Department Of Electronics and Telecommunications

PhD. Program in
"Telematics and Information Society"
(24^{th} cycle)

Instituted by the Italian University Consortium among
University of Florence
University of Siena

Signal Processing Techniques Applied to Automatic Music Transcription

A thesis submitted for the degree of
Doctorate of Philosophy

Candidate

Dr. Gianni Pantaleo

Coordinator:

Prof. Dino Giuli

Supervisors:

Prof. Paolo Nesi
Prof. Alessandro Fantechi

ING-INF/05:

Academic Year 2011/12

Acknowledgements

I would like to thank all the people who supported me in these past three years of work, without the help of whom this thesis would not have been possible. First of all my family, then I would like to thank Prof. Dino Giuli, the coordinator of the Doctoral School "Telematics and Information Society", my tutors Alessandro Fantechi and especially Paolo Nesi, who gave me the opportunity to conduct an extremely stimulating and noteworthy research activity in a field which have always engaged my interest, as a musician.

I owe a sincere and friendly gratitude to Prof. Fabrizio Argenti, for his constant technical and human support, which definitely helped me and contributed to the development of this thesis. A special thanks to my kind and skilled colleagues, who made my work experience at the DISIT Lab a motivating and really comfortable experience so far.

I cannot forget to show all my appreciation to my music teachers, through all these years: Marino Tomasini, Franco Lacava, Elena Mercuri, Alessandro Galati and Francesca Taranto, the Associazione Musicale Fiorentina *(where my childish pupil career has finally turned into a gratifying teaching activity), and the legendary years at* Centro Attività Musicali (CAM) Andrea del Sarto.

Many grateful thanks to all my music mates and colleagues, with whom I shared hard work and inspirations, so many hours of rehearsals, shows, jams, and long recording sessions... I had so much fun, great satisfaction and a lot, I say a lot of good music; thanks especially to those who showed a great interest in my work, always supporting me, sharing opinions, listening to my doubts, asking for news or simply "how does it work?"*; my dearest friends, gospel criminals a.k.a.* Stand!Bab Trio*, the big* Taddeo & Sons *family, my jazzy groovin'* Quadra Quintet*, the* For Joy Gospel Choir*, the Brazilian feel of* Iramar & Gota de Agua *band, the good old, passed by* Qui-Pro-Quartet*,* Danish Butter Quintet *and the legendary* Charivari*, all the great musicians I enjoyed playing with, in the snug and twilight atmosphere of Florence clubs and bars by night, most of all* La Cité, CPA *and the legendary* Chiodo Fisso *and* Dolly Pub*.*

Finally, last but not least (at all!), I wish to show all my gratitude and love to ALL *my dear friends, without whom I would have never become the person I am now (hopefully with an overall positive meaning!). They always encouraged me in undertaking this experience; my dearest friends, near and far.*

Contents

List of Figures

Preface

In these three years attending the Doctoral School in *Telematics and Information Society* at University of Florence, I mainly focused my research activity on the topic which gives the title to this Ph.D. thesis: study, design and validation of signal processing techniques applied to automatic music transcription.

Automatic transcription of music (AToM) is a difficult problem, which remains still unresolved in some of its application contexts (such as polyphonic music and multi-instrumental transcription). This task refers to the analysis of a digital acoustic or synthesized musical signal, in order to write down pitch, onset time, duration, intensity and source of each sound that occurs in it.

In many other research areas such as computer vision and semantic web indexing, efforts are made to automatize several types of human cognitive processes. Computer vision, for example, deals with identifying techniques and strategies for acquiring, processing and understanding images from the real world, in order to produce symbolic/numeric information or decisions rules, which is a daily trivial operation for most of the people. The distinctive aspect, which makes automatic transcription of music such a challenging task, is that its outcome result is a hardly achievable goal not only for ordinary people at large, but even for expert and well trained musicians. This fact can be partially explained by the high degree of perceptual fusion characterizing the human auditory system, according to which we perceive simultaneous and multitimbral sounds as a single entity. Furthermore, the lack of knowledge on human brain processes underlying this complex activity (though the functioning of inner ear transcoding mechanisms have been pervasively studied and understood), justifies the large variety of methods and approaches proposed, ranging from signal processing techniques to higher-level musicological models.

In fact, in addition to this predominant doctoral activity, I had the opportunity to cover other topics, and to undertake research in the following fields

and projects:

- *Study and modeling of educational and scientific archives management solutions.* In this research, a study of requirements and development of a multi-press platform for e-journals publishing and peer-review support have been conducted. The related implementation and validation has been performed by expanding the open source OJS (Open Journal System) as a multi-press and multi-journal platform. This process involved a deep reengineering of the originally distributed OJS architecture. PHP, PostgreSQL and Apache Server technologies have been used. The proposed solution refers to *Palamede* project [BNP11] and that has produced the experimental portal (http://palamede.fupress.com), and which it has been validated by a test experimentation of three Italian University Presses: the *Firenze University Press* (FUP), University of Parma and the *Forum Editrice* Press of University of Udine.

- *Development of an innovative system for detection of presence and number of people in an indoor secure access.* The identification of the human presence and/or counting of number of people are in the focus of many applications in the field of security. Specifically, automation of security systems has been a growing interest topic for controlling accesses in restricted areas such as banks, airports, railway stations and governmental accesses. The goal of this research has been the design and development of an automated solution for detecting the presence of more than one person in interlocked doors, adopted in many accesses. In most cases, the interlocked doors are small areas in which other information and sensors are placed, to detect the presence of guns, explosive, etc. The general goals and the environmental conditions required, allowed to implement a detection system at lower costs and complexity, with respect to other existing techniques, retrieved in the state of the art of related works. The system consists of a fixed array of microwave transceiver modules, whose received signals have been processed to collect information related to the volume occupied in the interlocked cabin door. A research has been conducted to study and identify volume measurement and image reconstruction techniques using microwave sensors. Statistical and predictive models have been applied to collected data and measurements to build stronger decision rules for detection. The solution proposed has been statistically validated against real experimental measures, and it has also

been implemented to be used in real time.

- *Collaborative and assisted SKOS generation and management.* I started to give my contribute to the *Open Space Innovative Mind* (OSIM) project [BCN11], a solution for assisting expert users in collaborative development and management of a SKOS knowledge. The SKOS production has been accelerated by crawling and exploiting different kinds of sources (in multiple languages and with several inconsistencies among them). The OSIM web based platform (http://openmind.axmedis.org) and tools support the experts in defining relationships among the most recurrent concepts, reducing the time of SKOS generation and allowing collaborative production. The main goal of the OSIM project is creating a portal to allow industries at posing semantic queries about potential competencies in a large institution such as the University of Florence.

In this thesis, an original system for automatic music transcription is described. The main goal of this research has been to investigate for novel techniques and solutions with respect to the ones proposed in the current state of the art, which has been carefully reviewed. The present document has been organized in the following chapters:

- **Chapter 1**: after a general explanation of automatic music transcription task and some basic concepts regarding audio signals and music notation, requirements and application areas are described, and a first, functional classification of the transcription techniques is presented.

