, string frictions, etc). The output of this procedure is a set of spectral peaks $\{a_i, f_i\}, i = 1 \ldots nPeaks$ having amplitude values a_i (linear value, that we named above $X(k_i^*)$) and frequency values f_i (according to the computed bin position k_i^*). This is the input information for the following steps of pitch class profile computation.

3.3 Reference frequency determination

The basis of pitch class profiles is to map frequency values to pitch classes following an equal-tempered scale and using a given reference frequency f_{ref}. As seen in Chapter 2, pieces are not always tuned to the standard reference frequency 440 Hz. In order to make the tonal features independent of the tuning frequency, it is necessary to estimate this reference frequency, that is the frequency used to tune the analyzed piece.

As reviewed in Section 2.6.3, the procedure for reference frequency determination can be performed in two different ways: some methods estimate the reference frequency before mapping frequency values to pitch class values; other algorithms shift the pitch class distribution features as a post-processing step in order to tune the features to the right reference frequency.

Our algorithm uses the first approach, performing an estimation of the reference frequency before computing the pitch class distribution vector. The reference frequency is estimated for each analysis frame by

analyzing the deviation of the spectral peaks with respect to the standard reference frequency 440 Hz. Then, a global value is obtained by combining the frame estimates. A quantitative evaluation of this procedure for reference frequency determinations is presented in Section 3.8.3.4.

3.3.1 Instantaneous reference frequency

As shown in Section 2.2.1, a note is composed of several harmonics whose frequencies f_n are multiple of the fundamental frequency f_0:

$$f_n = n \cdot f_0 \tag{3.14}$$

with $n = 1 \ldots N$. If the fundamental frequency is detuned with a given frequency factor α, then the frequencies f_n change to f'_n:

$$f'_n = n \cdot \alpha \cdot f_0 \tag{3.15}$$

We can compute the interval (expressed in semitones) between each harmonic frequency and the standard reference frequency (440 Hz) as follows:

$$\beta'_n = 12 \cdot \log_2(\frac{n \cdot \alpha \cdot f_0}{440}) = \beta_n + 12 \cdot \log_2 \alpha = \beta_n + d \tag{3.16}$$

The interval between each harmonic and the standard reference frequency 440 Hz ($\beta_n = 12 \cdot \log_2 \frac{n \cdot f_0}{440}$) in Equation 3.16) can be slitted into two terms: the interval between the fundamental frequency and 440 Hz and the interval between each harmonic and the fundamental frequency, so that we obtain the following expression for β'_n:

$$\beta'_n = 12 \cdot \log_2(\frac{f_0}{440}) + 12 \cdot \log_2 n + d \tag{3.17}$$

$$\beta'_n = \beta_0 + dev_n + d \tag{3.18}$$

We define the following term for each harmonic:

$$d_n = d + dev_n \tag{3.19}$$

The term $dev_n = 12 \cdot \log_2 n$ depends on the harmonic number. We can fold the value of this deviation into one single octave, as follows:

$$dev'_n = 12 \cdot (\log_2 n - round(\log_2 n)) \tag{3.20}$$

The harmonics $2f_0, 4f_0, 8f_0, \ldots$ are exactly in tune with the fundamental frequency (i.e. $dev_n = 0$),

as the formed interval is an equal-tempered octave. We can compute this detuning factor for the different harmonics of a perfect harmonic set as presented in Table 3.1.

Harmonic	**Frequency**	dev_n **(cents)**
1	f_0	0
2	$2 \cdot f_0$	0
3	$3 \cdot f_0$	1.95
4	$4 \cdot f_0$	0
5	$5 \cdot f_0$	-13.69
6	$6 \cdot f_0$	1.95
7	$7 \cdot f_0$	-31.17
8	$8 \cdot f_0$	0

Table 3.1: Detuning factor, measured in cents, between the first 8 harmonics of a complex tone and its fundamental frequency f_0.

Figure 3.7 shows an histogram of the detuning (in cents) of the first 50 harmonics with respect to its fundamental, where all the harmonics have the same weight. We observe that the most frequent value is around 0 cents. Based on this fact, our approach for reference frequency determination assumes that the majority of frequency components of a harmonic series are in tune with the fundamental frequency, i.e. they share the same detuning factor. This can be extended to polyphonic sounds, formed by the combination of several harmonic series, which is the focus of our study. In a polyphonic situation, the different fundamental frequencies are usually related by different intervals, as introduced in Section 2.2.1.

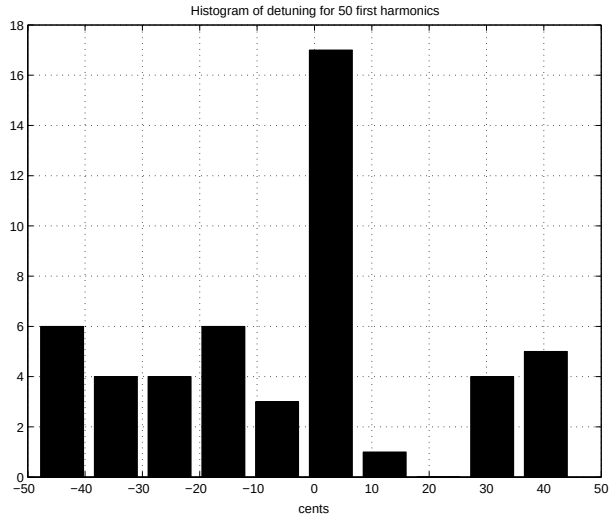


Figure 3.7: Histogram of detuning factor for the first 50 harmonics (f_n, $n = 1 \ldots 50$) with respect to its fundamental frequency f_0.

The goal of the reference frequency computation procedure is to estimate the detuning factor d in a real (monophonic or polyphonic) spectrum, where the prominent peaks in the spectrogram mainly correspond to

the harmonics of one or more notes. Then, the detuning factor d can be estimated by analyzing the deviation of the spectral peaks with respect to the standard reference frequency. Let's consider the ensemble of spectral peaks $\{a_i, f_i\}, i = 1 \ldots nPeaks$ having amplitude values a_i and frequency values f_i. The interval between a spectral peak and the standard reference frequency, measured in semitones, β_i is defined as:

$$\beta_i = 12 \cdot \log_2(\frac{f_i}{440}) \tag{3.21}$$

The deviation of this peak related to the quantized semitone, d_i, is given by:

$$d_i = \beta_i - round(\beta_i) \tag{3.22}$$

where $d_i \epsilon [-0.5, 0.5)$. The deviation, measured in semitones, estimated for this frame d is computed as an statistics on the deviations of each of the peaks. We build an histogram of 1 semitone width ($[-0.5, 0.5)$) of the deviations for each of the peaks:

$$hist(n) = \sum_{i, d_i + 0.5 \in [(n-1)r, nr]} a_i \tag{3.23}$$

where $n = 1, 2, \ldots, r$ is the histogram resolution measured in semitones and a_i weights the contribution of each spectral peak for tuning frequency computation, which is equal to its magnitude.

From this histogram, d is computed of its maximum value:

$$d = -0.5 + r \cdot argmax_n(hist(n)). \tag{3.24}$$

The reason for building an histogram instead of simply computing the average of d_i is the following one. If the deviation is near 0.5 semitones, the histogram has approximately 50% of values near -0.5 and 50% near +0.5, so that one of these values is chosen to compute d. If we would consider average, this would lead to 0 deviation.

From this deviation, d, measured in semitones, we can compute the reference frequency for a given analysis frame:

$$f_{ref} = 440 \cdot 2^{\frac{d}{12}} \tag{3.25}$$

3.3.2 Global reference frequency

The goal of this step is to obtain a single reference frequency value for each musical piece. We make then the assumption that the tuning frequency is usually constant for each musical piece. A global tuning frequency measure is obtained by performing some statistics of the frame values. Figure 3.8 shows and example of the instantaneous evolution of the deviation d measured in cents[1]. We can observe that in the case of a detuning

[1] These samples correspond to the following sounds included in Appendix A: Tune1-0.mid.wav, Tune1-20.mid.wav and Tune1-60.mid.wav

of 60 cents, the values fluctuate between ± 50 cents.

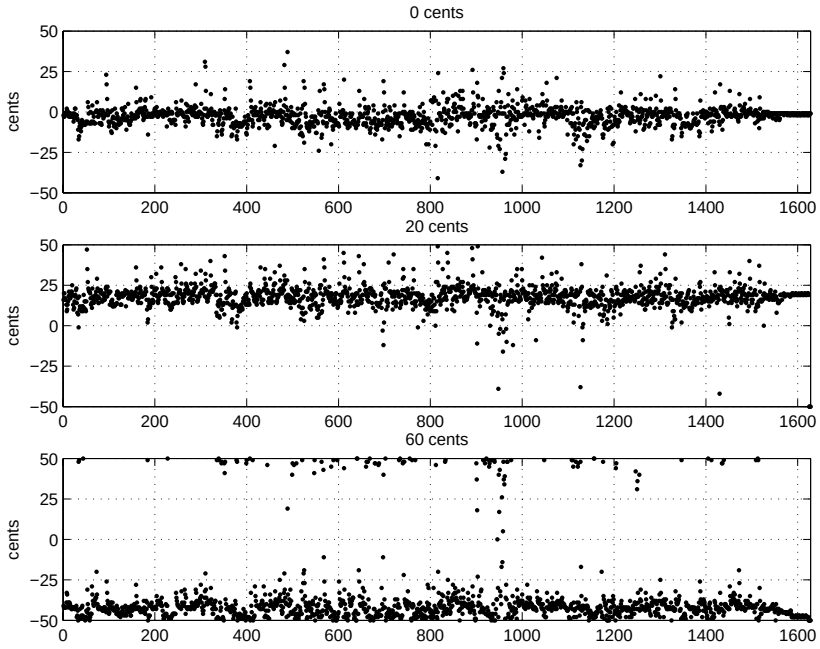


Figure 3.8: Frame deviations (in cents) with respect to 440, computed for a piece perfectly in tune (up), with a deviation of 20 cents (middle) and 60 cents (down).

Let's consider d_f the deviation computed for a given frame, and e_f the energy of that frame, given by the amplitude of its spectral peaks:

$$e_f = \sum_{i=1}^{nPeaks} a_i^2 \tag{3.26}$$

As for computing statistics on frames, a 1 semitone histogram (centered around 0 $[-0.5, 0.5)$) is built from all frames within a certain segment. We choose a segment of 30 seconds in the middle of the piece to compute a global measure, as it is more likely to find all instruments than in the beginning of the piece. Each frame contribution is determined by its energy:

$$hist(n) = \sum_{i, d_f + 0.5 \in [(n-1)r, nr]} e_f \tag{3.27}$$

where $n = 1, 2 \ldots$, and r is the histogram resolution measured in semitones. From this histogram, the global deviation d is computed of its maximum value:

$$d = -0.5 + r \cdot argmax_n(hist(n)). \tag{3.28}$$

The reason for building an histogram instead of simply computing the average of d_i is the same than for instantaneous tuning frequency computation. If the deviation is near 0.5 semitones, the histogram has approximately 50% of values near -0.5 and 50% near +0.5, so that one of these values is chosen as global detuning factor. If we would consider average, this would lead to 0 deviation.

From this deviation, d, measured in semitones, we can compute the reference frequency for the entire piece:

$$f_{ref} = 440 \cdot 2^{\frac{d}{12}} \tag{3.29}$$

3.4 The Harmonic Pitch Class Profile

The Harmonic Pitch Class Profile (HPCP) is a pitch class distribution feature based on the Pitch Class Profile (PCP) proposed by Fujishima (1999) in the context of a chord recognition system. As explained in Section 2.6.3, this vector measures the intensity of each of the twelve semitones of the diatonic scale, an it is obtained by mapping each frequency bin of the spectrum to a given pitch class. The HPCP introduces some modifications with respect to the PCP computation proposed by its author: first, we introduce a weight into the feature computation; second, we consider the presence of harmonics; third, we use a higher resolution in the HPCP bins (decreasing the quantization level to less than a semitone). As mentioned above, we only consider those spectral peaks whose frequency belongs to the interval $f_i \epsilon [100, 5000]$ Hz, not considering high frequencies in our analysis. The HPCP vector is then defined by the following formula:

$$\begin{aligned} HPCP(n) = \sum_{i=1}^{nPeaks} w(n, f_i) \cdot a_i^2 \\ n = 1...size \end{aligned} \tag{3.30}$$

where a_i and f_i are the linear magnitude and frequency values of the peak number i, $nPeaks$ is the number of spectral peaks that we consider, n is the HPCP bin, $size$ is the size of the HPCP vector (i.e. number of bins: $12, 24, 36, \ldots$), and $w(n, f_i)$ is the weight of the frequency f_i when considering the HPCP bin n.

3.4.1 Weighting function

The weighting function $w(n, f_i)$ is defined as follows. Instead of contributing to a single HPCP bin (as for instance the closest one), each frequency f_i contributes to the HPCP bin(s) that are contained in a certain window around this frequency value, as shown in Figure 3.9. For each of those bins, the contribution of the peak i (the square of the peak linear amplitude $|a_i|^2$) is weighted using a cos^2 function around the frequency of the bin. The value of the weight depends on the frequency distance between f_i and the center frequency of the bin n, f_n, measured in semitones, as follows:

Let the center frequency of the n bin be:

$$f_n = f_{ref} \cdot 2^{\frac{n}{size}} \qquad n = 1...size \tag{3.31}$$

Let the distance in semitones between the peak frequency f_i and the bin center frequency f_n be:

$$d = 12 \cdot log_2(\frac{f_i}{f_n}) + 12 \cdot m \tag{3.32}$$

where m is the integer that minimizes the module of the distance $|d|$.

Then, the weight is computed as follows:

$$w(n, f_i) = \begin{cases} cos^2(\frac{\pi}{2} \cdot \frac{d}{0.5 \cdot l}) & \text{if } |d| \leq 0.5 \cdot l \\ 0 & \text{if } |d| > 0.5 \cdot l \end{cases} \tag{3.33}$$

where l is the length of the weighting window. This value is a parameter of the algorithm, and we have set it empirically to $\frac{4}{3}$ semitone. Let's consider a HPCP $size = 36$. Then, the bin resolution will be $\frac{1}{3} \cdot semitone$. If we consider $l = \frac{4}{3}\ semitone$, each spectral peak will contribute to 4 different HPCP bins with different weights, as illustrated in Figure 3.9.

This weighting procedure minimizes the estimation errors that we find when there are tuning differences and inharmonicity present in the spectrum. Those factors can induce errors when mapping frequency values into HPCP bins.

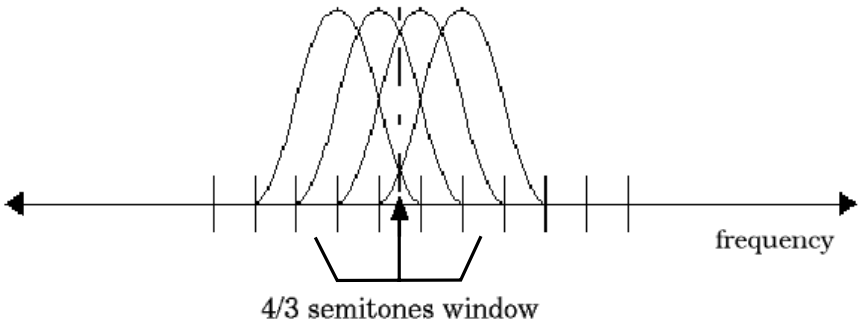


Figure 3.9: Weighting function.

3.4.2 Consideration of harmonic frequencies

As explained in Section 2.2.1, the spectrum of a note is composed of several harmonics, whose frequencies are multiples of the fundamental frequency (f, $2 \cdot f$, $3 \cdot f$, $4 \cdot f$, etc.). When we play a single note, spectral

components appear at frequencies of the different harmonics. This fact affects the HPCP values, where the value increase in different bins. If we consider a temperate scale, the value i_n associated to the n^{th} harmonic of a note (we can call it the n^{th} *harmonic pitch class*) can be computed as follows:

$$i_n = mod[(i_1 + 12 \cdot log_2(n)), 12]; \tag{3.34}$$

where i_1 is the pitch class that corresponds to the note fundamental frequency.

In order to make harmonics contribute to the pitch class of its fundamental frequency, we introduce a weighting procedure: each peak frequency f_i has a contribution to the frequencies having f_i as harmonic frequency ($f_i, \frac{f_i}{2}, \frac{f_i}{3}, \frac{f_i}{4}, \ldots, \frac{f_i}{nHarmonics}$). We make this contribution decrease along frequency using the following function:

$$w_{harm}(n) = s^{n-1} \tag{3.35}$$

where $s < 1$, in order to simulate that the spectrum amplitude decreases with frequency (Rossing (1989) pp. 125-132, Fletcher and Rossing (1991) and Morse (1983)). This function is shown in Figure 3.10 (where s is set to 0.6) and illustrated in Table 3.2. Ideally, the value of s should depend on the instrument timbre. However, in our experiments we have used a value of 0.6. Further experiments should be devoted to better study the influence of this parameter.

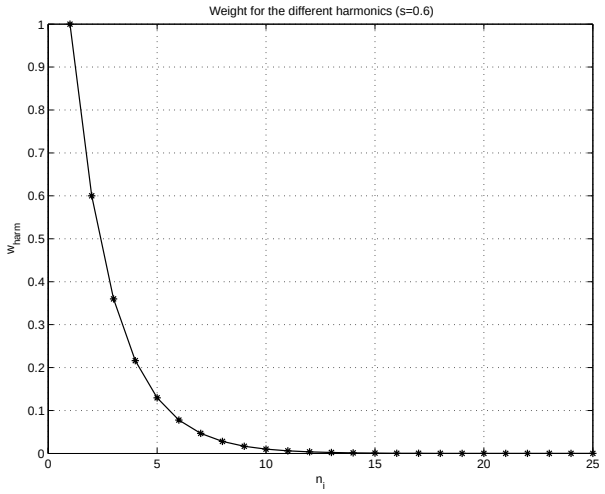


Figure 3.10: Weighting function for the contribution of harmonic frequencies.

3.4.3 Spectral whitening

The weighting function for each frequency should be adapted to the spectral envelope of the analyzed signal. In order to do so, we insert a pre-processing step where the spectrum is normalized according to its spectral

Table 3.2: Contribution for the first 6 harmonics of a note

n	Frequency	Factor
1	f	1
2	$2 \cdot f$	s
3	$3 \cdot f$	s^2
4	$4 \cdot f$	s^3
5	$5 \cdot f$	s^4
6	$6 \cdot f$	s^5

envelope, in order to convert it to a flat spectrum. Using this timbre normalization, notes on high octaves contribute equally to the final HPCP vector than those on low pitch range, and the results are not influenced by different equalization procedures. Schwarz and Rodet (1999) discuss about different methods to estimate the spectral envelope of a sound.

3.5 Post-processing

3.5.1 Normalization

For each analysis frame, the HPCP values are normalized with respect to its maximum value, in order to store the relative relevance of each of the HPCP bins.

$$HPCP_{normalized}(n) = \frac{HPCP(n)}{Max_n(HPCP(n))} \qquad n = 1...size \tag{3.36}$$

This normalization process, together with the peak detection stage, provides independence to dynamics and overall volume, as well as to the presence of soft noise. First, only spectral peaks having a magnitude higher than a threshold are selected, so that if the energy is very low there are no detected peaks and the computed HPCP vector is flat. Second, if the spectral amplitude is multiplied by a factor, the resulting HPCP vector is also scaled. The influence of this scale factor is then eliminated by the normalization process.

3.6 Segment features

Statistics of HPCP descriptors computed for each analysis frame are computed in order to characterize a given segment. The main statistic studied in this dissertation is the average. If the temporal boundaries of a given segment are represented by its beginning and end frame indexes, $f = b$ and $f = e$, and $HPCP_f$ represents the normalized HPCP vector computed for the frame with index f, the average is given by the following formula:

$$HPCP_{av}(n) = \frac{1}{e-b+1} \sum_{f=b}^{f=e} HPCP_f(n) \qquad n = 1...size \tag{3.37}$$

We analyze in Chapter 4 how the average of the HPCP is used for key estimation over a certain audio excerpt. Depending on the segment duration, the estimation corresponds to the global key or to a local estimation of the key. When the segment duration is small, this average represents the played chord, as shown in Section 4.3.6. Section 4.3.4 studies the differences in average profile for different musical genres. Chapter 5 studies how the average profile can be exploited to compute similarity between pieces.

3.7 The Transposed Harmonic Pitch Class Profile

We have defined a feature derived from HPCP, which is invariant to transposition: a *transposed* version of the HPCP. The THPCP is computed as a shifted version of the HPCP according to a certain index $shift$:

$$THPCP(n) = HPCP(mod(n - shift, size)) \qquad n = 1, \ldots, size \tag{3.38}$$

where $shift$ can be defined in different ways. We see in Chapter 4 that it can be fixed to the index corresponding to the annotated key in order to analyze the tonal profile given by the HPCP features. In order to compute this vector in an automatic way, the value of $shift$ can be also set either to the index of the estimated key or the index corresponding to the maximum value of the HPCP vector. This feature is invariant to transposition. Experiments have shown that interval is an important element for melody recognition (see Dowling (1978)), so that contour representations are used for melodic retrieval. We extend this idea to tonal similarity in Chapter 5, where pieces are considered to be similar if they share the same tonal profile regardless of its tonic.

3.8 Evaluation

The goal of this section is to illustrate how the proposed features fulfill the requirements mentioned in Section 2.6.3: representation of the pitch class distribution of both monophonic and polyphonic signals, to consider the presence of harmonic frequencies, robustness to noise, independence of timbre and played instrument, independence of dynamics and independence of tuning.

As mentioned in the introduction of this chapter, the evaluation of pitch class distribution features is a difficult task. The first reason for that is the lack of ground truth of pitch class distribution features from polyphonic audio. The only available data is a set of pitch class profiles computed from symbolic representation of classical music, that we employ in Section 3.8.5. Second, it is difficult to find other implementations of these features in order to verify how they perform compared to the proposed one. Thanks to some researchers

working on the same field, Dan Ellis and Alexandre Sheh (Sheh and Ellis (2003)), Hendrik Purwins (2005) and Chris Harte (Harte and Sandler (2005)), we have been able to compare our features to similar approaches, as presented in Section 3.8.5.

We have tackled the problem of evaluation by presenting some partial evaluation and case studies. We first present some examples of visualization of HPCP features. We also perform some quantitative evaluation of some aspects of the procedure and we finally compare our approach with similar ones dealing with audio and symbolic notation. A quantitative evaluation of the utility of these features for high-level tonal description, focusing on key estimation from audio recordings, is given in the following chapters.

3.8.1 Case study

In this section we show an example of the level of the description that is achieved and coded in the HPCP features. We have chosen this excerpt because it includes different aspects of the music that can be observed in our low-level tonal descriptors. We can analyze the characteristics of this audio excerpt based in the HPCP representation, as follows.

Figures 3.11, 3.12, 3.13, 3.14 and 3.15 show different aspects that can be observed by analyzing the evolution of the short-term HPCP features for a fragment of the song *Imagine*, by John Lennon (see Appendix A for further description and to listen to the audio sample). In Figure 3.11 there is a plot of the temporal evolution (horizontal axis) of the HPCP features (vertical axis) when considering an interval resolution of $\frac{1}{10}$ semitones (10 cents or $size = 120$ bins). The figure shows that there appear the pitch classes A, B, C, D, E, F and G.

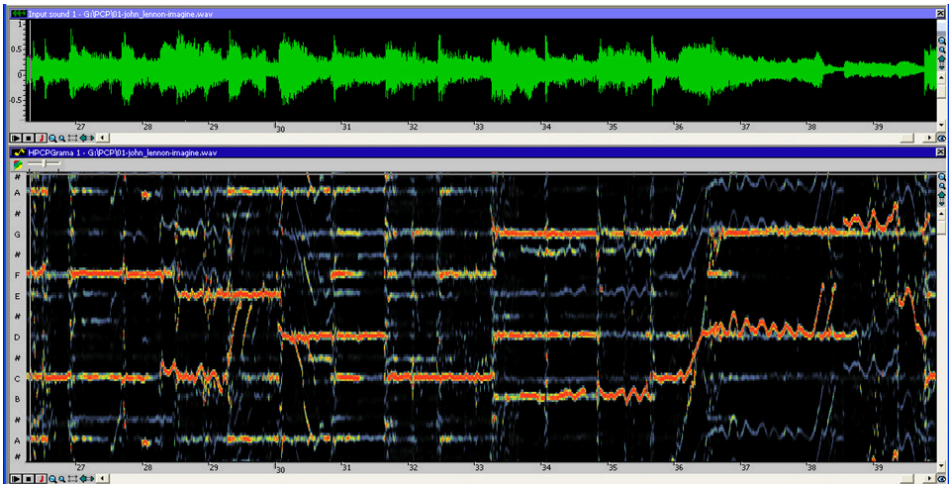


Figure 3.11: HPCP short-term evolution, where we identify the pitch classes C, D, E, F and G. See *Imagine.wav* in Appendix A for further details on the sound excerpt.

We can identify the different pitch classes played by the different instruments: in Figure 3.12 the bass line is marked, while in Figure 3.13 the singing voice. Second, some applied processing is shown in the

representation, as the echo applied to the voice shown in Figure 3.13.

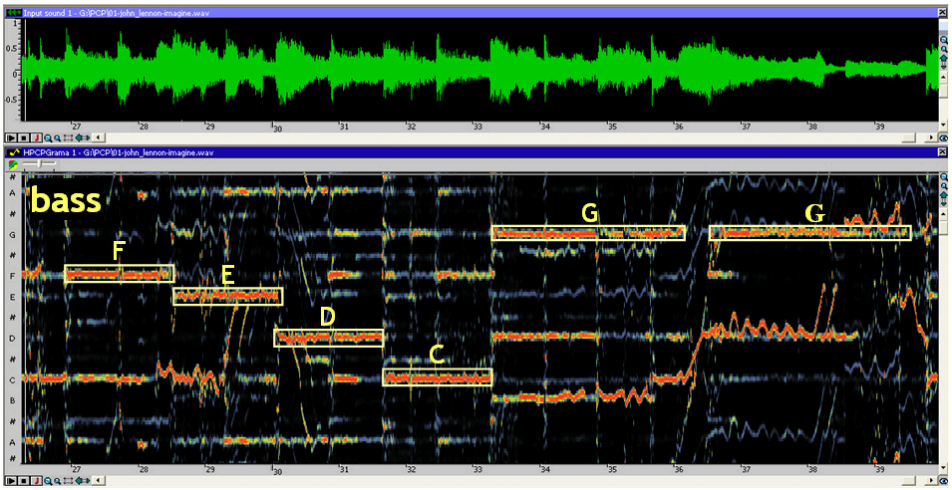


Figure 3.12: HPCP evolution with marked bass line of an excerpt of the song *Imagine*, by John Lennon. We identify the following melody: F-E-D-C-G. See *Imagine.wav* in Appendix A for further details on the sound excerpt.

Finally, Figure 3.15 shows that it is possible to observe overlapping notes in the HPCP values in the case that one of the sounds exhibits vibrato or frequency movement.

As a conclusion of observing similar examples of visualization of HPCP, we can say that these features are visually informative as a representation of the pitch class content of the analyzed fragment. This observation make us think that just a visualization of HPCP features can be useful for music analysis tasks.

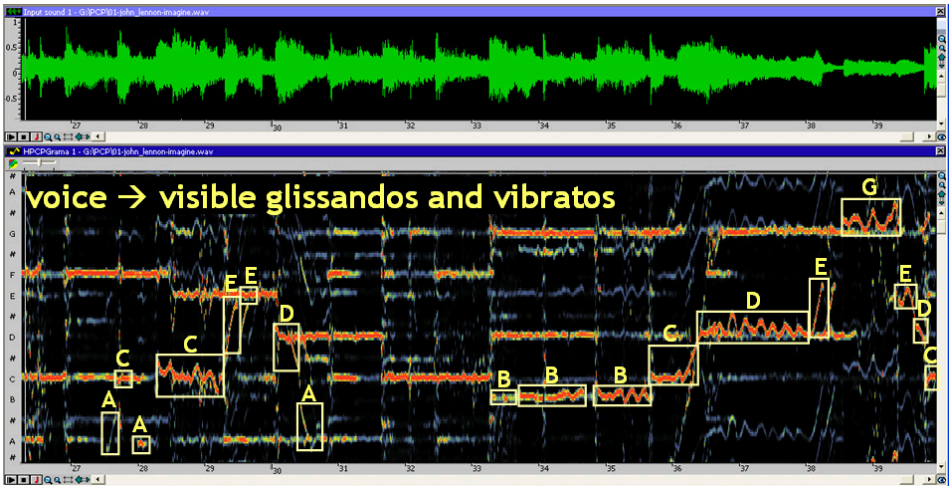


Figure 3.13: HPCP evolution with marked leading voice of an excerpt of the song *Imagine*, by John Lennon. We can follow here the main melody, as well as small fundamental frequency variations due to vibrato and glissandi. See *Imagine.wav* in Appendix A for further details on the sound excerpt.

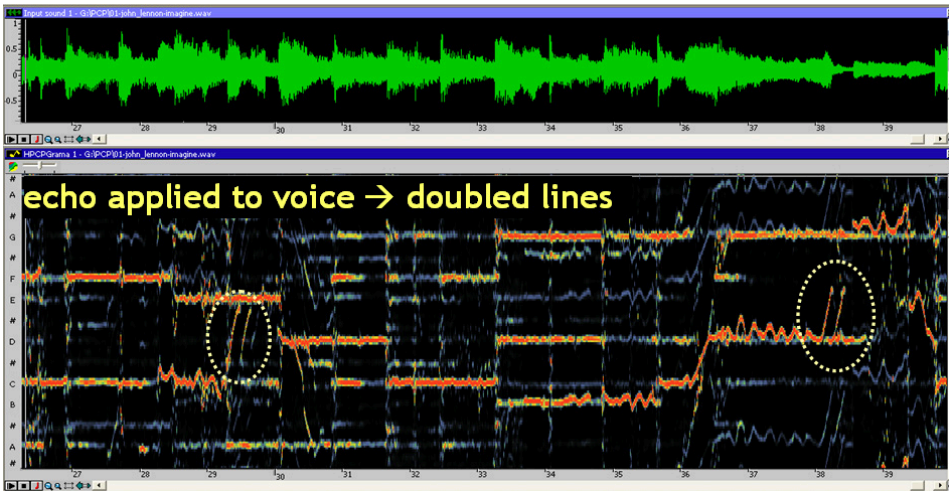


Figure 3.14: HPCP evolution of an excerpt of the song *Imagine*, by John Lennon. We can identify here some echo effect applied to the voice. See *Imagine.wav* in Appendix A for further details on the sound excerpt.

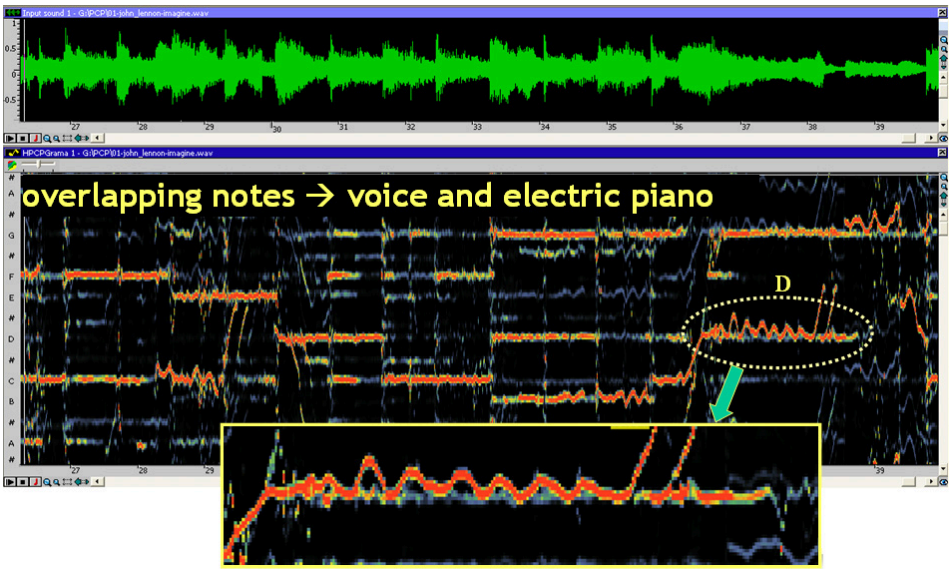


Figure 3.15: HPCP evolution of an excerpt of the song *Imagine*, by John Lennon. We have marked overlapping notes. See *Imagine.wav* in Appendix A for further details on the sound excerpt.

3.8.2 Influence of analysis parameters

There some analysis parameters used for HPCP computation that have an important influence in the behavior of these low-level tonal features. Interval (frequency) and temporal resolution, as well as frequency band used for analysis are factors that influence the information coded in the HPCP vector. We give some examples of how these parameters affects the final result.

3.8.2.1 Interval resolution

Figure 3.16 shows an example of the influence of the parameter $size$ of the HPCP vector when analyzing a fragment of a popular song (See *Donde-Estas-Yolanda.wav* in Appendix A for further details on the sound excerpt).

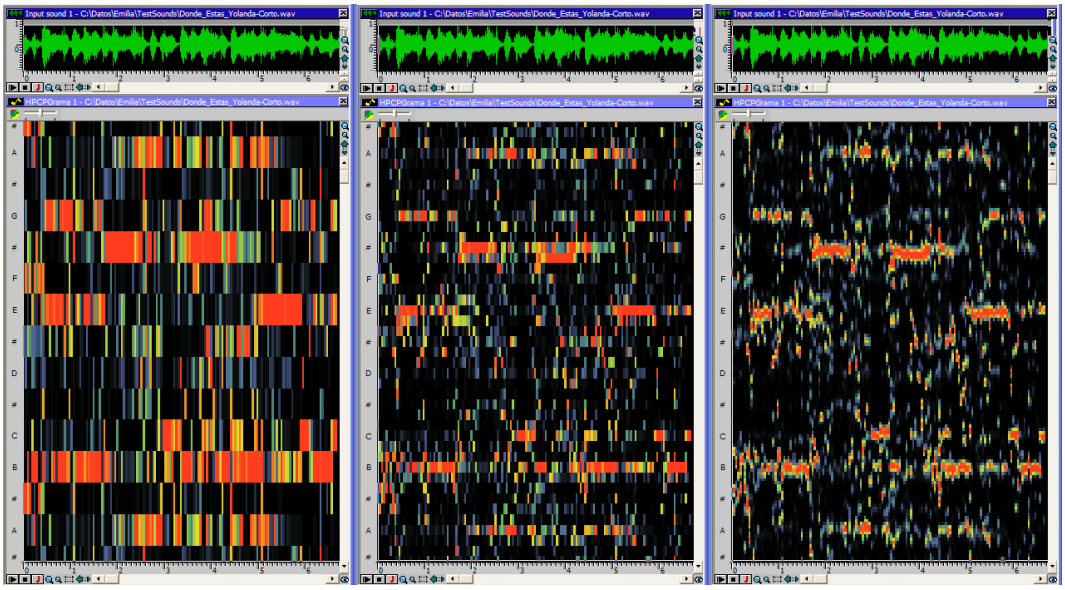


Figure 3.16: Sound file analyzed using different resolution: $size = 12$, i.e. 1 semitone or 100 cents resolution (left), $size = 36$, i.e. $\frac{1}{3}$ semitone or 33.3 cents resolution (middle) and $size = 120$, i.e. $\frac{1}{10}$ semitone or 10 cents resolution (right). See *Donde-Estas-Yolanda.wav* in Appendix A for further details on the sound excerpt.

This parameter determines the number of divisions to the tempered scale that are considered: one semitone or 100 cents ($size = 12$), one third of semitone or 33 cents ($size = 36$) or even 10 cent resolution ($size = 120$). This parameter influences the frequency resolution of the HPCP vector. We see that as the interval resolution increases, it is easier to distinguish frequency details as for instance vibrato, glissando and to differentiate voices in the same frequency range. It would be desirable to use a high frequency resolution when analyzing expressive frequency evolutions. On the other hand, increasing the interval resolution also increases the quantity of data and the computation cost. Figure 3.17 shows the evolution of the time required

to compute the HPCP vector with respect to its size. This computation time is normalized by the duration of the excerpt and an average is computed for 75 musical excerpts (see WTC collection described in 4.3.1.1). We see that the computation time increases, and the speed of computation decreases: 120 times faster than real-time for one semitone resolution and 20 times faster than real-time for a resolution of 10 cents. According to Figure 3.17, the computation time is asymptotically linearly dependent on the *size* parameter. The machine used for this experiment is a LINUX machine Intel(R) Pentium(R) 4 CPU 2.00GHz.

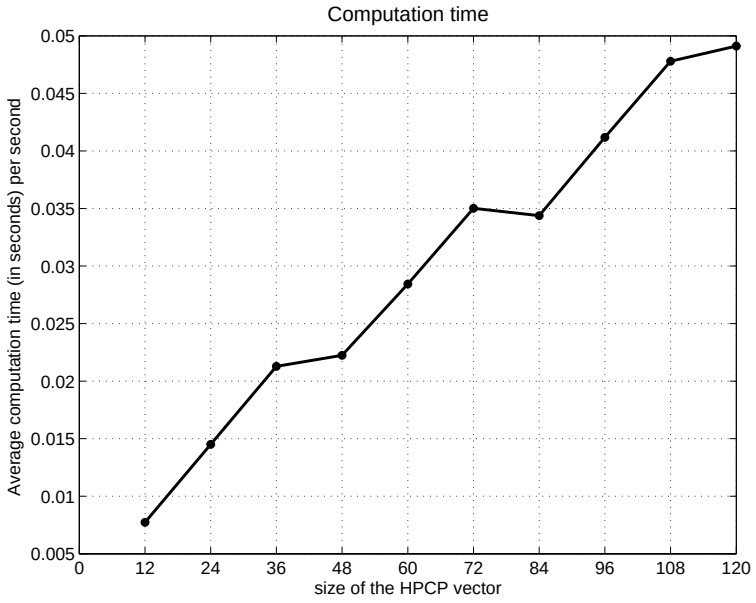


Figure 3.17: Variation of the time required for computing the HPCP vector with respect to its size. This computation time is normalized by the duration of the excerpt and an average is computed for 75 musical excerpts (see WTC collection described in 4.3.1.1).

In some situations such as symbolic transcription or tonality estimation, the resolution required for the output is always equal to one semitone. Using higher resolution improves the robustness to tuning and to deviations of harmonics with respect to the tempered scale.

3.8.2.2 Temporal resolution

It is also important to consider temporal resolution, which is partially determined by the window size used for analysis and also by the hop size considered between consecutive frames. As we have seen in Section 3.2.3, in the context of tonal description it is necessary to get a good frequency resolution, so that large frames are usually defined. We use a frame size of $N_{frame} = 4096$ samples, that is, $T = 93$ ms for a sample rate of 44.1 KHz. Consecutive analysis frames are spaced N_{hop} samples. We define a hop size of $N_{hop} = 512$ samples (that is approximately $T_{hop} = 11$ ms) for tonality description. But if we change the separation between consecutive frames, N_{hop}, for HPCP computation, we see that the temporal resolution varies. An example is

shown in Figure 3.18.

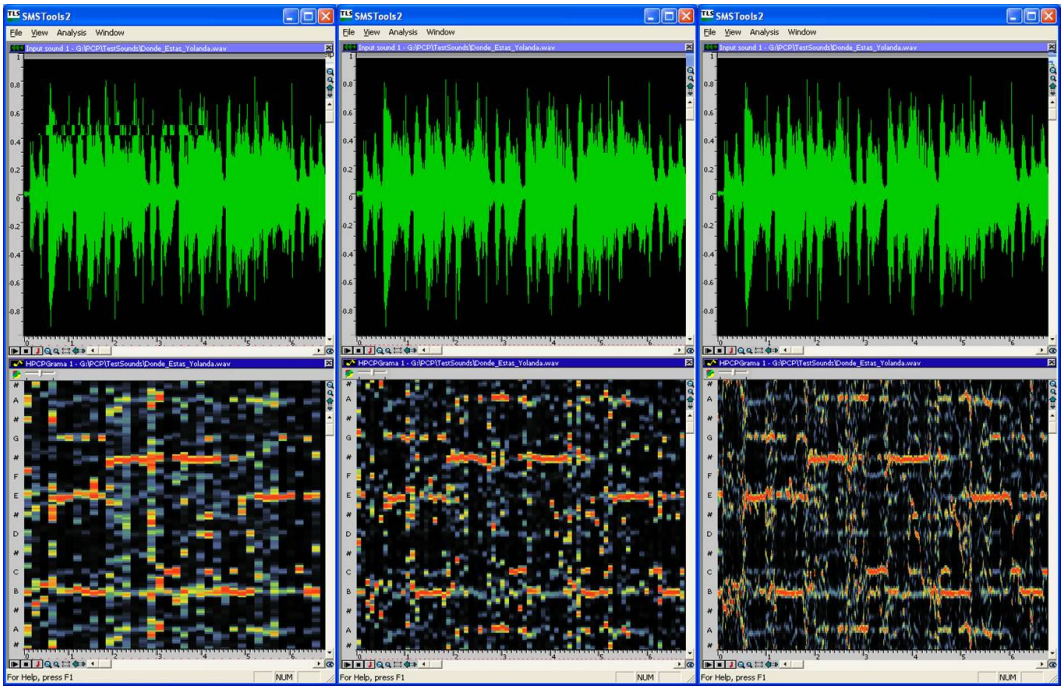


Figure 3.18: Sound file analyzed using different temporal resolution: 5.38 frames/second (left), 10.76 frames/second (middle) and 43.06 frames/second (right). See *Donde-Estas-Yolanda.wav* in Appendix A for further details on the sound excerpt.

As the temporal resolution increases, it is easier to distinguish transitions and fast temporal variations (e.g. arpeggios). On the other hand, the quantity of data and the computation cost increases proportionally to the frame rate used for analysis.

3.8.2.3 Frequency band

The frequency band considered for analysis is also an important parameter, which is determined by the frequency range of the considered spectral peaks. This frequency band affects the range of the selected frequencies and can be used to focus on certain instruments (as leading voice, bass line, etc), in a way similar as proposed by Goto (1999). Figure 3.19 shows and example on the features computed in low frequencies (from 32 to 261 Hz) for the same excerpt of popular music which was represented in Figure 3.16.

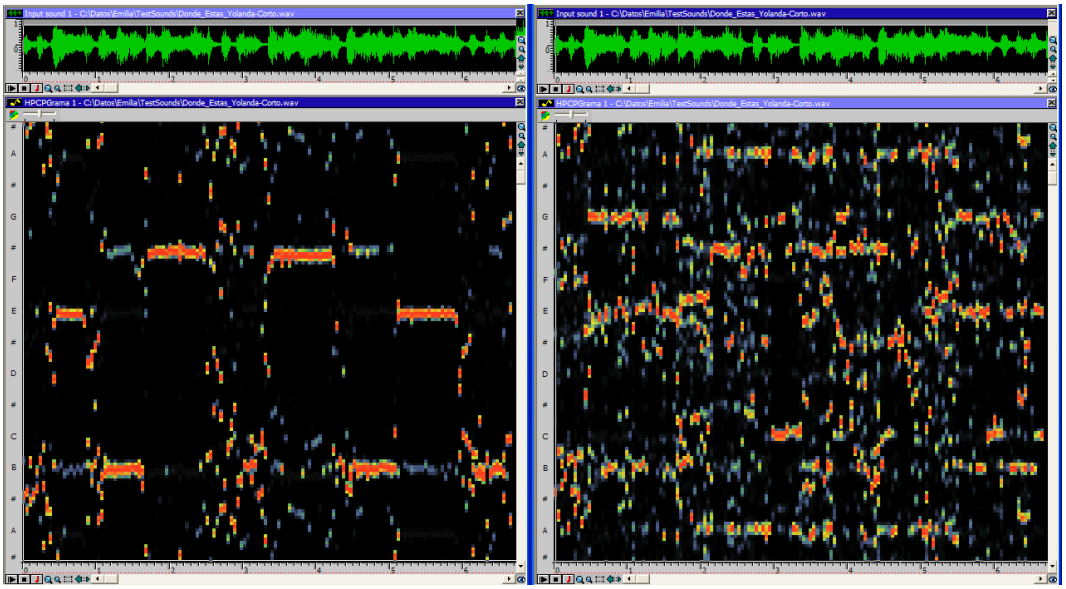


Figure 3.19: Sound file analyzed using different frequency bands: from 32.7 to 162.6 Hz corresponding to low frequencies (left) and mid and high frequencies (162.6 to 5000 Hz) (right). See *Donde-Estas-Yolanda.wav* in Appendix A for further details on the sound excerpt.

3.8.3 Robustness

3.8.3.1 Robustness to noise

Pitch class distribution features should also be robust to noise. This is the case for white noise, whose energy is equally distributed in the frequency range. Figure 3.20 represents the HPCP of an excerpt of white noise. We observe that the energy is equally distributed over the HPCP values, so that this type of noise will not affect higher level tonal descriptors.

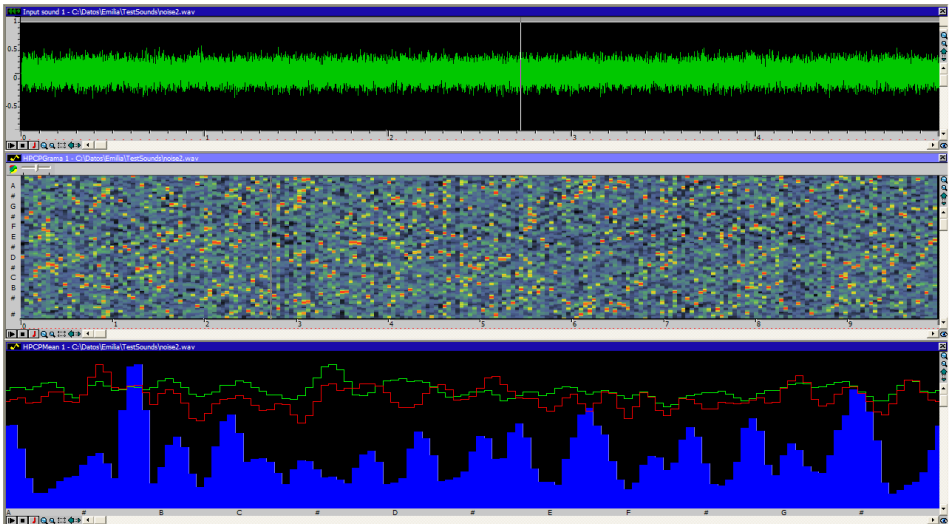


Figure 3.20: HPCP of an excerpt of white noise. The bottom panel shows the instantaneous HPCP (filled line), its average over the whole excerpt (green line) and its average over a sliding window of 1 second duration (red line). See *Noise.wav* in Appendix A for further details on the sound excerpt.