- **Chapter 2**: the current state of the art of automatic music transcription is deeply and carefully reviewed. A big effort has been made to compare the features of all the most quoted transcription systems in literature, since the first pioneering works in the late 70s up to the most recent solutions.

- **Chapter 3**: In Section 3.1, a mathematical theory of higher-order spectra is recalled, and definitions of bispectrum, main properties and computational models are provided. In Sections 3.2 and 3.3, the Constant-Q analysis and its application to bispectral signal representation for multi-pitch detection are described. Finally, Sections 3.4 and 3.5 deal with a detailed case study of bispectral nonlinearity implications in multi-pitch

detection procedure, which is consequently compared with traditional spectrum based estimation techniques.

- **Chapter 4**: the general architecture of the proposed transcription system is described first, followed by a detailed outline of the main modules for pitch detection and note duration tracking.

- **Chapter 5**: some experimental results are reported. The proposed system has been validated against some excerpts of the standard *Real World Computing* (RWC) database. Results of the system performance at the MIREX 2009 international contest are also reported, for further validation.

- **Chapter 6**: this final chapter is left for conclusions and discussion on future work guidelines.

Chapter 1

Automatic Music Transcription: An Introduction

Music Information Retrieval (MIR) multidisciplinary research field has revealed a great increment in academic interest in the last decades, although yet barely comparable to the commercial involvement grown around speech recognition. It must be noticed that music information is much more complex than speech information, both from a physical (range of frequency analysis) and a semantic (big number, high complexity and many abstraction levels of the possible queries) point of view.

Automatic music transcription is a specific task within MIR: it is defined as the process of converting a musical audio recording into a symbolic notation (a musical *score* or *sheet*) or any equivalent representation, usually concerning event information associated with *pitch*, note *onset times*, *durations* (or equivalently, *offset times*) and *intensity*. This task can be accomplished by a well ear-trained person, although it could be quite challenging for experienced musicians as well; besides, it is difficult to be realized in a completely automated way. This is due to the fact that human knowledge of musicological models and harmonic rules are useful to solve the problem, although such skills are not easy to be coded and wrapped into an algorithmic procedure. Complete automatic transcription of real world musical signals can be very hard or even theoretically impossible in some cases; so the goal is usually redefined in annotating as many of the concurrent sounds as possible, or in transcribing only some specific and well-defined parts, for example the melody or some prominent melodic or rhythmical figures, like bass lines or drum sounds (in this case, the process is intended as a partial transcription).

1.1 Representation of Sound

Sound is a physical phenomenon produced by the propagation of a sequence of waves of pressure through compressible media, generated by the vibration of an elastic body. Consequently, an audio signal is composed of a single or a mixture of approximately periodic and locally stationary acoustic waves. The present work is not intended to provide an exhaustive study of physics of sound and acoustics. For this aim, the interested reader can refer to the large documentation available (e.g., [Hal91], [BS04]).

According to the Fourier representation, any finite energy signal $x(t)$ with period T_0 is represented as the sum of an infinite number of sinusoidal components, weighted by appropriate amplitude coefficients:

$$x(t) = \frac{a_0}{2} + \sum_{n=1}^{+\infty} a_n \sin(2\pi n f_0 t + \phi_n) \tag{1.1}$$

where:

$$a_n = |A_n| \qquad , \qquad \phi_n = \angle A_n$$

and:

$$A_n = \frac{2}{T_0} \int_0^{T_0} x(t) e^{-j2\pi n f_0 t} dt$$

An acoustic wave is a particular case in which, ideally, frequency values of single harmonic components are integer multiples of the first one, called *fundamental frequency* (which is the perceived pitch). Harmonic components are called *partials* or simply *harmonics*. Since the fundamental frequency of a sound, denoted as $F0$, is defined to be the greatest common divisor of its own harmonic set (actually, in some real cases, the first spectral component corresponding to $F0 = 1/T_0$ can be missing), the task of music transcription, that is, the tracking of the partials of all concurrent sounds, is practically reduced to a time periodicities search, which is equivalent to looking for energy maxima in the frequency domain. Thus, every single note can be associated with a fixed and distinct comb-pattern of local maxima in the amplitude spectrum, which appears like the one shown in Figure 1.1. The distances between energy maxima are expressed as integer multiples of $F0$ (top) as well as in semitones (bottom): the latter are an approximation of the natural harmonic frequencies in the well-tempered system (see Section 1.3 for reference of some basic elements of music notation).

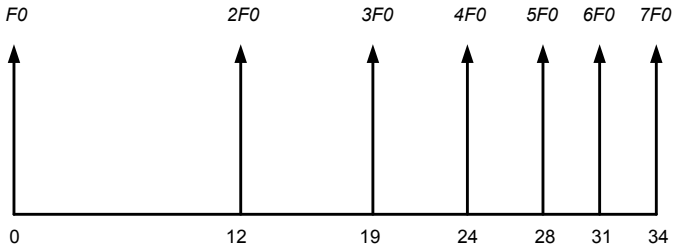


Figure 1.1: Fixed comb-pattern representing the harmonics set associated with every single note. Seven partials (fundamental frequency included) with the same amplitude have been considered. The distances are also expressed (bottom) as semitones.

1.2 Monophonic and Polyphonic Music

A major distinctive cue in music transcoding is given by the number of voices a music piece consists of: there can be only one voice playing at each time; these cases are treated as a monophonic transcription task. On the contrary, if several voices are played simultaneously, we deal with a polyphonic transcription process.

The mixture of two or more sounds present a degree of consonance (and equivalently, a degree of dissonance) which depends on harmonic relationship between their pitches. In this regard, it is convenient to recall the following proposition by Klapuri [Kla98]: let R ed S be two interfering sounds; if the relationship between their fundamental frequencies, f_{0R} e f_{0S} is a rational number, i.e.:

$$\frac{f_{0R}}{f_{0S}} = \frac{m}{n} \qquad , \qquad \text{con } n, m \geq 1, \tag{1.2}$$

then each n-th partial of R overlaps to each m-th partial of S.

Low values of n e m imply a high degree of consonance between R e S. It is worthy to be noticed that if m/n is an integer, the overlapping of the two sounds' partials is complete.

Automatic transcription of monophonic music is currently considered as a resolved problem, while transcription of polyphonic music is still far from being successfully settled, and additional difficulties arise in presence of multi-instrumental contexts. Development of techniques for monophonic pitch detection has received a greater attention and deeper interest for speech analysis,

rather than for music, even in quite recent literature. In Figure 1.2, some examples of the spectral content of typical audio signals are shown.

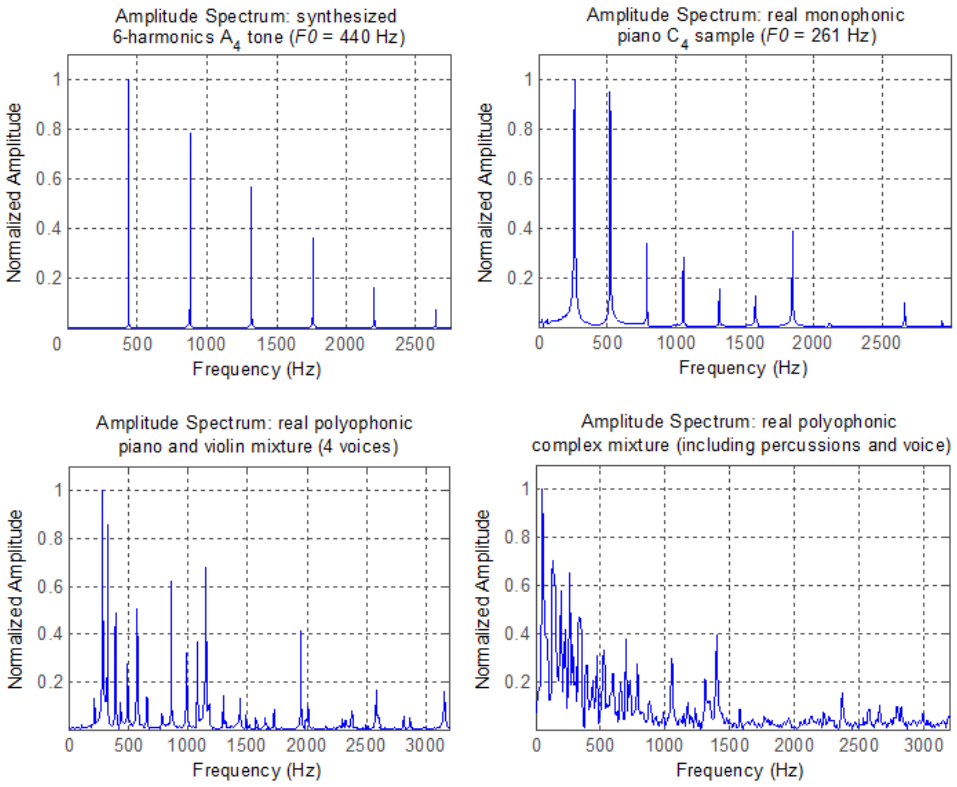


Figure 1.2: Amplitude spectrum representation of some typical audio signals. Noteworthy is the increasing complexity of the spectral content, as the number of concurrent playing voices increases.

Difficulties arise in polyphonic music transcription since two or more concurrent sounds may contain partials which share the same frequency values. This is one of the main reasons why simple amplitude spectral analysis is considered inadequate, if not joined to other signal processing techniques or a priori knowledge resources.

1.3 Music Notation

The purpose of illustrating the principles of theory of music is beyond this work; however, it is convenient to present some theoretical preliminaries, which are

necessary to understand the topics presented in this thesis; actually, the intent is to cover only those music notation aspects and issues which are necessary for a complete comprehension of the following dissertation. For a more detailed overview of music theory, please refer to [Sor95] or to one of the many manuals related to this topic.

The seven notes (in addition to the usual staff sheet or score representation), can be named after the first seven alphabetical letters, from A to G. The octave number is indicated as a subscript. In the following, the lowest piano octave is associated with number 0; thus, middle C, at 261 Hz, is denoted with C_4, and A_4 (which is commonly used as a reference tone for instruments tuning) univocally identifies the note at 440 Hz.

The distance between two notes is defined *interval*: it can be measured in frequency, fundamental frequencies ratio, or scale step (or degrees).

In western music, adopting the *well-temperament* as the standard tuning system, a *semitone* is defined to be the smallest audible interval between two generic notes. If f_1 and f_2 represent the pitches of two notes separated by one semitone interval, then $f_2 = f_1 \cdot 2^{1/12}$. An interval of one octave is characterized by $f_2 = 2f_1$, and it is composed of 12 semitones. Other examples of intervals between notes are the perfect fifth ($f_2 = 3/2\ f_1$), the perfect fourth ($f_2 = 4/3\ f_1$), and the major third ($f_2 = 5/4\ f_1$). The following symbols, $\sharp$ (*sharp*) and $\flat$ (*flat*), known as *accidentals*, are used to raise or lower, respectively, a note by a semitone. This implies that, in western music notation, notes with different nomenclature have the same pitch (see Figure 1.3).

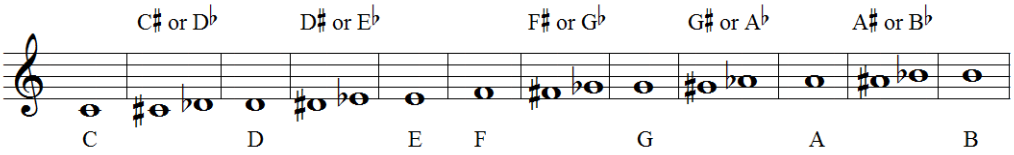


Figure 1.3: Music staff, notation and nomenclature. In this example, the seven notes, and corresponding accidentals, of the diatonic scale on C in the 4-th octave are represented.

Retaining the parallel between speech and music, music notation is mainly a set of instruction for a musical performance, rather than a representation of a musical signal [Kla04a]; in the same way, written text is to be considered

as the equivalent for speech. The main difference is that music information is much more multi-faceted, since it includes many different levels of information (note pitch, harmonic and rhythmic information, indications for expression, dynamics).

1.4 Requirements and Application Areas

At the present time, many music transcription system have been developed, both for academic research purpose and commercial distribution. This has produced a large variety of different features and tasks to be accomplished. Generic requirements for a music transcription system can be classified into the following:

- **Representation of the transcoded output**: the input audio source is usually converted into a MIDI file, a Piano-roll representation (like the one depicted in Figure 1.4), or simply into a list of note information (onset time, pitch or frequency, duration or offset time, loudness etc ...).
- **General transcription features**: an advanced music transcription system, dealing with real world audio signals and recordings, should allow the user to choose between monophonic or a polyphonic transcription, as well as supporting single and multi-instrumental recognition.
- **Level of automation**: the very large assortment and combination of musical instruments, genres, recording background conditions make the task of music transcription very hard to be fulfilled in a completely automated way, for instance using a black box algorithm solution. Some user defined parameters are often introduced (e.g.: energy threshold values, a priori knowledge like the number of voices in a polyphonic mixture or more complex instrumental models).
- **Processing time**: according to the computing performances, usually two main categories are identified: real-time and offline transcription systems. Real-time solution is generally achieved for a monophonic transcription process, whereas it generally results out of reach for the complexity of the polyphonic transcription task.