We will also show in Section 4.2.3 that the HPCP features obtained for a white noise or a percussive sound are not correlated to a particular tonal profile, as it is the case for tonal music.

3.8.3.2 Robustness to dynamics

In order to analyze what happens when there are some changes in dynamics, we present an example on the effect in the HPCP of applying a volume envelope (represented in Figure 3.21) to an audio signal.

Figure 3.22 shows the influence of this dynamic envelope into the HPCP features. We compare the HPCP features of two different excerpts. In the second one (*ImagineVolumeEnvelope.wav*), the volume envelope shown in Figure 3.21 has been applied to the original excerpt (*Imagine.wav*). Figure 3.23 displays the error between the HPCP values, as expressed in the following equation:

$$err(i,j) = abs(HPCP_1(i,j) - HPCP_2(i,j)) \tag{3.39}$$

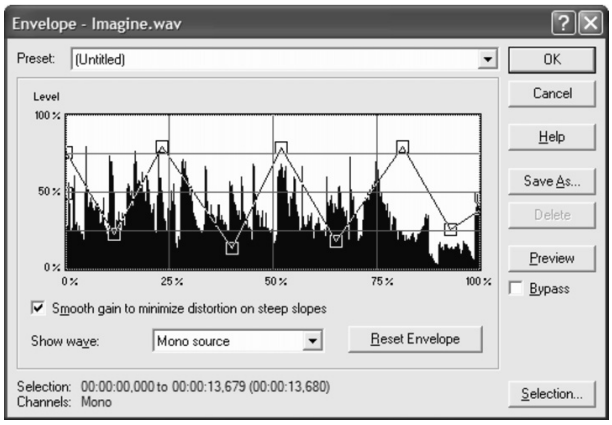


Figure 3.21: Volume envelope.

where $HPCP_1$ represents the HPCP values computed from the original audio excerpt and $HPCP_2$ represents the values of the excerpt after applying a volume envelope. These HPCP vectors were normalized with respect to its maximum value, as expressed in Equation 3.36, so that its maximum value is equal to 1. The index i indicates the frame index, with $i = 1 \ldots nFrames$, while the index j represents the HPCP bin, with $j = 1 \ldots size$. An example of the error $e(i, j)$ for this audio excerpt is shown in Figure 3.23. We can compute the average absolute error as follows:

$$averr = \frac{1}{nFrames \cdot size} \sum_{i=1}^{nFrames} \sum_{j=1}^{size} err(i, j) \tag{3.40}$$

We obtain a value of $averr = 0.001$. As a conclusion, the normalization process presented in Section 3.5.1 makes the features not to be influenced by this envelope. The small errors might be due to the appearance of new spectral peaks or the elimination of existing spectral peaks as the volume changes. It is also influenced by the spectral peak magnitude threshold parameter of the peak detection method presented in Section 3.2.4.

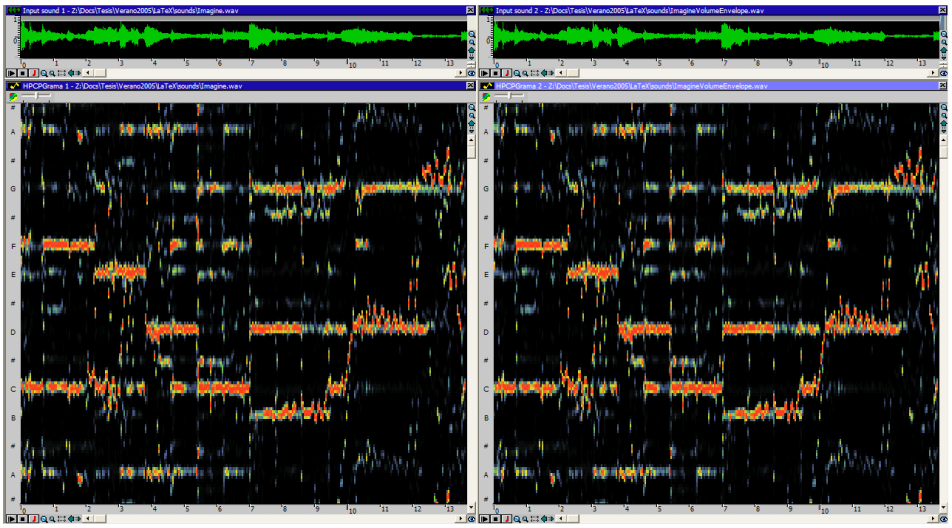


Figure 3.22: Influence of a volume envelope on the HPCP features. Audio before (left) and after (right) applying a volume envelope (see *Imagine.wav* and *ImagineVolumeEnvelope.wav* in Appendix A).

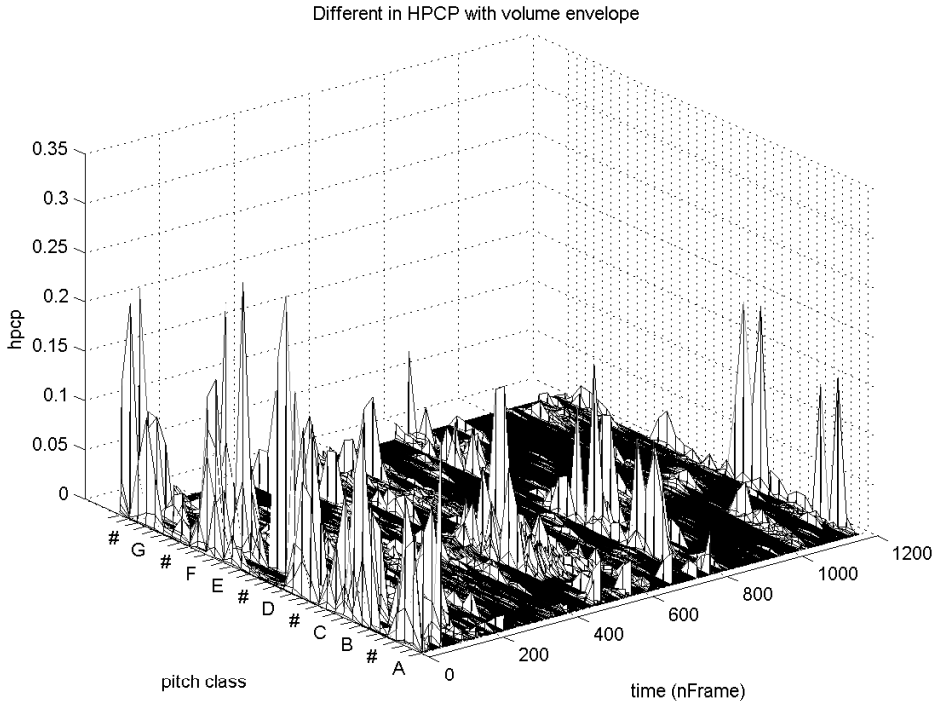


Figure 3.23: Influence of a volume envelope on the HPCP features. Absolute value of the different between the HPCP values (see *Imagine.wav* and *ImagineVolumeEnvelope.wav* in Appendix A).

3.8.3.3 Robustness to timbre

In order to verify the robustness of these low-level features to the played instrument, we could visualize the descriptors of two versions of the same piece played by different performers and instruments, as proposed in Purwins (2005). This comparison is also related to the robustness to dynamics, noise and post-processes, as the different versions have some dissimilarities in dynamics (expression) and recording conditions. Figure 3.24 represents the HPCP vector ($size = 36$) for three different versions of the first Prelude in C Major of Bach's Well-Tempered Clavier (WTC), the two first ones played by a harpsichord and the third one played by a piano. Figure 3.25 represents the HPCP vector ($size = 36$) for three different versions of the second Prelude in C Minor of Bach's WTC, the two first ones played by a harpsichord and the third one played by a piano. These figures show the similarity between the pieces.

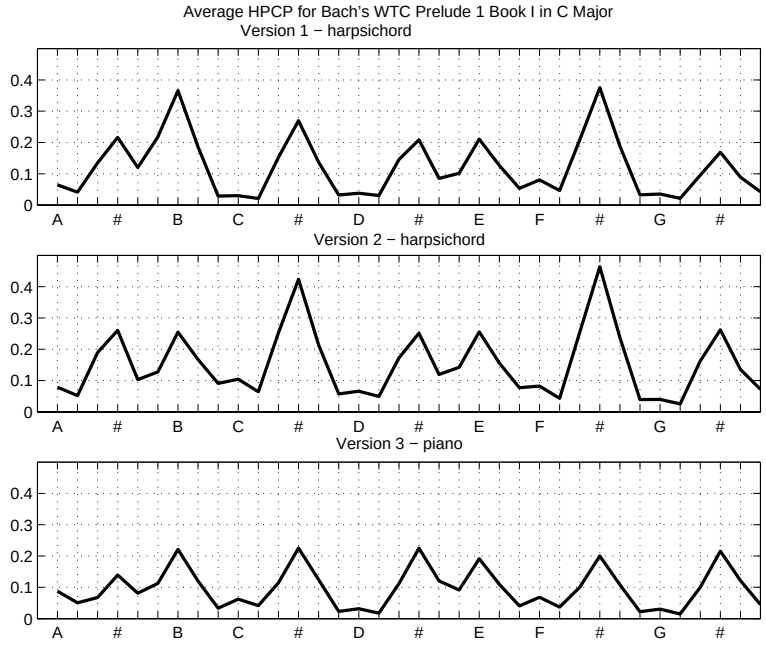


Figure 3.24: Influence of timbre on the HPCP features. Average HPCP for three versions of Bach's WTC First Prelude in C Major, the first two played by a harpsichord and the third one by a piano (see *Book-I-Prelude-1-C-Version1Jaccottet.wav, Book-I-Prelude-1-C-Version2Leonhardt.wav* and *Book-I-Prelude-1-C-Version3GlennGould.wav* in Appendix A).

The correlation coefficient r between two different HPCP vectors, $HPCP_1$ and $HPCP_2$, with expected values μ_{HPCP_1} and μ_{HPCP_2} and standard deviations σ_{HPCP_1} and σ_{HPCP_2}, is given by the following formula:

$$r = \frac{E[(HPCP_1 - \mu_{HPCP_1}) \cdot (HPCP_2 - \mu_{HPCP_2})]}{\sigma_{HPCP_1} \cdot \sigma_{HPCP_2}} \tag{3.41}$$

If we compute the correlation of the three versions for the 24 Fugues and Preludes, we obtain the results

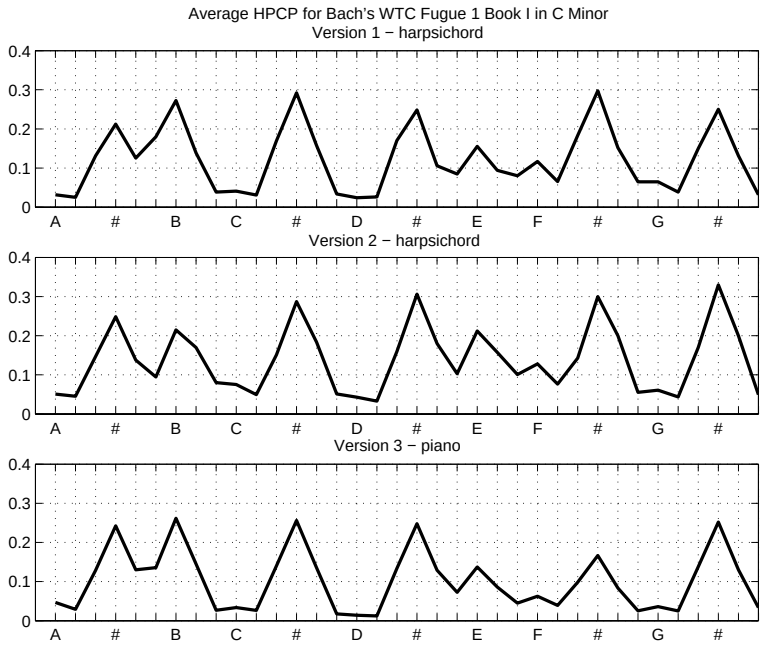


Figure 3.25: Influence of timbre on the HPCP features. Average HPCP for three versions of Bach's WTC Second Prelude in C Minor, the first two played by a harpsichord and the third one by a piano (see *Book-I-Prelude-2-c-Version1Jaccottet.wav, Book-I-Prelude-2-c-Version2Leonhardt.wav* and *Book-I-Prelude-2-c-Version3GlennGould.wav* in Appendix A).

presented in Figure 3.26. Table 3.3 shows the average and variance of these correlations. We can see that the correlation is around 0.9 for the different versions. Although the difference of the performances is restricted to a change of instruments, these results shows the robustness of these features within this particular situation. We will further study in Chapter 5 how these profiles correlate when the difference in timbre and performance is higher, as it is the case for different versions of popular music not strictly following a score.

Versions	**Average correlation**	**Variance of correlation**
Preludes 1 (harpsichord) vs 2 (harpsichord)	0.912	0.001
Fugues 1 (harpsichord) vs 2 (harpsichord)	0.921	0.001
Preludes 1 (harpsichord) vs 3 (piano)	0.852	0.005
Fugues 1 (harpsichord) vs 3 (piano)	0.921	0.001
Preludes 2 (harpsichord) vs 3 (piano)	0.822	0.005
Fugues 2 (harpsichord) vs 3 (piano)	0.91	0.0016

Table 3.3: Statistics of correlation coefficient of the average HPCP vector for three different versions of Bach's Well-Tempered Clavier. The first and the second one are played by a harpsichord and the third one is played by a piano.

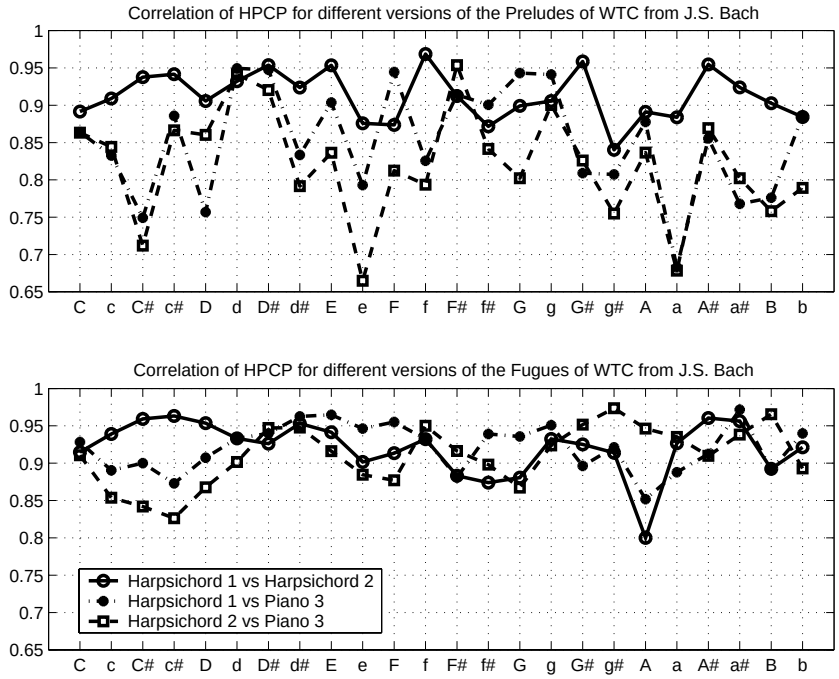


Figure 3.26: Correlation of HPCP average vector for three different versions of the 24 Preludes and Fugues of Bach's Well-Tempered Clavier. The first two played by a harpsichord and the third one by a piano. The x-axis represents the piece index (indicated with the key of the piece, where small caps represent minor mode).

3.8.3.4 Robustness to tuning

In order to evaluate the procedure for reference frequency estimation, we have built an audio database of 3854 files. This database is built using the first 30 seconds or 94 classical pieces in MIDI format and transposing them using 41 different intervals. The selected MIDI pieces correspond to the training set provided for symbolic key finding algorithms that participated to the ISMIR 2005 evaluation exchange [2]. This training set include MIDI representations of classical pieces from different instruments (piano, orchestra, etc).

We first apply to each of the pieces of the MIDI collection a transposition factor that varies from -100 to 100 cents, defining a detuning interval between consecutive pieces equal to 5 cents $dev = (-100, -95, -90, \ldots, 0, \ldots, 90, 95, 100)$. We obtain then a total of 41 MIDI files per piece. Each MIDI file is converted to audio (wav format 16 bits and 44.1 KHz sampling rate) using a software synthesizer[3].

The algorithm is run over the evaluation database in order to test if the estimated reference frequency (i.e. deviation with respect to $440Hz$) is equal to the applied deviation. We measure deviations between -50 and 50 cents. For instance, an applied frequency deviation of 100 cents correspond to a tuning deviation of 0 cents

[2] http://www.music-ir.org/mirexwiki
[3] Timidity++ free software synthesizer http://timidity.sourceforge.net

(i.e. one semitone higher with the same tuning frequency). Figure 3.27 represents the estimated deviation (in cents) with respect to the transposition factor applied to the MIDI file. We compare the use of different resolution to define the histograms for instantaneous and global estimation: $r = 1, 5$ or 10 cents.

This figure shows that the tuning frequency is correctly estimated. In order to compare the variation of performance when using different histogram resolution, some error measures are presented in Table 3.4. We provide the measure of % of correct estimation with different tolerance values (1, 5 and 10 cents), and we can conclude that similar results are obtained for the different values of histogram resolution.

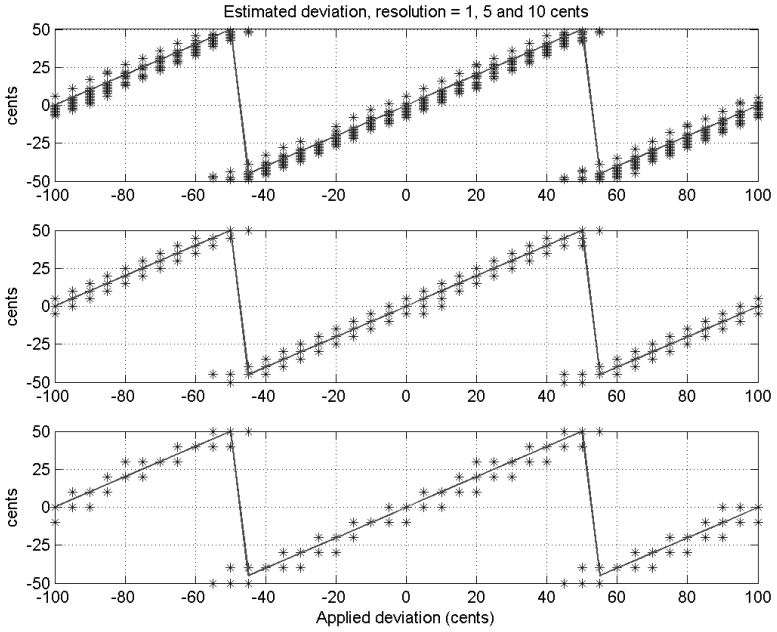


Figure 3.27: Estimated deviation (asterisks), expressed in cents, with respect to the applied one (filled line), for different histogram resolution: 1 cent (top), 5 cents (middle) and 10 cents (bottom). We observe that the tuning frequency is robustly estimated.

Res.	**% Correct**			**Abs. Error**	**Abs. Error**	**Square Error**
(cent)	1 cent	5 cents	10 cents	Mean (cent)	Var (cent)	Mean
1	58.32	81.49	99.97	2.19	6.02	10.82
5	76.4	99.87	100	1.19	4.59	6
10	49.26	98.04	100	2.63	7.21	14.16

Table 3.4: Accuracy measures used to evaluate the algorithm for tuning frequency estimation: % of correct estimation, average and variance of the absolute error and average of the square error. These statistics are computed for all the analyzed files. Results are presented when using 1, 5 or 10 cent resolution.

3.8.4 Monophonic vs polyphonic

As an another requirement, the HPCP should also be suitable to analyze monophonic signals. Figure 3.28 represents the HPCP for a piano phrase, where the played notes are identified. In the case of polyphonic signals, as in Figure 3.11, where we identify all the notes played by the different instruments.

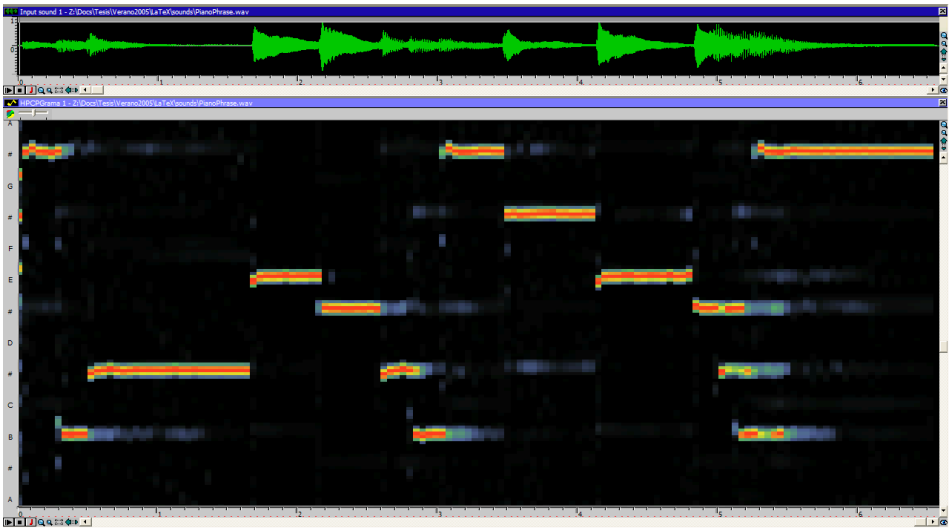


Figure 3.28: HPCP of a monophonic phrase from piano (see *PianoPhrase.wav* in Appendix A for further details on the sound excerpt).

3.8.5 Correspondence with pitch class distribution

In this Section, we analyze how the proposed HPCP features corresponds to pitch class distribution features computed from symbolic representation, and how these features are related to other chromagram implementations.

As a musical material, we choose the 24 fugues of Book II from Johann Sebastian Bach's Well-Tempered clavier (WTC). The audio versions are interpreted by Glenn Gould. We compare the automatically extracted pitch distribution features with the set of pitch histograms provided by the Scores from the Ohio State University Cognitive and Systematic Musicology Laboratory Huron (1994), denoted as Muse Data. We compare different pitch class distribution features obtained from audio (average over the whole piece): CQ-Profiles from Hendrik Purwins's PhD thesis, presented in Purwins (2005), the proposed HPCP and the original PCP presented by Fujishima (1999) in an implementation proposed by Sheh and Ellis (2003).

Figure 3.29 shows the individual results for each of the 24 analyzed pieces. The correlation coefficient is computed according to Equation 3.41. Figure 3.30 shows some details for high correlation values of these results. We can see that the correlation is high for all the implementations, although the best results are found for the proposed features.

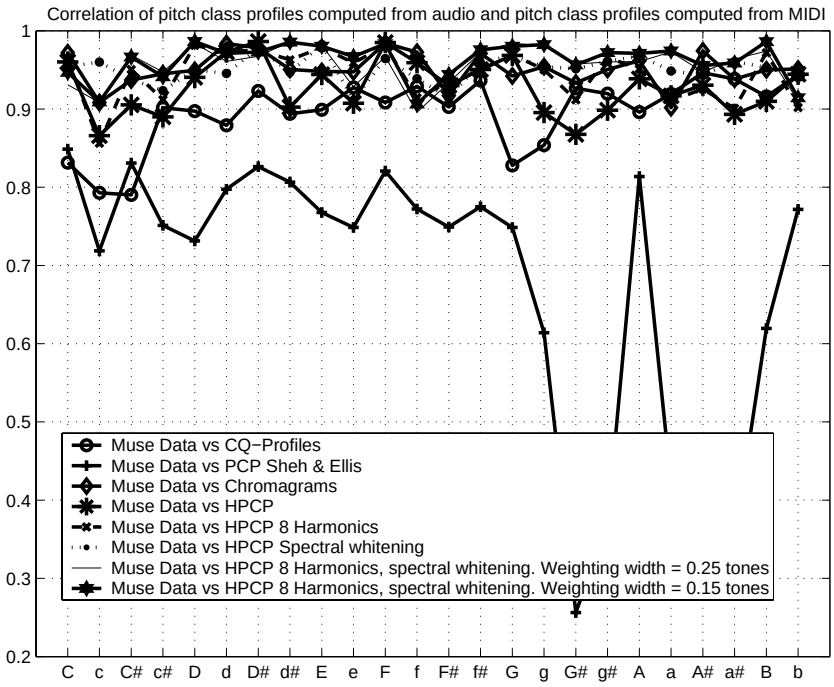


Figure 3.29: Correlation of different pitch class profiles computed from audio and pitch class profiles computed from expressive MIDI. The x-axis represents the piece index (indicated with the key of the piece, where small caps represent minor mode).

The overall results of the correlation between the different pitch class distribution features computed from audio and the ones computed from expressive symbolic notation are shown in Table 3.5. We can see that the results are worse for the original PCP, implemented by Sheh and Ellis (2003), and very different in some particular pieces (Fugues 17 and 18 in G# major and minor). The correlation for HPCP are higher than 0.9 for all the pieces and improves the results of the C-Q profiles. It is slightly better than the implementation of Chromagrams in Harte and Sandler (2005).

We also compare different HPCP implementations: the first one do not consider the presence of harmonic frequencies ; the second ones includes 8 harmonic frequencies ($nHarmonics = 8$) with a fixed decreasing exponential contribution, as explained in Section 3.4.2; finally the third one incorporates a previous step where we perform spectral whitening, normalizing the spectrum with respect to the spectral envelop in order to convert it into a flat spectrum. In this way, the instrument timbre does not affect the final result.

Increasing the number of harmonics does not improve the results. Using this spectral whitening procedure, notes at high octaves contribute equally to the final HPCP vector than those one on low pitch range, and the results are not influenced by different equalization procedures. Finally, we can slightly improve the performance by diminishing the window size used for weighting l. Eliminating the weighting procedure decreases the results.

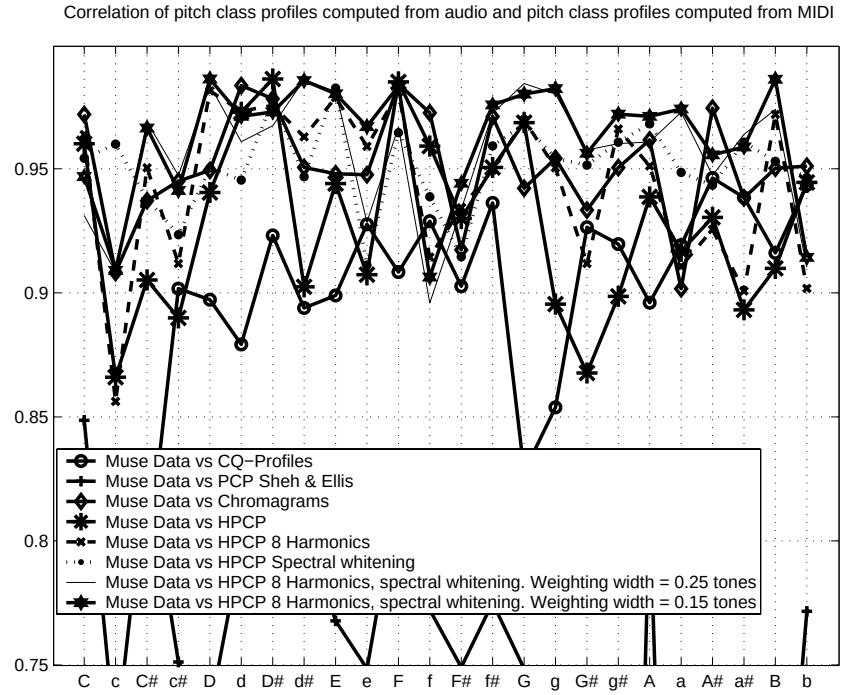


Figure 3.30: Details on the correlation of different pitch class profiles computed from audio and pitch class profiles computed from expressive MIDI. Only correlation values between 0.75 and 1 are shown. The x-axis represents the piece index (indicated with the key of the piece, where small caps represent minor mode).

We can see that the inclusion of harmonic frequencies make the HPCP to correlate better with pitch class distribution features extracted from symbolic representation, as well as the normalization step considering the spectral envelope.

3.9 Conclusions

In this chapter, we have proposed and evaluated a vector of low-level tonal features representing the pitch class distribution of a piece of music in audio format.

The following characteristics have been studied: representation of the pitch class distribution in monophonic and polyphonic recordings, consideration of harmonic frequencies, and robustness to dynamics, tuning and timbre. Some of these properties have been evaluated in a more quantitative way, such as the evaluation of the tuning frequency estimation method. We have also compared the current features with the results obtained using symbolic notation and other approaches for pitch class distribution features computation. We have justified the utility of all the steps for feature computation, and analyzed the influence of the different analysis parameters.

Method	**Average**	**Variance**
Muse Data vs CQ-Profiles Purwins (2005)	0.89583	0.0020176
Muse Data vs PCP Sheh and Ellis (2003)	0.68137	0.030971
Muse Data vs Chromagrams Harte and Sandler (2005)	0.95092	0.00048227
Muse Data vs HPCP not considering harmonic frequencies	0.92761	0.0012062
Muse Data vs HPCP $nHarmonics = 8$	0.94382	0.0010743
Muse Data vs HPCP including spectral whitening	0.95089	0.00030347
Muse Data vs HPCP $nHarmonics = 8$ and spectral whitening. $l = \frac{4}{3}$ semitones	0.95596	0.00065629
Muse Data vs HPCP $nHarmonics = 8$ and spectral whitening. $l = \frac{1}{2}$ semitones	0.96196	0.00057632

Table 3.5: Correlation values.

Although a further evaluation of the utility of these features for tonality estimation will be carried out in the next Chapter, we have shown that the HPCP features take into account all the requirements mentioned in Section 2.6.3 for pitch class distribution features.

Chapter 4

Tonality estimation: from low-level features to chords and keys

4.1 Introduction

In this chapter, we present our approach for tonality estimation using the pitch class distribution features that have been presented in Chapter 3. If we consider the schema presented in Figure 4.1, we tackle here the adaptation and similarity computation blocks of the diagram: how to adapt a tonality model to audio features and how to match this tonal model to the computed audio descriptors. Then, the overall schema for key estimation from audio is shown in Figure 4.1.

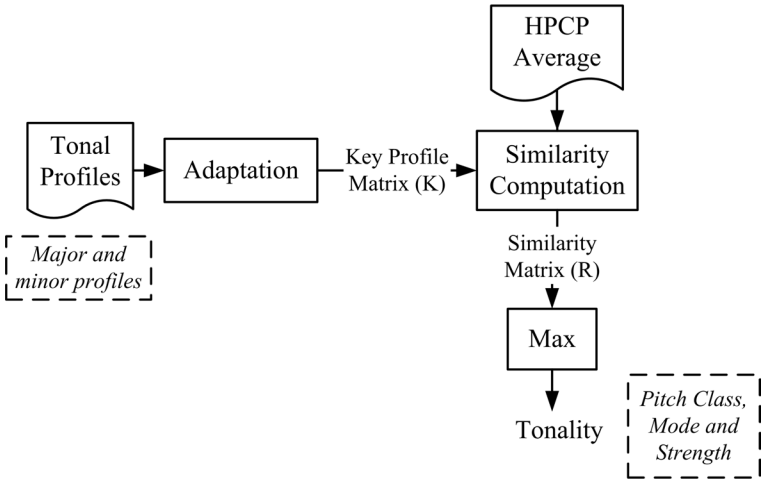


Figure 4.1: Block diagram for key computation using HPCP

We first explain how the tonal profile proposed for melodies is adapted to the proposed low-level features by considering melodies vs harmonies and the presence of harmonics. We perform a quantitative evaluation of the proposed approach for key estimation from polyphonic signals considering Different musical genres. We compare the use of different tonal profiles reviewed in Section 2.3 and the consideration of different segment duration of the analyzed piece. We finally compare the performance for different values of algorithm parameters, analyzing the influence of changing the distance measure, the interval resolution and the use of some preprocessing and postprocessing steps explained in Chapter 3.

We consider that when we estimate the key in a very short segment, where only one chord is played, we can estimate the played chord. Although it is not the main focus of this PhD thesis, we also present some evaluation results for chord estimation from polyphonic audio.

As an alternative to the use of tonal models applied to pitch class distribution features, we propose the use of some machine learning techniques, and we make a comparison of the improvement that these techniques provide in the context of the proposed tonal description system.

As mentioned in Section 2.4, a piece rarely maintains a single tone center through its entire duration. We have seen that there is not much research devoted to locate modulation and to perform an evaluation of methods for tonality tracking. We propose a sliding window method and a multiresolution approach for tonality tracking which will be evaluated in this Chapter.

The main questions that we try to answer in this chapter are related to the influence of different aspects on the performance of the proposed key finding method:

- Which is the influence of the different tonal models?
- Which is the influence of analyzing all the piece or only a fragment?
- Which is the influence of including a learning stage from labelled data?
- Which is the influence of the music genre and its relation with the tonal model under consideration?
- Can this approach for key estimation be extended to chord estimation?

In order to answer the mentioned questions, we perform the set of quantitative evaluation experiments mentioned above.

4.2 Adaptation of tonal models to audio

The goal of this section is to explain how the tonal profiles designed to work with MIDI representations are adapted to the proposed HPCP features, which corresponds to the adaptation block in Figure 4.1. In order to estimate the key of a piece we compute the similarity of the HPCP vector with a matrix of HPCP profiles corresponding to the different keys K. This matrix has three dimensions, so that its size is equal to $2 \times 12 \times size$ (i.e. 2 possible modes, 12 possible tonics, and $size$ represents the size of the HPCP vector).

In order to measure similarity we propose the use of a correlation function (see Section 2.9 for a comparison with alternative similarity measures). As it appears in Formula 4.1, we then obtain a correlation value $R(i,j)$ for each tonic:

$$R(i,j) = r(HPCP, K(i,j)) \tag{4.1}$$

$K(i,j)$ is the key profile. $i = 1, 2$, where 1 represents the major profile and 2 the minor profile. $j = 1...12$ for the 12 possible tonics. Both vectors ($HPCP$ and $K(i,j)$) have $size$ elements. The correlation between these two vectors $HPCP$ and $K(i,j)$, having expected values μ_{HPCP} and $\mu_{K(i,j)}$ and standard deviations σ_{HPCP} and $\sigma_{K(i,j)}$, is given by:

$$R(i,j) = r(HPCP, K(i,j)) = \frac{E[(HPCP - \mu_{HPCP}) \cdot (K(i,j) - \mu_{K(i,j)})]}{\sigma_{HPCP} \cdot \sigma_{K(i,j)}} \tag{4.2}$$

The maximum correlation value corresponds to the estimated tonic and mode (represented by the indexes i_{max} and j_{max}). We use the maximum correlation value $R(i_{max}, j_{max})$ as a measure of the *"tonalness"*, degree or tonality or key strength.

$$R(i_{max}, j_{max}) = \max_{i,j}(R(i,j)) \tag{4.3}$$

To construct the key profile matrix, we use a tonal profile approach based on Krumhansl (1990) and further developed by Temperley (1999). This collection of methods are used to estimate the key from MIDI representations. As reviewed in Section 2.3, this technique considers that tonal hierarchy may be acquired through internalizing the relative frequencies and durations with which tones are sounded. The original method estimates the key from a set of note duration values, extracted from MIDI, measuring how long each of the 12 pitch classes of an octave (C, C#, etc) have been played in a melodic line. In order to estimate the key for a melodic line, the vector of note durations is correlated with a set of key profiles or probe tone profiles. These profiles represent the tonal hierarchies of the 24 major and minor keys. Each of them contains 12 values, which are the ratings of the degree to which each of the 12 chromatic scale tones fit a particular key. The original profiles were obtained by analyzing human judgements with regard to the relationship between pitch classes and keys, as explained in Krumhansl (1990), pp. 78-81.

We have shown in Chapter 3 that the HPCP value for a given pitch class $HPCP(i)$ represents the relative intensity of this pitch class for the audio excerpt in which it is computed. In the same way than the relative duration of each pitch class, we assume that this intensity is also related to the relative importance of the pitch class within the established tonal context.

Then, in our work we adapt the original technique in two aspects: first to work with HPCP instead of note duration values, and second to consider polyphony instead of melodic lines. In a polyphonic situation, the notes belonging to the same chord sound at the same time.

Let us consider $T_M(i)$ and $T_m(i)$ as the original profiles for major (represented by the subscript M) and minor (represented by the subscript m) scales, where $i = 1...12$. These profiles represent the strength of the

pitch class i in a given major or minor key. The index i corresponds to the pitch class that is $i-1$ semitones higher than the tonic. Then, in a given major key, $i=1$ will correspond to the tonic (that is the maximum value), $i=8$ to the dominant, etc.

As reviewed in Section 2.3, several major and minor profiles are found in the literature: flat profiles are represented in Figure 2.7, the probe tones proposed by Krumhansl (1990) are shown in Figure 2.12 in Section 2.3 and some modifications proposed later in Temperley (1999) are represented in Figure 2.13. Some modifications are applied to these profiles, which are explained in the following sections, in order to obtain the applied profiles T_{Mp} and T_{mp}.

Finally, the profile matrix for the 24 different keys, K (equation 4.1), is built as follows. We consider that the tonal hierarchy (related to major and minor keys) is invariant with respect to the chosen tonic. This means, for instance that the profile for B major will be the same as A major but shifted two bins (corresponding to the two semitones between A and B). We perform a linear interpolation between bins to compute $size$ bins from the original 12 values in $T_{T_{Mp}}$ and $T_{T_{mp}}$.

As explained above (and shown in Figure 4.1), a correlation value is computed between the global HPCP value and each of the 24 key profiles. The one with the highest correlation value is chosen as the estimated tonality.

4.2.1 Melodies vs chords

Many of the proposed key finding approaches have been defined considering unaccompanied melodies. We have seen in Chapter 2 that one of the most popular collections for evaluating key estimation algorithms is formed by the Fugue subjects from J.S. Bach's Well-Tempered Clavier. In fact, one may find excellent and almost arbitrarily complex examples of unaccompanied melodies in composers such as J.S Bach, as mentioned in Auhagen and Vos (2000). When dealing with harmonized music, it seems necessary to define which sequences of single notes in the melodies carries the intended key. We claim in this study that the models defined for melodies can be exploited for the analysis of chord sequences, as it was also found in Krumhansl (1990).

The tonal hierarchy within a melodic line should be maintained in a polyphonic situation, where the melodic line is converted into a chord sequence. In this way, we can consider that T_M and T_m represent the strength of the chord i (tonic, subdominant, dominant chord, etc.) in a given key. Given this assumption, we should consider all the chords containing a given pitch class when measuring the relevance of this pitch class within a certain key. For instance, the dominant pitch class ($i=8$) appears in both tonic ($i=1$) and dominant ($i=8$) chords (among others), so that the profile value of the dominant pitch class will add to the contribution of the tonic and the dominant chords of the key. This can be expressed by the equation $T_{Mp}(8) = T_M(1) + T_M(8) + \ldots$, where p stands for *polyphonic*.

We compute the weight of a given pitch class in a given major key $T_{Mp}(i)$, as the weighted sum of the contribution of each of the chords (of a major key) this pitch class belongs to:

$$T_{Mp}(i) = \sum_{j=1}^{12} \alpha_M(i,j) \cdot T_M(j) \;\; i = 1...12 \tag{4.4}$$

In the same way, we compute the weight of a given pitch class in a given minor key $T_{mp}(i)$, as the weighted sum of the contributions of the chords (of a minor key) this pitch class belongs to:

$$T_{mp}(i) = \sum_{j=1}^{12} \alpha_m(i,j) \cdot T_m(j) \;\; i = 1...12 \tag{4.5}$$

These equations can be also expressed as follows:

$$\begin{aligned} T_{Mp} &= \alpha_M \cdot T_M^t \\ T_{mp} &= \alpha_m \cdot T_m^t \end{aligned} \tag{4.6}$$

where T_{Mp}, T_M, T_{mp} and T_m are vectors 1×12, and α_M, α_m are square matrices 12×12. Within a major key, $\alpha_M(i,j)$ represents the weight of the pitch class i when considering the chord having pitch class j as a root. The matrix $\alpha_m(i,j)$ represents the weight of the pitch class i when considering the chord whose root is pitch class j within a minor key. After performing some comparative tests (see Section 4.3), we only consider the three main triads of the key as the most representative ones: tonic ($i = 1$) dominant ($i = 8$) and subdominant ($i = 6$), defining the following matrices for major and minor keys:

$$\alpha_M = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \end{pmatrix} \tag{4.7}$$

$$\alpha_m = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \end{pmatrix} \tag{4.8}$$

4.2.2 Fundamental frequency and harmonics

As explained in Section 2.2.1 and Section 3.4.2, the spectrum of a note is composed of several harmonics, whose frequencies are multiples of the fundamental frequency (f, $2 \cdot f$, $3 \cdot f$, $4 \cdot f$, etc.). When a note is played, we observe that the spectral magnitude increases for these frequency values. This fact appears on the HPCP values, as we have discussed in Chapter 3. If we consider a scale in equal temperament, the index i_n associated to the n^{th} harmonic of a note (we can call it the n^{th} *harmonic pitch class*) can be computed according to Equation 3.34.

The presence of harmonics can be considered either when computing the HPCP vector (as explained in Section 3.4.2) or in the definition of the tonal profiles. The procedure is similar in both cases.

The profile values for a given pitch class i ($T_{Mp}(i), T_{mp}(i)$) are equal to the sum of contributions of all the pitch classes containing i as harmonic pitch class. That means that each note of the considered chords (associated to each of the "1" values within α_M and α_m matrices) contribute to the profile values of its i_n harmonics (n=1,2,..., nHarmonics). We make the contribution decrease along frequency using a decreasing function $f(n) = s^{n-1}$ with a certain slope s, in order to simulate that the spectrum amplitude decreases with frequency. This is illustrated in Figure 3.10 and Table 3.2. The spectral decay factor s has been empirically set to 0.6, and we have considered the first four harmonics for computation ($nHarmonics = 8$). If a harmonic is located between two different pitch classes, we use the same weighting scheme as for HPCP computation.

After considering a polyphonic situation and the presence of harmonic frequencies, we obtain the final profiles T_{Mp} and T_{mp}. The final profiles T_{Mp} and T_{mp} obtained from the original Kruhmansl profiles (Figure 2.12) are shown in Figure 4.2. We observe that the dominant has the maximum value both in major and minor mode, followed by the tonic.

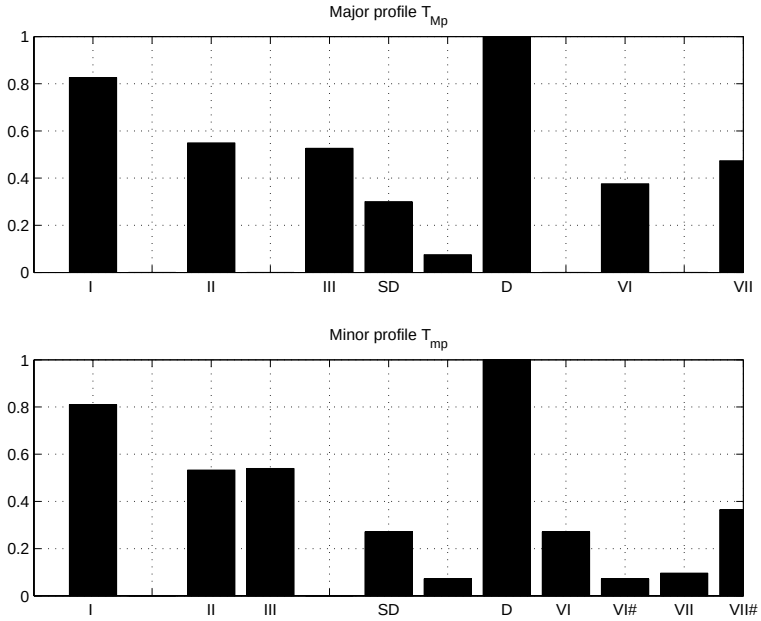


Figure 4.2: T_{Mp} and T_{mp} profiles adapted from Krumhanl's profiles T_M and T_m

4.2.3 Case study

In this section, we present some examples of how the algorithm works. Figure 4.3 shows the results of the analysis of a polyphonic excerpt in C major. Average HPCP values, as well as correlation with major and minor profiles (from Figure 4.2), are shown. In the HPCP profile, we can identify peaks in the tonic C and dominant G notes, as well as in the major third E. The correlation shows a maximum value 0.84 in C major. Other peaks appear at the key located a 5th above within the circle of fifths (G major and its relative minor E minor), as well as in C minor, which shares the dominant and subdominant chords (considering harmonic minor scale). In both examples we can check that the features represent some relationships between close keys.

Figure 4.4 shows the results of the analysis of a polyphonic excerpt in F# minor. As well as for last example, mean HPCP values, as well as correlation with major and minor models, are shown. In the HPCP profile, we can identify peaks in the tonic F# and dominant C# notes, as well as on the minor third A. The correlation shows a maximum value 0.79 in F# minor. Other peaks appear a 5th below within the circle of fifths (which is D major and its relative minor B minor), as well as in F# major, which shares the dominant and subdominant chord (considering harmonic minor scale).

Figure 4.5 shows the results of the analysis for a polyphonic percussion excerpt, where we cannot identify any key. In the HPCP profile there is no clear peak, while the correlation peaks stay around 0.5. In this case, the low correlation values indicate that the key is not well defined.

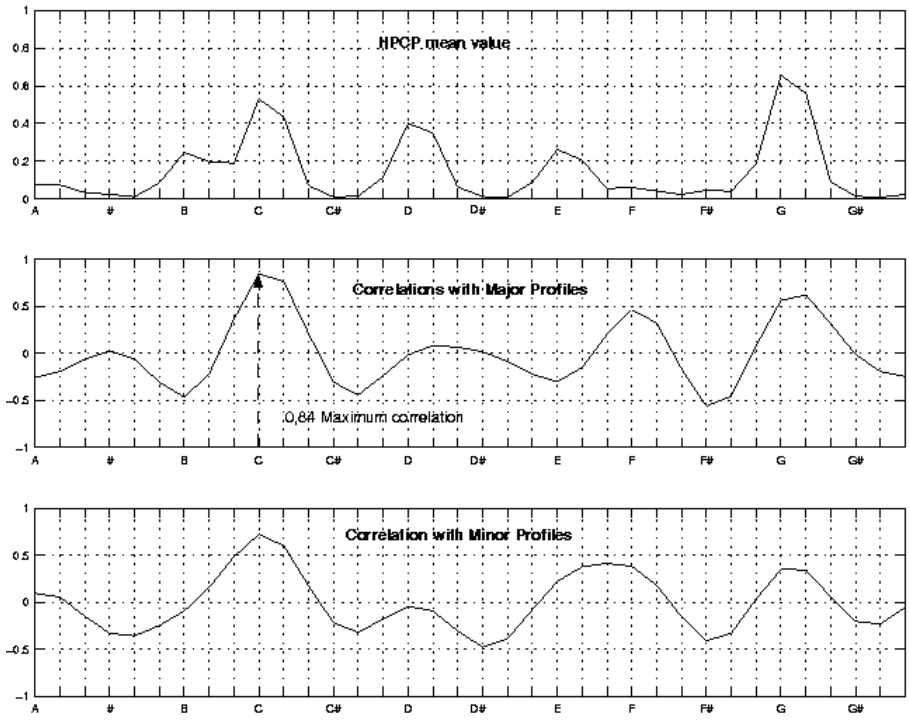


Figure 4.3: Example of HPCP profile and key correlation for an audio excerpt in C major key. We refer to *Organ-C-Major.wav* in Appendix A for further details on the audio excerpt. Lowest values of HPCP correspond to the tonic and the dominant degrees.

Figure 4.6 shows the results of the analysis for a polyphonic excerpt from String Quartets Op. 30 I Moderato, from Arnold Schoenberg, which is an example of atonal music. We observe that it is not possible to define which is the key of the fragment. In the HPCP profile there is no clear peak, while the peaks on the correlation with major and minor tonalities stay below 0.32. In this case, the low correlation values indicate that the key is not well defined. As a future work, we intend to study if it is possible to experimentally define a threshold for deciding atonality.

4.3 Evaluation of the use of tonal models and definition of a global key

In this section, we present a quantitative evaluation of the proposed approach based on the adaptation of a tonal model to perform global tonality estimation. We first compare the use of different tonal models, studying the dependency with the analyzed musical genre. Then, we analyze where the main tonality is established in a musical piece. We finally verify that the proposed approach can be extended to chord estimation.

Our starting point to set a baseline for key estimation is based on the finding that university music majors are able to sing, though not name, the scale of the key in which a piece is written correctly in 75% of the

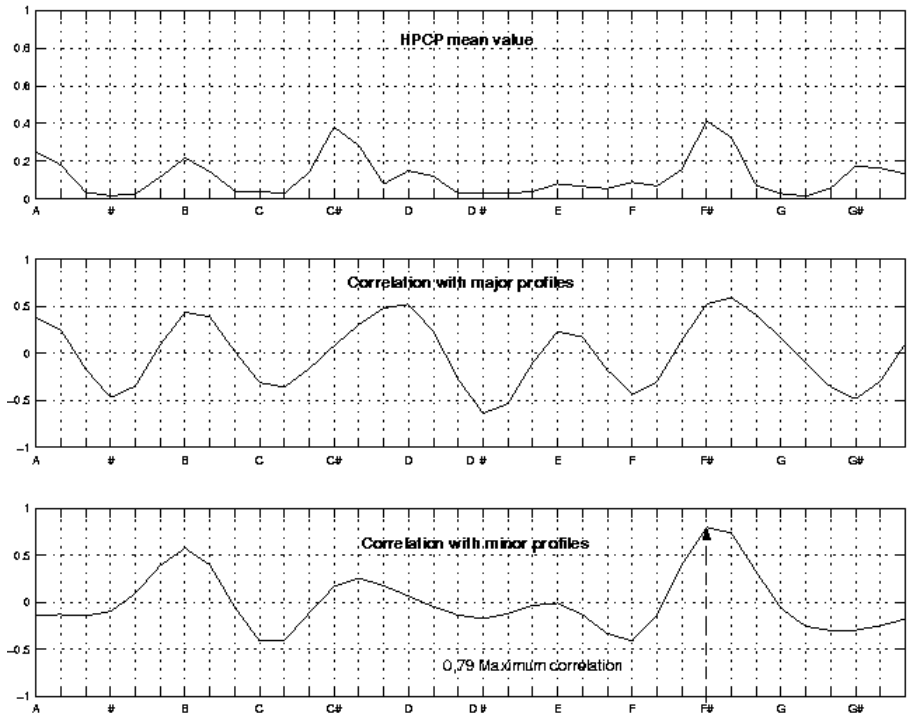


Figure 4.4: Example of HPCP profile and key correlation for a F# minor key audio excerpt. See *Piano-F-sharp-minor.wav* in Appendix A for further details on the audio excerpt.

times after hearing the first measure (Cohen (1977), cited by Krumhansl (1990) and Pauws (2004)).

4.3.1 Evaluation strategy

4.3.1.1 Evaluation material

In order to evaluate the proposed approach, we have built a music collection that includes four different sub-collections with various difficulty levels and musical styles:

1. **WTC**: Fugue subjects of the Well-Tempered clavier by J.S. Bach played by two different performers. A total of 76 audio files. This first collection has been chosen because it always appears in the literature as a standard collection to evaluate tonal description algorithms. The timbre complexity of this collection is very low, as the pieces are played by a single instrument.

2. **Contest**: Training data of the second ISMIR contest (2005), made by synthesizing the first 30 second of MIDI classical pieces. A total of 96 audio files. This collection has been chosen in order to represent a direct conversion from MIDI to audio. This collection also presents a low difficulty when considering timbre and recording conditions.

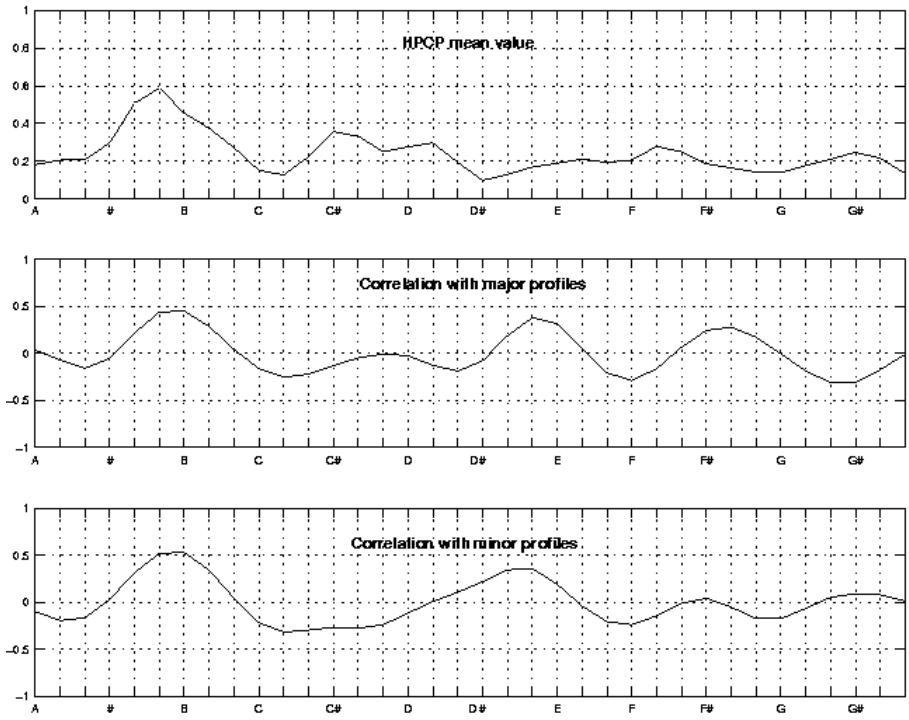


Figure 4.5: Example of HPCP profile and key correlation for a percussive sound (no clear tonality). See *Percussion.wav* in Appendix A.

3. **Classical DB**: 902 classical music pieces segmented by track and labelled by title (e.g., *Mozart, Flute Concerto No 1 K313 G Major Andante non troppo*). They include composers such a Mozart, Chopin, Scarlatti, Bach, Brahms, Beethoven, Handel, Pachelbel, Tchaikovsky, Sibelius, Dvorak, Debussy, Telemann, Albinoni, Vivaldi, Pasquini, Glenn Gould, Rachmaninoff, Schubert, Shostakovich, Haydn, Benedetto, Elgar, Bizet, Listz, Boccherini, Ravel, Debussy, etc. We also included some jazz versions of classical pieces (e.g. Jacques Lousier, The Swingle Singers). Most of the included pieces were first movements (in the case that the piece is a multi-movement form such as a sonata or symphony). All the tonic and mode annotations were taken from the FreeDB database [1]. Some additional manual corrections were done because of incorrect FreeDB meta-data, although systematic checking was not performed. We assumed that the key is constant within the whole piece. That means that the modulations we find do not modify the overall tone center of the piece. This collection is quite complex if we look at the instrumentation: some orchestra works, piano performances, and different instrumentation, as well as recording conditions. As we also include jazz versions of classical pieces, we include some additional jazz instrumentation.

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

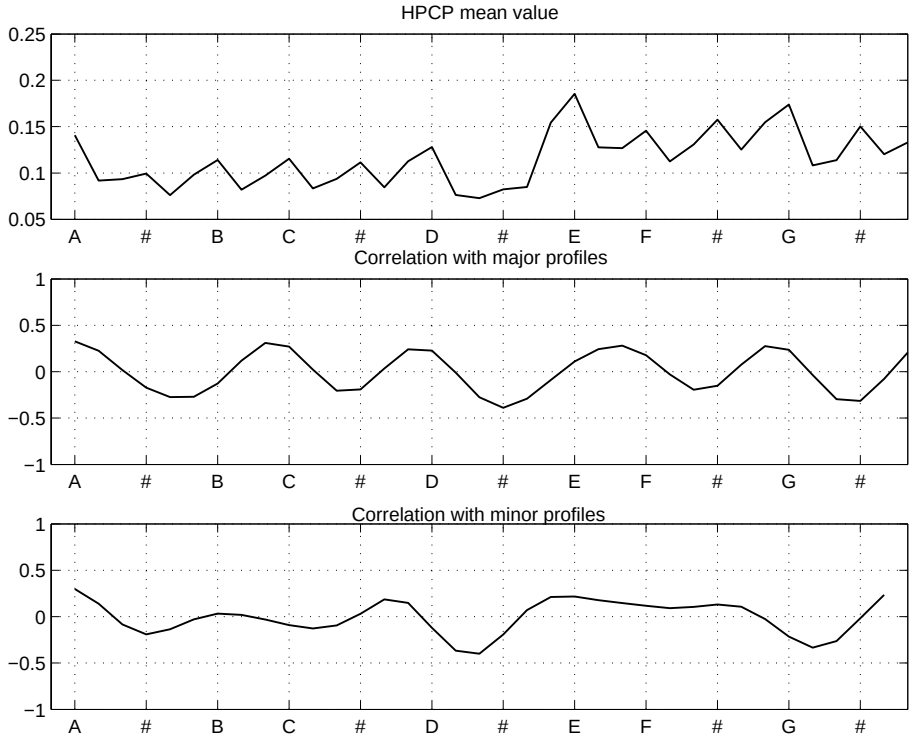


Figure 4.6: Example of HPCP profile and key correlation for an excerpt of String Quartet Op. 30 I Moderato, from Arnold Schoenberg. See *Schoenberg.wav* in Appendix A for details on the audio excerpt

4. **The Beatles collection**: 179 songs from the Beatles. The key of the pieces where taken from Pollack (1999) and cross annotated by a musicologist. In this collection the complexity increases, as there appear percussive sounds and different post-production effects. We evaluate in this collection the performance of different tonal profiles when considering popular music.

5. **Magnatune**: first minute of 108 pieces of different musical genres from Magnatune [2], annotated by an expert musicologist. The proportion of different musical genres for this collection is shown in Figure 4.7. Here, we introduce different musical genres in order to analyze the performance of the tonal model for the different genres. The difficulty of increasing the number of items of this last collection is the unavailability of manual key annotations.

6. **MIREX2005**: we also introduce here the collection used to compare our approach with existing ones. This comparison took place in the context of ISMIR 2005, Audio and Symbolic Key task of MIREX [3]. This collection consisted on the first 30 seconds of 1.252 audio files synthesized from MIDI. Two databases were used in order to have a certain timbre variation: Winamp synthesized audio (w) and

[2] Mp3 music and music licensing, open music record label. http://www.magnatune.com
[3] http://www.music-ir.org/evaluation/mirex-results/audiokey/index.html

Timidity with Fusion soundfonts synthesized audio (t). Each database is approximately 3.1 gigabytes for a total of 6.2 gigabytes of audio files. Unfortunately, this collection was not available for further experiments and statistics.

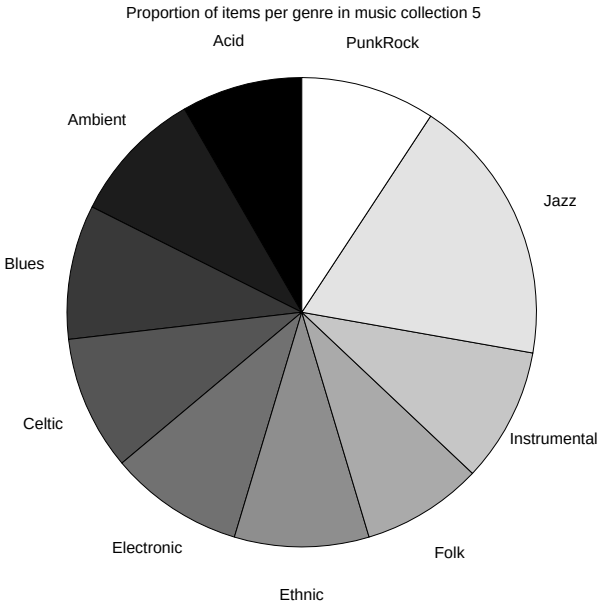


Figure 4.7: Proportion of musical genres for the magnatune collection to evaluate the behavior of the proposed tonality estimation method.

The overall proportion of modes and key is given in Figure 4.8 and 4.9. We see that the majority of the tonics are in the diatonic C major scale.

The proportion for each database is given in Figure 4.10 and 4.11. We observe in Figure 4.11 that in the Beatles database, most of the songs are in major tonality. For the rest of the collections the proportion is similar for minor and major keys.

4.3.1.2 Evaluation measures

In order to evaluate the approach for key estimation, different accuracy measures have been used. In addition to the percentage of accurate estimations, we have measured the relationship between the estimated key and the correct one, including weights for the different close tonalities (in *ContestMeasure*). We have also analyzed the distribution of the errors, in order to find which are the common ones for each of the studied tonal models.

- **ContestMeasure**: a relevant accuracy measure is the one proposed during the Audio and Symbolic Key contest for the evaluation of key estimation algorithms that took place during the ISMIR 2005

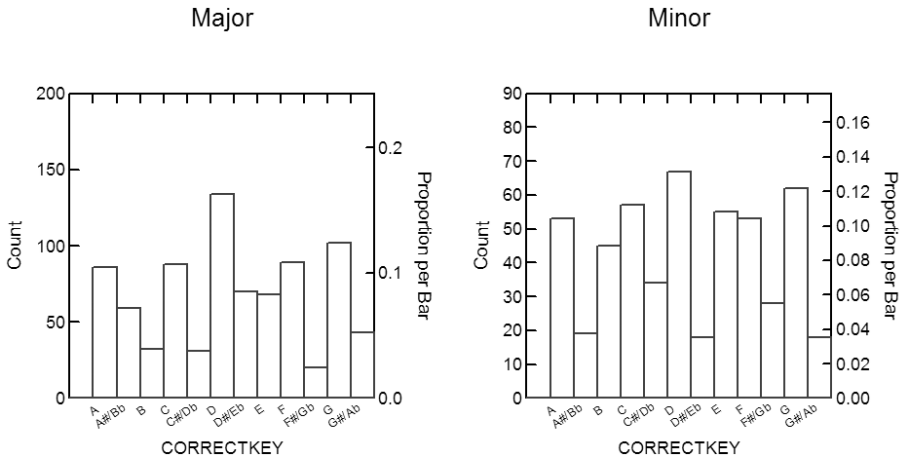


Figure 4.8: Proportion of key notes (or tonics) in the evaluation collection (excluding MIREX2005) for both major and minor modes.

conference [4]. In this evaluation measure, we take into account how close is the identified key to the correct one. Keys are considered as *close* if they have one of the following relationships: distance of fifth, relative minor and major and parallel (having the same tonic but different mode) (see Chapter 2 for further explanation). A correct key assignment is given a full score, but incorrect key assignments are allocated factors of the full score according to Table 4.1.

Relation to correct key	**Score (in %)**
Equal	100
Perfect fifth	50
Relative major/minor	30
Parallel major/minor	20

Table 4.1: Contest evaluation measure.

- **nCorrect**: percentage of correct key estimation as a measure of the algorithm accuracy.
- **nCorrectMode**: percentage of correct mode estimation as a measure of the algorithm accuracy for mode estimation (major and minor discrimination).
- **nTuningError**: percentage of tuning errors, measuring the confusion between a given key and a key which is one semitone higher or lower. This tuning error may be due to a different in the reference frequency used for tuning the instruments or to some detuning of harmonics.
- **nFifthError**: percentage of confusion with the key forming an interval of perfect fifth with the correct one.

[4] http://www.music-ir.org/mirexwiki

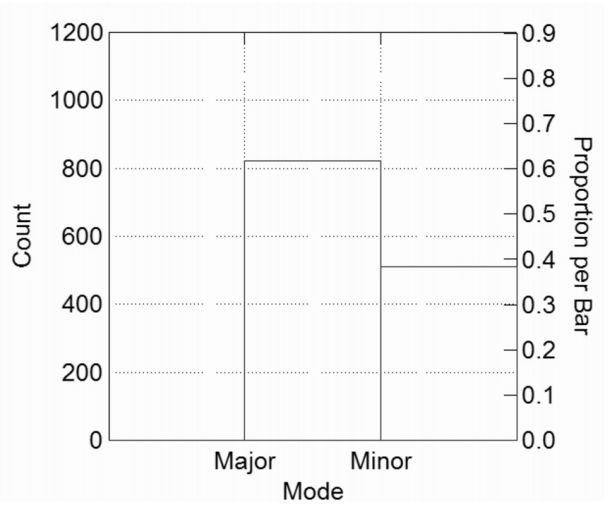


Figure 4.9: Proportion of modes in the evaluation collection (excluding MIREX2005).

- **nParallelError**: percentage of confusion with the parallel major/minor key.
- **nRelativeError**: percentage of confusion with the relative major/minor key.
- **nDistantError**: percentage of confusion with a distant key, which is not one of the other cases. This evaluation measure should be minimized.

4.3.2 Comparison with existing approaches for audio key finding

We first present some evaluation made in the context of the Audio and Symbolic Key contest for the evaluation of key estimation algorithms that took place during the ISMIR 2005 conference [5]. Here, we compare the proposed approach with a set of existing approaches. We use this evaluation information to improve our preliminary system as it will be shown in further sections of this chapter. The results for this approaches were obtained using 36 bins for the HPCP vector and the tonal major and minor profiles from David Temperley shown in Figure 2.13, as it is explained in Gómez (2005).

As explained in Section 4.3.1.1, the MIREX2005 evaluation database consisted on the first 30 seconds of 1.252 audio files synthesized from MIDI. Results are shown in Table 4.2.

[5] http://www.music-ir.org/evaluation/mirex-results/audiokey/index.html

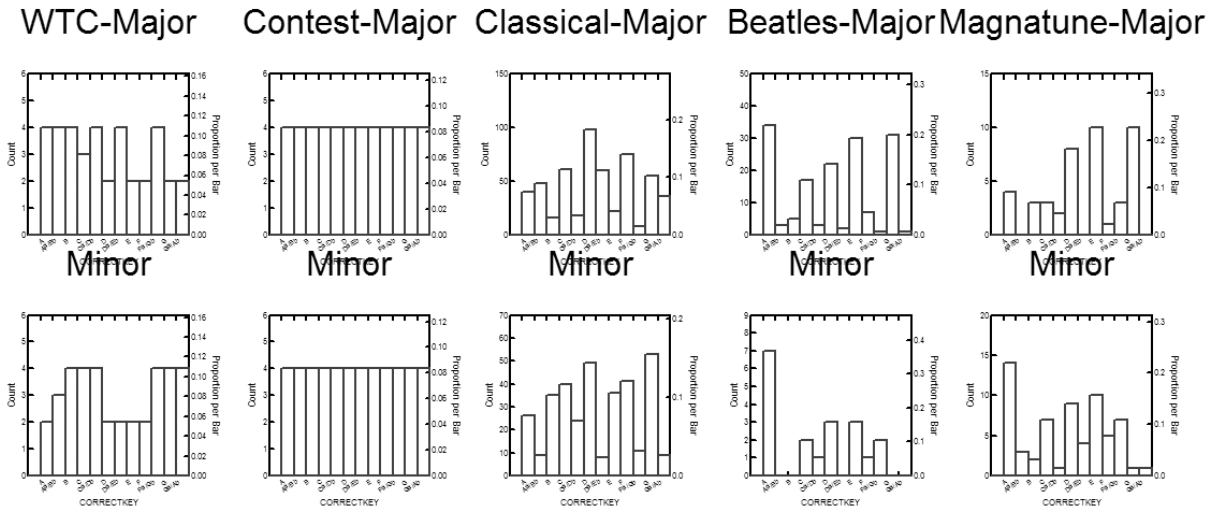


Figure 4.10: Proportion of key notes (tonics) for both major and minor modes in each of the music collections used for evaluation.

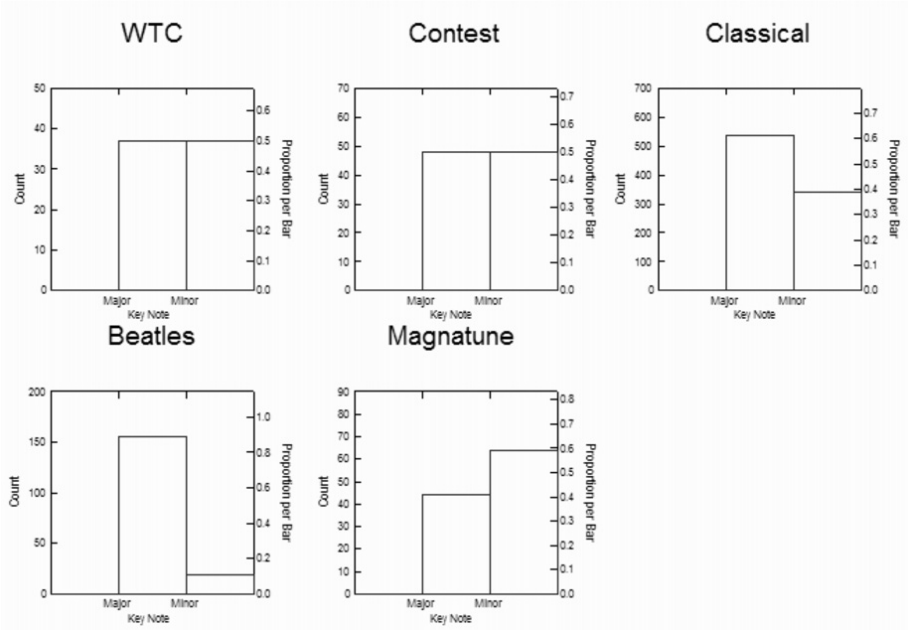


Figure 4.11: Proportion of modes in each of the music collections used for evaluation.

Rank	Participant	Composite % Score	Total Score		% Score		Correct keys		Perfect 5th Errors		Relative Errors		Parallel Errors		Other Errors		Runtime (s)	
			w	t	w	t	w	t	w	t	w	t	w	t	w	t	w	t
1	Izmirli	89.55	1188.8	1122.9	89.4	89.7	1086	1089	36	42	38	31	17	18	75	72	15284	16354
2	Purwins and Blankertz	89	1122.4	1106.5	89.6	88.4	1090	1060	44	72	24	21	16	21	78	78	45003	44232
3	Gómez, first 15 s	86.05	1081.9	1072.9	86.4	85.7	1048	1034	35	44	38	43	25	20	106	111	1560	1531
4	Gómez, all 30 s	85.9	1076.1	1073.8	86.0	85.8	1019	1015	69	73	62	59	20	23	82	82	2041	1971
5	Pauws	85	1055.1	1072.8	84.3	85.7	1019	1034	20	23	67	69	30	33	116	93	503	507
6	Zhu	83.25	1066.2	1017.7	85.2	81.3	1034	964	38	66	28	47	24	33	128	142	25233	24039
7	Chuan and Chew	79.1	1002.3	977.3	80.1	78.1	937	905	83	95	66	68	20	22	146	162	3299	3468

Table 4.2: Evaluation results for the ISMIR contest (2005) in audio key finding. The composite % score represents the overall accuracy for each methods. w (winamp) and t (timidity) represents the results for the two evaluated databases. The number of errors with close tonalities (perfect 5th, relative, parallel) is also given, as well as the computation time.

We also compared the analysis of the first 15 seconds and the first 30 seconds of each analyzed piece. We can see that the proposed preliminary algorithm ranked third (only analyzing the first 15 seconds) and fourth. Unfortunately, this database is not available to make further studies. There is also a close relationship (same musical datasets) between this contest and the Symbolic Key Finding contest, where the best algorithm (Temperley (2005)) obtained a global accuracy of 91.4%, only 4.9% more than our preliminary approach. These results gave use some hints to set up some experiments which will be the core of our study. Our goal is to study the following problems:

- What is the influence of the different tonal models when using the same set of features?
- What is the influence of analyzing all the piece or only a fragment?
- What is the influence of including a learning stage from labelled data? We see that models using machine learning and some statistical analysis of the data are the ones that perform better in the comparison, Izmirli (2005) and Purwins (2005).

Other questions that we could not analyze with the contest data set and that we address in this dissertation using other music collections are the following ones:

- Which is the influence of the music genre and its relation with the tonal model under consideration? We have only analyzed in this evaluation classical music, but an important issue is how these approaches perform with other musical genres.
- Can our approach be extended to chord estimation? If we estimate the tonality within a small segment, can we get an estimation of the played chord?

We try to answer to all these questions along the next sections of this chapter.

4.3.3 Comparison of tonal models

In this Section we study the influence on the system performance of different tonal models when applied to the proposed low-level tonal features (HPCP). We focus on the use of tonal models which can be expressed into a set of tonal profiles, mainly a major and minor profile. Some papers have made an effort to study the behavior of using different tonal models, such as Chuan and Chew (2005) and Izmirli (2005). We propose here a quantitative comparison of the different profiles used in the literature:

1. **Diatonic**: flat diatonic profiles shown in Figure 2.7. This tonal model is only based on music analysis and do not assign weights into the different degrees of the scale. There are different minor scales (e.g. harmonic or melodic). We have selected the harmonic minor scale (minor sixth and major seventh) as considered in Temperley (1999). Minor sixth is rated higher by humans in Krumhansl (1990) (see Figure 2.7) and Temperley (1999) increased the weight of the raised seventh in minor mode (see Figure 2.13).

2. **Krumhansl**: Krumhansl's profiles shown in Figure 2.12, obtained from human ratings.
3. **Temperley**: Temperley profiles shown in Figure 2.13, derived from the ones by Krumhansl.
4. **Chai**: Profiles proposed by Wei Chai by analyzing MIDI files of folk music, represented in Figure 2.16. These profiles are obtained after a statistical analysis.
5. **Triad**: Flat tonic triad profile, only having 1 values in the pitch class of the tonic triad, in order to investigate the influence of the tonic triad in tonizitation when considering different styles of music, as mentioned in Section 2.5. This basic profile do not include any weight into the different degrees of the tonic triad, as in diatonic profile.
6. **Temperley2**: Temperley profiles obtained empirically using statistical analysis over a corpus of excerpts taken from the Kostka and Payne music theory textbook (see Kostka and Payne (1995)), as shown in Figure 2.18.
7. **Average**: we observe that profiles 1, 2, 3 and 5 are fixed and determined by direct application of music theory (in case of profiles 1 and 5) or after analysis of human ratings (profile 2 and 3). On the other side, profiles 4 and 6 have been obtained after a statistical analysis of pieces with labelled key. We will then evaluate the usability of a learning stage. According to Auhagen and Vos (2000) (pp.424), *the probe-tone profiles are not always coherent with the statistical distribution of notes, mainly in minor keys*. In order to analyze this fact, we included a profile which is obtained by analyzing our labelled data, computing the average THPCP for each piece (i.e. HPCP vector shifted according to its tonic) and finally averaging this THPCP vector for all the pieces in major key to obtain the major profile and for all the pieces in minor tonality to obtain the minor profile. This profile is labelled as *Average*, and it is shown in Figure 4.12. Contrary to the previous ones, the size of the *Average* profile is equal to 36, because it has been computing using an interval resolution of $\frac{1}{3}$ of semitone. We can check that the dominant gets the highest value of the profile, verifying the modifications proposed in Section 4.2. We also see peaks on the tonic, second, third, fourth, sixth and seventh degrees. Regarding the minor profile, we also observe the importance of the dominant degree, in addition to the tonic, as well as the minor third for the tonic triad chord. We can see that the natural minor scale gets a higher value than the harmonic scale.

In addition to the original profiles proposed by the authors, we evaluate the use of the modified profiles proposed in Section 4.2 with respect to the original ones by using three different configurations for each profile:

1. Original profile, as proposed by the authors.
2. Modified profile using the main three chords, as described in Section 4.2.
3. Modified profile using all the chords, where we consider the contribution of triads built from all the degrees of the major or minor scales, as described in Section 4.2.

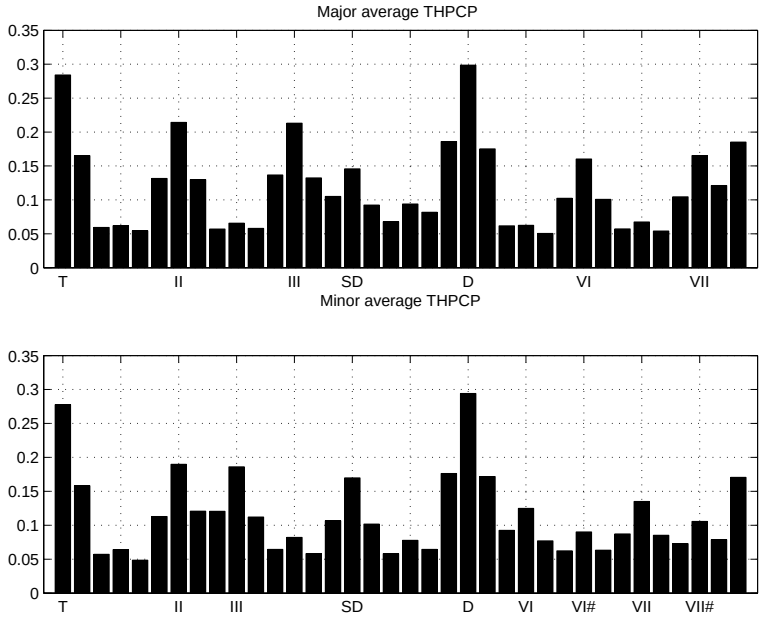


Figure 4.12: Average profile computed from the evaluation collection for major and minor modes, using $pcpsize = 36$.

These modifications do not make much sense when using the average profile, which is computed directly from HPCP values. We will verify this fact below.

The global accuracy obtained for the different profiles and configuration is summarized in Figure 4.13, and the detailed measures are found in Table B.1 in Appendix B.

The highest value of the contest measure is equal to 76.15% and is obtained using the average profile. This result was expected, as this average profile is computed directly using statistics of the THPCP data over the evaluation material. For this profile, the percentage of correct estimations is equal to 70.35%, the contest measure is equal to 76.15%, and the mode is correctly estimated in the 86.11% of the cases. The majority of the estimation errors are with distant keys (9.2% of the instances), followed by fifth errors (7.69%).

The second-best performance is obtained using the tonal model proposed in Temperley (1999), which is cognition oriented, with the proposed adaptation (using the three main chords of the key). The estimation accuracy is equal to 69.9%, the contest measure is equal to 75.61%, and the mode is correctly estimated in 84.76 % of the instances. We observe that there is only a small decrease in performance of 0.45% of correct estimations. In fact, Temperley profiles depend on certain (simplifying) assumptions on the used chords and have been adapted to HPCP features (as presented in Section 2.7). It was then expected that their performance for classification would be weaker than using directly statistics over the HPCP features.

This tonal model is followed by other tonal profiles such as Krumhansl (contest measure equal to 75.41 %) and diatonic (contest measure equal to 74.75 %), both of them including the proposed adaptation. We observe

that this difference in accuracy between Temperley, Krumhansl and diatonic modified profiles (using 3 main chords) is very small, as shown in Figure 4.13. Then, we find the profiles which are obtained empirically using statistics on pieces, such as Temperley's empirical profiles (contest measure equal to 74.43 %), which do not need to be adapted. Then, we obtain a contest measure equal to 69.79 % when using tonic triad profile without adaptation. The worst results is found for Chai's statistical profiles including the proposed adaptation (contest measure equal to 61.68 %).

We can see that the performance is higher when using the modified version considering the 3 main chords of the key (middle plot in Figure 4.13 and configuration equal to 2 in Table B.1) for diatonic, Krumhansl, Temperley and Chai profiles.

The global error distribution is also shown in Figure 4.14. Using the original profiles, we get more fifth errors than using the original profiles. Chai profiles, used as originally designed, yield some errors with parallel tonalities. Distant errors are more common when using tonic triads in the case of modified 3 chords profiles.

Fifth errors are found when using modified triad profiles (25.84 %), Krumhansl's original profiles (23.08 %) and Temperley's original profiles (12.32 %). These fifth errors when using Diatonic, Krumhansl's or Temperley's original profiles are solved when using modified profiles.

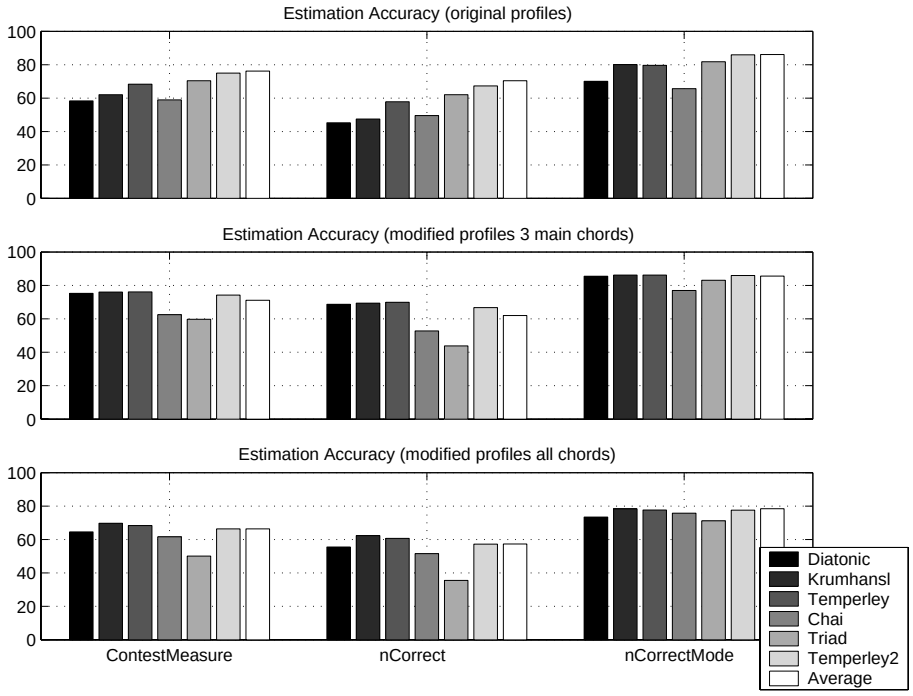


Figure 4.13: Global accuracy for the different tonal profiles.

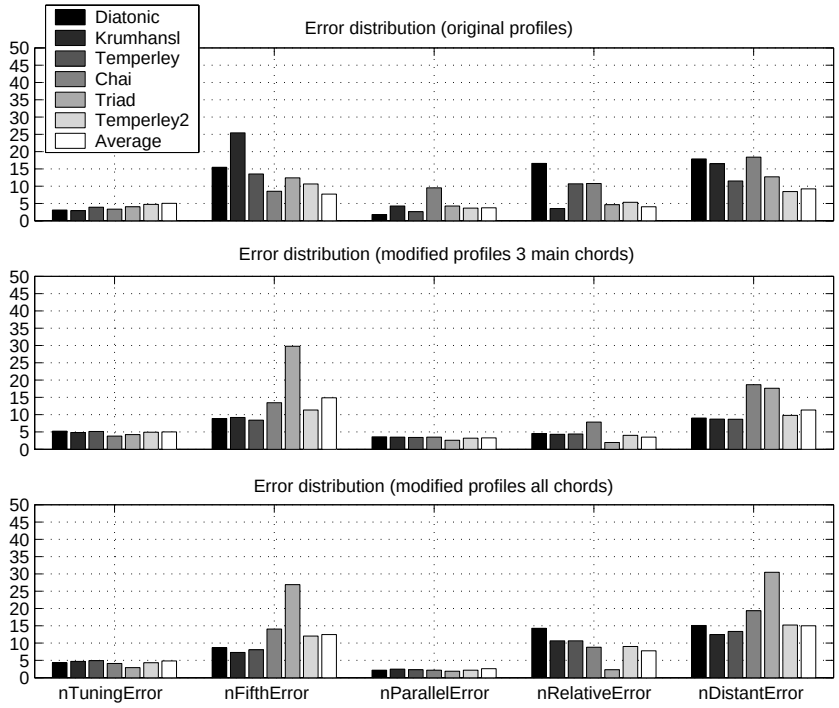


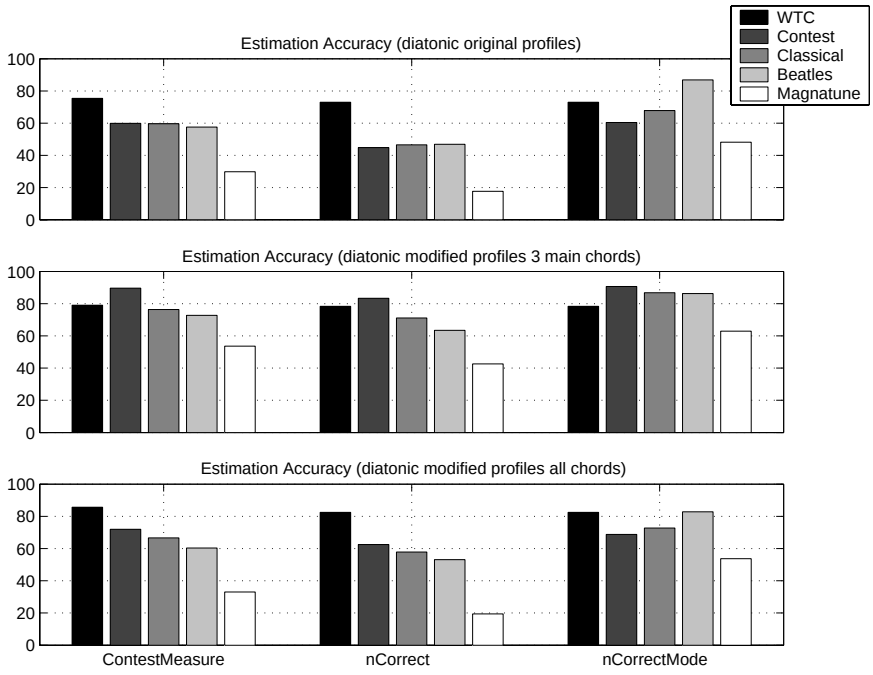
Figure 4.14: Global error distribution for the different tonal profiles.

We can also analyze the results for each of the databases of the evaluation material for each of the profiles (see Figures 4.15 for the diatonic profile, Figure 4.16 for Krumhanl's profiles, Figure 4.17 for Temperley's profiles, Figure 4.18 for Chai's profiles, Figure 4.19 for tonic triad profiles, Figure 4.20 for Temperley's empirical profiles and Figure 4.21 for THPCP averages profiles). The results are also detailed in Table B.1. In general, the performance is worse for the different genres (collection ID equal to 5), compared to the classical database (collection ID equal to 1, 2 and 3) and the Beatles's collection (collection ID equal to 4). Using diatonic original profiles, for instance, the performance for the fifth database is 15%, while for the Beatles collection the accuracy is around 47%. When considering 3 chords in the diatonic modified profiles, the performance increases to around 43%.

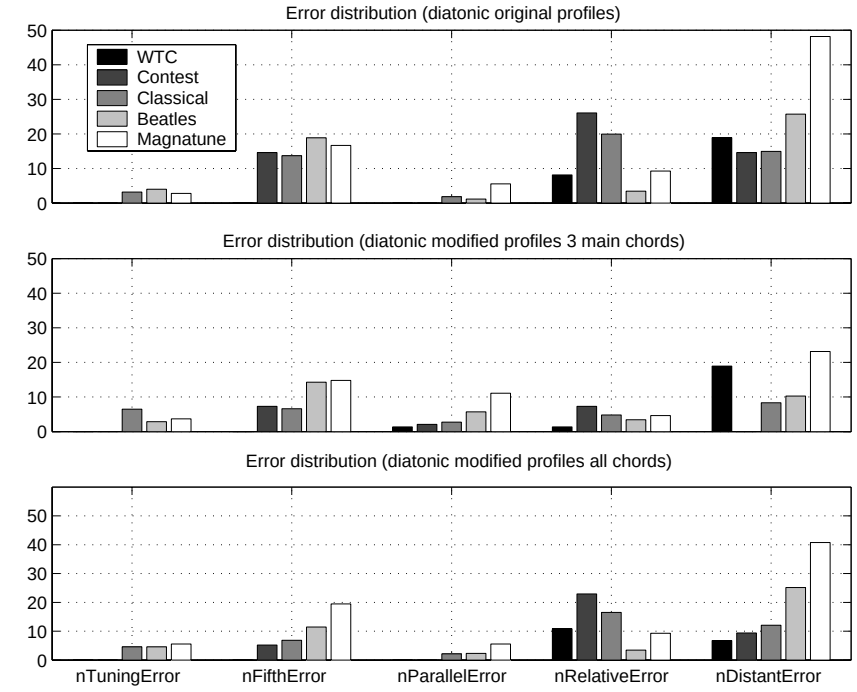
The use of all the chords for the modified profiles causes less accuracy for all the databases except for Bach's Well-Tempered Clavier. When using Krumhansl's profile (2), we see also that the performance decreases for the last database. The accuracy is around 45% for popular database (5), and for the rest more than 60% (85% in the contest training data). The best performance is also obtained when using modified profiles with 3 main chords. We obtain here a 15% of fifth errors, mainly for popular music, which reveals a confusion between the tonic and the dominant. Using the third profile proposed by Temperley, we obtain the best results for all the evaluation databases.

Finally, the best results for the popular music is obtained using a profile for the diatonic triad, overall accuracy of around 55%. This reveals that the tonal profile which is more suitable for one style of music may not be optimal for all musical genres and it verifies the importance of the tonic triad in popular music which was pointed by Temperley and commented in Section 2.5. A complete list of the experimental results is found in Table B.1 in Appendix B.

As a conclusion of this study, we see that tonal models which are obtained after psychological studies (as the ones derived from Krumhansl) with the adaptations proposed in this work seem to work better than the ones obtained through statistics, with the exception of the average profiles computed from THPCP values. A second conclusion is the dependency on musical genre. These tonal models have been designed to work for classical western music, and we verify that these models should be adapted to other musical genres.

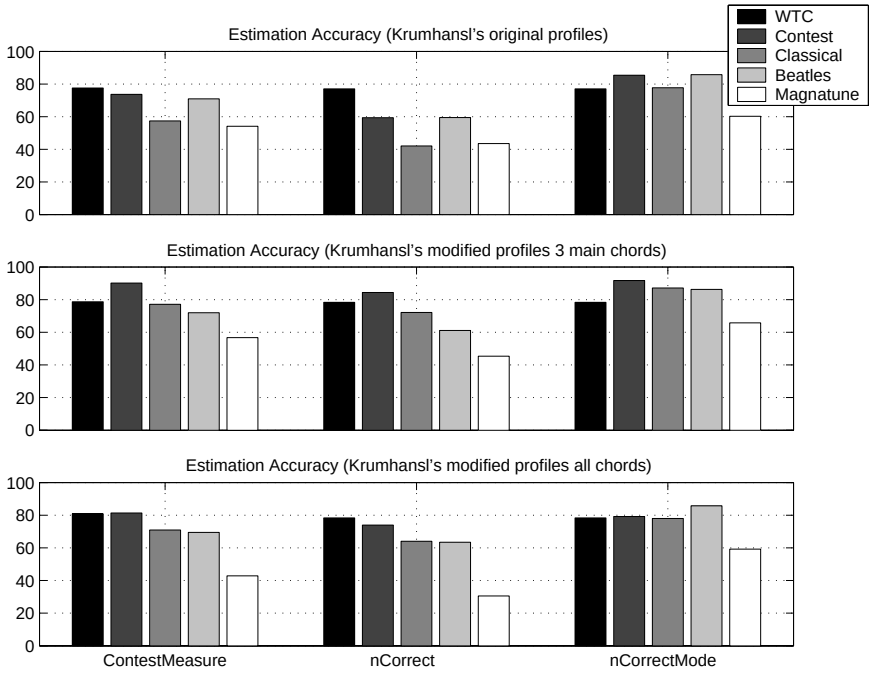


(a)

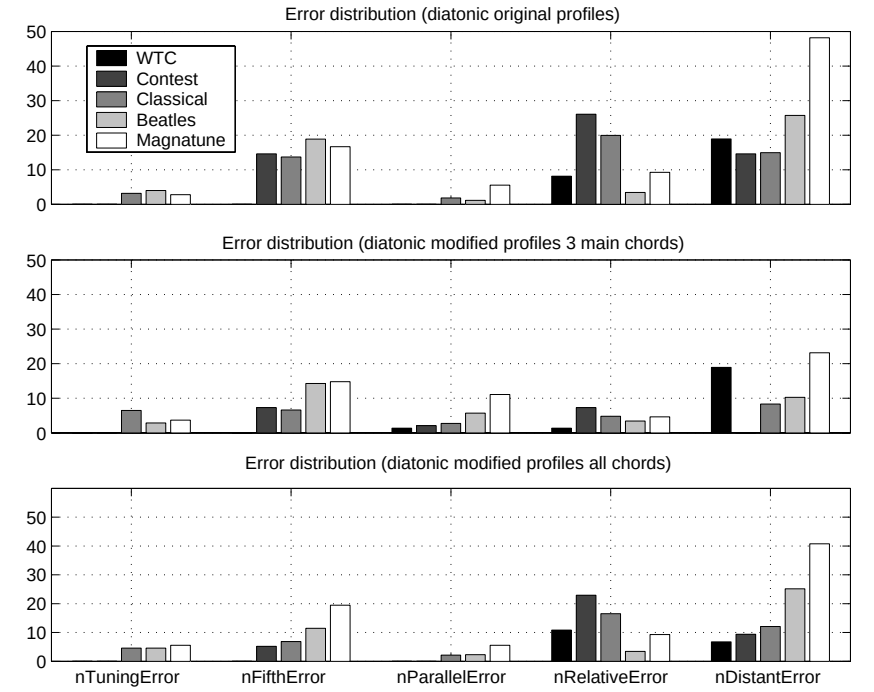


(b)

Figure 4.15: Estimation accuracy (contest measure, percentage of correct key and percentage of correct mode estimation) (a) and error distribution (b) for the different evaluated collections using diatonic profiles.