Automatic transcription of music can be a key task for many application fields, for instance: educational music frameworks; interactive computer music equipment for generating accompaniment for soloists; sound resynthesis

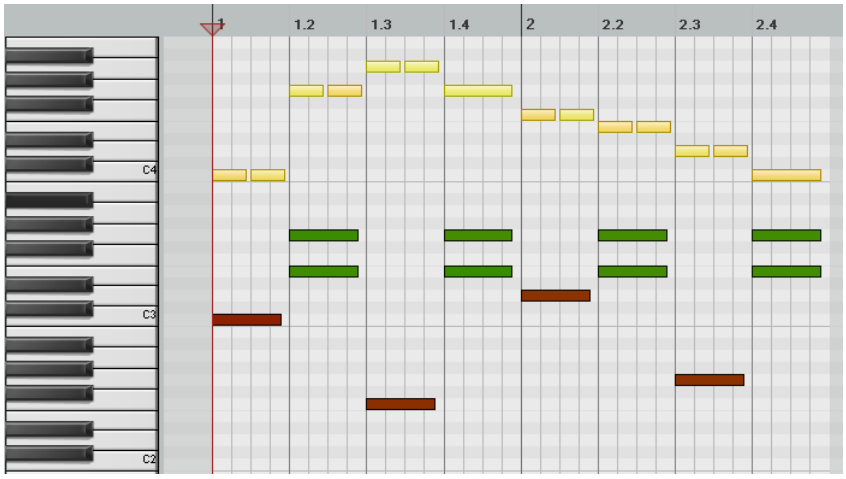


Figure 1.4: Piano-roll representation of music: in abscissa, the time (here expressed in subgroups of musical bars); in ordinate, the note pitch (represented with the piano keys).

for preservation or restoration of old and historical recordings; musicological analysis of improvised and ethnic music for which musical notations do not exist etc. Regarding some applications in the educational environment, some future *desiderata* are the achieving of a more robust score representation of the transcription systems output, that can be helpful for monitoring the musician execution as well as for real-time transcoding of any musical performance.

1.5 Classification of Music Transcription Systems

Many efforts have been made to realize exhaustive reviews and to provide classification models for automatic transcription methods. Remarkable works are the ones by Rabiner [Rab77] for monophonic transcription, and by Bello [Bel03] Klapuri [Kla04b], [Kla04a], Brossier [Bro06] and Yeh [Yeh08] also for polyphonic transcription. Citing a statement by Klapuri, "it is difficult to categorize multiple-$F0$ estimation methods (and music transcription methods in general) according to any single taxonomy because the methods are complex and typically combine several processing principles and procedures. As a consequence, there is no single dimension which could function as an appropriate

basis for categorization"[Kla04a].

This aspect suggests a decomposition of the problem as an efficient processing approach. Quite recently, some specialized sub-areas of this research field have been developed, dealing with more limited transcription tasks, such as the extraction of melody or bass lines within a polyphonic mixture of sounds. Besides, modularity is a similar aspect observed also in the human brain [Kla04b], [PC03]. The human auditory system (the inner hear, together with the part of the brain appointed to music cognition) results to be the most reliable acoustic analysis tool [Kla04b]. Actually, an expert musician can accomplish the task of music transcription, relying also on a set of knowledge sources (musicological models, harmonic rules, experience). Such skills are difficult to be coded and wrapped into an algorithmic procedure.

Human capability to achieve the comprehension of music transcription problem is understood as the sum of two different attitudes: the bottom-up and the top-down processing. This suggests a first boundary of classification, given by the following approaches:

- The **bottom-up processing**, or data-driven model, starts from low level elements (the raw audio samples) and it uses processing blocks to analyze and cluster these elements in order to gather the required information.
- The **top-down processing**, or prediction-driven model, starts from information at a higher level (based on external knowledge) and it uses such information to understand and explain elements at lower hierarchy levels (physical stimuli).

We have considered this, reported by Bello [Bel03], as the most general categorization criterion for the music transcription problem, since these two approaches are non-mutual-exclusive, and contain ideally all the other fields of codification we intend to review in the following. There are many reviews of automatic music transcription methods in literature, and most of them present their own criteria, upon which the different front ends, used to obtain a useful mid-level representation of the audio input signal, are grouped together. One of the most commonly used criterion (adopted by Gerhard [Ghe03], Brossier [Bro06] and Yeh [Yeh08]) is based on a differentiation at signal analysis level:

- **Time domain analysis**: systems belonging to this category process the audio waveform in order to obtain information about pitches (peri-

odicities of the audio signal) or onset times. In general, this family of methods is suitable for monophonic transcription.

- **Frequency domain analysis**: methods belonging to this class vary from spectral analysis (FFT, cepstrum, multi-resolution filtering, Wavelet transform and related variants) to auditory models developed in the first 90s within the Computational Auditory Scene Analysis (CASA) framework [SL90], [Ell96], [MO97], as well as many spectral matching or spectral features extraction techniques.

Another classification concept is reported by Yeh [Yeh08], for whom music transcription methods can be catalogued into two different approaches:

- **Iterative estimation**: such principle refers to all the methods which iteratively estimate predominant $F0$, and subsequently cancel the residual harmonic pattern of estimated notes from the observed spectrum, processing the residual until a stop criterion is met; usually, a condition related to residual energy is adopted. The block diagram of this architecture is shown in Figure 1.5.

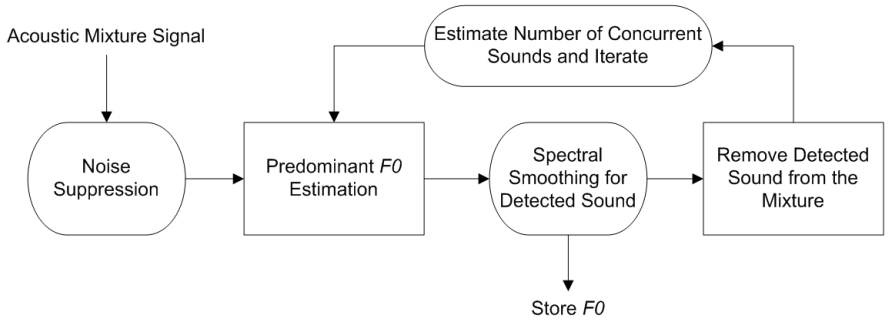


Figure 1.5: Amplitude spectrum representation of some typical audio signals. Noteworthy is the increasing complexity of the spectral content, as the number of concurrent playing voices increases.

- **Joint estimation**: under this approach we find algorithms that jointly evaluate many hypotheses on $F0$ estimation, without involving any cancelation. These solutions include the use of salience functions or other knowledge source, in order to facilitate spectral peak-picking, and other

frameworks like Martin's Blackboard architecture [Mar96a]. This name comes from the metaphor of a group of researchers standing in front of a blackboard, working to find out the solution to a problem. This framework is a problem-solving model, which integrates knowledge from different sources and allows the interaction of different parts of the model. An expert musical knowledge, integrated with signal processing and other physical, engineering or mathematical frameworks, is considered useful to accomplish the task of automatic music transcription. Another subgroup belonging to the Joint Estimation category is the spectral matching by parametric/non parametric models, like Non-negative Matrix Approaches (NMA) including Non-negative Matrix Factorization (NMF), frequently used in recent literature [Vir07], [VBB08].