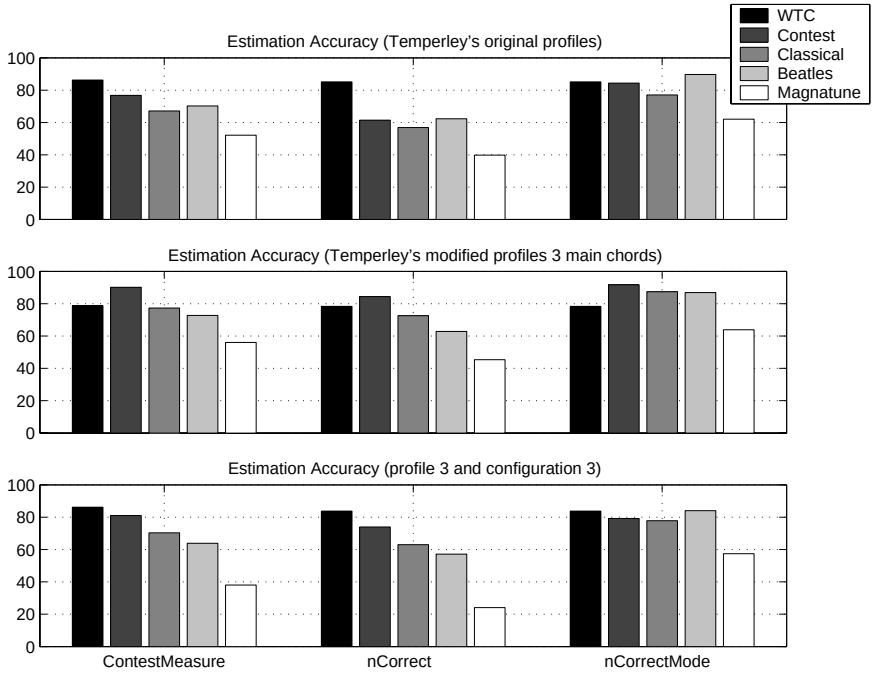


(a)

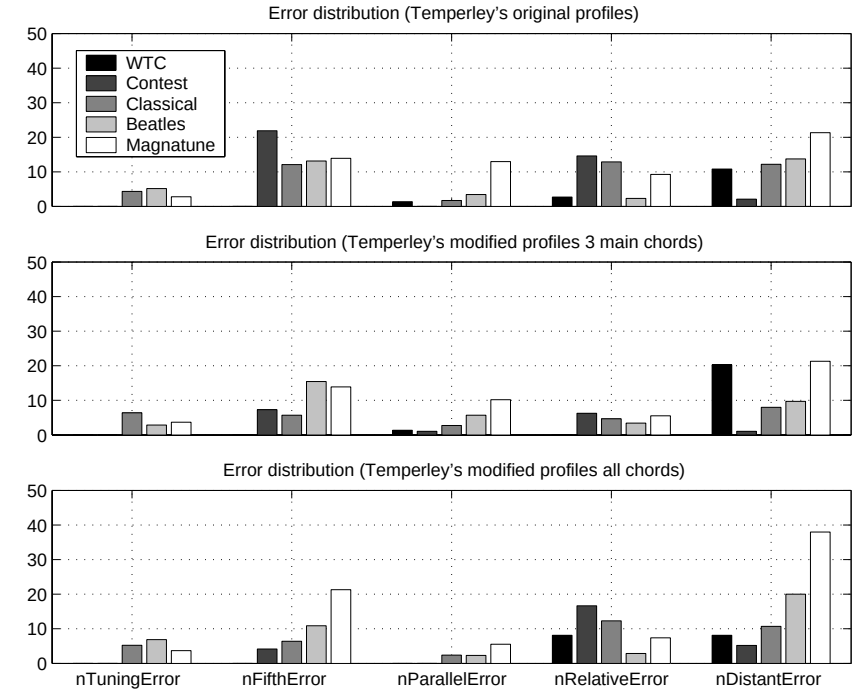


(b)

Figure 4.16: Estimation accuracy (contest measure, percentage of correct key and percentage of correct mode estimation) (a) and error distribution (b) for the different evaluated collections using Krumhansl's profiles.

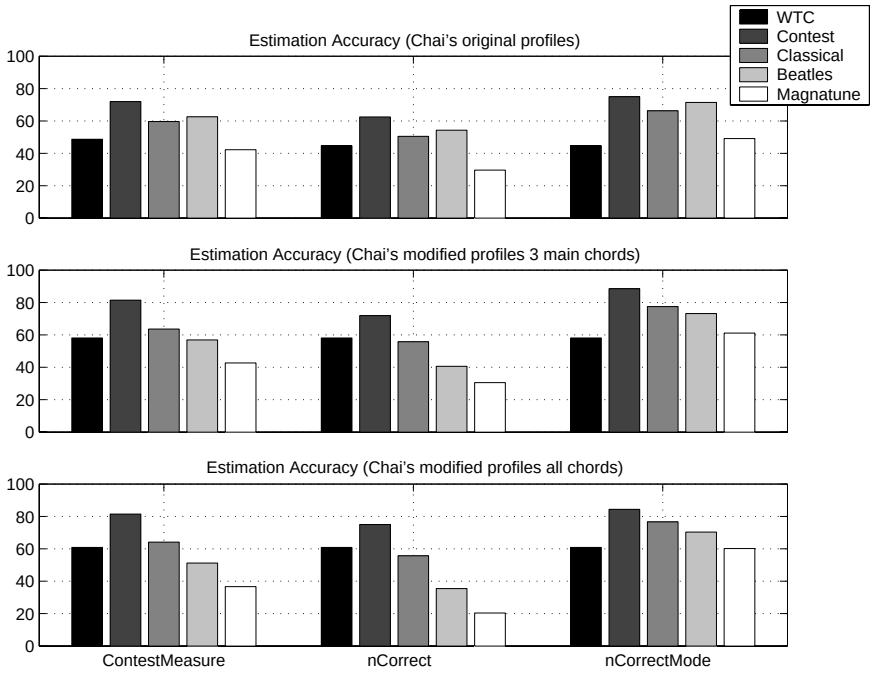


(a)

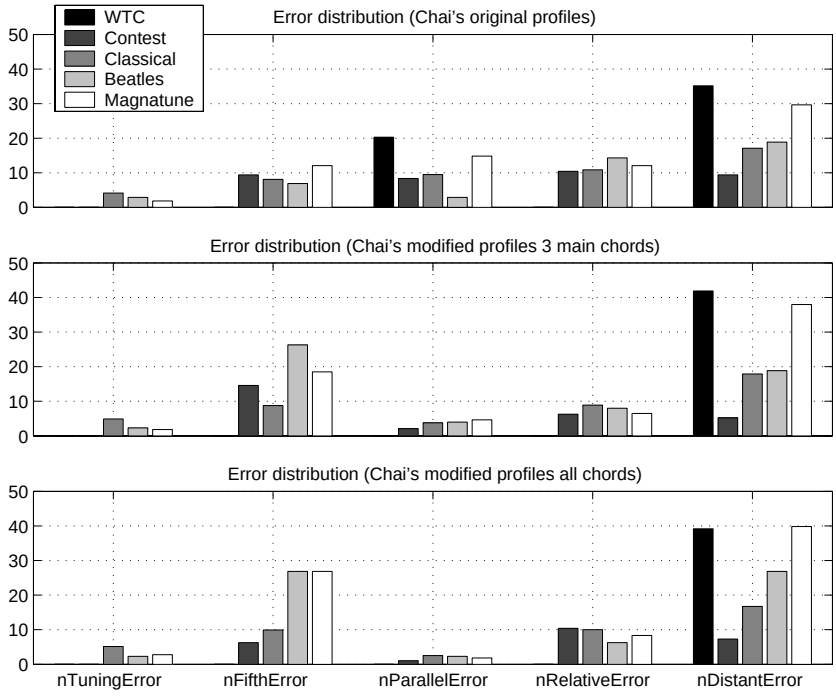


(b)

Figure 4.17: Estimation accuracy (contest measure, percentage of correct key and percentage of correct mode estimation) (a) and error distribution (b) for the different evaluated collections using Temperley's theoretical profiles.

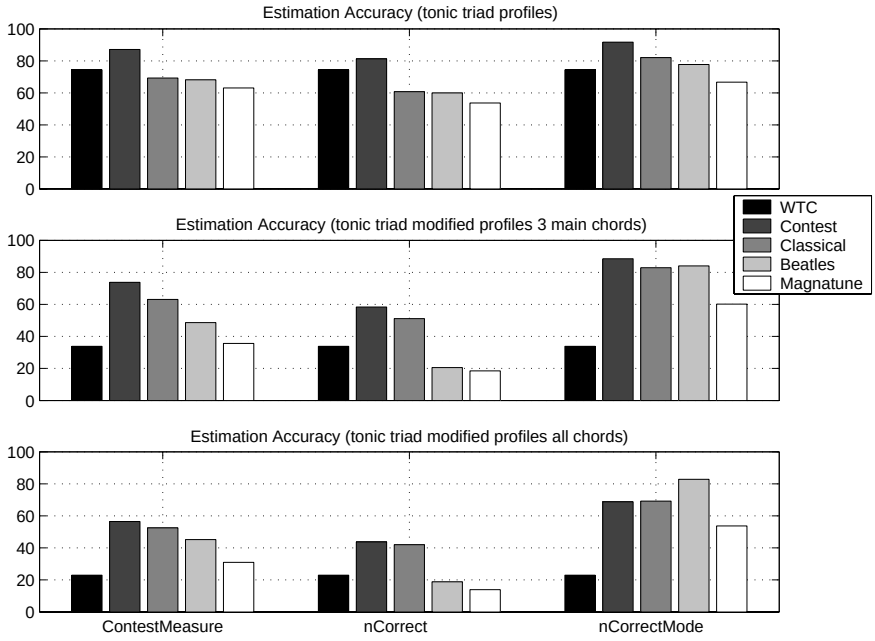


(a)

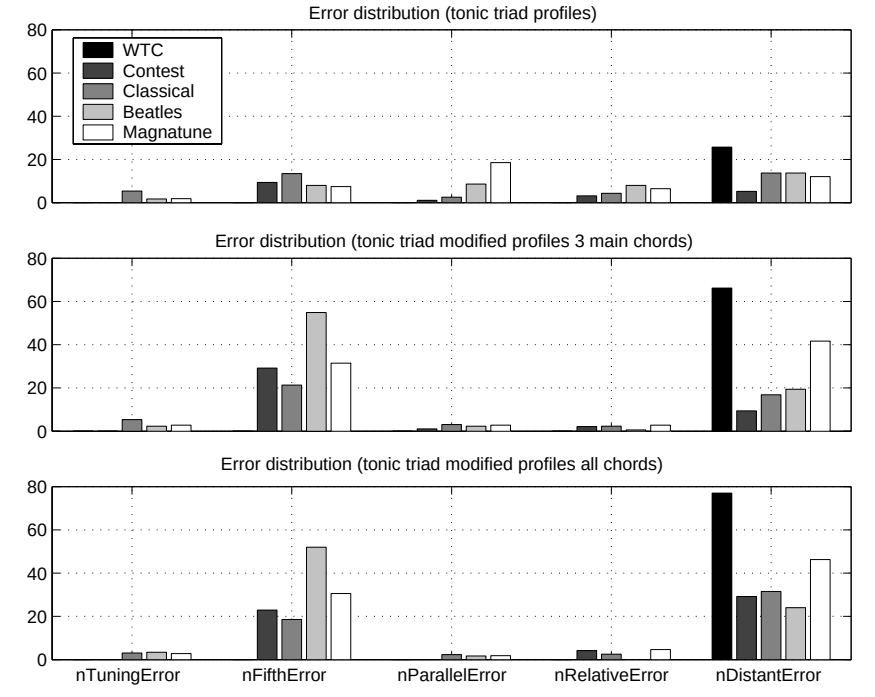


(b)

Figure 4.18: Estimation accuracy (contest measure, percentage of correct key and percentage of correct mode estimation) (a) and error distribution (b) for the different evaluated collections using Chai's profiles.

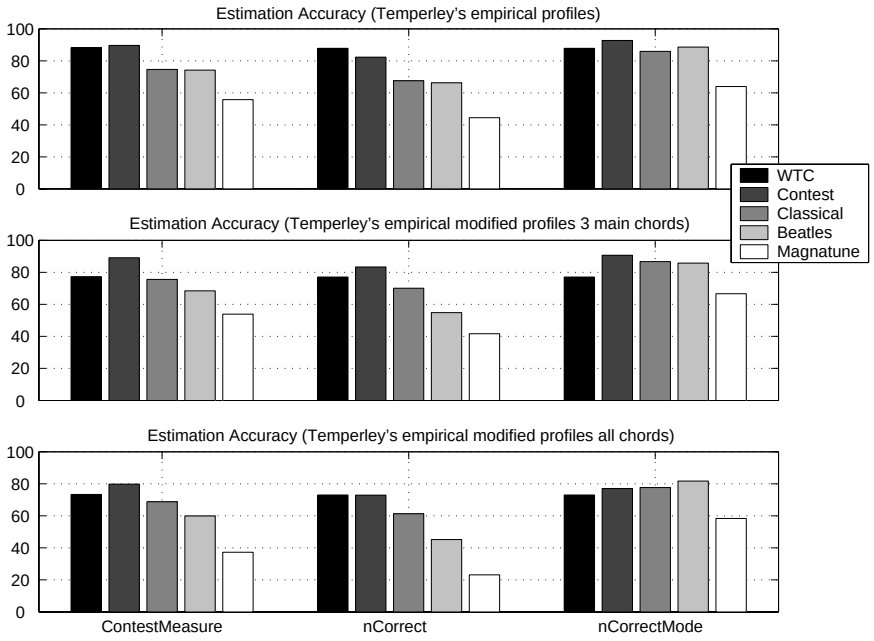


(a)

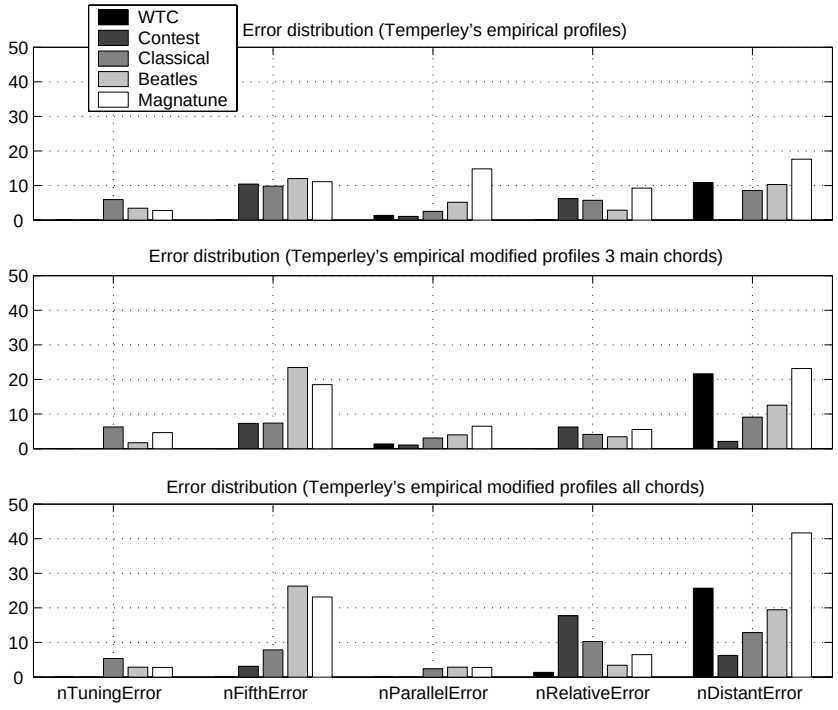


(b)

Figure 4.19: Estimation accuracy (contest measure, percentage of correct key and percentage of correct mode estimation) (a) and error distribution (b) for the different evaluated collections using tonic triad profiles.

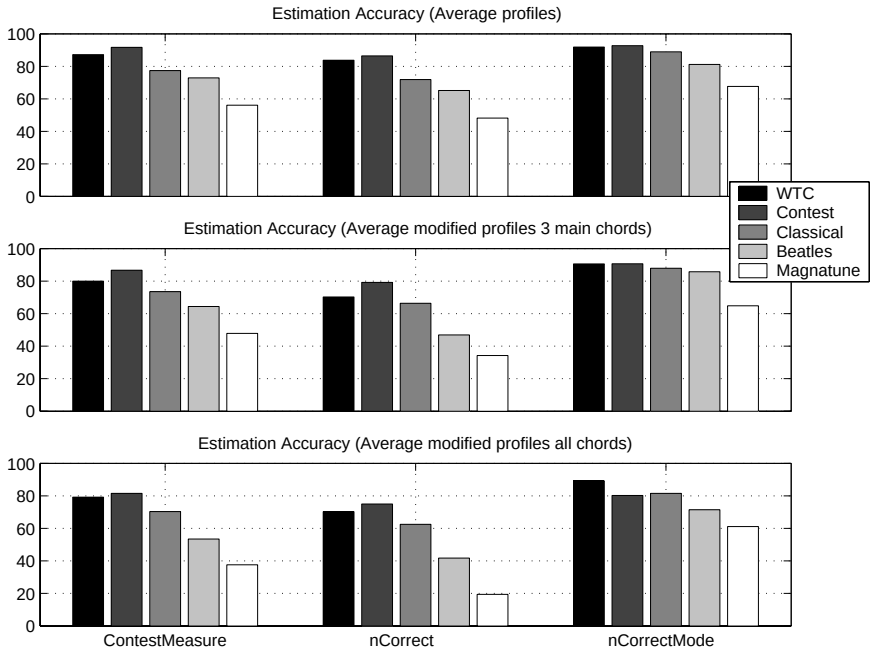


(a)

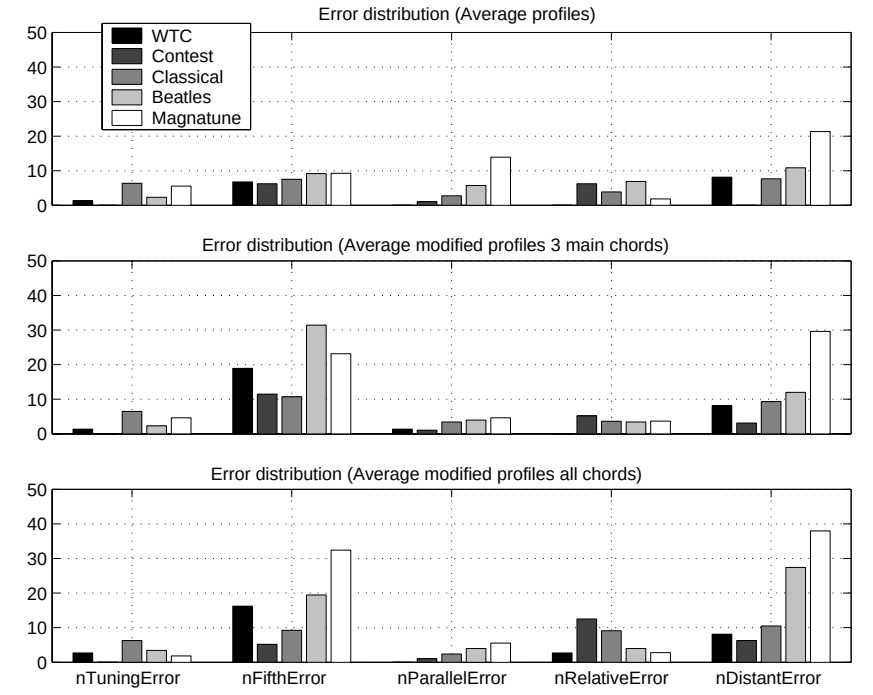


(b)

Figure 4.20: Estimation accuracy (contest measure, percentage of correct key and percentage of correct mode estimation) (a) and error distribution (b) for the different evaluated collections using Temperley's empirical profiles.



(a)



(b)

Figure 4.21: Estimation accuracy (contest measure, percentage of correct key and percentage of correct mode estimation) (a) and error distribution (b) for the different evaluated collections using THPCP average profiles.

4.3.4 Tonal models and musical genres

We have seen that most of the tonal models perform worse for other musical genres than classical music. This can be, in part, due to the presence of percussive sounds and post-production effects, but also to the difference in tonal models. This fact has been already noticed in Temperley (2001).

In this section, we make a descriptive study of how the tonal profile, represented by the average HPCP vector, varies for different musical genre. Figure 4.22 shows the system accuracy for different musical genres for two different profiles: Temperley's theoretical profiles (Temperley (1999)) and tonic triad profiles. We can see that using a simple tonic triad profile makes the results improve for many of the musical genres, such as jazz, punk rock ambient, blues and instrumental. Pieces in jazz belong to various styles, and most of them have a strong presence of the tonic triad. These results reveal the importance of the tonic triad for some musical genres, as pointed out in Temperley (2001).

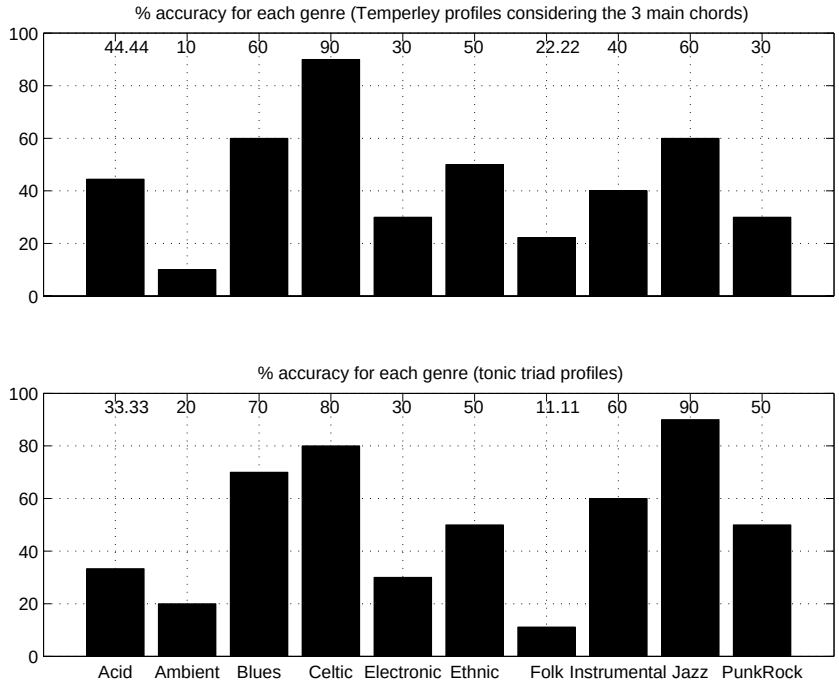


Figure 4.22: Estimation accuracy for each musical genre of Temperley profiles versus Tonic triad profiles.

Figures 4.23 and 4.24 show the average and standard deviation of the THPCP for minor and major pieces within the different musical genres. This THPCP vector is computed as explained in Chapter 3, by shifting the average or standard deviation HPCP over the whole piece with respect to the index corresponding to the annotated tonality. Although this statistical analysis has been made with a limited number of pieces (108 for the different genres), we can still make some observations. First, there are significant differences between the different musical genres.

For acid and electronic genres, for instance, we did not find major pieces in this music collection. We can clearly identify the major and minor diatonic degrees for classical music (Figures 4.23 and 4.24), which is the genre where we have a higher number of instances.

Regarding the major THPCP average, shown in Figure 4.23, we can see the predominance of the dominant for most of the genres, and it is not so clear in pop and Celtic music. We also see that ambient music lacks of a clear tonal profile compare to the other genres. This is also visible for punk rock music, where the profile is flatter than the others. The diatonic degrees are also clearly found in other labelled genres such as folk, Celtic, ethnic and pop. This is not the case for others such as ambient, blues and punk.

When looking at the minor THPCP average, shown in Figure 4.24, we can also observe the predominance of the dominant for most of the genres. We also see that acid music lacks of a clear tonal profile compared to the other genres. The main reason is that these pieces mainly include speech and electronic unpitched sounds. This fact is also visible for punk rock music, where the profile is flatter than the others. Further experiments should be devoted to extensively analyze these differences with genre using a bigger annotated collection.

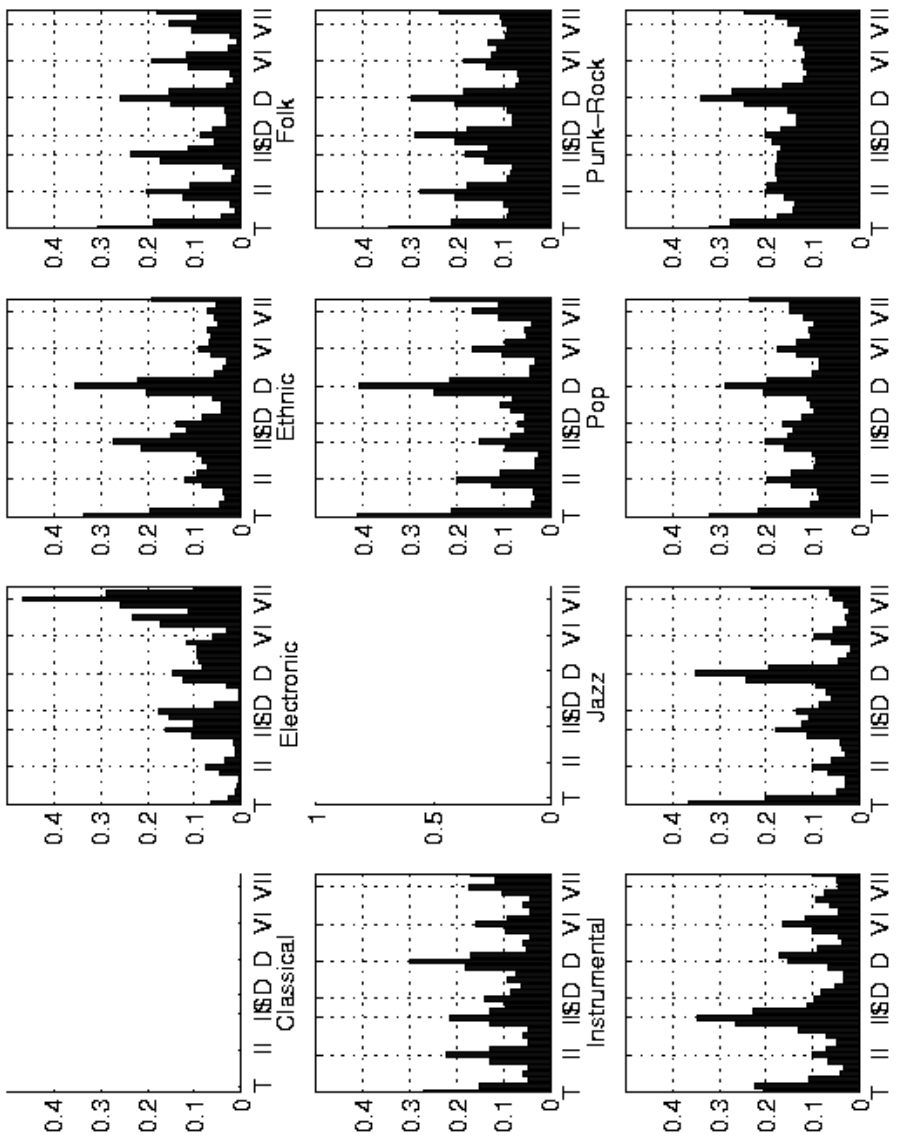


Figure 4.23: Average of THPCP vector (normalized by the annotated key) for major tonalities per each analyzed musical genre.

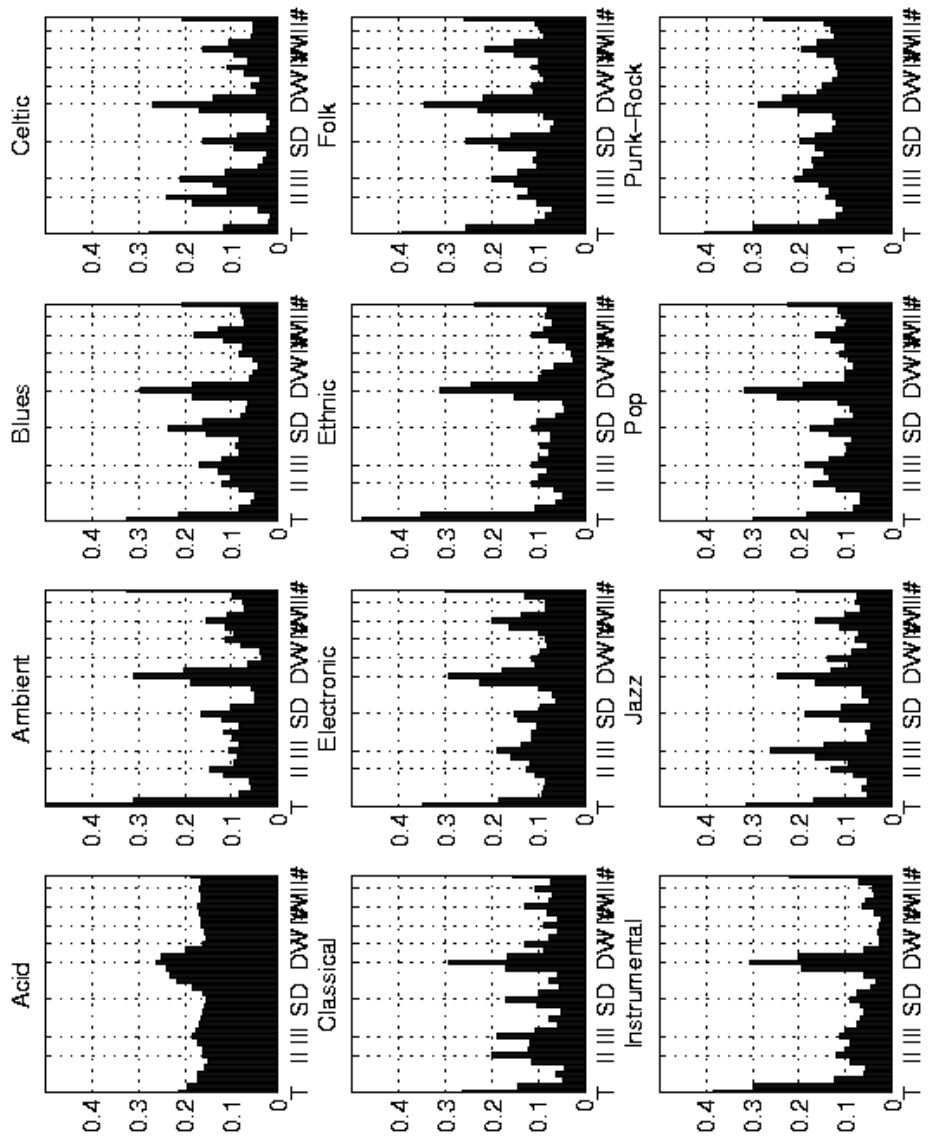


Figure 4.24: Average of THPCP vector (normalized by the annotated key) for minor tonalities per each analyzed musical genre.

4.3.5 Location of the main tonality within a piece

We saw in section 4.3.2 how a preliminary version of our approach for audio key finding performed with respect to other methods. In this evaluation, we submitted two different algorithms, differing in the analyzed excerpt for tonality estimation (15 or 30 first seconds). Also, for some approaches such as the one from Purwins (2005) (ranked the second), only the first 15 seconds are used. The results showed that best performance was obtained when analyzing the first 15 seconds of the piece, which can indicate that the main key of a piece is established at the beginning of the composition.

In this experiment, we analyze this fact, trying to answer to where the main key is located within a given piece. The main key corresponds to the annotated one. This experiment is also done in Pauws (2004) for piano sonatas (30 seconds at the beginning, middle and end of the piece). As proposed in Pauws (2004) with piano sonatas, we have analyzed the influence of segment duration into global key estimation. As mentioned in Pauws (2004), there are some observed tendencies: a simple time measure of 4 beats at a tempo of 100 beats per minute lasts 2.4 seconds, and the end of the performances is often played by slowing down and sustaining the closing chord until it dies out.

For this experiment we use the Classical DB collection, and we estimate the global tonality by analyzing different segment durations at the beginning, middle and end of the piece. We have estimated the global key using 2.5, 5, 10, 15, 20, 25, 30, 35, 40, 45, 50, 55 and 60 seconds at the beginning, middle and end of the piece. The profiles chosen for this experiment are the theoretical ones proposed by Temperley (1999) (profile 3) modified by using the three main chords of the key (second configuration). These profiles yielded the best results on the evaluation presented in Section 4.3.3 after the average profiles. As the average profiles are computed from the evaluation collection, we prefer to use Temperley's profiles, which are obtained independently from the test set. The results for different segment durations are summarized in Figure 4.25.

As a result, the global accuracy slightly decreases from 72.55% to 67.084% when considering a short segment, so that to obtain the best possible judgement of the main key of a piece it is necessary to analyze the whole piece. Nevertheless, the analysis of the beginning or ending segments provides a good estimation of the overall tonality of the composition. The accuracy gets almost constant when considering more than 15 seconds at the beginning and at the end of the piece.

As it was expected, the middle of the piece does not seem to be in the global tonality, as the accuracy decreases a lot. This finding agrees with the fact that there are often modulations in the middle of the piece.

4.3.6 Chord estimation

In order to study how this approach performs the task of chord estimation, we have set up the following evaluation experiment in which we try to estimate the root and the mode of a chord. The evaluation measures are analogous to the ones used for key estimation, but using chord root and mode. We have set a small database of 148 isolated chords, including guitar (110), keyboards (19) and strings (24). There are major and minor chords, some of them including seventh and ninth degrees.

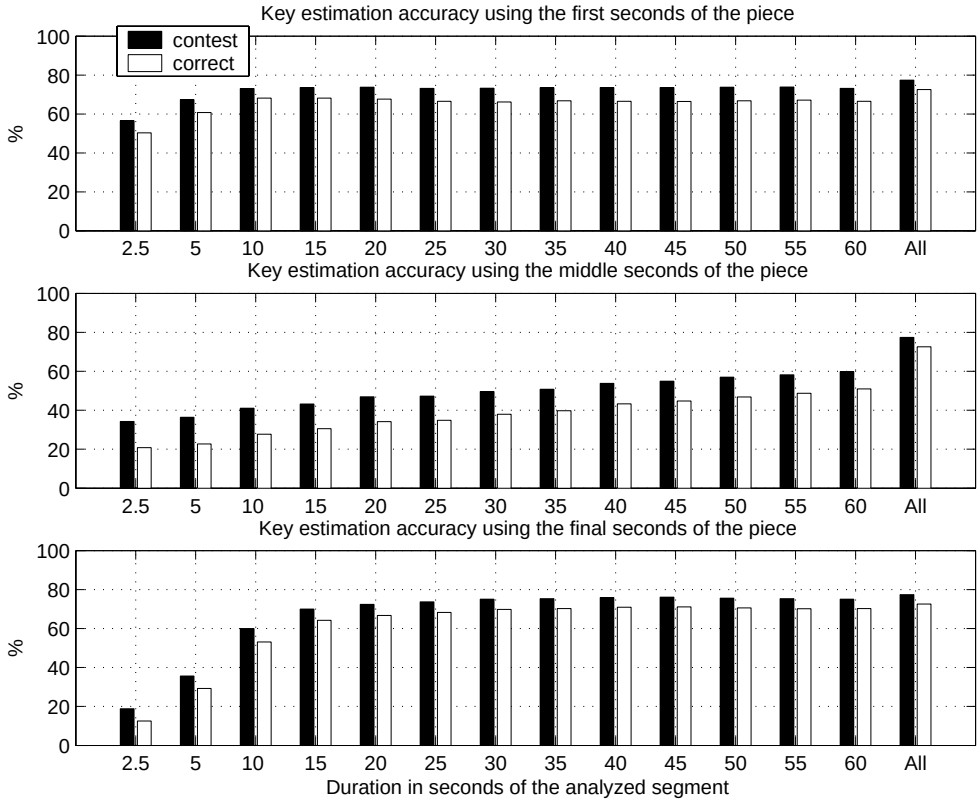


Figure 4.25: Estimation accuracy when analyzing n seconds at the beginning, middle and end of the piece in order to estimate the global key.

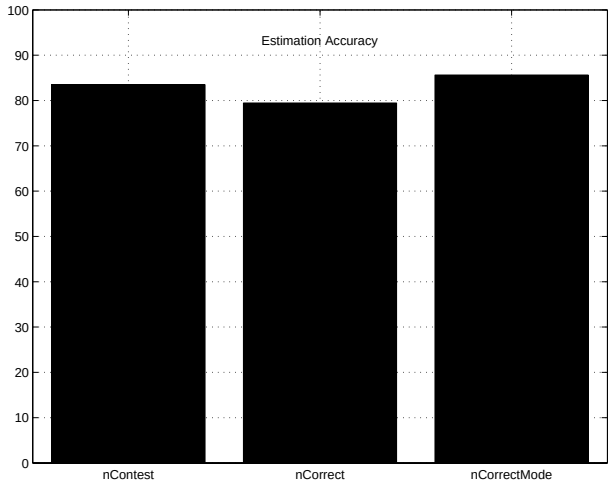
The results obtained when using the proposed approach for key estimation and a tonic triad profile are shown in Figure 4.26 and in Table 4.3:

If we look at the errors, 36% of them are from chords with 7th degree, where the algorithm estimates a close triad chords (for instance A Minor 7 is estimated as C Major instead of A Minor) and 20% are from chords with 9th degree having the same problem. Only 43% of the errors correspond to distant chords, that means a total of 8.84% of estimation error. If we consider errors with close triad chords as correct, the estimation accuracy is equal to 91.16%. We could introduce other profiles devoted to detect different chord configurations other than tonic triads. For instance, there should be prototypes for 7th, 9th, augmented and diminished chords as well. We have contributed to a text representation for musical chord symbols presented in Harte et al. (2005), which can be parsed with a computer program and converted into different chord templates.

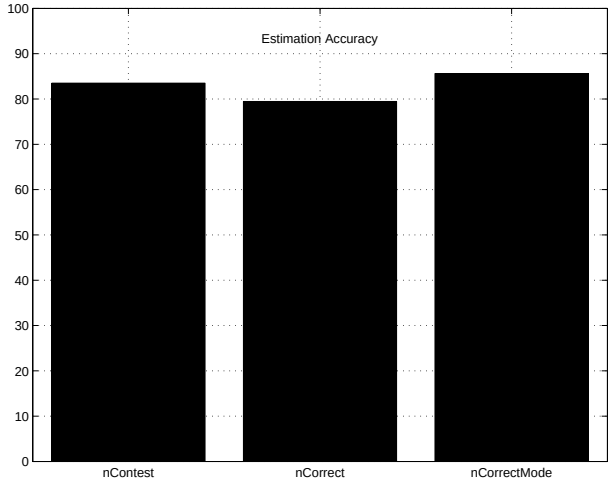
In the same way, the system should be evaluated considering chords in real pieces. We refer to Fujishima (1999); Sheh and Ellis (2003); Harte and Sandler (2005) as some approaches for chord estimation using pitch class distribution features.

Measure	**%**
Correct	79.45
Contest measure	85.62
Correct Mode	85.62
Tuning Errors	1.37
Fifth Errors	3.42
Parallel Errors	1.37
Relative Errors	6.85
Far keys errors	7.53

Table 4.3: Chord estimation evaluation results.



(a)



(b)

Figure 4.26: Estimation accuracy (contest measure, percentage of correct key and percentage of correct mode estimation) (a) and error distribution (b) for chord estimation using tonic triad profiles ($pcpsize = 36$).

4.4 Machine learning techniques for tonality estimation

We compare here how the use of machine learning techniques, combined with the studied tonal profiles, can improve the system for tonality estimation from audio. This work has been carried out in collaboration with Perfecto Herrera. Initial results of this work were already reported in Gómez and Herrera (2004), where different experiments, involving comparisons between the most usual learning strategies like binary trees, Bayesian estimation, neural networks, support vector machines, boosting, and bagging were summarized. The main conclusions were that modest improvements in performance were achieved by machine learning techniques and by the combination of machine learning techniques and the algorithm presented in Section 4.3.3.

Now, in this section, we report very succinctly new results. The experiments were carried out using Weka[6], a collection of machine learning algorithms for data mining tasks implemented in java. The algorithms implemented in Weka are applied directly to a dataset, and the software package includes tools for data pre-processing, classification, regression, clustering, association rules, and visualization.

4.4.1 Methodology

The database used for evaluating this machine learning approach consisted in nearly 1400 pieces, as presented in Section 4.3.1.1, excerpt for the MIREX2005 collection, which was not available.

For this experiment, we define 24 different classes, one for each key. We see in Figure 4.8 and 4.9 that all classes do not have equal size, for instance there are more pieces in major mode than in minor mode and there are some classes with a small number of instances (the less represented classes are D#/Eb minor and G#/Ab minor with 18 instances and A#/Bb minor with 19 instances). As the number of instances is very high, we assume that this pattern might represent the distribution of pieces that a system would find in the real world.

The system gets as input parameters the average HPCP vector computed over the whole piece ($size = 36$), as well as the estimated key using Temperley's theoretical profiles from Temperley (1999). These profiles provided the second-best performance in previous experiments (see Section 4.3.3, including the proposed adaptation (using the three main chords of the key). As seen below, the estimation accuracy of the Temperley-based HPCP is equal to 69.9%, the contest measure is equal to 75.61%, and the mode is correctly estimated in 84.76 % of the instances. We did not use *Average* profiles, providing better results (70.35% of correct estimation) because they are define as the average of THPCP for major and minor pieces, calculated using the evaluation database and the key annotations. On the other hand, Temperley's adapted profiles are entirely independent from the test collection.

We define a 10-fold cross-validation procedure in order to define the training and the test set. In this procedure, we divide the data into 10 subsets of approximately equal size and we train the model 10 times, each time leaving out one of the subset from training and using only this omitted subset to compute the accuracy measure.

[6] http://www.cs.waikato.ac.nz/ml/weka

4.4.2 Results

The best results to estimate the key from HPCP values were obtained using a Support Vector Machine (SVM) with a Radial Basis Function Kernel, and a systematic adjustment of several of its parameters. Support Vector Machine are classifiers that, for the simplest case of solving two-class linear classification problems, separate the two classes with a hyperplane (a generalization of a plane in three dimensional space to more than three dimensions) in the feature space such that: (a) the "largest" possible fraction of instances of the same class are on the same side of the hyperplane, and (b) the distance of either class from the hyperplane is maximal. The prediction of a SVM for an unseen instance $\mathbf{z}$ is $y+ = 1$ or $y- = -1$, given by the decision function:

$$pred(\mathbf{z}) = sgn(\mathbf{w} * \mathbf{z} + b) \tag{4.9}$$

where b is an unregularized bias term, and the hyperplane $\mathbf{w}$ is computed by maximizing a vector of Lagrange multipliers α in:

$$W(\alpha) = \sum_{i=1}^{l} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{l} \alpha_i \alpha_j y_i y_j K(\mathbf{x_i}, \mathbf{x_j}) \tag{4.10}$$

subject to:

$$0 \leq \alpha_i \leq C;\ i = 1, \ldots, l\ and\ \sum_{i=1}^{l} \alpha_i y_i = 0 \tag{4.11}$$

where C is a parameter set by the user to regulate the effect of outliers and noise, i.e. it defines the meaning of the word "largest" in (a). After calculating the coefficients α_i in Equation 4.10, the decision function becomes:

$$pred(\mathbf{z}) = sgn(\sum_{i=1}^{l} \alpha_i y_i K(\mathbf{x_i}, \mathbf{z}) + b) \tag{4.12}$$

An instance x_i for which α_i is not zero is called a Support Vector (SV). Note that the prediction calculated above uses the support vectors only. As such, the support vectors are those instances that are closest to the decision boundary in feature space. The function K is a kernel function and maps the features in T, called the input space, into a feature space defined by K in which then a linear class separation is performed. For Linear Support Vector Machines this mapping is a linear mapping:

$$K(\mathbf{x_i}, \mathbf{x_j}) = \mathbf{x_i} * \mathbf{x_j} \tag{4.13}$$

For the Radial Basis Function (RBF) Kernel, the mapping is defined by:

$$K(\mathbf{x_i}, \mathbf{x_j}) = exp(-\gamma ||\mathbf{x_i} - \mathbf{x_j}||^2),\ \gamma > 0 \tag{4.14}$$

The RBF kernel nonlinearly maps samples into a higher dimensional space, so that it can handle the case when the relation between class labels and attributes is nonlinear. It also has less parameters to be adjusted and less numerical specificities (or difficulties) than other available Kernels. The SVM implementation included in Weka is based on Platt's Sequential Minimal Optimization, presented in Platt (1998), and solves multi-class problems using pair-wise classification (i.e., combining 2-class SVMs solving each pair-wise decision that has to be taken).

Using a Support Vector Machine with a RBF Kernel, plus a systematic adjustment of several of its parameters, the obtained precision was equal to 73.704%. This performance is 3.8% higher than the one obtained using Temperley's theoretical profiles. The mode is correctly estimated in 85.49% of the cases, which is only a 0.73% better. As the classes that have generated more errors are mostly those having the smallest number of instances (e.g. D#/Eb minor with a 58.3% of correct estimation), so that it could be possible to improve our results by simply doubling the amount of instances for these classes.

Table 4.4 provides the confusion matrix, where we observe that many errors are made with close keys: parallel (e.g. confusion between D#/Eb Major and minor), relative (e.g. G Major and E minor, G#/Ab Major and F minor) or related by an interval of fifth (e.g. errors between D Major and A Major, D# Major and A# Major, F Major and C Major and G Major and C Major), and some of them are tuning errors (e.g. C# Major confused with D Major and B Major with C Major). Considering these errors, the contest measure (as explained in Section 4.3.1.2) is equal to 79.1%, which is 3.49% higher that the one obtained using Temperley's theoretical profiles.

As it is the case in some meta-learning approaches, the combination of two different algorithms can improve the performance provided that both generate different error patterns. Our experiments using the estimated key from the algorithm explained in Section 4.3.3 (Temperley's theoretical profiles) as an additional input of the algorithm has yielded no improvement to the presented results.

4.4.3 Discussion

Comparing an approach based on the correlation with fixed major and minor profiles to the machine learning techniques that we can consider as "tools of the trade", slight improvements in performance can be achieved by the latter alone. The combination of both approaches, by embedding the estimation based on profile correlations as an additional feature for the Support Vector Machine, did not lead to a better performance than using SVM alone. A more detailed discussion is planned for a future publication.

A	a	A#	a#	B	b	C	c	C#	c#	D	d	D#	d#	E	e	F	f	F#	f#	G	g	G#	g#	
62	3	0	0	1	0	0	0	0	0	8	1	0	0	1	0	0	0	0	3	1	0	5	1	A
5	32	0	0	0	0	4	0	0	0	0	1	0	0	2	4	1	0	0	0	3	0	0	1	a
2	0	45	0	0	0	0	1	0	0	2	0	6	0	0	0	2	0	0	0	0	1	0	0	A#
0	0	1	10	0	0	0	0	0	1	3	0	1	0	0	0	0	2	0	0	1	0	0	0	a#
2	0	1	0	25	2	0	0	0	0	0	0	0	0	2	0	0	0	0	0	0	0	0	0	B
0	0	0	1	1	32	0	1	0	0	0	0	1	0	0	3	0	0	0	3	3	0	0	0	b
0	3	0	0	6	0	68	0	0	0	0	1	0	0	0	0	6	0	0	0	4	0	0	0	C
0	0	0	0	0	1	4	41	0	0	0	0	4	0	0	0	0	2	0	0	1	1	3	0	c
0	0	0	2	0	0	1	0	21	2	2	0	0	0	0	0	0	0	0	1	1	0	1	0	C#
0	0	0	0	1	0	0	1	2	26	0	0	0	0	4	0	0	0	0	0	0	0	0	0	c#
8	1	0	0	0	1	0	0	6	0	111	0	0	0	1	0	0	0	0	0	6	0	0	0	D
0	3	1	0	0	0	2	0	0	2	2	50	0	0	0	0	1	0	0	1	1	4	0	0	d
1	0	2	0	2	6	0	4	2	0	2	0	47	0	0	1	0	0	0	0	0	1	2	0	D#
2	0	0	0	0	0	0	0	0	0	0	0	6	7	0	1	0	0	0	0	1	0	1	0	d#
6	1	1	0	0	0	0	0	0	0	1	0	1	0	52	4	0	0	0	1	0	0	0	1	E
0	2	0	0	1	2	0	0	0	1	0	2	0	4	5	31	0	0	0	0	6	0	0	1	e
1	5	1	0	0	1	0	0	1	0	0	3	0	0	2	0	72	3	0	0	0	0	0	0	F
0	1	0	0	0	0	0	0	0	1	0	0	0	0	0	2	3	40	0	0	0	0	6	0	f
0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	18	1	0	0	0	0	F#
2	0	0	0	0	0	0	0	0	0	1	1	0	0	2	0	0	1	2	18	0	0	0	1	f#
0	3	0	0	0	0	3	0	1	0	1	2	0	0	4	1	0	0	4	0	82	1	0	0	G
0	3	0	0	0	0	0	2	1	0	0	4	2	0	0	0	0	0	1	3	3	43	0	0	g
1	0	0	1	0	0	0	0	0	0	0	0	1	1	0	0	0	2	0	0	0	0	37	0	G#
0	0	0	0	2	0	0	0	0	3	0	0	0	0	2	0	0	0	0	0	0	0	0	11	g#

Table 4.4: Confusion matrix for key estimation using SVM. Small caps represent minor keys and we consider enharmonic equivalence.

4.5 Tonality tracking

In this section, we will study different ways of visualizing the evolution of the tonal content of a piece of music in audio format. Labelling a piece with a single key is often too simplistic in terms of tonal description. A musical piece rarely maintains a single tonality throughout its duration. There are also some pieces where there is no clear tonality, and the tonal center is constantly moving. The instantaneous evolution of the tonality of a piece and its strength can give a more powerful tonal description of it. Applications of this description include structural description, genre classification and music similarity. We have reviewed some of the approaches to locate modulations in Section 2.4.

We present here some algorithms that we have developed along this dissertation, which are intended to address the problem of analyzing the evolution of the tonal content of a piece of music. This research was made in collaboration with Jordi Bonada (see Gómez and Bonada (2005)). Our work is inspired in the work by Craig Stuart Sapp on harmonic visualization (see Sapp (2001)) and other approaches for tonality tracking of from symbolic, such as Chew and François (2003); Janata et al. (2002); Toiviainen and Krumhansl (2003), and acoustic representations Leman (1994, 1995a). We extend these ideas to the analysis of audio signals. Working directly with audio avoids the need of score transcriptions, being suitable for pieces where the score is unknown, as it often occurs. We introduce a set of additional visualizations which are specific to the analysis of audio.

As mentioned in Section 2.4, less research has been devoted to locate modulations than to estimate the global tonality of a piece. There have been some attempts, but it still remains a difficult task and quite hard to evaluate. The first problem to solve when trying to segment a piece according to its key is how to correctly identify regions of stable key centers and regions of modulations. Some approaches apply a sliding analysis window to the piece to generate a set of localized key estimations Chew (2004). This set of measures gives a good description of the key evolution of the piece, but calls for the setting of a suitable window size, which normally depends on the tempo, musical style and the piece itself.

4.5.1 Sliding window approach for tonality tracking

We can track the correlation of the average HPCP in a given temporal window with the possible minor and major keys, using the same frequency resolution than for computing the HPCP. This idea was considered by Sapp with a representation of the 'clarity' of the key, as explained in Sapp (2001). However, it only represented the relationship within the second most correlated key without explicitly indicating both keys. In our approach, we consider the correlation of the average HPCP within a certain analysis window with a set of tonal profiles, which are the ones presented in Section 4.3.3. This representation allows to compare the key estimation in a certain temporal window with the global key estimation. This is illustrated in Figure 4.27. The window size is a parameter that can be changed through the user interface and that may depend on the analyzed piece (e.g. on the tempo or the speed of modulation). This view is useful to compare different tonal profiles and distance measures. We can also sort the keys, and hence the coloring scheme, according to the

circle of fifths, as shown in Figure 4.27 (bottom).

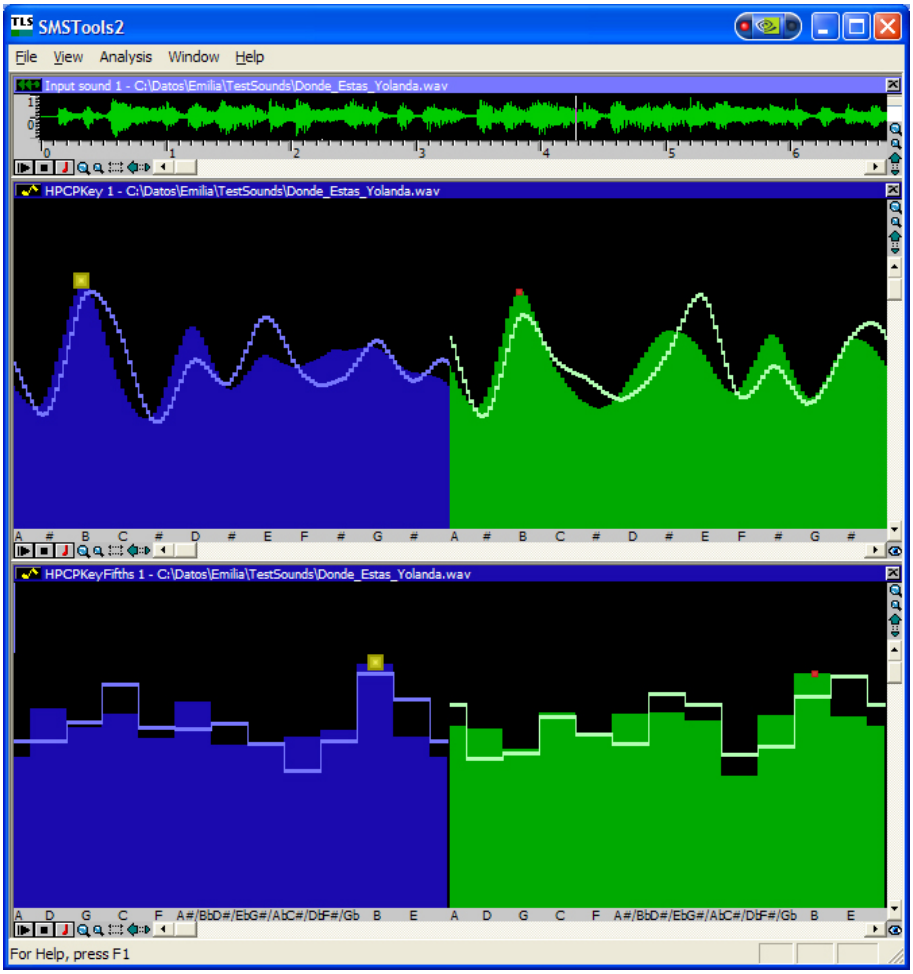


Figure 4.27: Key Correlation with major (left) and minor (right) keys. The horizontal axis represents the pitch class in chromatic scale from A to G# (top) and in the circle of fifths from A to E (bottom). Filled bars indicate the estimation on the current instant, while non filled bars represent the global estimation. The square represents the maximum correlation value, equivalent to the estimated key (B Major in this example).

We can also track the key correlation values using a rectangular representation, where the points for the 24 major and minor keys are located on the surface of a torus (as proposed in Krumhansl (1990) pp. 46). Figure 4.28 shows an example. This representation provides a musical meaning, given that each key is located near its close tonalities (relatives, parallel and neighbors in the circle of fifths) in this space.

The previous representations do not provide information about the temporal evolution of the tonality. For this purpose we have developed the KeyGram, shown in Figure 4.29. The KeyGram is inspired in Leman (1994) and represents the temporal evolution of the key correlations. The KeyGram is very relevant to track

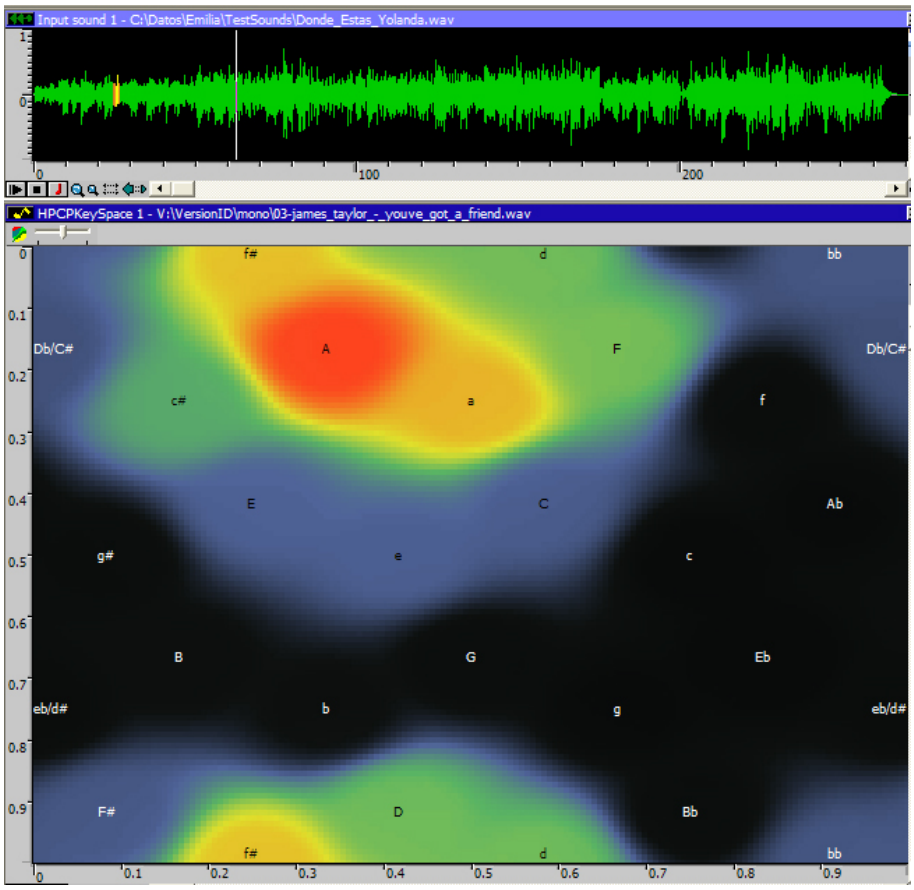


Figure 4.28: Key Correlation in the surface of a torus, as proposed in Krumhansl (1990) pp. 46.

modulations within a piece.

This evolution can also be displayed in the surface of a torus (as in Figure 4.28), defining a trajectory for tonality evolution. An example is shown in Figure 4.30. The size of the sliding window used for key estimation can be also changed depending on the piece, providing a way to navigate through different temporal scopes. This view is useful to study the evolution of the key and its strength along the piece, to determine the most suited window size for each piece, and to investigate different distance measures and tonal profiles.

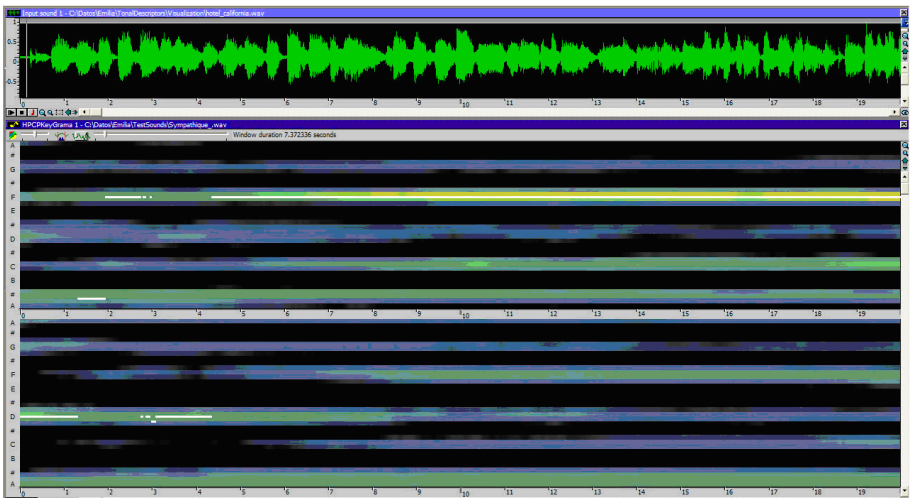


Figure 4.29: KeyGram. Instantaneous correlation with major (top) and minor (bottom) keys. White color indicates the highest correlation.

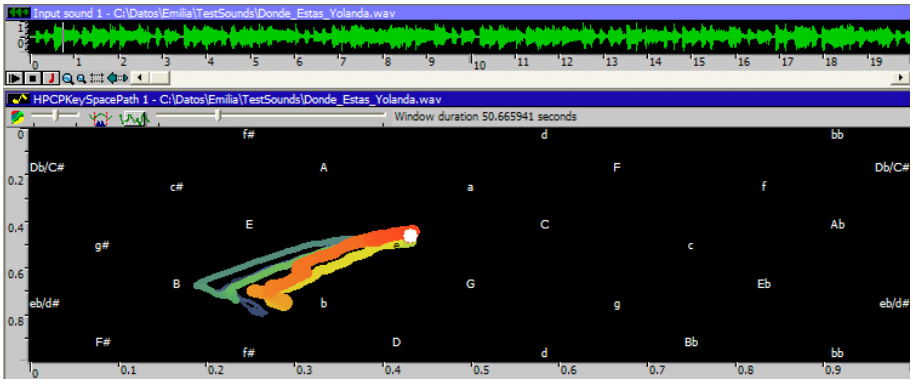


Figure 4.30: KeyGram in the surface of a torus.

4.5.2 Multiresolution description

As mentioned in Section 2.4, many approaches apply a sliding analysis window to the piece to generate a description of the key evolution of the piece (Shmulevich and Yli-Harja (2000) and Chew (2004)). In these approaches, the most problematic issue is the setting of a suitable window size, which normally depends on the tempo, musical style and the piece itself. According to Shmulevich and Yli-Harja (2000), there seems not to be a general rule for choosing the length of the sliding window.

We introduce here a multi-resolution representation of the estimated tonality, in order to visualize the evolution of the key (presented above) within different temporal scopes. To display tonal data in a compact visual manner, each key is mapped to a different color. We added to the original color representation in Sapp (2001) the distinction between major and minor keys by assigning different brightness (brighter for minor keys). The colors are represented in Figure 4.31. It is also possible to order the keys, and hence the coloring scheme, according to the circle of fifths.

Figure 4.31: Colors used to represent Major (top) and minor (down) tonalities.

The multiresolution diagram displays the tonality using different temporal scales. Each scale is related to the number of equal-duration segments in which the audio signal is divided to estimate its tonality. This way, the top of the diagram shows the overall tonality, the middle scales identify the main key areas present in the piece and the bottom scale displays chords. An example is presented in Figure 4.32.

4.5.3 Tonal contour

Pitch intervals are preferred to absolute pitch in melodic retrieval and similarity applications, given the assumption that melodic perception is invariant to transposition. We extend this idea to tonality, after observing that different versions of the same piece share the same tonal evolution but can be transposed. This is usually made to adapt the song to a singer or instrument tessitura. This view provides a relative representation of the key evolution. The distance between consecutive tonalities is measured in the circle of fifths: a transition from C major to F major is represented by -1, a transition from C major to A minor by 0, a transition from C major to D major by +2, etc. Figure 4.33 represents the tonal contour of the beginning of the song 'Imagine' by John Lennon, using a sliding window of 1 s. The estimated chord is moving from C major to F major, giving contour values equal to 0 and -1. A drawback of this contour representation is that relative major and minor keys are considered equivalent, which results in these modulations not being shown. This problem can be solved by using a representation of the torus KeyGram display shown in Figure 4.30 centered in the graphic.

The size of the sliding window can be adjusted, providing a way to navigate through different temporal scopes. It is possible then to visualize, for instance, chord progressions or key progressions. The goal of this

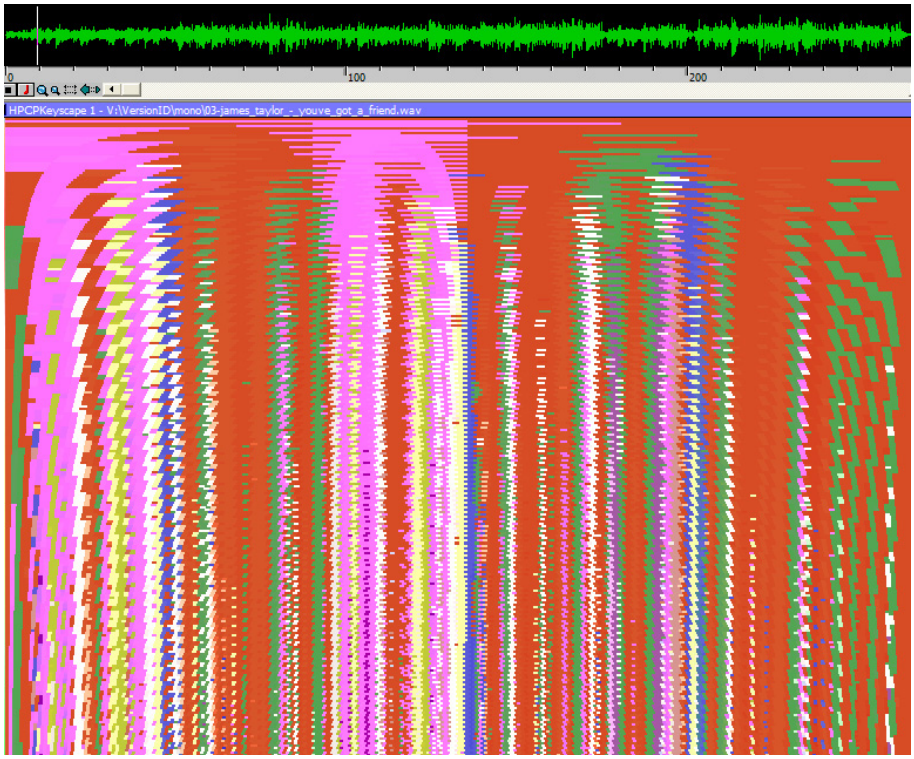


Figure 4.32: KeyScape from the song *You've got a friend* by James Taylor. The global tonality is A major.

view is to analyze the relative evolution of the key as a valid descriptor for tonal similarity between pieces.

4.5.4 Case study

The evaluation of key tracking systems is a hard task. First, it is very difficult to find a corpus of annotated material with musical analysis in terms of modulations, specially for other genres than classical music. As mentioned in Section 2.4, the annotation and segmentation process can be very difficult, as it is difficult to correct identify regions of stable key centers and regions of modulations. Some pieces may also present regions of ambiguous key. Second, it is hard to define a set of parameters that work well in different situation, as for instance the size of the sliding window used for key estimation (mentioned in Shmulevich and Yli-Harja (2000)). A quantitative evaluation of the accuracy of our approach for audio key tracking remains a future work of this dissertation.

We show in this section how these different visualization are useful when analyzing a piece of music, taking as an example a song from The Beatles: *When I am sixty four*, composed in 1996. This piece is analyzed by an expert musicologist in terms of key and chord evolution in Pollack (1999).

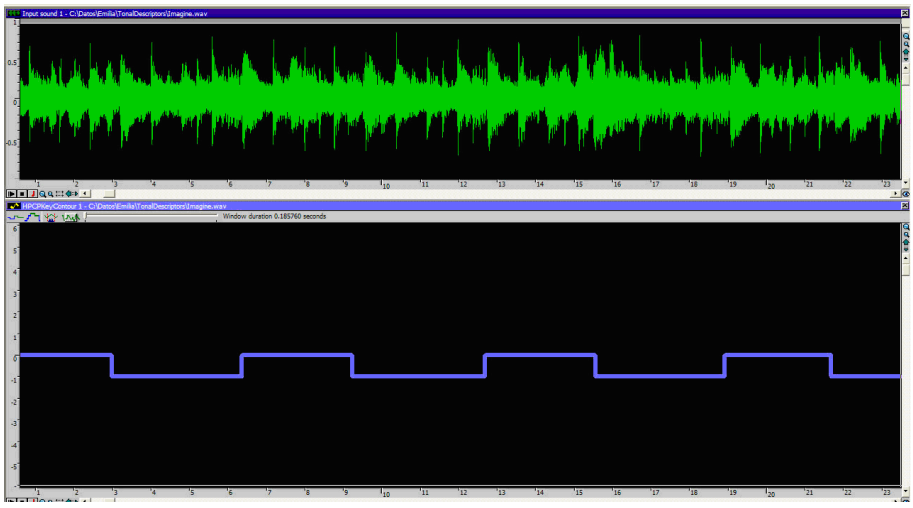


Figure 4.33: Tonal contour.

According to Pollack, *this song is Paul's first official foray into the carefully put-on nostalgic-cum-vaudeville stylization that would become a stock part of his compositional arsenal for the remainder of his career as a Beatle*. In terms of tonality and chords, the piece is in Db Major tonality, including some modulations to its relative minor key. Pollack proposes the following labelling for chords and key changes, presented in Figure 4.34 and Table 4.5. This tonal description has been checked by the author of this thesis and by Chris Harte, from QMUL (personal communication). The analysis outputs the results seen in Figures 4.35, 4.36, 4.37, 4.38 and 4.39, where we can see that the analysis agrees with the manual annotation of the key evolution.

Begin (seconds)	**End (seconds)**	**Key**
0	0.34	None
0.34	38.41	Db Major
38.41	58.72	Bb Minor
58.72	94.49	Db Major
94.49	114.67	Bb Minor
114.67	156.8	Db Major
156.8	157.62	None

Table 4.5: Annotated key evolution of the song *When I am sixty four* by The Beatles.

4.6 Conclusions

In this chapter, we have verified the performance of the proposed approach and the validity of the adaptation and comparison steps, obtaining an accuracy around 76.1% for a varied set of styles, instrumentation and

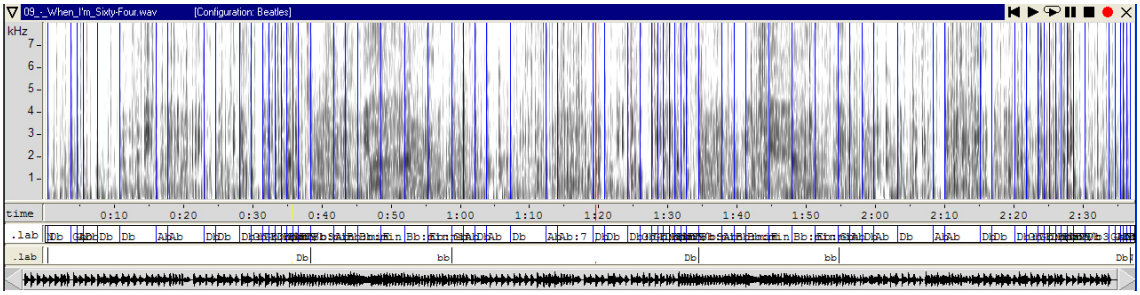


Figure 4.34: Annotated chord (top) and key (bottom) evolution within the song *When I am sixty four* by The Beatles.

recording conditions. We could argue that current results are comparable to human ratings, if we contrast this number with the studies made by Cohen (1977), in which university music majors are able to sing (though not name) the key with a 75% accuracy (see Chapter 1 for further details).

As a conclusion of the study presented in Section 4.3.3, we see that tonal models which are obtained after psychological studies (as the ones derived from Krumhansl) with the adaptations proposed in this work seem to work better than empirically obtained ones, with the exception of the statistical average profiles computed from THPCP values. This indicates that tonal profiles obtained after psychological studies should be adapted to work with HPCP features in a polyphonic situation. A second conclusion is the dependency of musical genre. These tonal models have been designed to work for classical western music, and we verify that these models should be adapted to other musical genres.

We have also studied which is the influence of the tonal model used within the different musical styles in Section 4.3.4, verifying the importance of the tonic triad for popular music mentioned by Temperley (2001). In Section 4.3.5, we have verified that the main key of a piece is often located at the beginning or the end of a piece, not in the middle, and that it is usually necessary to only analyze a segment of 15 or 20 seconds instead of the whole piece.

Finally, in Section 4.4, we have studied the benefits of using machine learning tools to learn the tonal model from annotated data. We saw that comparing an approach based on the correlation with fixed major and minor profiles to the machine learning techniques that we can consider as "tools of the trade", slight improvements in performance can be achieved by the latter alone. The combination of both approaches, by embedding the estimation based on profile correlations as an additional feature for the machine learning algorithm, did not lead to a better performance. As it is pointed out by Krumhansl, the tonal descriptors we have considered are severely restricted, in the sense that they do not capture any musical time structure. There features take into account neither order information nor the chords' position in the event hierarchy, as for instance, its place in the rhythmic or harmonic structure (see Krumhansl (1990), pp. 66). In fact, some of the estimation errors may be caused by tonality changes that affect the overall key measurement and labelling.

We considered then that a single key measurement is often too simplistic as a tonal description, and for

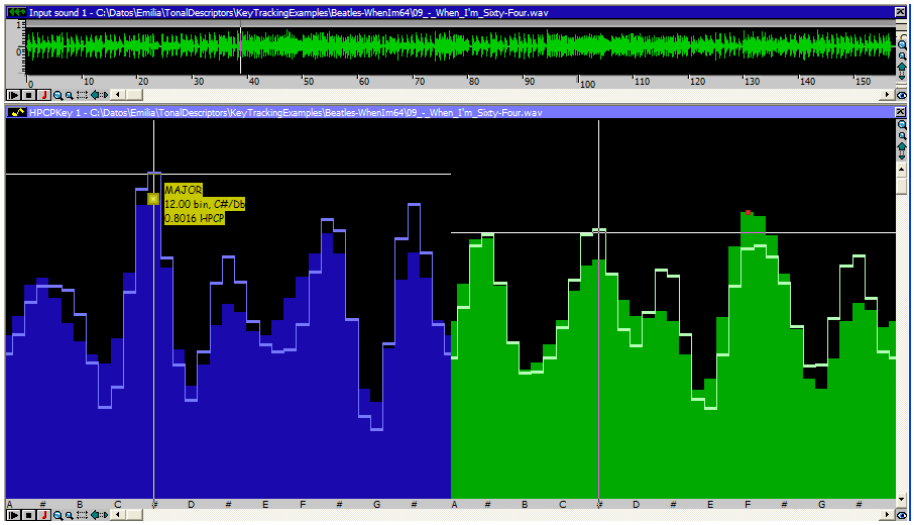


Figure 4.35: Global (empty line) and instantaneous (filled line) key estimation for song *When I am sixty four* by The Beatles. The top figure displays the audio signal, and the bottom figure represents the correlation with major (left) and minor (right) keys. The squared point represent the maximum correlation value, which corresponds to the estimated key (Db major).

this reason we have presented different ways of visualizing the tonal content of a piece of music by analyzing audio recordings. These diverse views can be combined in varied ways depending on the user needs. We offer also the possibility of visualizing two different pieces simultaneously, which can serve to study tonal similarity and define distances between pieces based on tonal features. One possible application that we will discuss in Chapter 5 is how these descriptors can be used to measure similarity between audio recordings and to organize digital music collections.

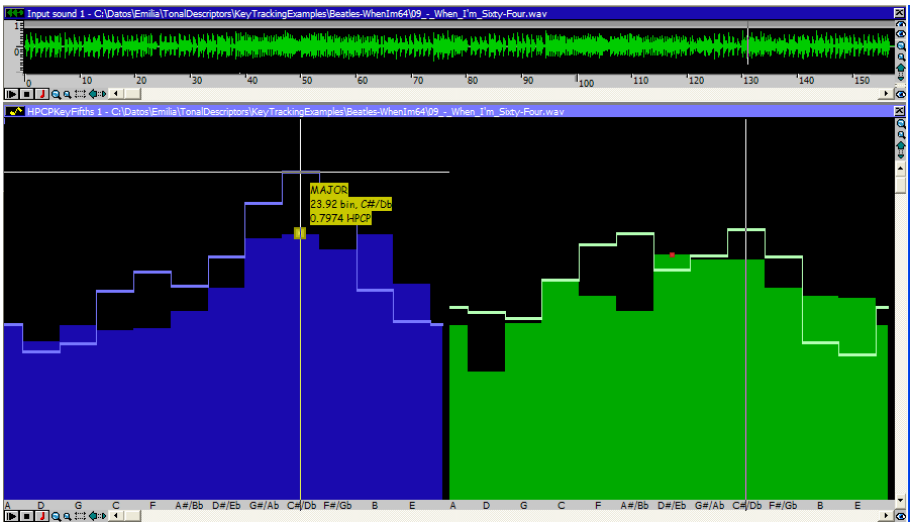


Figure 4.36: Global key estimation in the circle of fifths for song *When I am sixty four* by The Beatles.

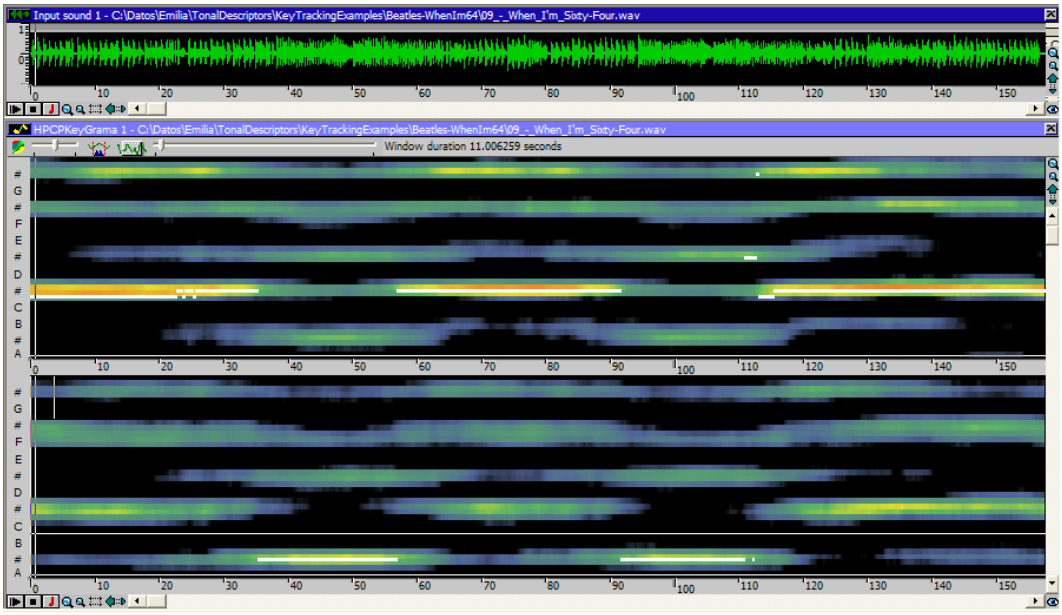


Figure 4.37: Estimated key evolution for the song *When I am sixty four* by The Beatles. The figure represents the audio signal (top), and the temporal evolution of the correlation with major (middle) and minor (bottom) keys. The white line indicates the maximum correlation value, which corresponds to the estimated key

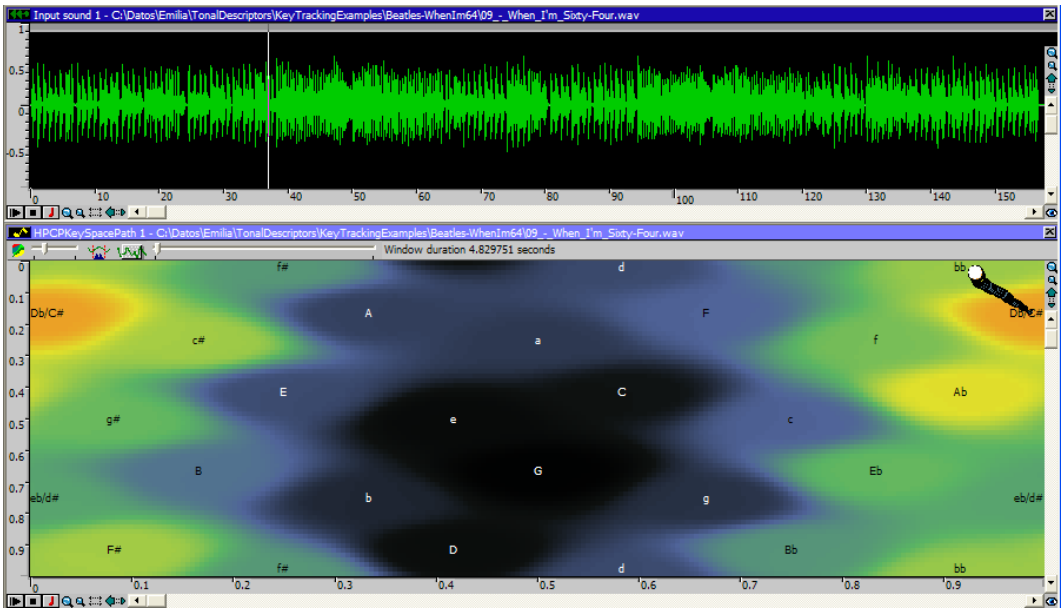


Figure 4.38: Estimated key evolution in the surface of a torus for the song *When I am sixty four* by The Beatles.

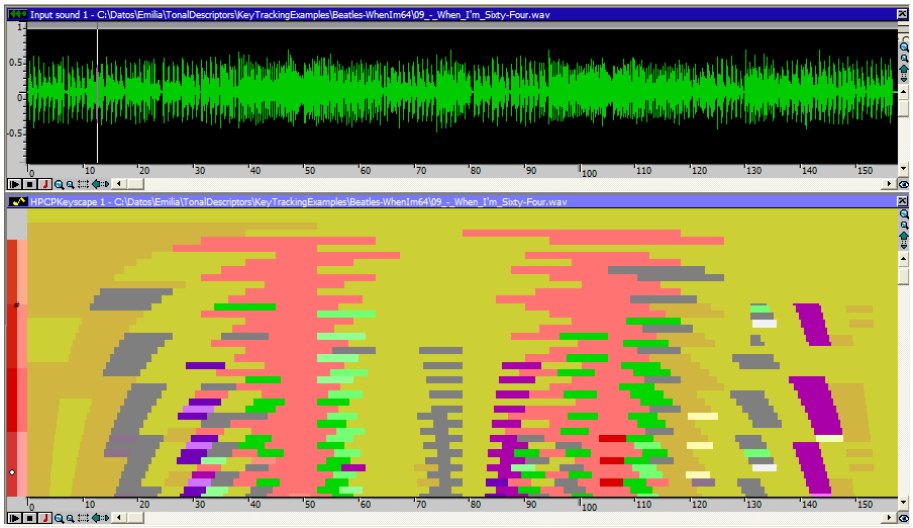


Figure 4.39: Estimated multiresolution key evolution for the song *When I am sixty four* by The Beatles.

Chapter 5

Tonality for music similarity and to organize digital music collections

5.1 Introduction

Along Chapter 3 and 4, we have first proposed a set of low-level tonal descriptors (HPCP) representing the pitch class distribution of a musical piece. Then, we have studied how these tonal descriptors are useful to estimate the key of a music piece in audio format. In this Chapter, we investigate some other uses of tonal description of audio. Here, we present two application contexts in addition to key estimation that can benefit from the proposed tonal description. We analyze how the tonal features proposed in previous chapters can be used for two different purposes: first, to compare two different musical pieces from the analysis of audio, and second to organize and navigate through compilations of audio files. For this study, we present a small experiment where tonal features can be used to measure similarity between two different pieces, and we show how these descriptors are integrated in a system for organizing and exploring digital music collections in audio format.

This chapter is divided in two main blocks. Section 5.2 focuses on the analysis of tonal similarity and its application to the identification of different versions of the same piece. We study here in which sense version identification is a complex problem that requires a multifaceted and multilevel audio description (as presented in Section 1.3), and we present a small experiment showing that tonal descriptors by themselves can be helpful for this task.

Section 5.3 presents an example of the integration of the proposed tonal descriptors proposed in a system for music organization and recommendation based on the analysis of audio signals. We present some examples on how it is possible then to organize a music collection according to key, tonal profile or key strength. Here, we would like to point the difficulty of evaluating this type of systems. Even though an evaluation of the usefulness of these descriptors by means of human ratings has not been carried out (remaining as a future

work of this dissertation), this chapter contributes to justifying that tonal description is valid to index musical collections and perform search by similarity, which is one of the goals of this dissertation.

5.2 Tonal similarity and version identification

The possibility of finding "similar" pieces is one of the most attractive features that a system dealing with large music collections can provide. Similarity is an ambiguous term, and music similarity is one of the most complex problem and most difficult to quantify in the field of MIR. Music similarity is a property which may depend on different musical, cultural and personal aspects. Many studies in the music information retrieval literature have been trying to define and evaluate the concept of *similarity*, that is, when two pieces are considered to be *similar*. There are many factors involved in this problem, and some of them (maybe the most relevant ones) are difficult to measure.

There have been some proposals on how to compute similarity from audio signals. Many approaches are based on timbre similarity, such as Pampalk (2004) and Aucouturier and Pachet (2004). They propose the use of some similarity measures from audio which are related to timbre low-level features, mainly MFCCs. Pampalk et al. (2005) include the use of some low-level features which are related to spectral similarity (based on MFCCs) and fluctuation patterns, representing loudness fluctuations in different frequency bands.

Other approaches focus on the study of rhythmic similarity. Foote et al. (2002) proposes some similarity measures based on the "beat spectrum", including Euclidean distance, a cosine metric or inner product. Tempo is also used for measuring similarity between songs in Vignoli and Pauws (2005).

The evaluation of similarity measures is a hard task, given the difficulty of gathering ground truth data for a large quantity of material. Logan and Salomon (2001) present the evaluation of a similarity measure based on spectral features. An "objective" evaluation experiment is based on the assumption that songs from the same style, by the same artist and on the same album are similar. This is also the approach in Pampalk et al. (2003). Berenzweig et al. (2003) focus only on the problem of artist similarity. They gather ground truth data from different sources: music experts (through The All Music Guide [1]), a survey (intended to rank similarity between artists) and co-occurrence of artists in playlists and user collections. A direct way to measure the similarity between songs is to gather ratings from users (see Vignoli and Pauws (2005)), which is a difficult and time-consuming task. Vignoli and Pauws (2005) allow the user to weight the contribution of timbre, genre, tempo, year and mood to the final similarity measure.

Tonality has not been much studied within the music similarity literature, mainly because it might be not so clear as timbre or rhythmic similarity for people not having a musical background. Up to our knowledge, there is no study on how tonal descriptors are relevant for audio-based music similarity. We focus here on this aspect of music similarity, analyzing how tonal descriptors can be used to measure similarity between pieces. We restrict then our consideration to tonal similarity, which means the following: two pieces are *tonally* similar if they share a similar tonal structure, related to the evolution of chords (harmony) and key. We see

[1] http://www.allmusic.com

below that *tonal similarity* is not equivalent to *similar key* if we consider that the cognition of melodies and harmonies is not absolute but interval dependent. As mentioned in Chapter 3, experiments have shown that interval is an important element for melody recognition (see Dowling (1978)), so that contour representations are used for melodic retrieval. We extend this idea to tonal similarity, where pieces are considered to be similar if they share the same tonal contour. We analyze how only tonal descriptors can help to locate different versions of the same piece in audio format. Here, we consider that versions of the same piece share the same overall harmonic structure or *tonal contour*.

When dealing with huge music collections, version identification is a relevant problem, because it is usual to find different versions of the same song. We can identify different **situations** where a song is versioned in the context of mainstream popular music, which are listed here:

1. **Re-mastered track** obtained after digital re-mastering of an original version, where we can find some slight differences between the original and the re-mastered piece.

2. **Karaoke version** or a version of a song translated to a different language. Usually, the instrumental tracks are preserved, and then the tempo. There is only a change on the leading voice.

3. **Recorded live track**: a recorded live track is a song or audio sequence recorded from live performances. Here, there is a change on the recording conditions, as there is typically the noise of hall cheering. The original score is here preserved. In principle, the instrumentation, rhythm and harmony are usually kept, even though there can be some differences in the tempo and structure of the piece (for instance, repetitions, intro, etc). This modification is usually performed by the same original band. Although recorded versions of songs are sometimes redundant and inaccurate with respect to original studio, they have quite a bit of popularity, and even sometimes some studio recorded tracks try to imitate being live tracks.

4. **Acoustic track**: in some situations, the piece is played using different instrumentation than the original song. Due to this fact, there are changes in timbre. Finally, but not so usually, we can have some variations of the pitch range, in order to adapt the piece to different instruments. An extension of this situation could be an **extended track** or a **disco track** by the same band but having just a different drum pattern.

5. **Cover version**: in some situations, a given artist performs a song from a different one, as explained in Witmer and Marks (2006). Pop musicians sometimes play covers as a tribute to the original performer or group, or in order to win audiences who like to hear a familiar song, to increase their chance of success by using a proven hit or to gain credibility by its comparison with the original song. In other situations, some companies re-edit cover versions of know songs in order to avoid paying the performers' copyright. One can see the relevance of cover songs by looking at the *Second Hand Songs* database[2], which already contains around 37000 cover songs. In this situation, there can be different

[2]http://www.secondhandsongs.com

levels of musical similarity between pieces. There may be changes on instrumentation, tempo, structure and even harmony. This is the case, for instance, of some jazz versions of classical pieces, where there are also changes on the harmonization of the piece.

6. **Remix**: this term stands for a recording produced by combining sections of existing tracks with a new structure and new material. As explained in Fulford-Jones (2006), it can also be a radical transformation of the original work, leaving very little of the original recording.

A song can be versioned in different ways, reaching different degree of disparity between the original and the versioned tune. The musical characteristics of the piece that are usually modified can be divided in the following categories:

1. Alteration of acoustic characteristics: due to modifications on the recording conditions.
2. Tempo adjustment.
3. Change of instrumentation: leading voice, different instruments, added drum track, etc.
4. Change of structure, by introducing an instrumental part, intro or adding repetitions of the chorus.
5. Transformation on the pitch range i.e. transposed version.
6. Harmonic differences, which can consist on slight variations keeping the same harmonic structure or a completely different harmonization of the piece.

The degree of disparity on the different aspects establishes a vague boundary between what is considered a ”version” and a ”different song”. This frontier is difficult to define, and it is an attractive topic of research from the perspective of intellectual property rights and plagiarism. This complexity makes it difficult to develop algorithms to automatically identify versions with absolute effectiveness.

In this dissertation, we propose an approach to automatically identify versions of the same song using only tonal features. Our assumption is that the main tonal evolution of a piece is kept from a song to its versions. Variations are present in tempo, tessitura and instrumentation. Some structural and slight harmonic variations are also considered. Our approach is based on the analysis of polyphonic audio recordings for the extraction of a set of tonal features. We will present some experimental results, and we will evaluate it using a small database of 90 polyphonic pieces, which are versions of 29 different music titles of mainstream popular music.