Another categorization to be highlighted is often included in frequency analysis or joint estimation classes in the above mentioned review works: statistical versus non statistical framework. The statistical-inference approach generally aims at jointly performing $F0$ estimation and tracking of temporal parameters (onsets and durations) from a time-frequency representation of the input signal. In these models, the quantities to be inferred are considered as a set of hidden variables. The probabilistic model relates these variables to the observation variable sequence (the input signal or a mid-level representation) by using a set of properly defined parameters. Statistical frameworks frequently used for automatic music transcription are Bayesian networks [KNKT95], [CKB06] or Hidden Markov Models (HMM) [RK05], [YRR$^+$08]. Finally, another pivotal aspect is the evaluation of the transcription systems proposed so far. The absence of formalized paradigms to compare different methods, the necessity of commonly accepted evaluation criteria, and finally the difficulties to collect large enough databases (often due to intellectual property rights restrictions, which is another important difference with the speech recognition research area) led the IMIRSEL (International Music Information Retrieval Systems Evaluation Laboratory) community to create, in 2005, the MIREX (Music Information Retrieval Evaluation eXchange) evaluation framework. In few editions, MIREX has already become a worldwide accepted, standard reference for the evaluation of submitted methods and algorithms aimed at resolving several MIR proposed tasks , including polyphonic pitch estimation and note tracking [MIRb]. The tasks, the evaluation material and conditions, as well as many other elements of the MIREX architecture are

defined and discussed within the whole community, thus reflecting its own interests and accomplishing the necessity of formality and repeatability.

Chapter 2

State of the Art

In literature, a large variety of methods for both monophonic and polyphonic music transcription has been realized. Monophonic transcription solutions were the first to be proposed, starting from the second half of the 60s, in parallel with the initial development of the newly-born speech processing; in fact, monophonic pitch detection was basically applied for speech recognition purposes. Some of these methods were based on time-domain techniques like Zero Crossing Rate (ZCR) [Mil75], or on autocorrelation function (ACF) in the time-domain [RRM76], as well as parallel processing [GR77] or Linear Predictive Coding (LPC) analysis [Mar72].

First attempts of performing polyphonic music transcription started in the late 1970s, with the pioneering work of Moorer [Moo77] and Piszczalski and Galler [PG77]. As time went by, the commonly-used frequency representation of audio signals as a front-end for transcription systems has been developed in many different ways, and several techniques have been proposed. Klapuri [Kla03], [Kla05] performed an iterative predominant $F0$ estimation and a subsequent cancelation of each harmonic pattern from the spectrum; Nawab [NAW01] used an iterative pattern matching algorithm upon a constant-Q spectral representation. In the early 1990s, other approaches began to develop, based on applied psycho-acoustic models and also known as Computational Auditory Scene Analysis (CASA), from the work by Bregman [Bre90], started to be developed. This framework was focused on the idea of formulating a computational model of the human inner ear system, which is known to work as a frequency-selective bank of passband filters; techniques based on this model, formalized by Slaney and Lion [SL90], were proposed by Ellis [Ell96], Meddis and O'Mard [MO97], Tolonen and Karjalainen [TK00] and Klapuri

[Kla08]. Marolt [Mar01], [Mar04] used the output of adaptive oscillators as a training set for a bank of neural networks to track partials of piano recordings. A systematic and collaborative organization of different approaches to the music transcription problem is the mainstay of the idea expressed in the Blackboard Architecture proposed by Martin [Mar96a]. More recently, physical [OBCQTG05] and musicological models, like average harmonic structure (AHS) extraction in [DZZS08], as well as other a priori knowledge [KaNiSa07], and possibly temporal information [BDS06] have been joined to the audio signal analysis in the frequency-domain to improve transcription systems performances. Other frameworks rely on statistical inference, like hidden Markov models [Rap02], [RK05], [YRR+08], Bayesian networks [KNKT95], [CKB06] or Bayesian models [GDI06], [DD07]. Others systems were proposed, aiming at estimating the bass line [KR07], or the melody and bass lines in musical audio signals [Got00] [Got04]. Currently, the approach based on non-negative matrix approximation [ROS07] (in different versions like nonnegative matrix factorization of spectral features [SB03], [Vir07], [VBB08]) has received much attention within the MIR community. Recently, Higher Order Spectra Analysis (HOSA) has been applied to multipitch estimation [Abe04] and automatic music transcription [ANP11a].

2.1 Methods Overview and Comparison

In this section, a comparative review of some of the most important and cited music transcription systems is describe, as proposed in [ANP11a]. This review is not meant to be as an exhaustive and omni-comprehensive work, although it covers large part of the literature, starting from the first pioneering methods, realized at the end of the 70s, until nowadays. The aim is to illustrate the evolution of the state of the art, which is supposed to run in parallel with the development of technology in the fields of signal processing and computational elaboration power. In Figure 2.1, a functional block diagram related to the general architecture of an automatic music transcription system, is shown.

A Pre-Processing module is generally assigned to segment the input signal into frames, and to compute the mid-level representation (spectral analysis, auditory model based representation etc.). The retrieval of pitch information and note temporal parameters is performed usually by dedicated modules, referred to as Pitch Estimation and Time Information Estimation in Figure 2.1. To achieve better transcription accuracies, additional Knowledge Sources

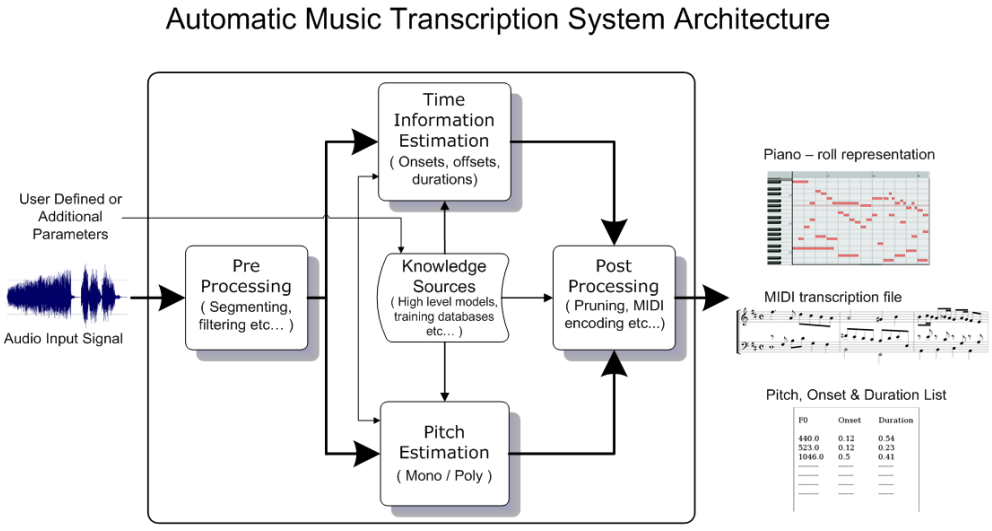


Figure 2.1: General architecture of an automatic music transcription system.