There is also a small amount of literature in the field of music information retrieval related to the identification of versions of the same piece using polyphonic audio. This is due partly to the complexity of establishing in a general way which makes a certain piece be a version instead of a different composition. This complexity already appears in the context of comparing scores, where timbre is not considered. Cheng Yang (2001) proposed an algorithm based on spectral features to retrieve similar music pieces from an audio database. This method considers that two pieces are similar if they are fully or partially based on the same

score, even if they are performed by different people or at different tempo. A feature matrix was extracted using spectral features and dynamic programming. Yang evaluated this approach using a database of classical and modern music, with classical music being the focus of his study. 30 to 60 second clips of 120 music pieces were used. He defined five different types of ”similar” music pairs, with increasing levels of difficulty:

- Type I. Identical digital copy.
- Type II. Same analog source, different digital copies, possibly with noise (differences in recording conditions).
- Type III. Same instrumental performance, different vocal components. Here, the considered difference in instrumentation was related to the singing voice.
- Type IV. Same score, different performances (possibly at different tempo). The considered difference in rhythm was only related to tempo.
- Type V. Same underlying melody, different otherwise, with possible transposition.

The proposed algorithm performed very well (90% accuracy) in the first 4 types, where the score is the same and there are also some tempo modifications. This is the most common situation for different performances of classical pieces. On the same idea, Purwins et al. (2000) calculate the correlation of Q-profiles for different versions of the same performance: different performers and different instruments (piano and harpsichord). In this dissertation, we have also showed (see Chapter 3) how the proposed low-level tonal features perform well for classical music. Now, we try to extend this study to more complex situations that appear when analyzing mainstream popular music.

5.2.1 Similarity using global tonal descriptors

As mentioned above, we can find different situations of audio recordings being similar because they come from the same *root* musical piece. We consider that the *root* piece is the first recorded version of a musical work, which is more close to the original composition. The different situations where versions are generate can be analyzed with respect to the musical features which are transformed from one root piece to its version. In addition to the four types proposed in Yang (2001), we add a situation that establishes a dissimilarity baseline, as well as two additional situations (Type V and VI) where the key of the compared pieces are different. Following this schema, each category refers to one musical aspect which is changed when elaborating a version of the piece (noise, instrumentation, tempo, harmonization, key and structure).

We try to isolate in the following examples the main musical characteristics of a piece which are modified from a piece to is version. We will see below that these modifications usually happen together in versions from popular music pieces.

- **Type 0: different pieces**. We analyze here which is the baseline for similarity comparison, i.e. how the proposed low-level tonal features correlate for different pieces. Figure 5.1 shows the HPCP average

vectors for two distant pieces, which is is equal to 0.0069. This small value indicates the dissimilarity between the profiles, and will be considered as a baseline for tonal similarity.

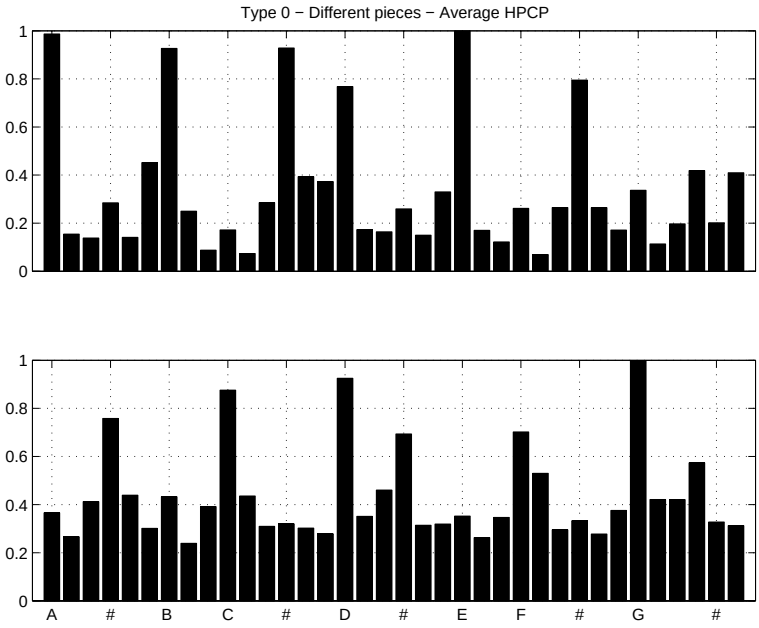


Figure 5.1: HPCP average for two distant pieces (*e43* top and *e114* bottom). Type 0. See *e114.wav* and *e043.wav* in Appendix A for further details on the sound excerpts.

- **Type I: identical digital copy**: in this case, as our approach for feature extraction is deterministic, the computed tonal features are identical and the similarity measures is equal to 1, its maximum value. We can consider that the proposed tonal descriptors can help to identify a given piece, useful for audio identification purposes. Cano et al. (2002) present some concepts and applications of audio fingerprinting.

- **Type II: noise**: in this situation, the analog source is the same, but there are different digital copies, possibly with noise. One example of this situation is a re-mastered track. We have studied in Chapter 3 how the proposed pitch class distribution features are robust to noise. The correlation between the HPCP average vectors, showed in Figure 5.2, for two pieces which differ in some noise (as defined by Yang (2001)), is equal to 0.9809. Both of them get the same estimated key, A major.

- **Type III: instrumentation**: in this situation, there is the same piece, with the same tempo, but with different vocal components and/or instruments. Here, the musical structure, the tempo and the overall score is kept, while there are differences in the vocal components or the instruments which are playing the piece. One example would be the situation of a karaoke or a disco version. We take as an example two pieces of popular music sung by different performers. The correlation between the HPCP average

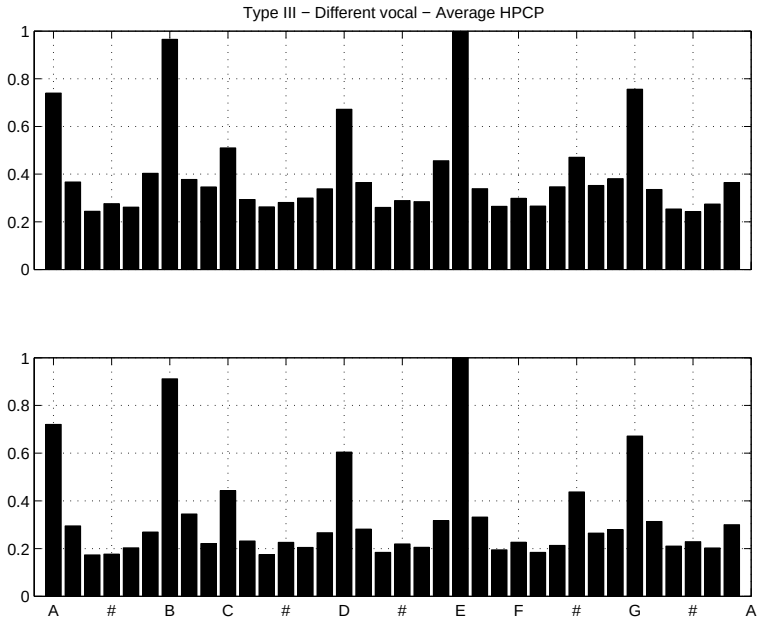


Figure 5.2: HPCP average for two digital copies from the same analog source (Type II). See *e043.wav* and *e108.wav* in Appendix A for further details on the sound excerpts.

vectors, showed in Figure 5.3, is equal to 0.9869. Both of them get the same estimated key, E minor. This situation can be also extended to different instruments playing the same piece. A change of instrumentation is sometimes linked to a change of tempo.

- **Type IV: tempo**: here, the piece is the same, as well as the instrumentation. The difference between pieces is the tempo in which they are performed. We compare in this situation two excerpts of classical pieces played at different tempi. The correlation between the HPCP average vectors, showed in Figure 5.4, for two pieces of classical music played at different tempi, is equal to 0.9540.

- **Type V: same underlying melody, different otherwise, with no transposition**. In this situation, the similarity between the pieces is lower, and there can be noise (Type II), change of instrumentation (Type III) and change of tempo (Type IV). We show here an example of the same piece arranged in two different styles, which may carry slight variations in harmonization, instrumentation and tempi. The correlation between the HPCP average vectors, showed in Figure 5.5, is equal to 0.9183. Both of them get the same estimated key, C minor.

- **Type VI: transposition**: in some of the situations, mainly in cover versions, the piece is transposed to a different key. This is usually done to adapt the pitch range to a different singer or instrument. In this type, we include the modifications mentioned above but with a possible transposition. We take as an example 6 versions of the first phrase of the song *Imagine*, by John Lennon, played by different

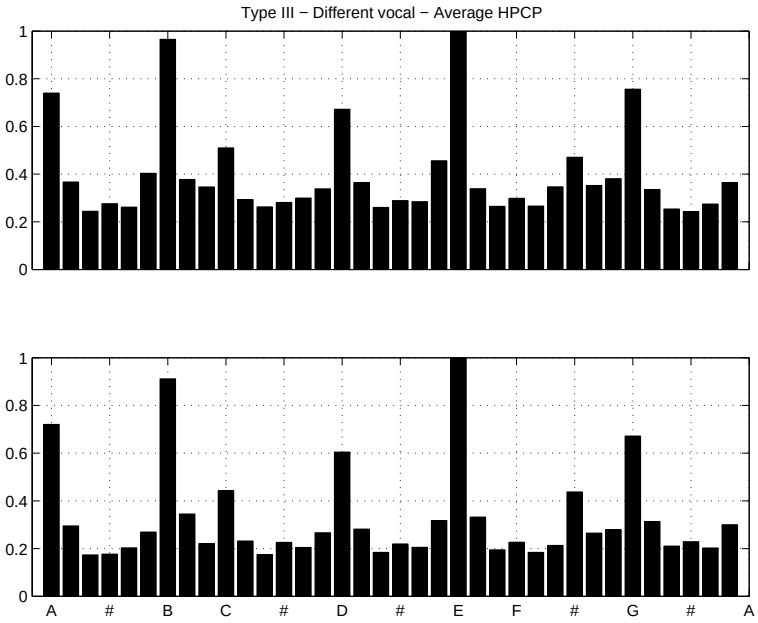


Figure 5.3: HPCP average for two versions of the same piece with different vocal components (Type III). See *e116.wav* and *e117.wav* in Appendix A for further details on the sound excerpts.

performers (1. John Lennon, 2. Instrumental, guitar solo, 3. Diana Ross, 4. Tania Maria, 5. Khaled, 6. Noa). Versions 1, 2 and 4 are in C major key, versions 3 is transposed to F major, and finally versions 5 and 6 are transposed to Eb major. We analyze only the first phrase so that all the excerpts have the same structure. Structural changes are studied below. HPCP average vectors are showed in Figure 5.6

The correlation matrix R between the average HPCP vectors for the different versions is equal to:

$$R = \begin{pmatrix} 1 & 0.97 & 0.82 & 0.94 & 0.33 & 0.48 \\ 0.97 & 1 & 0.86 & 0.95 & 0.31 & 0.45 \\ 0.82 & 0.86 & 1 & 0.75 & 0.59 & 0.69 \\ 0.94 & 0.95 & 0.75 & 1 & 0.18 & 0.32 \\ 0.33 & 0.31 & 0.59 & 0.18 & 1 & 0.95 \\ 0.48 & 0.45 & 0.69 & 0.32 & 0.95 & 1 \end{pmatrix} \tag{5.1}$$

We can see that there are some low values of correlation between versions, mainly for the ones which are transposed to Eb major (version 5 and 6), as this tonality is not as close to C major as F major is (for version 3). We could consider that some of the pieces are transposed with respect to some others. We can normalize the HPCP vector with respect to the considered key by ring-shifting the HPCP vectors in order to obtain its transposed version, THPCP (as described in Chapter 3). THPCP average vectors are

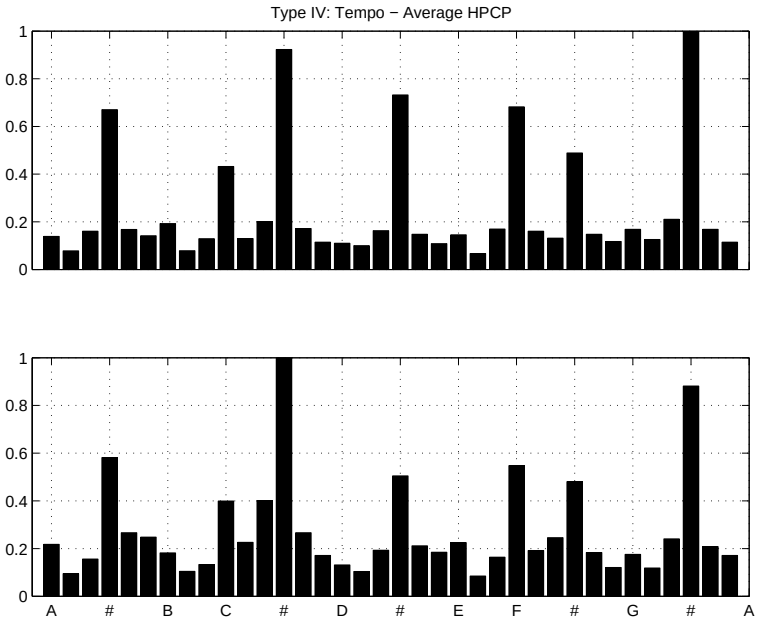


Figure 5.4: HPCP average for two versions of the same piece with different tempo (Type IV). See *e071.wav* and *e107.wav* in Appendix A for further details on the sound excerpts.

showed in Figure 5.7. The correlation matrix R_t between the average THPCP vectors for the different versions is equal to:

$$R_t = \begin{pmatrix} 1 & 0.97 & 0.97 & 0.94 & 0.94 & 0.97 \\ 0.97 & 1 & 0.98 & 0.95 & 0.91 & 0.98 \\ 0.97 & 0.98 & 1 & 0.92 & 0.95 & 0.99 \\ 0.94 & 0.95 & 0.92 & 1 & 0.86 & 0.94 \\ 0.94 & 0.91 & 0.95 & 0.86 & 1 & 0.95 \\ 0.97 & 0.98 & 0.99 & 0.94 & 0.95 & 1 \end{pmatrix} \tag{5.2}$$

This correlation matrix shows high values for the different versions, all of them higher than 0.8.

- **Type VII: structure**. This type represents possible structural changes. Until now, and in the work by Yang (2001), only audio excerpts were compared. When we compare complete songs in popular music, most of the versions have a different structure that the original piece, adding repetitions, new instrumental sections, etc. We take as an example 5 versions of the song *Imagine*, by John Lennon, played by different performers (1. John Lennon, 2. Instrumental, guitar solo, 3. Diana Ross, 4. Tania Maria, 5. Khaled and Noa). Contrary to Type VI, we analyze the whole piece and not just a phrase. Versions 1, 2 and 4 are in C major key, version 3 is transposed to F major, and finally versions 5 is

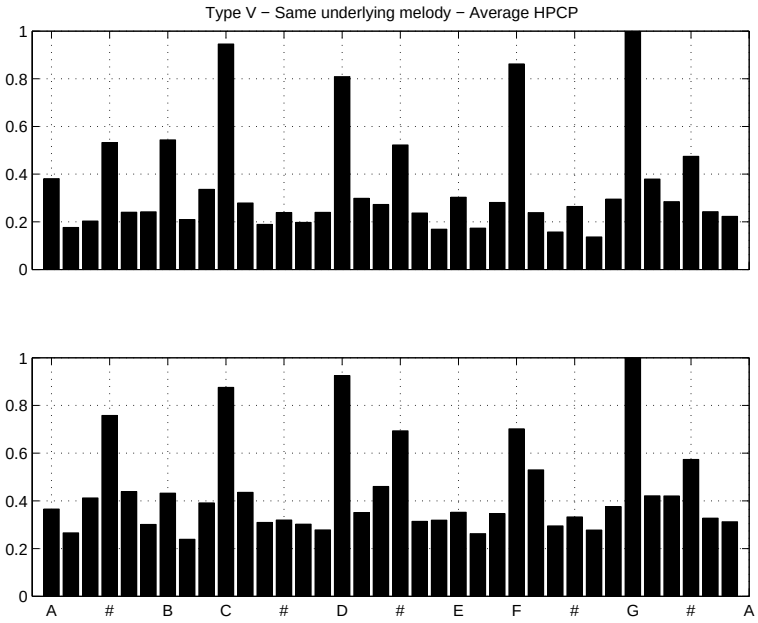


Figure 5.5: HPCP average for two versions of the same piece with the same underlying melody, different otherwise (Type V). See *e106.wav* and *e114.wav* in Appendix A for further details on the sound excerpts.

transposed to Eb major. The correlation matrix R between the average HPCP vectors for the different versions is equal to:

$$R = \begin{pmatrix} 1 & 0.99 & 0.83 & 0.96 & 0.45 \\ 0.99 & 1 & 0.86 & 0.95 & 0.45 \\ 0.83 & 0.86 & 1 & 0.79 & 0.65 \\ 0.96 & 0.96 & 0.79 & 1 & 0.35 \\ 0.45 & 0.45 & 0.65 & 0.35 & 1 \end{pmatrix} \tag{5.3}$$

We can see that the correlation is lower for the piece in distant key, which is version 5 in Eb major. We can again normalize the HPCP vector with respect to the key. THPCP average vectors are showed in Figure 5.8. The correlation matrix R_t between the average THPCP vectors for the different versions is equal to:

$$Rt = \begin{pmatrix} 1 & 0.99 & 0.98 & 0.96 & 0.98 \\ 0.99 & 1 & 0.99 & 0.95 & 0.98 \\ 0.98 & 0.99 & 1 & 0.95 & 0.99 \\ 0.96 & 0.95 & 0.95 & 1 & 0.95 \\ 0.98 & 0.98 & 0.99 & 0.95 & 1 \end{pmatrix} \tag{5.4}$$

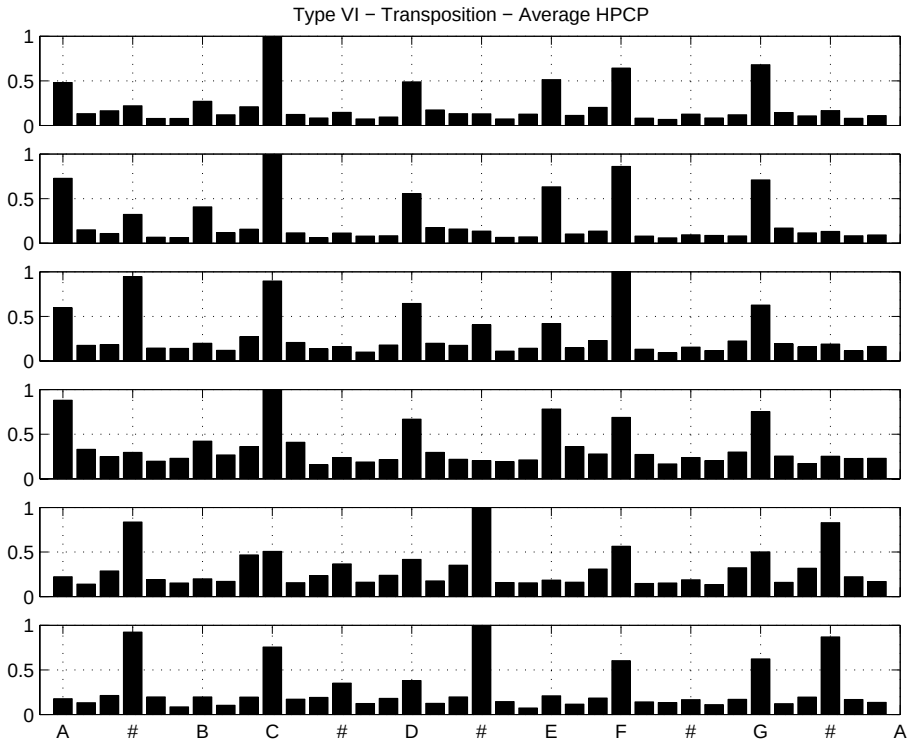


Figure 5.6: HPCP average for 6 different versions of the first phrase of the song *Imagine*, by John Lennon (Type VI). See *ImagineFirstPhrase1-JohnLennon.wav*, *ImagineFirstPhrase2-Instrumental.wav*, *ImagineFirstPhrase3-DianaRoss.wav*, *ImagineFirstPhrase4-TaniaMaria.wav*, *ImagineFirstPhrase5-Khaled.wav* and *ImagineFirstPhrase6-Noa.wav* in Appendix A for further details on the sound excerpts.

We observe that the correlation values increase for version 5. In this situation, it becomes necessary to look at the structure of the piece, which will be done in next section. When the pieces under study have different structures, we study the temporal evolution of tonal features, in order to locate similar sections. Table 5.1 summarizes the different defined types according to the musical feature which is mainly transformed. We have seen above that it is usual to find several musical features that change from a piece to its version.

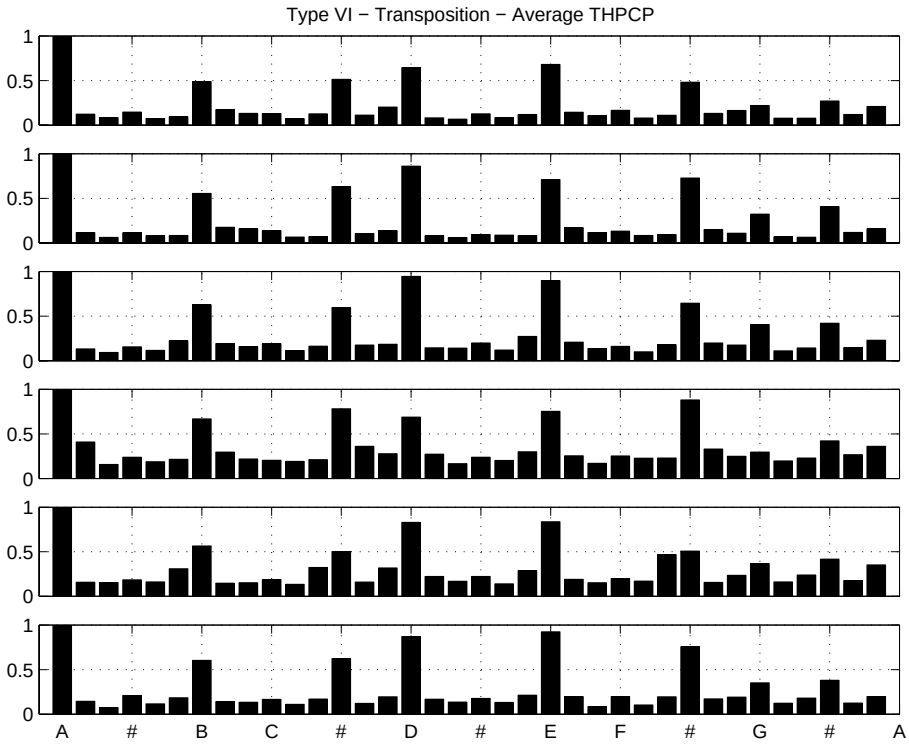


Figure 5.7: THPCP average for 6 different versions of the first phrase of the song *Imagine*, by John Lennon (Type VI). See *ImagineFirstPhrase1-JohnLennon.wav*, *ImagineFirstPhrase2-Instrumental.wav*, *ImagineFirstPhrase3-DianaRoss.wav*, *ImagineFirstPhrase4-TaniaMaria.wav*, *ImagineFirstPhrase5-Khaled.wav* and *ImagineFirstPhrase6-Noa.wav* in Appendix A for further details on the sound excerpts.

Type	**Main transformed musical feature**
0	All (different pieces)
I	None (identical pieces)
II	Noise
III	Instrumentation
IV	Tempo
V	Harmony
VI	Key
VII	Structure

Table 5.1: Classification of versions according to the musical feature which is mainly transformed. It is usual to find several features transformed at the same time.

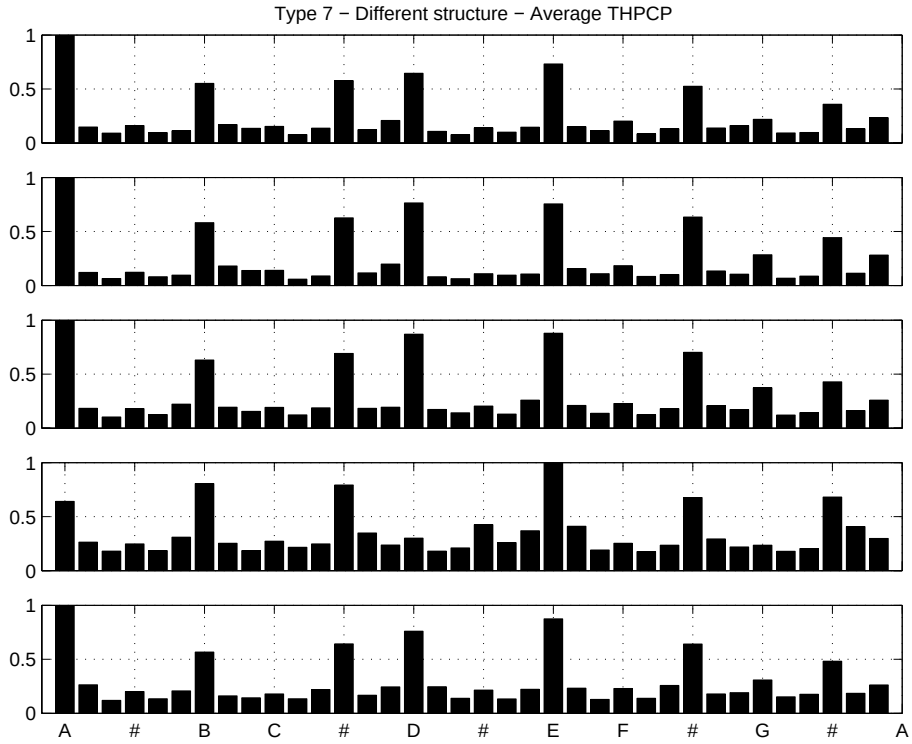


Figure 5.8: THPCP average for 5 different versions of the song *Imagine*, by John Lennon (Type VII). See *Imagine1-JohnLennon.wav*, *Imagine2-Instrumental.wav*, *Imagine3-DianaRoss.wav*, *Imagine4-TaniaMaria.wav*, *Imagine5-NoaKhaled.wav* and *ImagineFirstPhrase6-Noa.wav* in Appendix A for further details on the sound excerpts.

5.2.2 Similarity using instantaneous tonal descriptors

As seen in the last example, it becomes necessary to look at the structure of the piece when looking for similar songs. Structural description is a difficult problem, and many studies have been devoted to this issue. We refer here to some recent work by Chai (2005) and Ong and Herrera (2005). It is important to consider the structure when dealing with version identification. We analyze here the last example consisting on 5 different cover versions of the song *Imagine*, by John Lennon. These versions are played by different performers: 1. John Lennon (original version), 2. Instrumental (guitar as leading instrument), 3. Diana Ross, 4. Tania Maria (jazz version having a different melody and harmonization) and 5. Khaled and Noa. Versions 1, 2 and 4 are played in C major key, while version 3 is transposed to F major and version 5 is transposed to Eb major.

Foote (1999) proposed the use of self-similarity matrices to visualize music. Similarity matrices were built by comparing Mel-frequency cepstral coefficients (MFCCs), representing low-level timbre features. Here, we extend this approach to low-level tonal descriptors representing the pitch class distribution of the piece, being independent with respect to timbre. Figure 5.9 represents the self-similarity matrix for the original version of *Imagine* by John Lennon, using the proposed low-level tonal descriptors normalized with respect to the key, THPCP, as explained in Section 3.7. This similarity measure is computed using correlation between the average of the THPCP profile over a sliding window of 10 seconds. This window is large enough to include more than one chord. The time difference between consecutive windows is equal to 1 second, in order to track the chord evolution.

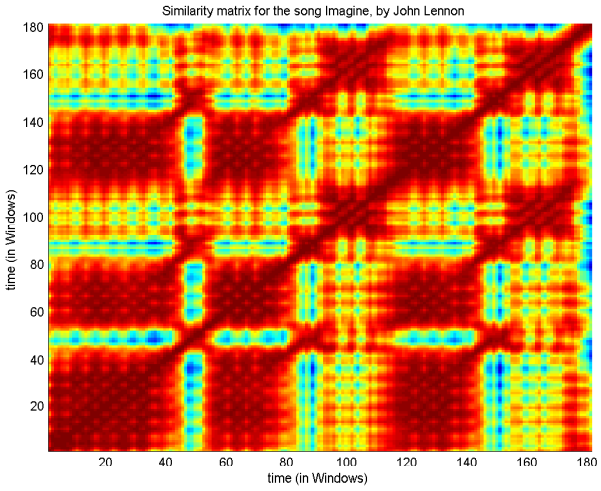


Figure 5.9: Similarity matrix for the song *Imagine*, by John Lennon, using THPCP. See *Imagine1-JohnLennon.wav* in Appendix A for further details on the sound excerpts.

In this self-similarity matrix, we verify that the diagonal is equal to 1, and we can identify the structure of the piece by locating side diagonals; this structure is the following: verse-verse-chorus-verse-chorus. We also observe that there is a chord sequence which is repeating along the verse (C-F), so that there is a high

self-similarity inside each verse.

Instead of computing a self-similarity matrix, we can extent this idea to the comparison of two different pieces. We then compute a similarity matrix between two different THPCP representations. Figure 5.10 shows the similarity matrix between the original song *Imagine* by John Lennon (1) and the instrumental version (2). In this figure, we also identify here the same song structure than above, which is preserved in

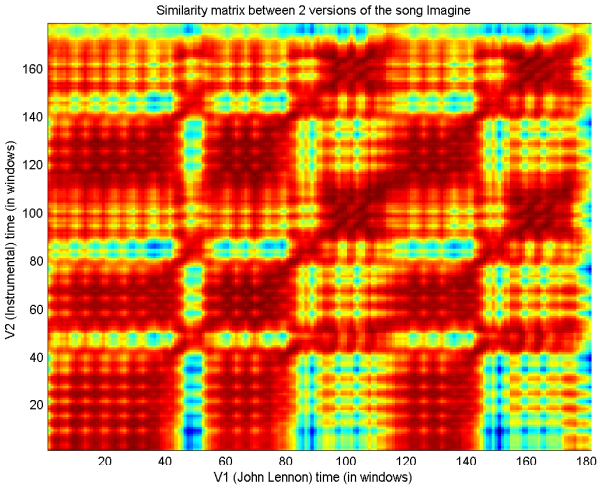


Figure 5.10: Similarity matrix between 2 different versions of the song *Imagine*, by John Lennon. See *Imagine1-JohnLennon.wav*, *Imagine2-Instrumental.wav* in Appendix A for further details on the sound excerpts.

the versioned song. We also see that the tempo is kept, as the diagonal is located so that the time index (1 window/second) remains the same in x and y axis. To see how the tempo affects the similarity matrix, we have performed a time stretch of version 2 with different time stretch factors: 50% of the original duration (twice faster), 70%, 100% (original duration), 130% and 160%. The time stretch algorithm used is the one proposed by Bonada (2000). The similarity matrices are shown in Figure 5.11.

We observe that the diagonal is kept, but its slope changes according to the tempo ratio:

$$slope = \frac{t_v}{t_o} \tag{5.5}$$

where t_v is the tempo of the versioned song and t_o the original tempo.

Now, we analyze what happens if the structure is modified. Figure 5.12 shows the similarity matrix between the song *Imagine*, by John Lennon and a version of it performed by Khaled and Noa. Here, the original tempo is more or less kept (there is just a slight tempo variation), but we can identity in the similarity matrix some modifications in the structure of the piece. With respect to the original song, version 5 introduces a new instrumental section (which is not included in the original piece) plus an additional chorus at the end of the piece.

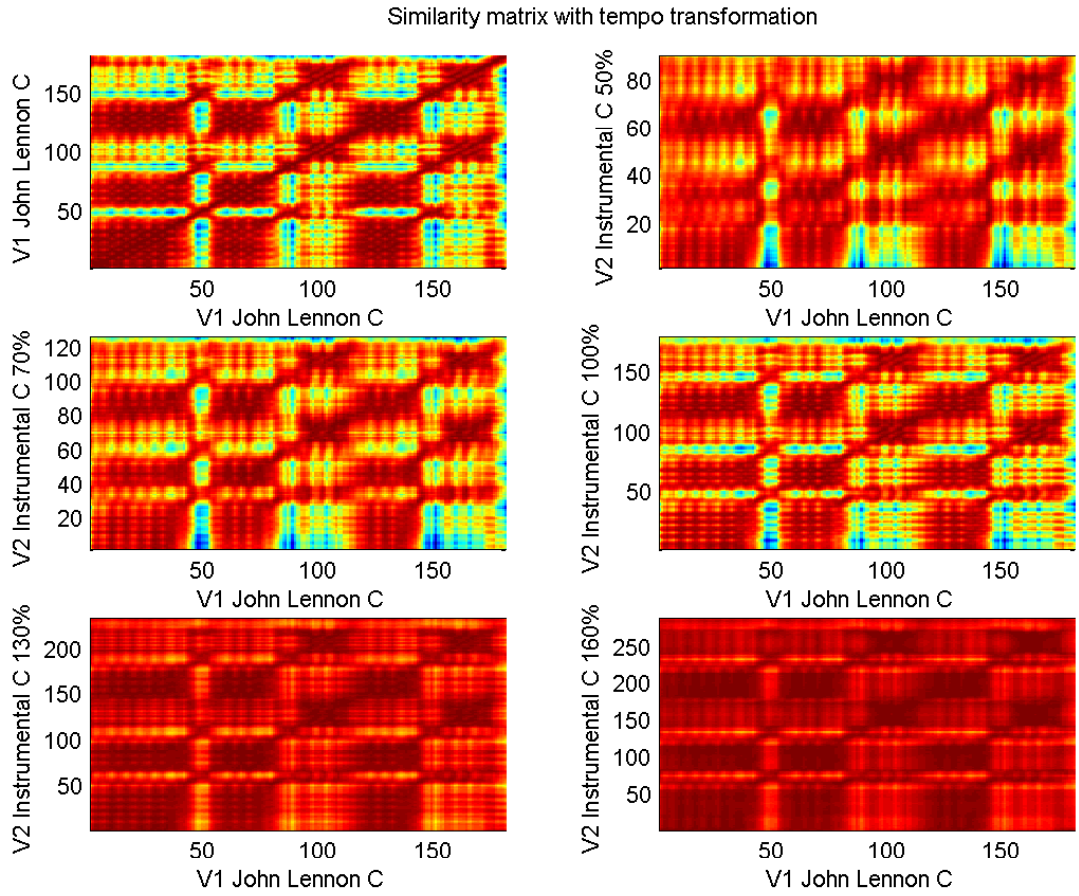


Figure 5.11: Similarity matrix between 2 different versions of the song *Imagine*, by John Lennon. See *Imagine1-JohnLennon.wav*, *Imagine2-Instrumental.wav* and its time stretched versions in Appendix A for further details on the sound excerpts.

Figure 5.13 represents the similarity matrix for each of the 5 cover versions and the self-similarity matrix of the original song. We can see that version 4 (Tania Maria) is the most dissimilar one, so that we can not distinguish clearly a diagonal in the similarity matrix. If we listen to both pieces, we can hear some changes in harmony (jazz), as well as changes in the main melody. These changes affect the THPCP features. In this situation, it becomes difficult to decide if this is a different piece or a version of the same piece. We should then study what determines the *identity* of a given piece, which is a difficult problem (Ong and Herrera (2004)). In Figure 5.13, we also present the similarity matrix with a different song, *Besame Mucho* by Diana Krall, in order to check that it is not possible to find a diagonal for different pieces if they do not share similar chord progressions.

As a conclusion to the example presented here and to the observation of 90 versions of different pieces,

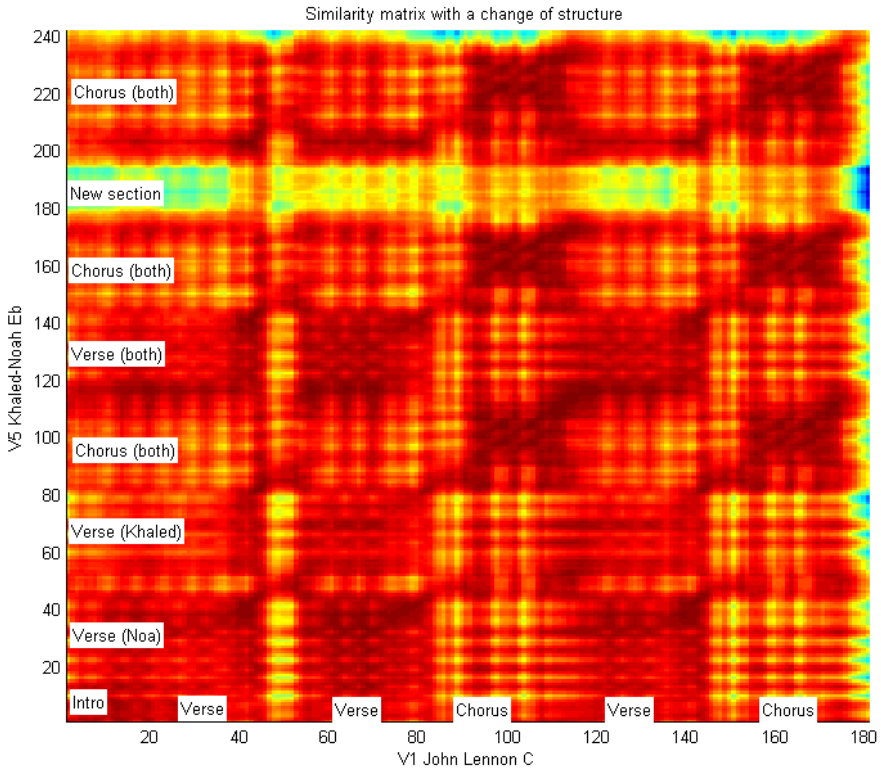


Figure 5.12: Similarity matrix between 2 different versions of the song *Imagine*, by John Lennon. See *Imagine1-JohnLennon.wav*, *Imagine5-KhaledNoa.wav* in Appendix A for further details on the sound excerpts.

we make the assumption that the instantaneous tonal similarity between pieces is represented by diagonals in the similarity matrix. The slope of the diagonal represents tempo differences between pieces. In order to track these diagonals, we will evaluate he use of a Dynamic Time Warping (from now DTW) algorithm. We have employed the one developed by Ellis (2005). The basic principle of DTW is to allow a range of *steps* in the temporal axis and to find the path that maximizes the local match between the aligned time frames, subject to the constraints implicit in the allowable steps. The total *similarity cost* found by this algorithm is a good indication of how well the two pieces match.

This algorithm estimates the minimum cost from one piece to the other one using the similarity matrix. We study in next section how this minimum cost can be used to measure similarity between pieces.

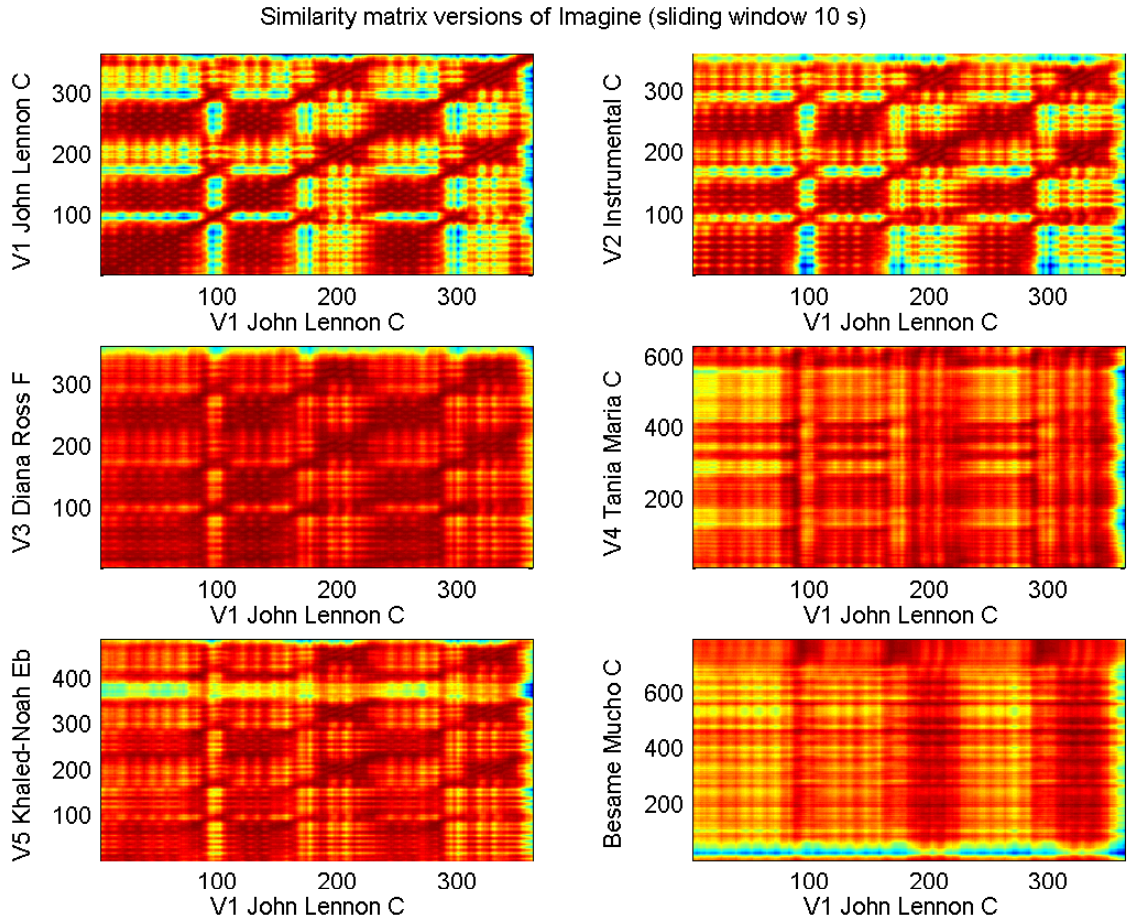


Figure 5.13: Similarity matrix for 5 different versions of the song *Imagine*, by John Lennon. See *Imagine1-JohnLennon.wav*, *Imagine2-Instrumental.wav*, *Imagine3-DianaRoss.wav*, *Imagine4-TaniaMaria.wav*, *Imagine5-NoaKhaled.wav* and *BesameMucho-DianaKrall.wav* in Appendix A for further details on the sound excerpts.

5.2.3 Evaluation

5.2.3.1 Methodology

The goal of this study is to evaluate how low-level tonal descriptors can be used for version identification. For this evaluation, we will compute a similarity measure between two different pieces, based on comparing low-level tonal features, i.e. HPCP values. We will compare four different similarity measures:

1. Correlation of average HPCP over the whole musical piece.
2. Correlation of average THPCP over the whole musical piece. This THPCP vector is computed as explained in Section 3.7 using the global key, which is labelled automatically and checked manually in order to isolate the problems of key finding and tonal similarity.
3. Minimum cost computed using DTW algorithm found in Ellis (2005) from HPCP values.
4. Minimum cost computed using DTW algorithm found in Ellis (2005) from THPCP values.

We have used precision and recall measures in order to compute the estimation accuracy. For each query, the recall is defined as the fraction of the relevant documents which has been retrieved (as explained by Baeza-Yates and Ribeiro-Neto (1999)). Given a song i ($i = 1 \dots N$), we consider that a song is relevant when it is a version of the same piece than the query. It is important to note that the system does not have any knowledge, neither on which is the original root song nor which are the cover versions.

$$Recall_i = \frac{nFoundItems_{Id(i)}}{nId(i) - 1} \tag{5.6}$$

where $nFoundItems_{Id(i)}$ is the number of found documents which have the same version Id than the query (i-th song), and $nId(i)$ is the total number of songs sharing the same version Id than the query (i-th song), including the query. The version Id represents an identification of each group of cover versions, as we show in Table 5.2. The precision is defined as the fraction of retrieved songs being relevant (as found in Baeza-Yates and Ribeiro-Neto (1999)):

$$Precision_i = \frac{nFoundItems_{Id(i)}}{n} \tag{5.7}$$

where n is the number of pieces returned by the algorithm. We finally compute an average of precision and recall measures, using as a query all the pieces in the collection.

$$Recall = \frac{1}{N} \cdot \sum_{i=1}^{N} Recall_i \tag{5.8}$$

$$Precision = \frac{1}{N} \cdot \sum_{i=1}^{N} Precision_i \tag{5.9}$$

Recall and precision measures, as defined originally, assume that all the documents in the answer set have been considered. We will compute recall and precision values for $n = 1 \ldots N-1$, where N is the number of pieces in the evaluation collection. It is usually desired to have a single value, instead of two different measures, and for this we use the F measure. The F measure can be considered to combine the information given by precision and recall, and is defined as follows:

$$F = \frac{2 \cdot Precision \cdot Recall}{Precision + Recall} \tag{5.10}$$

We can also try to establish a baseline for precision, in order to check if the proposed descriptors improve the result that we could obtain by randomly selecting pieces from the music collection. Lets consider that, given a query i from the collection ($i = 1 \ldots N$), we randomly chose a set of n pieces (different from the query) as most similar to the query. The number of correctly found pieces with the same version Id is given by the Newton formula:

$$nFoundItems_{Id(i)} = \sum_{k=1}^{\min(n,nId(i)-1)} k \cdot \frac{\begin{pmatrix} nId(i)-1 \\ k \end{pmatrix} \cdot \begin{pmatrix} N-nId(i) \\ n-k \end{pmatrix}}{\begin{pmatrix} N-1 \\ n \end{pmatrix}} = \frac{n \cdot (nId(i)-1)}{N-1} \tag{5.11}$$

The index n again represents the number of pieces randomly returned by the algorithm, $nId(i)$ is the total number of songs sharing the same version Id than the query (including the query), k is the amount of correct items (which can get values from 1 to the minimum value between n and $nId(i)$) and the scaling term represents the probability of getting k correct items. Then, we can compute the random precision and recall for the i-th query if we randomly select n items from the database as follows:

$$RandomPrecision_i = \frac{nFoundItemsId(i)}{n} = \frac{nId(i)-1}{N-1} \tag{5.12}$$

$$RandomRecall_i = \frac{nFoundItemsId(i)}{nId(i)-1} = \frac{n}{N-1} \tag{5.13}$$

We observe that the precision value is constant and it does not depend on the number of returned items. The average for all the possible queries is equal to:

$$RandomPrecision = \frac{1}{N} \cdot \sum_{i=1}^{N} RandomPrecision_i \tag{5.14}$$

The average for all the possible queries is equal to:

$$RandomRecall = \frac{1}{N} \cdot \sum_{i=1}^{N} RandomRecall_i \tag{5.15}$$

For the considered evaluation collection, Figure 5.14 represents the precision and recall curve. The constant value of $RandomPrecision = 3.196\%$, with a maximum value of the F measure equal to 0.0619. This is a very low value that this approach should surpass.

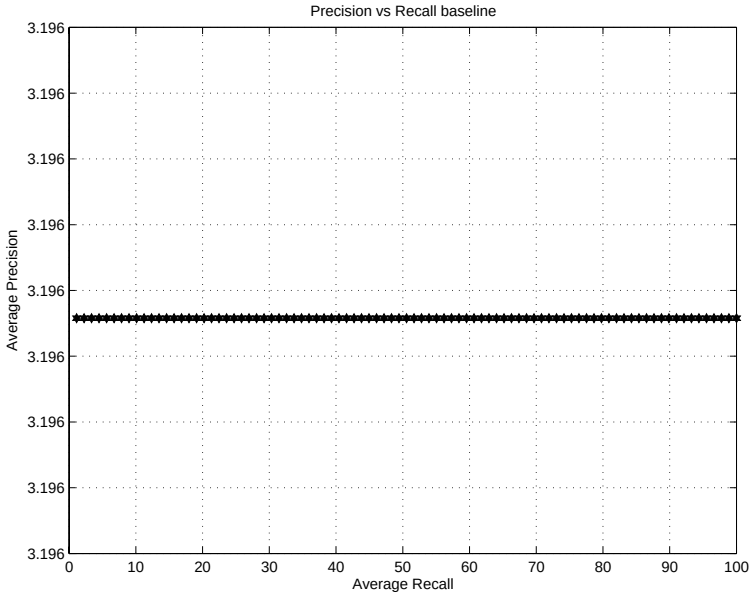


Figure 5.14: Precision vs recall baseline.

5.2.3.2 Material

The material used in this evaluation consists of 90 versions from 30 different songs taken from a music collection of popular music. The versions include different levels of similarity to the original piece, which are found in popular music: noise, modifications of tempo, instrumentation, transpositions and modifications of main melody and harmonization. The evaluation material is summarized in Table 5.2.

Some details about the similarity between the versions and the original pieces is found in appendix C. According to this, most of the versions belong to type V, VI or VII, explained above, and some of them include differences in harmony. We are then dealing with the most difficult examples, so that the evaluation can be representative of a real situation when organizing digital music collections.

Title	Number of items
Come Together (The Beatles)	3
Sgt. Pepper's Lonely Hearts Club Band (The Beatles)	2
Got to get you into my life (The Beatles)	2
Imagine (John Lennon)	5
My way (Frank Sinatra)	4
Here comes the sun (The Beatles)	4
Strawberry field forever (The Beatles)	3
Lucy in the sky with diamonds (The Beatles)	3
Maxwell's silver hammer	2
When I'm sixty four (The Beatles)	4
You've got a friend (James Taylor)	2
She came in through the bathroom (The Beatles)	3
Getting better (The Beatles)	2
Baby can I hold you (Tracy Chapman)	2
Oh! Darling (The Beatles)	3
Long and winding road (The Beatles)	3
Mean Mr Mustard (The Beatles)	2
Besame mucho	5
Fixing a hole (The Beatles)	2
I want you (She's so heavy) (The Beatles)	3
No woman no cry	4
She's leaving home (The Beatles)	2
Because (The Beatles)	2
Good morning good morning (The Beatles)	2
Being For The Benefit Of Mr. Kite! (The Beatles)	2
Golden Slumbers (The Beatles)	2
All of me	10
You never give me your money (The Beatles)	2
Get Back (The Beatles)	4
A day in the life (The Beatles)	2

Table 5.2: Evaluation Material.

5.2.3.3 Results

The evaluation results are presented in Figure 5.15.

When using the correlation of global average HPCP as a similarity measure between pieces, we observe that the maximum precision is very low, 20% with a recall level of 8% and a F measure of 0.145. If we consider the correlation of global average THPCP as a similarity measure between pieces, the maximum precision increases to 35.56%, around 15% higher than using HPCP. The recall level also increases from 8% to 17.6% and the F measure to 0.322.

Figure 5.15 also shows the average precision and recall measures using the DTW minimum cost computed from instantaneous HPCP as similarity measure. The maximum precision is equal to 23.35%, which is higher than using a global measure of HPCP. The recall level is slightly higher, equal to 10.37%, and the F value is

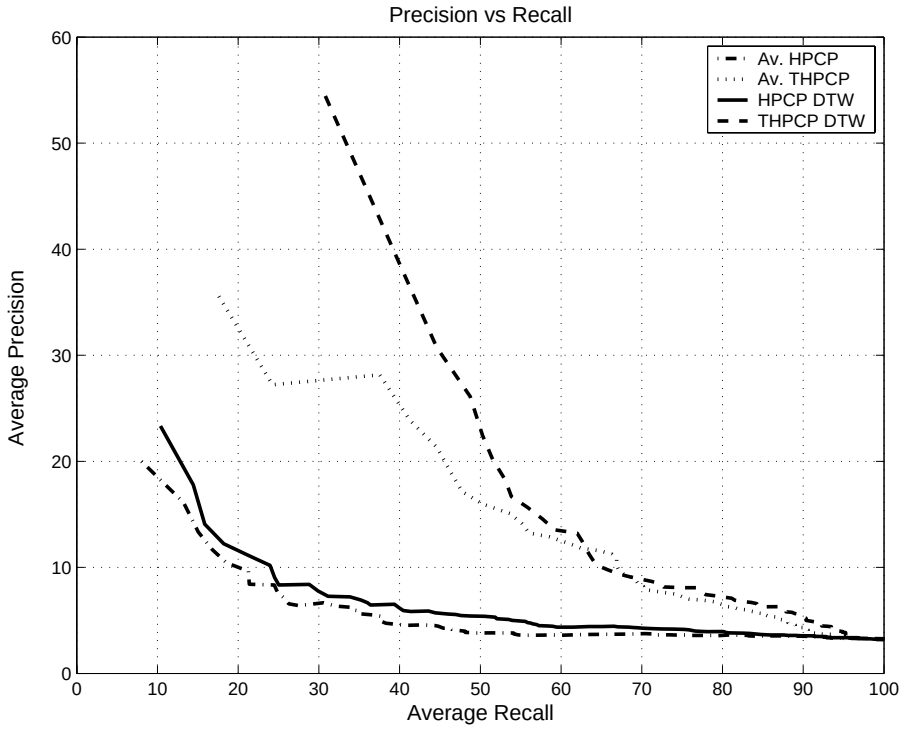


Figure 5.15: Precision vs recall for different descriptors.

equal to 0.159.

Finally, the use of the DTW minimum cost computed from instantaneous THPCP as similarity measure makes the maximum precision increases up to 54.5%, and the recall level is equal to 30.8%, with a F measure of 0.393.

5.2.3.4 Discussion

As a conclusion of this small-scale evaluation, we can see that relative descriptors (THPCP) seem to perform better than absolute pitch class distribution features, which is coherent with the invariability of melodic and harmonic perception to transposition. Also, it seems that it is important to consider the temporal evolution of tonality, which is sometimes neglected. The best accuracy is then obtained when using a simple DTW minimum cost computed from THPCP descriptors, and it is around 55% (recall level of 30.8% and F measure of 0.393).

Version identification is a difficult problem. As we mentioned above, our evaluation database represents a real situation of a database including cover versions, where even the harmony and the main melody is modified. This fact affects the pitch class distribution descriptors. Even in this situation, we see that only using low-level tonal descriptors and a very simple similarity measure, we can detect until 55% of the versions, with

a recall level of 30.8% and F measure of 0.393. These results overcome the baseline (F measure of 0.0619) and show that tonal descriptors are very important in music similarity. Further experiments will be devoted to improve the similarity measure, and to include other relevant aspects which are for instance structure analysis (e.g. determining the most representative segment), rhythmic description (in order to extract characteristic rhythmic patterns) and predominant melody estimation.

5.3 Characterizing music collections according to tonal features

As mentioned in Section 1.5, one of the main application contexts in which tonal description becomes relevant is to organize digital music collections. Tonal description, then, is complemented with descriptors related to other musical facets, such as rhythm or timbre.

The work carried out along this PhD thesis has been integrated into a system intended to organize and recommend music based on the analysis of audio features: the *MusicSurfer* (Cano et al. (2005b,a)) which is accessible on-line[3]. As seen by Cano et al. (2005a), the *MusicSurfer* is a music browsing and recommendation system based on a high-level music similarity metric and computed directly from audio data. This metric accounts for several perceptual and musically meaningful dimensions as those evidenced in music research.

Rhythm is of course an important musical aspect, which is represented with several descriptors computed from audio signals: tempo, meter, rhythm patterns and swing. Rhythmic descriptors integrated into the *MusicSurfer* system are based on the research by Gouyon (2005), reported in his PhD dissertation. In addition, the similarity measure also accounts for timbre. Based on spectral characteristics of the audio signals, this dimension represents aspects of the instrumentation (Herrera et al. (2004)) as well as post-production and sound quality characteristics. Another perceptual dimension which is considered is the perceived dynamics of audio signals, as presented in Sandvold and Herrera (2005). It is also important to automatically characterize music complexity in its different facets, as presented in Streich and Herrera (2004), Streich (2005) and Streich and Herrera (2005). Finally, we have seen in Section 5.2 the relevance of structural analysis for music similarity, which is considered in the work by Ong and Herrera (2004, 2005).

Complementary to these dimensions, the similarity metric accounts for tonal aspects of musical pieces. Using tonal descriptors, the user can search for pieces according to different criteria:

- **Key labels**: the labels which are provided are the key note and the mode (*major* or *minor*). Following this option, the user can ask, for instance, for pieces in minor key, or pieces in *C Major*.
- **Tonal profile**: pieces are compared using the similarity measures explained in Section 5.2.
- **Tonal strength**: we also provide the *tonal strength* measure, which is computed when performing audio key finding, as explained in Section 4.

In *MusicSurfer*, a global measure of similarity is defined by specific weights assigned to the diverse musical dimensions. This results in a specific representation of the musical space that users can explore. As an additional feature, users can adjust the similarity metric at will, giving particular emphasis to a specific musical dimension, and thus exploring a different, personalized, musical space.

When dealing with large music collections, it is necessary to optimize the feature computation task. The online performance of the mentioned system in query-by-example tasks on a musical repository of over a million songs is about tenths of a second, while extraction of the musical features runs 20 times faster than playing time. All the operations can then be carried out using a PC.

[3] http://musicsurfer.iua.upf.edu

In order to illustrate how the system works using tonal descriptors, let us discuss an example of the *tonal strength* descriptor. We also refer to the examples presented in Chapter 4 for atonal and percussive music. Figure 5.16 shows the distribution of these descriptors along the music collection presented in Section 4.3.1.1. We can see that most of the pieces have a high value of tonal strength, and that the distribution is centered around 0.75.

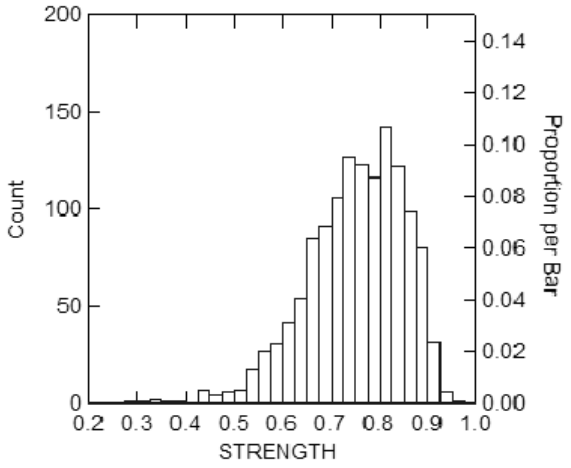


Figure 5.16: Distribution of the tonal strength measure.

We can order the pieces according to the tonal strength descriptor. Here, we present some examples. For high tonal strength values, Figure 5.17 shows an example of a piece in major key having a very high value of this feature (0.92). This example is a song called *You don't bring me flowers*, by Barbra Streisand. If we listen to the song, we verify that the piece is very tonal, in C major, with no modulations during its whole duration. Figure 5.18 shows an example of a piece in minor mode having also a high value of this feature (0.86). This example is a song called *Atrápame*, by Manny Manuel, a Latin artist. This song is a very danceable piece, which is very tonal, in A minor. There is no modulation throughout the piece, and the chords played along the piece are mostly the tonic and dominant triad chords.

Furthermore, we can look at the behavior of pieces with low value of tonal strength, Figure 5.19 shows an example of a piece in major key having a low value of this feature (0.367 for G minor). This song is an electronic piece called *Stop or I'll Shoot*, by Ryuichi Sakamoto. We can hear that this is a textural piece with no clear tonality.

After observing the variation of tonal strength for different pieces, we observe that tonal strength is related to the musical genre. It is usually low for musical genres where there is no clear tonality (e.g. electronic, ambient) or where the are many modulations (e.g. some jazz styles). Figure 5.20 shows an example of music genres having low tonal strength values. On the contrary, we observe that tonal strength is high for classical music, pop (e.g. The Beatles) or some folk music. Figure 5.21 shows an example of music genres having

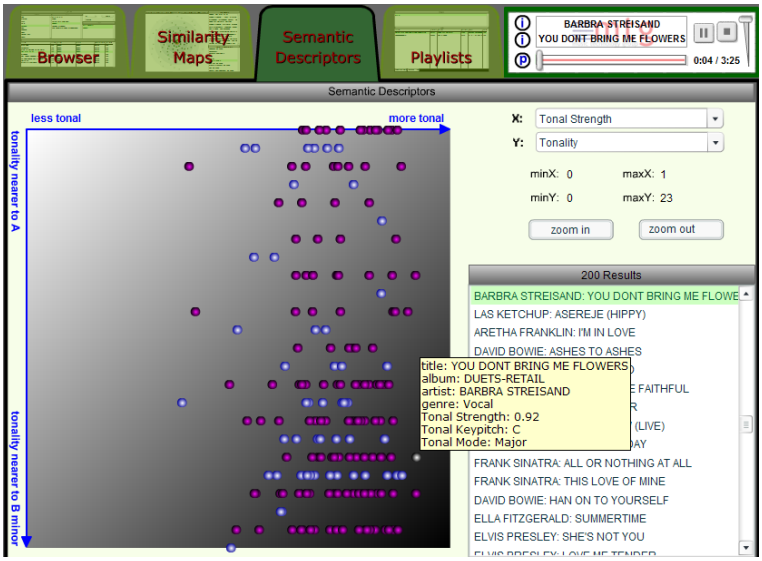


Figure 5.17: Example of a song in C Major with a high value of tonal strength (0.92). See *YouDontBringMe-Flowers.mp3* in Appendix A for further details on the sound excerpts.

high tonal strength values. These observations are related to the differences found when performing key estimation for different musical genres, as discussed in Chapter 4.

Evaluating the relevance of tonal descriptors in computing and exploiting music similarity is a difficult task, which can easily become subjective. A user study should be carried out in order to determine if tonal descriptors are relevant to compare pieces or to organize digital music collections.

5.4 Conclusions

In this Chapter, we have investigated some uses of tonal description of audio apart from audio key finding. We presented two application contexts that can benefit from the proposed tonal description: first, to compare two different musical pieces from the analysis of audio, and second to organize digital music collections, i.e. to navigate through compilations of audio files.

This chapter was divided in two main blocks. Section 5.2 focused on the analysis of tonal similarity and its application to the identification of different versions of the same piece. We studied here in which sense version identification is a complex problem that requires a multifaceted and multilevel audio description, and we presented a small experiment showing that tonal descriptors by themselves can be helpful for this task. There are some conclusions to this study. First, it is necessary to consider invariance to transposition when computing tonal descriptors for similarity tasks. Second, we should look at the tonal structure of the piece to yield relevant results. A simple approach for version identification based on Dynamic Time Warping of THPCP features has yielded a F measure of 0.393 (55% precision and 30.8% recall), which we believe is an

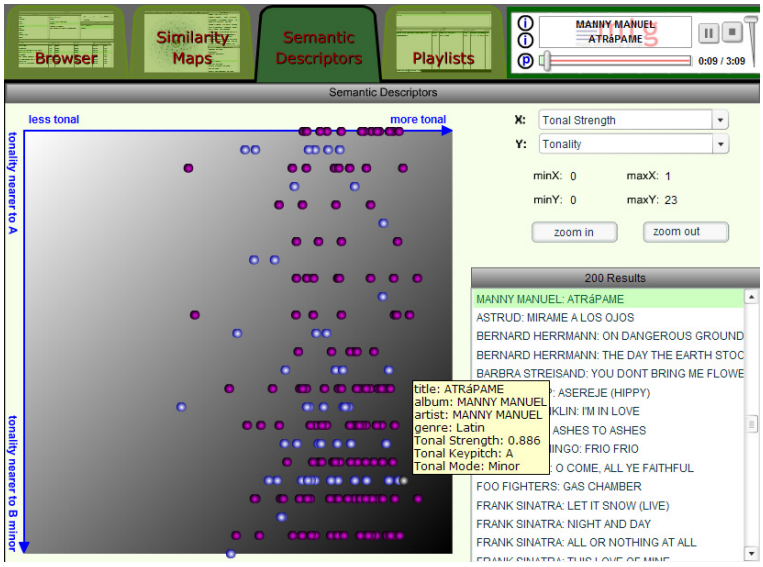


Figure 5.18: Example of a song in A Minor having a high value of tonal strength (0.886). See *Atrapame.mp3* in Appendix A for further details on the sound excerpts.

outstanding result to the problem.

Section 5.3 presents an example of the integration of the proposed tonal descriptors in a system for music organization and recommendation based on the analysis of audio signals. The integration of tonal description into the *MusicSurfer* system has allowed the testing of the proposed automatic tonal description in a real situation, dealing with over a million pieces, performing 20 times faster than real-time, and being found useful to search and recommend music. We presented some examples on how it is possible then to organize a music collection according to key, tonal profile or key strength. Here, we would like to point the difficulty of evaluating this type of systems. An evaluation of the usefulness of these descriptors by means of human ratings has not been carried out, and it remains as future work. This chapter contributes to the justification that tonal description is valid to index musical collections and perform search by similarity, which is one of the goals of this dissertation, mentioned in Section 1.6.

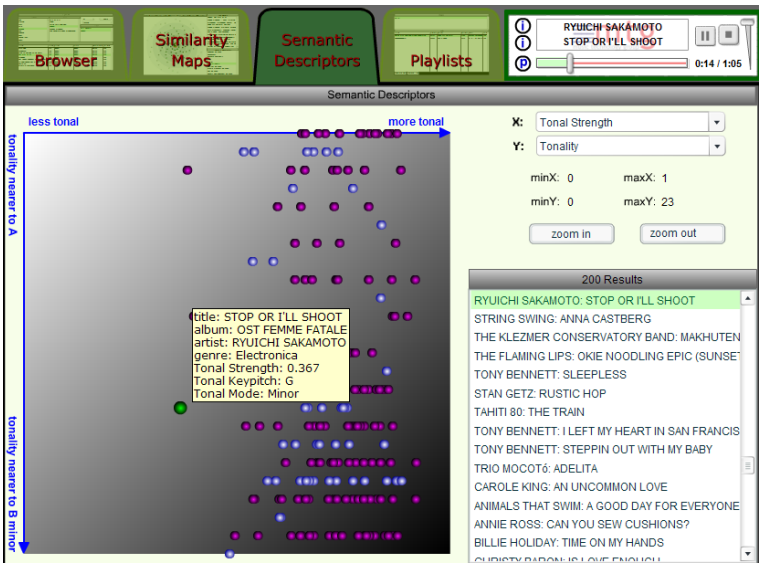


Figure 5.19: Example of a song in G Minor having a high value of tonal strength (0.367). See *StopOrIll-Shoot.mp3* in Appendix A for further details on the sound excerpts.

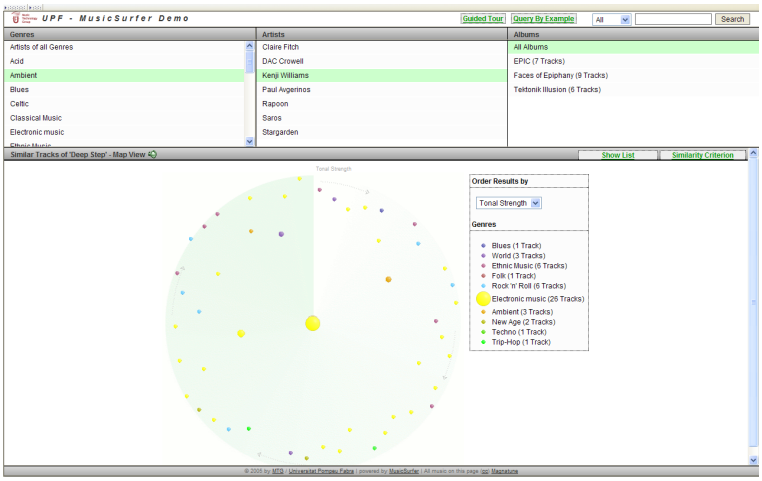


Figure 5.20: Example of music genres having low tonal strength values.

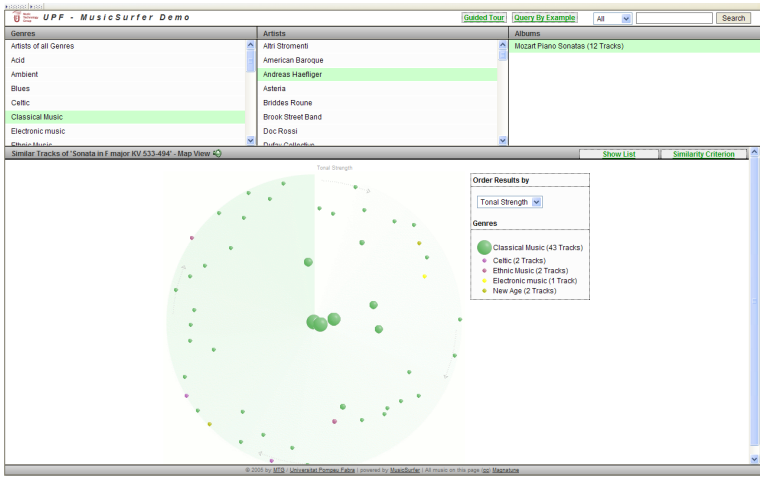


Figure 5.21: Example of music genres having low tonal strength values.

Chapter 6

Summary and future perspectives

6.1 Introduction

Along this dissertation, we have addressed several issues that appear when computers are asked to analyze a musical piece in audio format and describe its tonal content: the automatic computation of low-level tonal features representing its pitch class distribution, the estimation of its main key or the evolution of its tonal center along time, how to measure tonal similarity between two pieces and how to use tonal features to organize large collections of music. We have focused on some important aspects of tonal audio description, as the extraction of relevant low-level features from audio, the adaptation of tonal profiles to work with these features instead of symbolic data in order to estimate the key of a piece, and finally the use of these features to compare two different pieces or to organize digital music collections.

The goal of this chapter is first to summarize the contributions this dissertation makes to the research literature in computational methods for tonal description. After this summary, we go over the main conclusions of this dissertation, and we present some issues that we did not undertake, but could have done so, if it wasn't for time constraints. We also propose some ideas which are interesting for future work in this field, which we did not further pursue either because they implied very distant methodologies or simply because we did not have a clear idea about how to do it.

6.2 Summary of contributions

This dissertation makes some significant contributions which conforms to the current state of the art in tonal description of polyphonic audio. As discussed in Chapter 2, the problem of key estimation has been studied for more than two decades but, in contrast, tonal description from audio recordings has only gained interest over the last few years. In this sense, the publications generated by this dissertation (see Appendix D) belong to this recent literature which conforms the current state of the art in key estimation from audio. In addition

to this general remark, we think that this work has achieved the goals and hypotheses mentioned in Chapter 1, which we summarize here.

Review of tonal induction systems. Although there is much literature related to tonal description from symbolic data, this dissertation provides a multidisciplinary picture of the different functional blocks defined in systems for tonal description from audio. In Chapter 2, we have analyzed the current efforts in tonality induction from audio, studying the main similarities and differences between approaches according to the signal processing methods and the tonal models applied.

Comparison with other approaches for audio key finding. Along this PhD, we have had the opportunity to compare our approach with the current state of the art, in the context of the ISMIR 2005, where our preliminary system was ranked third. This accuracy rate was measured for audio signals generated from MIDI. The current proposed system increased this performance, being capable of automatically finding the key of real recordings (from varied styles, instruments and recording conditions) with an accuracy around 77%. There are some original features of our approach that complement the current state of the art. For pitch-class distribution features, we have evaluated our approach for automatic estimation of the tuning frequency. The system provides configurable interval and temporal resolution, and it considers the presence of harmonic frequencies and the timbre of the played instrument. For audio key finding, the proposed approach allows the use of a varied set of tonal profiles that can be adapted to different genres.

Requirements of low-level tonal features. This dissertation introduces the definition of the main requirements of pitch class distribution features from audio, as defined in Section 2.6.3. We have then tried to verify these requirements when proposing and evaluating the *Harmonic Pitch Class Profile*.

Description vs transcription. This dissertation supports the idea that some application contexts do not need a perfect audio to score transcription, as discussed in Section 2.6.2. We have seen that pitch class distribution features are extracted from audio signals with no lost of accuracy with respect to symbolic data. The use of pitch class distribution features overcomes the problem of automatic transcription. We can also verify this fact when comparing the performance of key finding methods working from symbolic data and from audio. According to the MIREX Audio key finding contest, the best accuracy obtained for symbolic data (Temperley (2005)) was only 1.85% higher than for audio (Izmirli (2005)). Although we do not find this to be a realistic comparison because the audio signals where synthesized from MIDI data, it provides an idea of the accuracy that could be obtained without transcription.

Bridging the gap between audio and symbolic-oriented methods. This dissertation contributes to bridge the gap between score analysis and audio analysis, by studying how current methods for tonal analysis from score can be applied to audio features.

Quantitative evaluation of different stages. We believe that this dissertation presents the most exhaustive and comprehensive evaluation carried out to date. We have provided a quantitative evaluation of the influence of different alternative solutions in various stages of the proposed system. We have designed a modular evaluation strategy, where a set of experiments are defined in order to study separately different issues: performance of the tuning frequency determination algorithm; performance of the HPCP features and

its relationship with pitch class distribution features obtained from symbolic data; influence of analysis parameters; influence of the use of different tonal profiles; comparison of fixed (obtained through experimental studies) versus machine learning modelling strategies; evaluation of the use of different segment durations; quantitative evaluation of the performance of tonal features for similarity measurements.

These quantitative evaluations have involved the creation of a music collection with annotated tonal information (mainly key and some chord annotations). This collection is composed of more than 1450 annotated pieces of different musical genres, and it is described in Sections 4.3.1.1 and 5.2.3.2.

From classical music to other musical genres. Most of the literature focuses on the analysis of classical music. In this dissertation, we have tried to extent this idea to a broader range of musical styles, analyzing the influence of the genre on the tonal profile (see Sections 4.3.4 and 5.3).

Tonal description for music retrieval. This dissertation contributes to prove that tonal description, in addition to the traditional interest related to musical analysis, provides a powerful tool for a musical meaningful indexing of music collections, and a way to measure similarity between musical pieces, and as a way to differentiate different styles of music. We have proved in Section 5 that tonal descriptors are relevant for measuring music similarity, and we have evaluated this usefulness to solve the complex problem of **version identification**.

Tonal description in different time scales. This work provides some tools to analyze and visualize the tonal content of a piece of music in audio format in different temporal scopes, from chord to global key, as presented in Chapter 4.

Optimization and visualization. The approach proposed in this dissertation is integrated into a working system to organize digital music collections dealing with a million music titles (Cano et al. (2005b,a)). The implementation of the methods described in this dissertation within a real situation has involved the optimization of the algorithm. A real-time implementation of the system has been developed, in order to provide real-time visualization of the tonal content of music (Gómez and Bonada (2005))

Towards high-level musical description. This work has also served for improving structural analysis of audio signals (Ong and Herrera (2005)), as well as for music visualization and musical analysis (Gómez and Bonada (2005)).

6.3 Future perspectives

6.3.1 On evaluation

There is a clear need for establishing benchmarks for comparing strategies for music descriptors in general. The music information retrieval community is currently discussing some problems related to evaluations (see for instance MIREX wiki[1] and Lesaffre et al. (2004); Gouyon (2005)). We focus here in the main issues we think should be faced when dealing with tonal features.

[1] http://www.music-ir.org/mirexwiki

The importance of annotations It becomes important to collect annotated music collections that can serve as a ground truth to develop tonal description systems, including a varied spectra of musical genres, instrumentation, etc. This task should join efforts from different disciplines, as manual annotation is a tough issue (see Lesaffre et al. (2004) for a detailed discussion on this issue). Herrera et al. (2005b) have recently proposed an environment for the generation and sharing of annotations related to different musical aspects.

Key ambigüity The concept of key is well defined in music theory, but in some cases can become difficult to annotate. Manual annotations can be sometimes ambiguous. In this sense, other disciplines could provide us with a set of ground truth annotations (i.e. related to cognition or to musical theory), which may depend on the user profile.

Towards modular evaluation It is difficult to compare systems that, even if they implement similar concepts, do not share any piece of code. The performance of each system depends on the performance of the different algorithms involved in the process of tonal description. For instance, for audio key finding, different algorithms are proposed for the different stages: preprocessing (FFT vs Constant-Q), tuning frequency computation, pitch class distribution feature extraction, key models, etc. Evaluation strategies should be devoted to measure the accuracy of these different modules, in order to analyze the performance of the different techniques in each step and analyze the influence of each of them in the final results.

Real audio and robustness tests MIDI synthesized audio as used in some evaluation strategies seems to be to simplistic for the problem of real recordings. Evaluation methodologies should include some robustness tests, for instance, to equalization, increasing level of noise (decreasing SNR) or different instrumentation.

User studies: another important aspect is to evaluate similarity measures by means of listening experiments. An evaluation of the usefulness of these descriptors by means of human ratings has not been carried out, and it remains as future work.

6.3.2 Aspects of tonal description

We believe that future work on tonal analysis should be devoted to improve the robustness of the features and to bridge the *semantic gap* between low-level audio descriptors and higher level concepts.

Tonal description for music content processing Although most of the literature has been focusing on timbre and rhythmic description of audio, recent literature shows that tonality is an important musical aspect which should be considered in music content processing systems. Tonality can provide some musical criteria to index music collections, which can be also useful for people lacking musical knowledge. This axis of description should be integrated with other musical aspects to arrive to a high-level abstract description of musical content.

Tonal description for music analysis We think that musicologists can benefit from automatic audio description tools to analyze material where the score is not available. Automatic analysis can also help to decrease the annotation effort, so that a larger corpus of material can be analyzed. More research should be devoted to this application context, in order to adapt features and visualization techniques.

Tonal description for version identification: as seen in Section 5.2, the problem of version identification is an extremely difficult task which is far from being solved in the current literature. There are many issues involved in this problem, and it seems simplistic to face this problem from the point of view of automatic audio description. A multifaceted study should be carried out, involving different aspects. Automatic structural analysis, for instance, can provide the location of the most repeated (Cooper and Foote (2002)) or most relevant segment of a musical piece (Ong and Herrera (2005)). Psychological experiments can help to clarify which are the features that characterize the "distinctiveness" of a given piece, as well as provide similarity ratings between different versions of the same song.

Temporal evolution rather than global features Until now, most of the literature is devoted to the evaluation of the performance of global key estimation systems, but it becomes clear that it is important to analyze the evolution of the tonal center within a piece. We have seen in Section 5.2 that the use of similarity matrices and a simple Dynamic Time Warping algorithm for similarity can increase the accuracy for version identification. We believe that more effort should be put into extracting and comparing temporal evolution of tonality within pieces. In this sense, key tracking is linked to tempo estimation and structural analysis, as shown in Section 5.2

Tonality and mood: one important aspect which should be included into automatic description systems is the relationship between tonal features and the mood which is induced by a given piece. Tonality has been shown to be linked with emotion, as for instance minor moods are related to sadness and major to happiness (see Juslin and Sloboda (2001)). Future systems should explore ways to provide an automatic description of *mood*. We think that further developments of this work, together with other music description axes (as timbre or rhythm), could contribute to this goal.

Tonality induction and musical training: according to Auhagen and Vos (2000), experimental studies for tonality induction do not agree on how to differentiate between *experienced* and *inexperienced* subjects. More research is necessary to study the difference between different musical training and cultural backgrounds, and to consider this knowledge when developing automatic tonal description systems.

We have studied how complex is the automatic characterization of the tonal content of audio recordings. There is a varied set of disciplines involved, such as signal processing, music theory or music cognition, and researchers working in this field must consider these different aspects when developing computational models.

There is a huge amount of approaches in the literature for musical analysis from score representations. This literature has been restricted until now to the available set of MIDI files. For this reason, music not having a score representation or transcription has not been much studied (e.g. popular or folk music). Signal processing tools are now able to automatically extract relevant information from audio, as for instance the distribution of pitches. This fact opens the chance of using current midi-oriented methods to analyze directly audio material.

There is a small amount of literature dealing with tonal description from audio in MIR, and most of the systems do not consider these features. Current systems are mainly based on timbre and less on rhythmic

descriptors. We see two main reasons for that. First, the concept of timbre (related to the played instrument) and rhythm (related to the speed of execution or the metric) are easier to understand by people not having a musical background. Second, some low-levels features related to timbre (e.g. spectral features) can be easily computed in an automatic way from polyphonic audio recordings. Only recent literature has focused on the problem of melody, harmony or key estimation from audio. We have seen that these descriptions are significant also to people not having any musical training, as tonal aspects are linked to music similarity and semantic aspects (e.g. mood). We think that the use of these descriptors will have a strong effect on MIR systems, and much research in the next future will be devoted to analyze and compare music according to this musical facet.

Bibliography

Amatriain, X., Bonada, J., Loscos, A., and Serra, X. (2002). Spectral processing. In Zölzer, U., editor, *DAXF-Digital Audio Effects*, pages 373–438. John Wiley & Sons.

Amatriain, X., Massaguer, J., García, D., and Mosquera, I. (2005). The clam annotator: a cross-platform audio descriptors editing tool. In *International Conference on Music Information Retrieval*, London, UK.

Anderson, E. J. (14-5-1997). Limitations of short-time fourier transforms in polyphonic pitch recognition. Technical Report Ph.D. qualifying project report, Department of Computer Science and Engineering in the Graduate School of the University of Washington. http://www.cs.washington.edu/homes/eric/Publications.html.

Aucouturier, J.-J. and Pachet, F. (2004). Tools and architecture for the evaluation of similarity measures: case study of timbre similarity. In *International Conference on Music Information Retrieval*, Barcelona, Spain.

Auhagen, W. and Vos, P. G. (2000). Experimental methods in tonality induction research: a review. *Music Perception, Special Issue in Tonality Induction*, 17(4):417–436.

Baeza-Yates, R. and Ribeiro-Neto, B. (1999). Retrieval evaluation. In *Modern Information Retrieval*, Series in Cognition and Perception, book chapter 3, pages 73–99. ACM Press, Pearson Addison Wesley, first edition.

Bartsch, M. A. and Wakefield, G. H. (2001). To catch a chorus: using chroma-based representations for audio thumbnailing. In *IEEE Workshop of Applications of Signal Processing to Audio and Acoustics*.

Berenzweig, A., Logan, B., Ellis, D. P., and Whitman, B. (2003). A large-scale evalutation of acoustic and subjective music similarity measures. In *International Conference on Music Information Retrieval*, Baltimore, USA.

Bharucha, J. J. (1999). Neural nets, temporal composites and tonality. In Deutsch, D., editor, *The psychology of music, second edition*, Series in Cognition and Perception, pages 413–440. Academic Press, second edition.

Bonada, J. (2000). Automatic technique in frequency domain for near-lossless time-scale modification of audio. In ICMA, editor, *International Computer Music Conference*, Berlin.

Bregman, A. S. (1998). Psychological data and computational auditory scene analysis. In Rosenthal, D. and Okuno, H. G., editors, *Computational auditory scene analysis*. Lawrence Erlbaum Associates, Inc.

Brown, J. C. (1991). Calculation of a constant-q spectral transform. *Journal of the Acoustic Society of America*, 89(1):425–434.

Brown, J. C. and Puckette, M. S. (1992). An efficient algorithm for the calculation of a constant-q transform. *Journal of the Acoustic Society of America*, 92:2698–2701.

Cano, P. (1998). Fundamental frequency estimation in the sms analysis. In *COSTG6 Conference on Digital Audio Effects (DAFX)*.

Cano, P., Batlle, E., Gómez, E., de C. T. Gomes, L., and Bonnet, M. (2002). Audio fingerprinting: concepts and applications. In *1st International Conference on Fuzzy Systems and Knowledge Discovery, Singapore*.

Cano, P., Koppenberger, M., Wack, N., Mahedero, J. P. G., Aussenac, T., Maxer, R., Masip, J., Celma, O., García, D., Gómez, E., Gouyon, F., Guaus, E., Herrera, P., Massaguer, J., Ong, B., Ramírez, M., Streich, S., and Serra, X. (2005a). Content-based music audio recommendation. In *ACM Multimedia*, Singapore.

Cano, P., Koppenberger, M., Wack, N., Mahedero, J. P. G., Masip, J., Celma, O., García, D., Gómez, E., Gouyon, F., Guaus, E., Herrera, P., Massaguer, J., Ong, B., Ramírez, M., Streich, S., and Serra, X. (2005b). An industrial-strength content-based music recommendation system. In *28th Annual International ACM SIGIR Conference*, Salvador, Brazil.

Celma, O., Ramirez, M., and Herrera, P. (2005). Foafing the music: a music recommendation system based on rss feeds and user preferences. In *International Conference on Music Information Retrieval*, London, UK.

Cemgil, A. T. (2004). *Bayesian Music Transcription*. Doctoral dissertation, Radboud University of Nijmegen.

Chai, W. (2005). *Automated analysis of musical structure*. Doctoral dissertation, MIT.

Chai, W. and Vercoe, B. (2005). Detection of key change in classical piano music. In *International Conference on Music Information Retrieval*.

Chew, E. (2000). *Towards a Mathematical Model of Tonality*. Doctoral dissertation, MIT.

Chew, E. (2004). Messiaen's regard iv: automatic segmentation using the spiral array. In *Proceedings of Sound and Music Computing*, Paris, France.

Chew, E. and François, A. (2003). Musart - music on the spiral array in real-time. In *Proceedings of the ACM Multimedia Conference*.

Chuan, C. and Chew, E. (2005). Polyphonic audio key finding using the spiral array ceg algorithm. In *IEEE International Conference on Multimedia and Expo*, Amsterdam, The Netherlands.

Cohen, A. J. (1977). Tonality and perception: musical scales prompted by excerpts from das wohltemperierte clavier of j.s. bach. In *Workshop on Physical and Neuropsychological Foundations of Music*, Ossiach, Austria.

Cohn, R. (1997). Neo-riemannian operations, parsimonious trichords, and their tonnetz representations. *Journal of Music Theory*, 41(1):1–66. http://www.leeds.ac.uk/icsrim/pub/.

Cooper, M. and Foote, J. (2002). Automatic music summarization via similarity analysis. In *International Conference on Music Information Retrieval*.

Dannenberg, R. B. (1993). A brief survey of music representation issues, techniques and systems. *Computer Music Journal*, 17(3):20–30. http://www-2.cs.cmu.edu/˜ music/papers.html.

Davy, M. and Godsill, S. (2003). Bayesian harmonic models for musical signal analysis. In *Cambridge Music Processing Colloquium*, Cambridge, UK.

de Cheveigné, A. (2005). Pitch perception models. In Plack, C. J., Oxenham, A. J., Fay, R. R., and Popper, A. N., editors, *Pitch: Neural Coding and Perception*, volume 24 of *Springer Handbook of Auditory Research*. Springer Verlag.

de Cheveigné, A. (2006). Multiple f0 estimation. In Wang, D. and Brown, G. J., editors, *Computational auditory scene analysis*, book chapter 1, in press. John Wiley and sons.

de Cheveigné, A. and Kawahara, H. (2002). Yin, a fundamental frequency estimator for speech and music. *Journal of the Acoustic Society of America*, 111:1917–1930.

Deutsch, D. (1999). The processing of pitch combinations. In Deutsch, D., editor, *The psychology of music, second edition*, Series in Cognition and Perception, pages 349–411. Academic Press, second edition.

Doval, B. and Rodet, X. (1991). Fundamental frequency estimation using a new harmonic matching method. In ICMA, editor, *International Computer Music Conference*, pages 555–558.

Dowling, W. J. (1978). Scale and contour: two components of a theory of memory for melodies. *Psychological Review*, 85(4):341–354.

Ellis, D. (2005). Dynamic time warp (dtw) in matlab. web publication. http://www.ee.columbia.edu/ dpwe/resources/matlab/dtw.

Fletcher, N. H. and Rossing, T. D. (1991). *The physics of musical instruments*. Springer-Verlag.

Foote, J. T. (1999). Visualizing music and audio using self-similarity. In ICMA, editor, *ACM Multimedia*, pages 77–84, Orlando, Florida, USA.

Foote, J. T., Cooper, M., and Nam, U. (2002). Audio retrieval by rhythmic similarity. In *International Conference on Music Information Retrieval*, Paris, France.

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

Fujishima, T. (2000). Apparatus and method for recognizing musical chords. U.S. Patent 6,057,502 Yamaha Corporation, Hamamatsu, Japan.

Fulford-Jones, W. (accessed March 2006). Remix. In Macy, L., editor, *Grove Music Online*. http://www.grovemusic.com.

Gómez, E. (2001). Fundamental frequency study report. Cuidado internal report, Music Technology Group, Universitat Pompeu Fabra.

Gómez, E. (2002). Melodic description of audio signals for music content processing. Technical report, PhD Research Work, Doctorat en Informática i Comunicació Digital, Music Technology Group, Universitat Pompeu Fabra.

Gómez, E. (2004). Tonal description of polyphonic audio for music content processing. In *Audio Engineering Society conference on metadata*.

Gómez, E. (2005). Key estimation from polyphonic audio. In *2005 MIREX Contest - Symbolic Key Finding*. http://www.music-ir.org/evaluation/mirex-results/audio-key/index.html.

Gómez, E. (2006). Tonal description of polyphonic audio for music content processing. *INFORMS Journal on Computing, Special Cluster on Computation in Music*, 18(3).

Gómez, E. and Bonada, J. (2005). Tonality visualization of polyphonic audio. In *International Computer Music Conference*.

Gómez, E. and Herrera, P. (2004). Estimating the tonality of polyphonic audio files: cognitive versus machine learning modelling strategies. In *International Conference on Music Information Retrieval*.

Gómez, E., Klapuri, A., and Meudic, B. (2003). Melody description and extraction in the context of music content processing. *Journal of New Music Research*, 32(1).

Gómez, E., Streich, S., Ong, B., Paiva, R. P., Tappert, S., Batke, J.-M., Poliner, G., Ellis, D., and Bello, J. P. (2006). A quantitative comparison of different approaches for melody extraction from polyphonic audio recordings. Technical report, MTG-TR-2006-01, Music Technology Group, Universitat Pompeu Fabra.

Godsmark, D. and Brown, G. J. (1999). A blackboard architecture for computational auditory scene analysis. *Speech Communication*, 27:351–366.

Gold, B. and Rabiner, L. (1969). Parallel processing techniques for estimating pitch periods of speech in the time domain. *Journal of the Acoustic Society of America*, 46:442–448.

Goto, M. (1999). A real-time music scene description system: Detecting melody and bass lines in audio signals. In *IJCAI Workshop on Computational Auditory Scene Analysis*, pages 31–40. http://www.etl.go.jp/ goto/PROJ/f0.html.

Goto, M. (2000). A robust predominant-f0 estimation method for real-time detection of melody and bass lines in cd recordings. In *IEEE International Conference on Acoustics Speech and Signal Processing*, pages 757–760. http://www.etl.go.jp/ goto/PROJ/f0.html.

Goto, M. and Muraoka, Y. (1999). Real-time beat tracking for drumless audio signals: chord change detection for musical decisions. *Speech Communication*, 27:311–335.

Gouyon, F. (2005). *A computational approach to rhythm description - Audio features for the computation of rhythm periodicity functions and their use in tempo induction and music content processing*. Doctoral dissertation, Universitat Pompeu Fabra.

Harris, F. J. (1978). On the use of windows for harmonic analysis with the discrete fourier transform. In *Proceedings of IEEE*, volume 66(1), pages 51–83.

Harte, C. and Sandler, M. (2005). Automatic chord identification using a quantised chromagram. In *118th Audio Engineering Society Convention*, Barcelona, Spain.

Harte, C., Sandler, M., Abdallah, S., and Gómez, E. (2005). Symbolic representation of musical chords: a proposed syntax for text annotations. In *6th International Conference on Music Information Retrieval*, London, UK.

Herrera, P. (2006). *Automatic classification of percussion sounds: from acoustic features to semantic descriptions*. Doctoral dissertation, in preparation, Universitat Pompeu Fabra.

Herrera, P., Bello, J., Widmer, G., Sandler, M., Celma, O., Vignoli, F., Pampalk, E., Cano, P., Pauws, S., and Serra, X. (2005a). Simac: Semantic interaction with music audio contents. In *2nd European Workshop on the Integration of Knowledge, Semantic and Digital Media Technologies*, Savoy Place, London, UK.

Herrera, P., Celma, O., Massaguer, J., Cano, P., Gómez, E., Gouyon, F., Koppenberger, M., García, D., Mahedero, J. P. G., and Wack, N. (2005b). Mucosa: a music content semantic annotator. In *International Conference on Music Information Retrieval*, London, UK.

Herrera, P., Sandvold, V., and Gouyon, F. (2004). Percussion-related semantic descriptors of music audio files. In *25th International Audio Engineering Society Conference*, London, UK.

Hess, W. (1983). *Pitch Determination of Speech Signals. Algorithms and Devices.* Springer Series in Information Sciences. Springer-Verlag, Berlin, New York, Tokyo, springer-verlag edition.