(harmonic/instrumental models, training databases) are often implemented in transcription systems, for many different purposes. Finally, a Post-Processing module groups all the detected note information and converts it into an appropriate output format (MIDI file, piano-roll or note parameters list). In the following, a multi-field classification is proposed through the use of a set of parameters which can be helpful to highlight the main characteristics and peculiarities of different algorithms, without forcing a strict categorization, not even focusing on specific parts of the processing framework. For this reason, the overview of each system includes information about all the different elements of the architecture: signal processing, pitch estimation and rhythm information extraction, I/O parameters and other computational aspects. The comparison summary is reported in Table 2.1. A tabular view has been chosen in order to maximize hint facilities, similarly to the one adopted by Klapuri in [Kla04b]. Systems are sorted by rows, in a chronological sequence. The columns report different fields describing the most interesting aspects of the architecture for the reviewed algorithms. They are defined as follows:

- **Reference**: this field contains the reference to the authors of each system. Where needed, the research group is specified. In the past years of automatic music transcription research activity, longer-term projects have been undertaken by Stanford university (in particular the Centre for

Computer Research in Music and Acoustics, referred to as CCRMA in Table 2.1), University of Michigan (U-M), University of Tokyo (UT), National Institute of Advanced Industrial Science and Technology (AIST), Massachusetts Institute of Technology (MIT), Queen Mary University of London (QMUL), University of Cambridge (CAM), Tampere/Helsinki University of Technology (TUT/HUT), and the Institut de Recherche et Coordination Acoustique/Musique (IRCAM) of Paris, France. Other names and abbreviations, not included in the above mentioned list, refer either to the name of the research projects, or to the commercial development of such systems (e.g., KANSEI, SONIC, YIN).

- **Year**: the year of publication of the referenced papers.
- **System Input / Output**: this field contains specifications, if they exist, on the input audio file, and it reports also the output format of the transcription process (e.g., MIDI file, list of pitches, onsets and durations), whether described in the referenced papers.
- **Pre-Processing and Mid-Level**: a list of the signal processing techniques, used to obtain a useful front end.
- **Real time / Offline**: this field specifies, if the system operates in real time or not.
- **Source Availability**: this specifies if the source code is available, directly or web-linked.
- **Mono / Poly**: this field shows if the system is mainly dedicated to monophonic or polyphonic transcription.
- **Time / Frequency**: indicates if the signal processing techniques used by the algorithm (which are listed in the Pre-Processing and Mid-Level categories described above) operates either in the time or in the frequency domain. Where needed, it is otherwise specified if a method uses a different transform domain (e.g., autocorrelation domain).
- **Pitch Estimation Knowledge**: a brief description about the approaches and the knowledge used to extract pitch information.
- **Rhythm Info Extraction**: in this field the techniques used to retrieve temporal information of estimated $F0$s (where this task is performed)

are summarized. It is divided into two sub-fields: *Onsets* and *Durations*, as they are often estimated with different strategies.

- **Evaluation Material**: this section shortly reports, where described, the type of the dataset used for evaluation and the number of test files / samples. Evaluation results are omitted. Actually, since different systems are tested against different databases and using different criteria, results reported in literature give a misleading outlook of overall transcription performances. For this reason, only MIREX results are reported, for all those algorithms which participated in the past editions. As to this topic, noteworthy is to highlight that a methodology for the evaluation of music transcription systems has not been firmly established yet. The transcription output (MIDI file or piano-roll usually) is compared with a reference ground truth of the audio source data; evaluation databases generally provide a reference MIDI file for each audio track or sample contained. Further work has often to be done, in order to check the correct alignment between the two representations. The procedure is as follows: a graphical comparison is commonly made, by using a dedicated GUI or other devices, between the audio signal spectrogram and the piano-roll of the reference MIDI; then a manual alignment is performed for the corresponding note events. An example of this graphical alignment is illustrated in Figure 2.2.

  Apart from defining the ground truth reference, evaluation criteria and parameters must be defined in order to design a comprehensive and well organized evaluation method. The MIREX framework proposes a validation approach which is becoming a standard reference in recent literature. For the evaluation of music transcription algorithms, two MIREX tasks are defined [MIRb]:

  1. **Multiple $F0$ Estimation (MF0E)** on a frame by frame basis. In this task, submitted systems are requested to report detected active pitches every 10 ms. A returned pitch is assumed to be correct (true positive, TP) if it is within a half semitone ($\pm 3\%$) of a ground-truth pitch for that frame. Otherwise, if a returned pitch is not present in the ground truth data, it is classified as a false positive (FP); finally, each not detected ground truth pitch is classified as a false negative (FN) Only one ground-truth pitch can be associated with

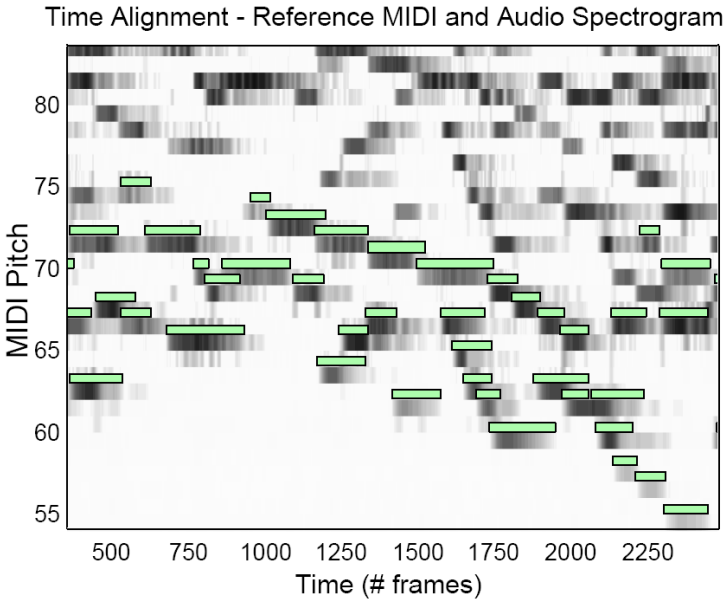


Figure 2.2: Example of graphical time alignment between input audio spectrogram and ground truth reference MIDI.

each returned pitch. Three performance measures are defined for this task:

- *Precision*: it is the portion of correct retrieved pitches for all the pitches retrieved for each frame:

$$Precision = \frac{TP}{TP + FP};$$

- *Recall*: it is the ratio of correct pitches to all the ground truth pitches for each frame:

$$Recall = \frac{TP}{TP + FN};$$

- *Accuracy*: it is an overall measure of the transcription system performance, given by:

$$Accuracy = \frac{TP}{TP + FP + FN};$$

2. **Note Tracking (NT)** task. A ground truth note is assumed to be correctly transcribed if the system returns a note that is within

a half semitone ±3% of that note AND the returned note's onset is within a 100ms range (± 50ms) of the onset of the ground truth note, and its offset is within 20% range of the ground truth note's offset. Again, one ground truth note can only be associated with one transcribed note. NT evaluation is further divided into the following subtasks:

a. Mixed Set Note Tracking;

b. Piano Only Note Tracking.

For this task, again Precision and Recall are reported. They are used to define a measure which is considered to indicate more correctly the balance between false positives and false negatives, that is:

$$F - Measure = 2\frac{Precision \cdot Recall}{Precision + Recall};$$

- **Additional Notes**: under this entry, any further noteworthy information, which can not be classified according to the defined categories, is recalled.