Hu, N., Dannenberg, R. B., and Tzanetakis, G. (2003). Polyphonic audio matching and aligment for music retrieval. In *IEEE Workshop of Applications of Signal Processing to Audio and Acoustics*, Barcelona, Spain.

Huron, D. (1994). Scores from the ohio state university cognitive and systematic musicology laboratory - bach, well-tempered clavier fugues, book ii. online. http://kern.ccarh.org/cgi-bin/ksbrowse?l=/osu/classical/bach/wtc-2.

Huron, D. and Parncutt, R. (1993). An improved model of tonality perception incorporating pitch salience and echoic memory. *Psychomusicology*, 12:152–169.

Izmirli, O. (2005). Template based key finding from audio. In *International Computer Music Conference*, Barcelona, Spain.

Janata, P., Birk, J., Horn, J. V., Leman, M., Tillmann, B., and Bharucha, J. (2002). The cortical topography of tonal structures underlying western music science. *Science*, 298:2167.

Jehan, T. (1997). *Musical Signal Parameter Estimation*. PhD thesis, CNMAT; IFSIC. http://www.cnmat.berkeley.edu/ tristan/Report/Report.html.

Juslin, P. N. and Sloboda, J. A. (2001). *Music and emotion*. Series in affective science. Oxford University Press, New York.

Kashino, K., Kinoshita, T., and Tanaka, H. (1995). Organization of hierarchical perceptual sounds: music scene analysis with autonomous processing modules and a quantitative information integration mechanism. In *International Joint Conference On Artificial Intelligence*, Montreal.

Klapuri, A. (2000a). Multiple fundamental frequency estimation by harmonicity and spectral smoothness. *IEEE Transactions in Speech and Audio Processing*, 11(6):804–816.

Klapuri, A. (2000b). Qualitative and quantitative aspects in the design of periodicity estimation algorithms. In *European Signal Processing Conference*.

Klapuri, A. (2004). *Signal processing methods for music transcription*. Doctoral dissertation, Tampere University of Technology.

Klapuri, A., Virtanen, T., Eronen, A., and Seppänen, J. (2001). Automatic transcription of musical recordings. In *Consistent & Reliable Acoustic Cues Workshop*.

Kohonen, T. (1984). *Self-organization and associative memory*. Springer-Verlag, Berlin.

Kostek, A. (2004). Automatic music transcription as we know it today. *Journal of New Music Research*, 33(3):269–282.

Kostka, S. and Payne, D. (1995). *Tonal Harmony*. McGraw Hill, New York.

Krumhansl, C. L. (1990). *Cognitive foundations of musical pitch*. Oxford University Press, New York.

Krumhansl, C. L. (2000). Tonality induction: a statistical approach applied cross-culturally. *Music Perception, Special Issue in Tonality Induction*, 17(4):461–481.

Krumhansl, C. L. (2004). The cognition of tonality - as we know it today. *Journal of New Music Research*, 33(3):253–268.

Lahat, A., Niederjohn, R. J., and Krubsack, D. A. (1987). A spectral autocorrelation method for measurement of the fundamental frequency of noise-corrupted speech. *IEEE Transactions on Acoustics, Speech and Signal Processing*, 35(6):741–750.

Laroche, J. (1995). Traitement des signaux audio-fréquences. Technical report, Ecole National Supérieure de Télécommunications.

Leman, M. (1991). *Een model van toonsemantiek: naar een theorie en discipline van de muzikale verbeelding*. Doctoral dissertation, University of Ghent.

Leman, M. (1994). Schema-based tone center recognition of musical signals. *Journal of New Music Research*, 23(2):169–204.

Leman, M. (1995a). A model of retroactive tone center perception. *Music Perception*, 12(4):439–471.

Leman, M. (1995b). *Music and schema theory: cognitive foundations of systematic musicology*. Number 31 in Information Science. Springer-Verlag, Berlin-Heidelberg.

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):481–509.

Leman, M. (2002). Musical audio mining. In *Dealing with the data flood Symposium*, Rotterdam.

Leman, M., Clarisse, L., Baets, B. D., Meyer, H. D., Lesaffre, M., Martens, G., Martens, J., and Steelant, D. V. (2002). Tendencies, perspectives, and opportunities of musical audio-mining. In *Forum Acusticum, session SS-MUS-01*, Sevilla.

Lerdahl, F. (2001). *Tonal pitch space*. Oxford University Press, New York.

Lerdahl, F. and Jackendoff, R. (1983). *A generative theory of tonal music*. MIT Press, Cambridge, Massachusetts.

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.

Lewin, D. (1987). *Generalized musical intervals and transformations*. Yale University Press, New Haven.

Logan, B. and Salomon, A. (2001). A music similarity function based on signal analysis. In *International Conference on Multimedia and Expo*, Tokyo, Japan.

Longuet-Higgins, H. C. and Steedman, M. J. (1971). On interpreting bach. *Machine Intelligence*, 6:221–241.

Maher, R. C. and Beauchamp, J. W. (1993). Fundamental frequency estimation of musical signals using a two-way mismatch procedure. *Journal of the Acoustic Society of America*, 95:2254–2263.

Manjuhath, B., Salembier, P., and Sikora, T. (2002). *Introduction to MPEG-7: Multimedia Content Description Language*. Wiley and Sons, New York.

Martens, G., Meyer, H. D., Baets, B. D., and Leman, M. (2004). Distance-based versus tree-based key recognition in musical audio. *Soft Computing (online)*.

Martens, G., Meyer, H. D., Baets, B. D., Leman, M., Martens, J.-P., Clarisse, L., and Lesaffre, M. (2002). A tonality-oriented symbolic representation of musical audio generated by classification trees. In *EUROFUSE workshop on Information Systems*, pages 49–54.

Martin, K. D. (1996). Automatic transcription of simple polyphonic music: Robust front end processing. In *Third Joint Meeting of the Acoustical Societies of America and Japan*.

McClelland, J. H., Schafer, R. W., and Yoder, M. A. (1998). *DSP First: A Multimedia Approach*. Prentice Hall.

Medan, J., Yair, E., and Chazan, D. (1991). Super resolution pitch determination of speech signals. *IEEE Transactions on Signal Processing*, 39(1).

Meddis, R. and Hewitt, M. J. (1991). Virtual pitch and phase sensitivity of a computer model of the auditory periphery: Pitch identification. *JASA*, 89(6):2866–2882.

Morse, P. M. (1983). *Vibration and sound*. American Institute of Physics for the Acoustical Society of America, second (paperback) printing edition.

Noll, A. M. (1967). Cepstrum pitch determination. *Journal of the Acoustic Society of America*, 41:293–309.

Ong, B. and Herrera, P. (2004). Computing structural descriptions of music through the identification of representative excerpts from audio files. In *25th International Audio Engineering Society Conference*, London, UK.

Ong, B. and Herrera, P. (2005). Semantic segmentation of music audio contents. In ICMA, editor, *International Computer Music Conference*, Barcelona, Spain.

Pampalk, E. (2004). A matlab toolbox to compute music similarity from audio. In *International Conference on Music Information Retrieval*, Barcelona, Spain.

Pampalk, E., Dixon, S., and Widmer, G. (2003). On the evaluation of perceptual similarity measures for music. In *International Conference on Digital Audio Effects*, London, UK.

Pampalk, E., Flexer, A., and Widmer, G. (2005). Improvements of audio-based music similarity and genre classification. In *International Conference on Music Information Retrieval*, London, UK.

Pauws, S. (2004). Musical key extraction from audio. In *International Conference on Music Information Retrieval*, Barcelona, Spain.

Piszczalski, M. and Galler, B. A. (1979). Predicting musical pitch from component frequency ratios. *Journal of the Acoustic Society of America*, 66:710–720.

Platt, J. (1998). Fast training of support vector machines using sequential minimal optimization. In Schoelkopf, B., Burgesa, C., and Smola, A., editors, *Advances in Kernel Methods - Support Vector Learning*. MIT Press.

Pollack, A. W. (1999). Notes on...series. *Soundscapes, Journal on Media Culture*, 1. http://www.icce.rug.nl/ soundscapes/DATABASES/AWP/awp-notes_on.html.

Purwins, H. (2005). *Profiles of Pitch Classes. Circularity of Relative Pitch and Key: Experiments, Models, Computational Music Analysis, and Perspectives*. Doctoral dissertation, Berlin University of Technology.

Purwins, H., Blankertz, B., and Obermayer, K. (2000). A new method for tracking modulations in tonal music in audio data format. *Neural Networks - IJCNN, IEEE Computer Society*, 6:270–275.

Purwins, H., Graepel, T., Blankertz, B., and Obermayer, K. (2003). Correspondence analysis for visualizing interplay of pitch class, key, and composer. In E. L. Puebla, G. M. and Noll, T., editors, *Perspectives in Mathematical Music Theory*. Verlag ep0s-Music.

Rabiner, L. R., Sambur, M. R., and Schmidt, C. E. (1975). Applications of a nonlinear smoothing algorithm to speech processing. *IEEE Transactions on ASSP*, 23(6).

Rabiner, L. R. and Schafer, R. W. (1978). *Digital Processing of Speech Signals*. Prentice-Hall.

Roads, C. (1996). Pitch and rhythm recognition in midi systems. In *The Computer Music Tutorial*, pages 503–531. The MIT Press.

Romero, J. and Cerdá, S. (1997). Uso del análisis multirresolución para calcular el pitch de señales en presencia de ruido. *Revista de Acústica (SEA)*, 28. http://www.ia.csic.es/Sea/.

Rossing, T. D. (1989). *The sience of sound*. Addison-Wesley, second edition.

Sadie, S., Tyrrell, J., and Kernfeld, B. (2005). Grove music online: the new grove dictionary of music and musicians, second edition, the new grove dictionary of opera and the new grove dictionary of jazz, second edition. online publication. http://www.grovemusic.com.

Sandvold, V. and Herrera, P. (2005). Towards a semantic descriptors of subjective intensity in music. In *International Computer Music Conference*, Barcelona, Spain.

Sapp, C. S. (2001). Visualizations of tonality: Key determination algorithm. online. http://ccrma.stanford.edu/ craig/keyscape/intro/algorithm.

Schmuckler, M. A. (2004). Pitch and pitch structures. In Neuhoff, J., editor, *Ecological Psychoacoustics*, pages 271–315. Elsevier.

Schwarz, D. and Rodet, X. (1999). Spectral envelope estimation and representation for sound analysis-synthesis. In *International Computer Music Conference*, pages 351–354, Beijing, China.

Serra, X. (1996). Musical sound modeling with sinusoids plus noise. In Poli, G. D., Picialli, A., Pope, S. T., and Roads, C., editors, *Musical Signal Processing*. Swets & Zeitlinger.

Sheh, A. and Ellis, D. (2003). Chord segmentation and recognition using em-trained hidden markov models. In *4th International Symposium on Music Information Retrieval*, Baltimore.

Shepard, R. N. (1982). Structural representations of musical pitch. In Deutsch, D., editor, *The Psychology of Music, First Edition*. Swets & Zeitlinger.

Shmulevich, I. and Yli-Harja, O. (2000). Localized key finding: algorithms and applications. *Music Perception, Special Issue in Tonality Induction*, 17(4):531–544.

Smeulders, A., Worring, M., Santini, S., Gupta, A., and Jain, R. (2000). Content-based image retrieval at the end of the early years. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22.

Solomon, L. (1996). Terms. http://www.azstarnet.com/ solo/glossary.htm.

Streich, S. (2005). *Automatic characterization of music complexity: a multi-faceted approach*. Doctoral pre-thesis work, Universitat Pompeu Fabra, Barcelona, Spain.

Streich, S. and Herrera, P. (2004). Toward describing perceived complexity of songs: computational methods and implementation. In *25th International Audio Engineering Society Conference*, London, UK.

Streich, S. and Herrera, P. (2005). Detrended fluctuation analysis of music signals: danceability estimation and further semantic characterization. In *118th Audio Engineering Society Convention*, Barcelona, Spain.

Talkin, D. (1995). Robust algorithm for pitch tracking. In Kleijn, W. B. and Paliwal, K. K., editors, *Speech Coding and Synthesis*. Elsevier Science B. V.

Temperley, D. (1999). What's key for key? the krumhansl-schmuckler key finding algorithm reconsidered. *Music Perception*, 17(1):65–100.

Temperley, D. (2001). Meter, harmony, and tonality in rock. In *The cognition of basic musical structures*, pages 237–264. The MIT Press.

Temperley, D. (2005). A bayesian key-finding model. In *2005 MIREX Contest - Symbolic Key Finding*. http://www.music-ir.org/evaluation/mirex-results/sym-key/index.html.

Terhardt, E. (1979). Calculating virtual pitch. *Hearing Research*, 1:155–182.

Terhardt, E., Stoll, G., and Seewann, M. (1981). Algorithm for extraction of pitch and pitch salience from complex tonal signals. *Journal of the Acoustic Society of America*, 71:679–688.

Tillmann, B. and Bigand, E. (2006). A comparative review of priming effects in language and music. In McKevitt, P., editor, *Language, visio and music*. Cognitive Science of Natural Language Processing CSNLP-8, John Benjamins, in press.

Tillmann, B., Janata, P., Birk, J., and Barucha, J. J. (2003). The costs and benefits of tonal centers for chord processing. *Journal of Experimental Psychology: Human Perception and Performance*, 29(2):470–482.

Toiviainen, P. and Krumhansl, C. L. (2003). Measuring and modeling real-time responses to music: the dynamics of tonality induction. *Perception*.

Tzanetakis, G. (2002). Pitch histograms in audio and symbolic music information retrieval. In *3rd International Symposium on Music Information Retrieval*, Paris.

Vignoli, F. and Pauws, S. (2005). A music retrieval system based on user-driven similarity and its evaluation. In *International Symposium on Music Information Retrieval*, London, UK.

Vos, P. G. (2000). Tonality induction: theoretical problems and dilemmas. *Music Perception, Special Issue in Tonality Induction*, 17(4):403–416.

Walmsley, P. J., Godsill, S. J., and Rayner, P. J. W. (1999). Bayesian graphical models for polyphonic pitch tracking. In *Diderot Forum*, Vienna.

Winograd, T. (1968). Linguistics and the computer analysis of tonal harmony. *Journal of Music Theory*, 12(1):Reprinted in Stephan Schwanauer and David Levitt, Eds., Machine Models of Music, MIT Press, 1993, pp. 113–153.

Witmer, R. and Marks, A. (accessed March 2006). Cover, cover version and cover record. In Macy, L., editor, *Grove Music Online*. http://www.grovemusic.com.

Yang, C. (2001). Music database retrieval based on spectral similarity. In *International Symposium on Music Information Retrieval (ISMIR)*.

Zhu, Y., Kankanhalli, M. S., and Gao, S. (2005). Music key detection for musical audio. In *11th International Multimedia Modelling Conference (MMM'05)*.

Zölzer, U. (2002). Introduction - fundaments of digital signal processing. In Zölzer, U., editor, *DAXF-Digital Audio Effects*, pages 2–28. John Wiley & Sons.

Appendix A

Audio samples

A.1 Audio samples in Chapter 3

These audio samples can be found in http://www.iua.upf.es/~egomez/thesis

1. *Tune1-0.mid.wav, Tune1-20.mid.wav* and *Tune1-60.mid.wav*: piece with a frequency deviation of 0, 20 and 50 cents. Figure 3.8 shows and example of the instantaneous evolution of the estimated deviation d for these files measured in cents.

2. *Noise.wav*: white noise generated by Sound Forge [1].

3. *PianoPhrase.wav*: excerpt of a monophonic piano phrase.

4. *Donde-Estas-Yolanda.wav*: excerpt of a polyphonic piece: *Donde est ás, Yolanda*, by Pink Martini.

5. *Imagine.wav*: excerpt of a polyphonic piece: *Imagine*, by John Lennon.

6. *Book-I-Prelude-1-C-Version1Jaccottet.wav*: Bach's WTC Prelude 1 in C Major played by Christiane Jaccottet (harpsichord), Vienna Master Series, PILZ, Germany.

7. *Book-I-Prelude-2-c-Version1Jaccottet.wav*: Bach's WTC Prelude 2 in C Minor played by Christiane Jaccottet (harpsichord), Vienna Master Series, PILZ, Germany.

8. *Book-I-Prelude-1-C-Version2Leonhardt.wav*: Bach's WTC Prelude 1 in C Major played by Gustav Leonhardt (harpsichord), Classical Edition, Deutsche harmonia Mundi, Germany.

9. *Book-I-Prelude-2-c-Version2Leonhardt.wav*: Bach's WTC Prelude 2 in C Minor played by Gustav Leonhardt (harpsichord), Classical Edition, Deutsche harmonia Mundi, Germany.

[1] http://www.sonicfoundry.com

10. *Book-I-Prelude-1-C-Version3GlennGould.wav*: Bach's WTC Prelude 1 in C Major played by Glenn Gould (piano), SONY Classical.

11. *Book-I-Prelude-2-c-Version3GlennGould.wav*: Bach's WTC Prelude 2 in C Minor played by Glenn Gould (piano), SONY Classical.

A.2 Audio samples in Chapter 4

1. *Organ-C-Major.wav*: Excerpt of organ in C Major.

2. *Piano-F-sharp-minor.wav*: Excerpt of a piano piece, Hungarian Dances from Johannes Brahms, in F# minor.

3. *Percussion.wav*: Drum loop.

4. *Schoenberg.wav*: Excerpt of String Quartet Op. 30 I Moderato, from Arnold Schoenberg, performed by Aron Quartet.

A.3 Audio samples in Chapter 5

1. Type 0: *e114.wav* and *e043.wav*: These sound samples are proposed in Yang (2001), and are found in http://www-db.stanford.edu/~yangc/musicir/.

2. Type II: *e043.wav* and *e108.wav*: These sound samples are proposed in Yang (2001), and are found in http://www-db.stanford.edu/~yangc/musicir/.

3. Type III: *e116.wav* and *e117.wav*: These sound samples are proposed in Yang (2001), and are found in http://www-db.stanford.edu/~yangc/musicir/.

4. Type IV: *e071.wav* and *e107.wav*: These sound samples are proposed in Yang (2001), and are found in http://www-db.stanford.edu/~yangc/musicir/.

5. Type V: *e106.wav* and *e114.wav*: These sound samples are proposed in Yang (2001), and are found in http://www-db.stanford.edu/~yangc/musicir/.

6. Type VI: first phrase of the song *Imagine*, by John Lennon, in 6 different versions: *ImagineFirstPhrase1-JohnLennon.wav*, *ImagineFirstPhrase2-Instrumental.wav*, *ImagineFirstPhrase3-DianaRoss.wav*, *ImagineFirstPhrase4-TaniaMaria.wav*, *ImagineFirstPhrase5-Khaled.wav*, *ImagineFirstPhrase6-Noa.wav*.

7. Type VII: the song *Imagine*, by John Lennon, in 5 different versions: *Imagine1-JohnLennon.wav*, *Imagine2-Instrumental.wav*, *Imagine3-DianaRoss.wav*, *Imagine4-TaniaMaria.wav*, *Imagine5-NoaKhaled.wav*.

8. Other: *Besame Mucho,* by Diana Krall: *BesameMucho-DianaKrall.wav*.
9. Time stretched versions: *Imagine2-Instrumental-50%.wav, Imagine2-Instrumental-70%.wav, Imagine2-Instrumental-130%.wav, Imagine2-Instrumental-160%.wav*.
10. High tonal strength examples: *YouDontBringMeFlowers.mp3,* by Barbra Streisand (Duets retail) and *Atrapame.mp3,* by Manny Manuel.
11. Low tonal strength example: *StopOrIllShoot.mp3,* by Ryuichi Sakamoto.

Appendix B

Details on the comparison of tonal models for key estimation

We present here the detailed results of the evaluation presented in Section 4.3.3. Different profiles are evaluated: diatonic, Krumhansl and Schmuckler profiles (Krumhansl (1990)) , Temperley's theoretical profiles from Temperley (1999), Chai (2005), tonic triad, Temperley's empirical profiles from Temperley (2005) and Average profiles. For each of the profiles, we have evaluated three different configurations: original values (1), modified profiles considering the three main chords of the key (2) and modified profiles considering all the chords of the key (3). The results are presented for five different musical collections: Fugue subject of Bach's Well-Tempered clavier (1), ISMIR 2005 contest training data (2), classical database (3), the Beatles's collection (5) and varied style collection (5). The evaluation measures presented are the following ones: ISMIR 2005 contest evaluation measure, percentage of correct estimation, percentage of correct mode estimation, percentage of tuning, fifth, parallel and relative errors (close tonalities), and finally percentage of errors with other tonalities. We refer to Chapter 4 for explanations.

Profile	Config.	CollectionID	Contest	nCorrect	nCorrectMode	nTuning	nFifth	nParallel	nRelative	nOther
Diatonic	1	1	75.41	72.97	72.97	0.00	0.00	0.00	8.11	18.92
Diatonic	1	2	59.90	44.79	60.42	0.00	14.58	0.00	26.04	14.58
Diatonic	1	3	59.65	46.47	67.77	3.19	13.67	1.82	19.93	14.92
Diatonic	1	4	57.54	46.86	86.86	4.00	18.86	1.14	3.43	25.71
Diatonic	1	5	29.81	17.59	48.15	2.78	16.67	5.56	9.26	48.15
Diatonic	1	global	57.50	45.26	68.04	2.84	13.82	1.79	16.58	19.72
Diatonic	2	1	79.05	78.38	78.38	0.00	0.00	1.35	1.35	18.92
Diatonic	2	2	89.58	83.33	90.63	0.00	7.29	2.08	7.29	0.00
Diatonic	2	3	76.36	71.07	86.79	6.49	6.61	2.73	4.78	8.31
Diatonic	2	4	73.20	64.00	86.86	2.86	14.29	5.14	3.43	10.29
Diatonic	2	5	53.61	42.59	62.96	3.70	14.81	11.11	4.63	23.15
Diatonic	2	global	**74.75**	**68.71**	**84.17**	4.93	7.92	3.58	4.56	10.31
Diatonic	3	1	85.68	82.43	82.43	0.00	0.00	0.00	10.81	6.76
Diatonic	3	2	71.98	62.50	68.75	0.00	5.21	0.00	22.92	9.38
Diatonic	3	3	66.66	57.86	72.78	4.56	6.83	2.16	16.51	12.07
Diatonic	3	4	60.34	53.14	82.86	4.57	11.43	2.29	3.43	25.14
Diatonic	3	5	33.06	19.44	53.70	5.56	19.44	5.56	9.26	40.74
Diatonic	3	global	64.16	55.49	72.37	4.03	7.92	2.17	14.26	16.13
Krumhansl	1	1	77.57	77.03	77.03	0.00	0.00	2.70	0.00	20.27
Krumhansl	1	2	73.65	59.38	85.42	0.00	26.04	3.13	2.08	9.38
Krumhansl	1	3	57.36	42.03	77.68	3.53	27.22	2.96	3.76	20.50
Krumhansl	1	4	71.37	60.00	86.29	2.86	18.86	4.57	3.43	10.29
Krumhansl	1	5	54.07	43.52	60.19	1.85	11.11	16.67	5.56	21.30
Krumhansl	1	global	60.87	47.42	77.45	2.84	23.08	4.26	3.51	18.89
Krumhansl	2	1	78.65	78.38	78.38	0.00	0.00	1.35	0.00	20.27
Krumhansl	2	2	90.10	84.38	91.67	0.00	7.29	1.04	6.25	1.04
Krumhansl	2	3	77.11	72.10	87.13	6.15	6.15	2.85	4.56	8.20

Krumhansl	2	4	72.46	61.71	86.86	2.29	17.14	5.71	3.43	9.71
Krumhansl	2	5	56.76	45.37	65.74	3.70	15.74	9.26	5.56	20.37
Krumhansl	2	global	**75.41**	**69.38**	**84.69**	4.63	8.07	3.51	4.33	10.08
Krumhansl	3	1	80.81	78.38	78.38	0.00	0.00	0.00	8.11	13.51
Krumhansl	3	2	81.35	73.96	79.17	0.00	4.17	0.00	17.71	4.17
Krumhansl	3	3	70.93	64.01	78.02	5.35	5.58	2.39	12.19	10.48
Krumhansl	3	4	69.43	63.43	85.71	4.57	9.14	2.86	2.86	17.14
Krumhansl	3	5	42.87	30.56	59.26	3.70	17.59	6.48	7.41	34.26
Krumhansl	3	global	69.34	62.36	77.15	4.41	6.57	2.46	10.68	13.52
Temperley	1	1	86.22	85.14	85.14	0.00	0.00	1.35	2.70	10.81
Temperley	1	2	76.77	61.46	84.38	0.00	21.88	0.00	14.58	2.08
Temperley	1	3	67.07	56.83	76.99	4.33	12.07	1.71	12.87	12.19
Temperley	1	4	70.69	62.86	90.29	5.14	13.14	2.86	2.29	13.71
Temperley	1	5	52.13	39.81	62.04	2.78	13.89	12.96	9.26	21.30
Temperley	1	global	67.69	57.80	78.04	3.73	12.32	2.61	10.68	12.85
Temperley	2	1	78.65	78.38	78.38	0.00	0.00	1.35	0.00	20.27
Temperley	2	2	90.10	84.38	91.67	0.00	7.29	1.04	6.25	1.04
Temperley	2	3	77.35	72.55	87.36	6.38	5.69	2.73	4.67	7.97
Temperley	2	4	73.20	63.43	87.43	2.86	15.43	5.14	3.43	9.71
Temperley	2	5	56.02	45.37	63.89	3.70	13.89	10.19	5.56	21.30
Temperley	2	global	**75.61**	**69.90**	**84.76**	4.85	7.39	3.44	4.41	10.01
Temperley	3	1	86.22	83.78	83.78	0.00	0.00	0.00	8.11	8.11
Temperley	3	2	81.04	73.96	79.17	0.00	4.17	0.00	16.67	5.21
Temperley	3	3	70.34	62.98	77.79	5.24	6.38	2.39	12.30	10.71
Temperley	3	4	63.89	57.14	84.00	6.86	10.86	2.29	2.86	20.00
Temperley	3	5	38.06	24.07	57.41	3.70	21.30	5.56	7.41	37.96
Temperley	3	global	68.12	60.64	76.92	4.63	7.62	2.32	10.68	14.12
Chai	1	1	48.65	44.59	44.59	0.00	0.00	20.27	0.00	35.14

Chai	1	2	71.98	62.50	75.00	0.00	9.38	8.33	10.42	9.38
Chai	1	3	59.64	50.46	66.29	4.10	8.09	9.45	10.82	17.08
Chai	1	4	62.74	54.29	71.43	2.86	6.86	2.86	14.86	18.29
Chai	1	5	42.22	29.63	49.07	1.85	12.04	14.81	12.04	29.63
Chai	1	global	58.56	49.51	64.60	3.21	7.84	9.48	10.75	19.19
Chai	2	1	58.11	58.11	58.11	0.00	0.00	0.00	0.00	41.89
Chai	2	2	81.46	71.88	88.54	0.00	14.58	2.08	6.25	5.21
Chai	2	3	63.61	55.81	77.45	4.90	8.77	3.76	8.88	17.88
Chai	2	4	57.20	40.57	73.71	2.29	26.86	4.00	8.00	18.29
Chai	2	5	42.69	30.56	61.11	1.85	18.52	4.63	6.48	37.96
Chai	2	global	**61.68**	**52.73**	**74.91**	3.66	11.80	3.51	7.84	20.46
Chai	3	1	60.81	60.81	60.81	0.00	0.00	0.00	0.00	39.19
Chai	3	2	81.46	75.00	84.38	0.00	6.25	1.04	10.42	7.29
Chai	3	3	64.16	55.69	76.65	5.13	9.91	2.51	10.02	16.74
Chai	3	4	51.20	35.43	70.29	2.29	26.86	2.29	6.29	26.86
Chai	3	5	36.67	20.37	60.19	2.78	26.85	1.85	8.33	39.81
Chai	3	global	60.92	51.53	73.71	3.88	12.62	2.17	8.81	20.99
Triad	1	1	74.32	74.32	74.32	0.00	0.00	0.00	0.00	25.68
Triad	1	2	87.08	81.25	91.67	0.00	9.38	1.04	3.13	5.21
Triad	1	3	69.23	60.71	82.00	5.35	13.44	2.51	4.33	13.67
Triad	1	4	68.57	60.57	78.29	1.71	8.00	8.00	8.00	13.71
Triad	1	5	63.06	53.70	66.67	1.85	7.41	18.52	6.48	12.04
Triad	1	global	**69.79**	**61.99**	**80.06**	3.88	11.13	4.26	4.63	14.12
Triad	2	1	33.78	33.78	33.78	0.00	0.00	0.00	0.00	66.22
Triad	2	2	73.75	58.33	88.54	0.00	29.17	1.04	2.08	9.38
Triad	2	3	63.09	51.14	82.92	5.35	21.30	3.08	2.28	16.86
Triad	2	4	48.91	20.57	84.57	2.29	55.43	2.29	0.57	18.86
Triad	2	5	35.65	18.52	60.19	2.78	31.48	2.78	2.78	41.67

Triad	2	global	57.79	43.76	78.49	4.03	25.84	2.61	1.94	21.81
Triad	3	1	22.97	22.97	22.97	0.00	0.00	0.00	0.00	77.03
Triad	3	2	56.46	43.75	68.75	0.00	22.92	0.00	4.17	29.17
Triad	3	3	52.52	42.03	69.25	3.08	18.56	2.28	2.51	31.55
Triad	3	4	45.49	18.86	83.43	3.43	52.57	1.71	0.00	23.43
Triad	3	5	30.93	13.89	53.70	2.78	30.56	1.85	4.63	46.30
Triad	3	global	48.19	35.55	66.84	2.69	23.15	1.87	2.32	34.43
Temperley empirical	1	1	88.11	87.84	87.84	0.00	0.00	1.35	0.00	10.81
Temperley empirical	1	2	89.58	82.29	92.71	0.00	10.42	1.04	6.25	0.00
Temperley empirical	1	3	74.65	67.54	85.88	5.92	9.79	2.51	5.69	8.54
Temperley empirical	1	4	74.17	66.29	88.57	3.43	12.00	5.14	2.86	10.29
Temperley empirical	1	5	55.74	44.44	63.89	2.78	11.11	14.81	9.26	17.59
Temperley empirical	1	global	**74.43**	**67.29**	**84.54**	4.56	9.63	3.66	5.30	9.56
Temperley empirical	2	1	77.30	77.03	77.03	0.00	0.00	1.35	0.00	21.62
Temperley empirical	2	2	89.06	83.33	90.63	0.00	7.29	1.04	6.25	2.08
Temperley empirical	2	3	75.59	70.05	86.67	6.26	7.40	3.08	4.10	9.11
Temperley empirical	2	4	68.40	54.86	85.71	1.71	23.43	4.00	3.43	12.57

Temperley empirical	2	5	53.89	41.67	66.67	4.63	18.52	6.48	5.56	23.15
Temperley empirical	2	global	73.51	66.69	84.17	4.71	9.93	3.21	4.03	11.43
Temperley empirical	3	1	73.38	72.97	72.97	0.00	0.00	0.00	1.35	25.68
Temperley empirical	3	2	79.79	72.92	77.08	0.00	3.13	0.00	17.71	6.25
Temperley empirical	3	3	68.76	61.28	77.68	5.35	7.86	2.39	10.25	12.87
Temperley empirical	3	4	59.89	45.14	81.71	2.86	26.29	2.86	3.43	19.43
Temperley empirical	3	5	37.22	23.15	58.33	2.78	23.15	2.78	6.48	41.67
Temperley empirical	3	global	65.69	57.21	75.88	4.11	10.68	2.17	9.04	16.80
Average	1	1	87.16	83.78	91.89	1.35	6.76	0	0	8.12
Average	1	2	91.67	86.46	92.71	0	6.25	1.04	6.25	0
Average	1	3	77.33	71.87	88.95	6.38	7.52	2.73	3.87	7.63
Average	1	4	72.91	65.14	81.14	2.28	9.14	5.71	6.86	10.85
Average	1	5	56.11	48.15	67.59	5.55	9.26	13.89	1.85	21.29
Average	1	global	**76.15**	**70.35**	**86.12**	5.01	7.69	3.73	4.03	9.19
Average	2	1	80	70.27	90.54	1.35	18.92	1.35	0	8.11
Average	2	2	86.66	79.17	90.62	0	11.46	1.04	5.21	3.12
Average	2	3	73.53	66.40	87.93	6.49	10.71	3.42	3.64	9.34
Average	2	5	64.4	46.86	85.71	2.28	31.43	4	3.43	12
Average	2	6	47.87	34.26	64.81	4.63	23.14	4.63	3.70	29.63
Average	2	global	71.13	61.99	85.59	5.00	14.86	3.29	3.51	11.35

Average	3	1	79.19	70.27	89.19	2.7	16.22	0	2.7	8.11
Average	3	2	81.56	75	80.21	0	5.21	1.04	12.5	6.25
Average	3	3	70.35	62.53	81.55	6.26	9.22	2.39	9.11	10.48
Average	3	4	53.43	41.71	71.43	3.43	19.43	4	4	27.43
Average	3	5	37.59	19.44	61.11	1.85	32.41	5.55	2.78	37.96
Average	3	global	66.37	57.28	78.42	4.85	12.47	2.61	7.77	15.01

Table B.1: Detailed results of the evaluation for different profiles.

Appendix C

Details on similarity between versions and original pieces

We present here some details about the music collection used to evaluate how tonal similarity is useful for version identification. We classify the pairs of pieces (version vs root) according to different modifications of musical features:

- Type I: identical digital copy
- Type II: noise
- Type III: instrumentation
- Type IV: tempo
- Type V: harmonization
- Type VI: transposition
- Type VII: structure

Song title	Version	Root	Type	Observations
Come together	Aerosmith	The Beatles	II,III,IV,VII	
Come together	Afónicos	The Beatles	II,III,IV,VI,VII	
Sgt. Pepper's lonely hearts club band	The cast	The Beatles	II,III,IV,V,VI,VII	Transposition in the different refrains
Got to get you into my life	Earth Wind and Fire	The Beatles	II,III,IV,V,VII	Completely different
Imagine	Tania Maria	John Lennon	II,III,IV,V,VII	
Imagine	Khaled and Noa	John Lennon	II,III,IV,VI,VII	
Imagine	Instrumental	John Lennon	II,III	
Imagine	Diana Ross	John Lennon	II,III,IV,VI,VII	

My way	Adrivalan Orchestra	Frank Sinatra	II,III,IV,VI,VII	Slight changes in harmonization
My way	Tom Jones	Frank Sinatra	II,III,IV,VI	Slight changes in harmonization
My way	Keely Smith	Frank Sinatra	II,III,IV,V,VI	
Here comes the sun	Sandy Farina	The Beatles	II,III,IV,VI,VII	Changes in the melody, tempo, structure, instrumentation
Here comes the sun	Johan Pizzarelli	The Beatles	II,III,IV,V,VI,VII	Completely different, jazz version
Here comes the sun	Carmen Cuesta	The Beatles	II,III,IV,V,VI,VII	
Strawberry fields forever	Sandy Farina	The Beatles	II,III,IV,V,VI,VII	Only the main melody is kept
Strawberry fields forever	Peter Gabriel	The Beatles	II,III,IV,V,VI,VII	Only the main melody is kept
Lucy in the sky with diamonds	Dianne Stienberg	The Beatles	II,III,IV,VI	Tempo slightly different
Lucy in the sky with diamonds	Elton John	The Beatles	II,III,IV,VI,VII	Tempo slightly different
Maxwell's silver hammer	Steve Martin	The Beatles	II,III,IV,V,VI,VII	Completely different version, even a different melody
Maxwell's silver hammer	John Pizzarelli	The Beatles	II,III,IV,V,VI,VII	Completely different harmonization, just the main theme is kept
When I'm sixty four	Benito Cabrera	The Beatles	II,III,IV,VI,VII	instrumental
You've got a friend	Paul Carrack	James Taylor	II,III,IV,VI	Slight different tempo and harmony
She came in through the bathroom window	Polythene pam	The Beatles	II,III,IV,VII	Slight different tempo and harmony
She came in through the bathroom window	Joe Cocker	The Beatles	II,III,IV,VII	Slight different tempo and harmony
Getting better	Peter Frampton and The Bee Gees	The Beatles	II,III	
Baby can I hold you	Byzone	Tracy Chapman	II,III,IV,VI	
Oh! Darling	John Pizzarelli	The Beatles	II,III,IV,V,VI,VII	
Oh! Darling	Robin Gibb	The Beatles	II,III,V,VI,VII	Slight changes in harmonization
The long and winding road	The Beatles (concert)	The Beatles	II	Concert version
The long and winding road	Peter Frampton	The Beatles	III,VII	Slight tempo variation, instrumental section added

Mean Mr Mustard	Frankie Howerd	The Beatles	II,III,IV,VI,VII	Different voice, very different melody. Very far from the original
Besame Mucho	Joao Gilberto	Leticia Herrero	II,III,IV,V,VII,VII	Only the main melody is kept
Besame Mucho	Isaac Turienzo	Leticia Herrero	II,III,IV,V,VII,VII	Only the main melody is kept
Besame Mucho	Bobby Morganstein	Leticia Herrero	II,III,IV,V,VII,VII	Only the main melody is kept
Besame Mucho	Diana Krall	Leticia Herrero	II,III,IV,V,VII,VII	Only the main melody is kept
Fixing a hole	George Burns	The Beatles	II,III	
I want you	The Bee Gees	The Beatles	II,III,IV,VI	
I want you	The Anthony Wilson Trio	The Beatles	II,III,IV	
No woman no cry	Bob Marley (concert)	Bob Marley	II,VII	Concert version, just one excerpt of the song
No woman no cry	Bob Marley (2)	Bob Marley	II	
She's leaving home	The Bee Gees	The Beatles	II,III,IV,VI	
Because	Alice Cooper and The Bee Gees	The Beatles	II,III,IV,VI,VII	
Good morning good morning	Peter Frampton and The Bee Gees	The Beatles	II,III,IV	
Being for the benefit of Mr Kite	Maurice Gibb and The Bee Gees	The Beatles	II,III,IV,V,VI	Changes in harmony
Golden Slumbers	Peter Frampton	The Beatles	II,III,IV,VI,VII	
All of me	Duke Ellington	Billie Holiday	II,III,IV,V,VI,VII	Sax vs voice, small changes in melody and harmonization jazz standard
All of me	Billie Holiday	Billie Holiday	II	Different remastering jazz standard
All of me	Johny Hartman	Billie Holiday	II,III,IV,V,VI,VII	Male vs female jazz standard
All of me	Johny Hartman	Billie Holiday	II,III,IV,V,VI,VII	Male vs female jazz standard
All of me	Canal Street Jazz Band	Billie Holiday	II,III,IV,C,VI,VII	Sax vs voice jazz standard
All of me	Jose Luis Cortés y NG La Banda	Billie Holiday	II,III,IV,V,VI,VII	Salsa version, different harmony and changes in main melody jazz standard
All of me	Orquesta de Músicos de Barcelona, Tete Montoliú	Billie Holiday	II,III,IV,V,VII	Instrumental jazz standard

All of me	Bobby Morganstein Productions	Billie Holiday	II,III,IV,V,VI,VII	Different voice and overall melody jazz standard
All of me	Frank Sinatra	Billie Holiday	II,III,IV,V,VI,VII	Male vs female jazz standard
You never give me your money	Paul Nicholas	The Beatles	II,III,IV,VI	
A day in the life	Barry Gibb and The Bee Gees	The Beatles	II,III,VI	

Table C.1: Classification of similarity evaluation Material.

Appendix D

Related publications by the author

In this annex, we provide a list of publications which are relevant to this PhD dissertation in which its author has participated. Abstracts and electronic versions of most of these publications, as well as a list of other publications from the author non related to this dissertation are available from http://www.iua.upf.es/mtg

D.1 Journal articles

- **Authors**: Emilia Gómez

  Title: Tonal description of polyphonic audio for music content processing

  Journal: INFORMS Journal on Computing, Special Cluster on Computation in Music, Vol. 18(3)

  Editor: Elaine Chew

  Year: 2006 (Accepted 2004)

  Related to Chapter 3 and 4

- **Authors**: Emilia Gómez, Anssi Klapuri and Benoit Meudic

  Title: Melody description and extraction in the context of music content processing

  Journal: Journal of New Music Research, Vol. 32(1)

  Editor: Xavier Serra

  Year: 2003

  Related to Chapter 2

D.2 Book chapters

- **Authors**: Fabien Gouyon, Xavier Amatriain, Jordi Bonada, Pedro Cano, Emilia Gómez, Perfecto Herrera and Alex Loscos

  Title: Content processing of musical audio signals

  Book title: Sound to sense, sense to sound: A state-of-the-art

Year: in press

Related to Chapter 2

D.3 Theses and reports

- **Authors**: Emilia Gómez, Sebastian Streich, Beesuan Ong, Rui Pedro Paiva, Sven Tappert, Jan-Mark Batke, Graham Poliner, Dan Ellis and Juan Pablo Bello

  Title: A quantitative comparison of different approaches for melody extraction from polyphonic audio recordings
  Type: MTG Technical Report (MTG-TR-2006-01)

  Institution: Universitat Pompeu Fabra

  Year: 2006

  Related to Chapter 2

- **Authors**: Emilia Gómez

  Title: Melodic description of audio signals for music content processing

  Type: Doctoral Pre-thesis

  Institution: Universitat Pompeu Fabra

  Year: 2002

  Related to Chapter 2 and 3

- **Authors**: Emilia Gómez

  Title: Fundamental frequency estimation study report

  Type: CUIDADO IST project internal report

  Year: 2001

  Related to Chapter 2

D.4 Presentations in conferences

- **Authors**: Emilia Gómez and Jordi Bonada

  Title: Tonality visualization of polyphonic audio

  Conference: Proceedings of International Computer Music Conference; Barcelona

  Year: 2005

  Related to Chapter 3, 4 and 5

- **Authors**: Emilia Gómez

  Title: Key estimation from polyphonic audio

  Conference: 6th International Conference on Music Information Retrieval, MIREX poster; London, UK

  Year: 2005

  Related to Chapter 3 and 4

- **Authors**: Chris Harte, Mark Sandler, Samer Abdallah and Emilia Gómez

  Title: Symbolic representation of musical chords: a proposed syntax for text annotations

  Conference: Proceedings of 6th International Conference on Music Information Retrieval; London, UK

  Year: 2005

- **Authors**: Pedro Cano, Markus Koppenberger, Nicolas Wack, Jose Pedro Garcia Mahedero, Jaume Masip, Oscar Celma, David García, Emilia Gómez, Fabien Gouyon, Enric Guaus, Perfecto Herrera, Jordi Massaguer, Beesuan Ong, Miquel Ramírez, Sebastian Streich and Xavier Serra

  Title: An industrial-strength content-based music recommendation system

  Conference: Proceedings of 28th Annual International ACM SIGIR Conference; Salvador, Brazil

  Year: 2005

  Related to Chapter 5

- **Authors**: Pedro Cano, Markus Koppenberger, Nicolas Wack, Jose Pedro Garcia Mahedero, Thomas Aussenac, Ricard Maxer, Jaume Masip, Oscar Celma, David García, Emilia Gómez, Fabien Gouyon, Enric Guaus, Perfecto Herrera, Jordi Massaguer, Beesuan Ong, Miquel Ramírez, Sebastian Streich and Xavier Serra

  Title: Content-based Music Audio Recommendation

  Conference: Proceedings of ACM Multimedia

  Year: 2006

  Related to Chapter 5

- **Authors**: Perfecto Herrera, Oscar Celma, Jordi Massaguer, Pedro Cano, Emilia Gómez, Fabien Gouyon, Markus Koppenberger, David Garcia, Jose Pedro García Mahedero and Nicolas Wack

  Title: Mucosa: a music content semantic annotator

  Conference: Proceedings of 6th International Conference on Music Information Retrieval; London, UK

  Year: 2005

  Related to Chapter 3 and 4

  Related to Chapter 2 and 4

- **Authors**: Emilia Gómez and Perfecto Herrera

  Title: Automatic extraction of tonal metadata from polyphonic audio recordings

  Conference: Proceedings of 25th International AES Conference; London, UK

  Year: 2004

  Related to Chapter 3 and 4

- **Authors**: Emilia Gómez and Perfecto Herrera

  Title: Estimating the tonality of polyphonic audio files: cognitive versus machine learning modelling strategies

  Conference: Proceedings of Fifth International Conference on Music Information Retrieval; Barcelona

  Year: 2004

  Related to Chapter 4

- **Authors**: Oscar Celma, Emilia Gómez, Jordi Janer, Fabien Gouyon, Perfecto Herrera and David García

  Title: Tools for content-based retrieval and transformation of audio using MPEG-7: the SPOffline and the MD-Tools

  Conference: Proceedings of 25th International AES Conference; London, UK

  Year: 2004

  Related to Chapter 2

- **Authors**: Emilia Gómez, Maarten Grachten, Xavier Amatriain and Josep Lluïs Arcos

  Title: Melodic characterization of monophonic recordings for expressive tempo transformations

  Conference: Proceedings of Stockholm Music Acoustics Conference; Stockholm, Sweden

  Year: 2003

  Related to Chapter 2

- **Authors**: Emilia Gómez, Fabien Gouyon, Perfecto Herrera and Xavier Amatriain

  Title: MPEG-7 for content-based music processing

  Conference: Proceedings of 4th WIAMIS-Special session on Audio Segmentation and Digital Music; London, UK

  Year: 2003

  Related to Chapter 2

- **Authors**: Emilia Gómez, Fabien Gouyon, Perfecto Herrera and Xavier Amatriain

  Title: Using and enhancing the current MPEG-7 standard for a music content processing tool

  Conference: Proceedings of Audio Engineering Society, 114th Convention; Amsterdam, The Netherlands

  Year: 2003

  Related to Chapter 2

- **Authors**: Emilia Gómez, Gilles Peterschmitt, Xavier Amatriain and Perfecto Herrera

  Title: Content-based melodic transformations of audio for a music processing application

  Conference: Proceedings of 6th International Conference on Digital Audio Effects; London, UK

  Year: 2003

  Related to Chapter 2

- **Authors**: Nicolas Durand and Emilia Gómez

  Title: Periodicity analysis using an harmonic matching method and bandwise processing

  Conference: Proceedings of MOSART Workshop on Current Research Directions in Computer Music; Barcelona, Spain

  Year: 2001

  Related to Chapter 2