When the value of a certain parameter is missing, or information about one of the defined fields is not available in the referenced paper, the abbreviation N.A. is used in Table 2.1. In Table 2.2, other acronyms used in Table 2.1 are defined.

2.2 Review of Some Music Transcription Systems

Moorer - 1977

Moorer was one of the first, in literature, to propose a system which attempted to separate simultaneous harmonic sounds in a polyphonic mixture [Moo77]. His system has been developed to track pitches of both synthesized and real duets, although it presents several strong limitations: sounds are supposed to be harmonic and characterized by constant amplitude (no vibrato or jitter is therefore allowed). In addition, the two voices should not cross in pitch, and the two fundamental frequencies should not be in an 1:N relationship, which is equivalent to a complete overlapping of the partials of the concurrent sounds. The frequency range of analysis is also limited. The mid-level spectral

[Reference] (Group)	Year		System Input / Output	Pre-Processing & Mid-Level	Real time	Source	Mono	Time	Pitch Estimation	Rhythm Info Extraction		Addiotional	Evaluation
					Offline	Avail.	Poly	Freq.	Knowledge	Onsets	Durations	Notes	Material
[Moo77] (CCRMA)	1977	I	N.A.	Band-pass filter bank (optimum comb filter)	N.A.	No	Poly	F	Time periodicity research by detecting sinusoidal components	No	No	Max # voices: 2 Limited freq. range FOs ratio can't have an integer relationship	Synthesized violin and real guitar duets
		O	N.A.	Short - time FFT									
[PG77] (U-M)	1977	I	N.A.	Spectral equalization to enhance partials	Offline	No	Poly	F	Evaluation of harmonic relations among spectral peaks	No	No	Robust with missing F0 & inharmonic partials	Synth. and real signals (carillon bells)
		O	F0s & amplitudes list										
[Fri79]	1979	I	Speech signal, fs=10 kHz	Band-Pass FIR Lo-Freq emphasis	Suitable for Real time		Poly	T	Zero-crossing rate detection upon processed waveform	No	No	Pitch estimation for speech	3-second speech sample
		O	24 ms pitch frames										
[Cl86] (CCRMA)	1986	I	Digital audio recording	Bounded Q Frequency Transform	N.A.	No	Poly	F	Grouping partials in sinusoidal analysis	Detect changes in spectral energy over time	No	Knowledge of source acoustic applied	N.A.
		O	Hihg-level MIDI score										
[Ki89] (KANSEI)	1989	I	N.A.	Time - frequency map obtained with the interpolation method using complex spectra.	N.A.	No	Poly	F	Peaks extraction in the frequency domain and matching procedure	No	No	Heuristic rules implemented to group detected frequency peaks into notes.	System developed for piano, guitar and shamisen. Results are not reported.
		O	N.A.										
[SL90]	1990	I	N.A.	Correlogram: cochlear model, 2nd order filter bank, HWR and Automatic Gain Control	N.A.	Partially	Poly	F	Periodicities research in the correlogram by use of the autocorrelation function	No	No	—	Acoustical Society of America Database Qulitative, not clear results
		O	Time-frequency pitch representation										
[Mah90]	1990	I	Digital audio signals < 20 s	Short-time FFT (512 and 1024 sample window) Hi-freq. pre-emphasys	Offline	No	Poly	F	McAulay-Quatieri sinusoidal & two-way mismatch analysis Several strategies to resolve colliding sinusoidal partials	No	No	Limited to duets, non ovrlapping frequency ranges; nearly harmonic sounds	Synth samples and real signals (bassoon/ clarinet and trumpet/tuba) Qualitative results
		O	Chains of peaks for partials tracking										

Table 2.1 - Comparison of Automatic Music Transcription Systems (1 of 5).

Reference (Group)	Year	System Input / Output		Pre-Processing & Mid-Level	Real time	Source	Mono	Time	Pitch Estimation Knowledge	Rhythm Info Extraction		Addiotional Notes	Evaluation Material
					Offline	Avail.	Poly	Freq.		Onsets	Durations		
[KT92] (UT)	1992	I	Monaural signals	A/D conversion and spectrogram representation	N.A.	No	Poly	F	Peaks - peaking in STFT (segregation); statistic rules for partials grouping (integration)	No	No	Timbre models to detect different instruments	Synthesized Vivaldi Concerto (op. 3, no. 6)
		O	multi-channel MIDI										
[KT93] (UT)	1993	I	Monaural signals (48 kHz / 16 bits)	Frequency Analysis: band-pass 2-order IIR filters	N.A.	No	Poly	F	Frequency content extraction by pinching planes thresholds and bottom-up clustering	No	No	Automatic timbre modelling based on perceptual rules	Synth. random chords Good recognition up to 3 voices
		O	multi-channel MIDI										
[Haw93] (MIT)	1993	I	N.A.	Short-time spectral analysis	N.A.	No	Poly	F	Spectral comb filtering for note identification	Hi-freq. energy content; bilinear time-domain filtering	No	—	Bach piano excerpts Non extensive tests
		O	N.A.										
[KNKT95] (UT)	1995	I	Monaural signals	STFT	N.A.	No	Poly	F	Frequency content extraction by pinching planes and Bayesian network integration	No	No	Many knowledge sources applied (timbre, chord type...)	2-3 voices synthesized MIDI chords with real instruments samples
		O	multi-channel MIDI										
[Mar96a] (MIT)	1996	I	CD quality audio input	STFT Blackboard architecture front-end	Offline	No	Poly	F	Knowledge-based source (KS) applied to sinusoidal track extraction	Peaks picking on squared and low-pass filtered signal energy	No	—	4-voices Bach Corales Bad in octave detection Good recognition in B2 - A4 notes interval
		O	Transcription in counterpoint style										
[Mar96b] (MIT)	1996	I	N.A.	Log-lag correlogram (Auditory model of pitch perception)	N.A.	No	Poly	F	Periodicities research in the correlogram (autocorrelation) Knowledge sources applied (Blackboard framework)	Energy maxima of signal envelope	No	Advantages of auditory models in detecting octave intervals	Mono & Poly test on Bach piano pieces Qualitative results
		O	MIDI file, symbloic score or piano-roll										
[FCCQ98]	1998	I	N.A.	Multi scale sinusoidal model (constant-Q) filter bank	Not true real time	No	Poly	F	Prominent harmonic pattern search in synth. spectrum (amplitudes of peaks are set after a *quality-of-fit* measure).	No	No	General source models Masking effect test Post processing to kill too short notes.	Not specified dataset High error rate for typical musical signals are revealed
		O	MIDI file										

Table 2.1 - Comparison of Automatic Music Transcription Systems (2 of 5).

Reference (Group)	Year	System Input / Output		Pre-Processing & Mid-Level	Real time		Source	Mono	Time	Pitch Estimation Knowledge	Rhythm Info Extraction		Addiotional Notes	Evaluation Material
					Offline		Avail.	Poly	Freq.		Onsets	Durations		
[TK00] (HUT)	2000	I	N.A.	Pre-whitening filtering two channels - filter bank (Hp & LP, crossover @ 1 kHz)	Real time		No	Poly	ACF Transf. domain	Periodicity estimation by the *summary autocorrelation* function on both channels	No	No	F0 estimation examples available on Web	2-4 Clarinet tones mixed to form various chords
		O	N.A.											
[Got00] [Got04] (AIST)	2000 2004	I	16-bit PCM signal fs = 16 kHz	STFT obtained with a multi-rate filter bank	Real time		No	Poly	F	Frequency-to-instantaneous frequency mapping	No	No	Melody and bass line detection from real-world audio signals	10 excerpts from commercial CD recordings.
		O	N.A.	Band-pass filter to split the spectrum in 2 freq. regions						Maximize probability function for each candidate F0				
[Mar01] (SONIC)	2001	I	PCM signal, fs=44.1 kHz	Auditory model 200 *gammatone* filters	N.A.		Yes	Poly	F	Network of adaptative oscillators	Multy-layer perceptron NN	Time observation of NN activity	Piano transcription Note range: A1-C8	120 synthesized piano pieces
		O	MIDI file											
[dCK02] (YIN)	2002	I	N.A.	Autocorrelation (ACF) Method	Suitable for real time		No	Mono	F	Difference function (similar to autocrrelation function) computed via FFT, to find periodicities in signals	No	No	Cumulative mean function and parabolic interpolation to reduce sub-harmonic errors	Speech databases Informal evluation on music
		O	N.A.											
[Rap02]	2002	I	N.A.	Fourier analysis of input audio signal	N.A.		No	Poly	F	Generative probabilistic HMM (Baum-Welch training)	Signal "burstiness" measures for attack, steady and silence states		Method for piano music transcription	Excerpts from Mozart's piano sonata 18, K570
		O	MIDI file											
[GD02] (CAM)	2002	I	PCM, 22.05 kHz	sinusoidal model differentiated for mono/poly	N.A.		No	Mono	T	Bayesian Models MCMC harmonic inference	Frame by frame F0 tracking		Audio examples on the Web	Solo flute extract (Debussy's *Syrinx*)
		O	frame by frame F0 list											
[Kla03] (TUT)	2003	I	16-bit PCM signal fs = 44.1 kHz	DFT on Hamming windowed signal frames	N.A.		Algorithm only	Poly	F	Analysis of harmonic relationships between partials on 18 overlapping bands	No	No	Iterative estimation and harmonic pattern cancellation from the spectrum	Mixed samples from McGill, Iowa and IRCAM database
		O	Frame by frame F0 list	Signal plus noise model										
[BN05]	2005	I	16 bits PCM, 44.1 kHz (mono or multi-ch.)	Patterson-Meddis auditory model	N.A.		No	Mono Poly	F	Neural Network tracking of pitches detected by the onset detection algorithm	Peak-Picking algorithm on signal envelope	Offset detected by Neural Networks	Different instrument models are used Training Mode' available	Piano, guitar and violin samples Bach corale excerpt
		O	List of note parameters											

Table 2.1 - Comparison of Automatic Music Transcription Systems (3 of 5).

Reference (Group)	Year	System Input / Output		Pre-Processing & Mid-Level	Real time	Source	Mono	Time	Pitch Estimation Knowledge	Rhythm Info Extraction		Addiotional Notes	Evaluation Material
					Offline	Available	Poly	Freq.		Onsets	Durations		
[RK05] (TUT)	2005	I	PCM stereo, 44.1 kHz	70 Channels band pass filter	N.A.	No	Poly	F	Comb filters bank estimates periodicity in Freq. domain	Positive changes in F0s strength	Note events tracking by HMM	Evaluated in MIREX 2007-08 Accuracy ≈ 61% (MF0E) F-measure ≈ 34% (NT)	Excerpts from RWC Database
		O	MIDI file	HWR, STFT for all bands									
[BDS06] (QMUL)	2006	I	PCM files fs = 22.05 kHz	STFT & spectral smoothing Signal frames modeled as	N.A.	No	Poly	F	Generate F0 hypotheses for revelant amplitude partials Heuristic rules for partials grouping	Temporal parameters estimated integratin frame estimations over time		Method for transcription of recorded piano music. Use of a hybrid method combining time-freq. info	Disklavier played piano MIDI files. Error rate increases for high harmonic reate chords
		O	N.A.	weighted sum of an internal database piano waveforms									
[CKB06]	2006	I	N.A.	Modified sinusoidal model (state space form)	Real time	No	Poly	T	Use of bayesian networks, switching Kalman filters and a generative model to estimate note parameters.	Onsets and offsets are detected by transitions of the states *mute-sound* of the generators		The approach used allows to remove the frame by frame assumption for audio analysis	Own recordings of 2-3 voices chords. Qualitative results, especially offset errors
		O	Piano-roll of note parameters	to obtain a piano-roll like representation									
[PE07]	2007	I	8 kHz sampled audio from MIDI files	STFT	N.A.	No	Poly	F	87 One-versus-all SVM classifier for each piano note trained using *Sequential Minimal Optimization*	Note tracking by two state (on/off) HMM	No	Method for piano music transcription; Evaluation test are available	Synthesized, recorded and Disklavier played MIDI files
		O	N.A.										
[KNS07] (UT)	2007	I	PCM files fs = 44.1 kHz	Multi-resolution power spectrum obtained via Gabor wavelet transform	Offline	No	Poly	F	Harmonic temporal clustering (HTC) model for source separation	Joint estimation by using HTC model		Evaluated in MIREX 2007, 2008 and 2009 editions: Accuracy ≈ 49% (MF0E) F-measure ≈ 32% (NT)	Excerpts from RWC-Classical and RWC-Jazz databases
		O	$F0_s$, onset & offsets list										
[Kla08] (TUT)	2008	I	N.A.	Auditory (*gammatone*) filter bank (two types of 2nd order IIR resonators) IHC model, HWR	Real time	No	Poly	F	Periodicities search in the *Summary spectrum*. Detect F0 as peaks of a salience function	Energy peaks detection on signal envelope	No	Iterative estimation and harmonic pattern cancellation from the summary spectrum	Mixed samples from McGill, Iowa and IRCAM database
		O	N.A.										

Table 2.1 - Comparison of Automatic Music Transcription Systems (4 of 5).

Reference (Group)	Year	System Input / Output	Pre-Processing & Mid-Level	Real time	Source Available	Mono	Time	Pitch Estimation Knowledge	Rhythm Info Extraction		Addiotional Notes	Evaluation Material
				Offline		Poly	Freq.		Onsets	Durations		
[DZZS08]	2008	I: PCM audio signals O: Piano-roll (F0 tracking)	STFT	N.A.	No	Poly	F	Maximum Likelihood F0 est. Average Harmonic Strucutre (AHS) extraction for source separation	No	No	The system performs bad F0 recognition for inharmonic sounds	Synthesized, real instruments & singing voice Non standard evaluation metrics used for F0 est.
[Pin08]	2008	I: PCM mono audio signals, fs = 44.1 kHz O: Sequence of MIDI notes	STFT Gaussian smoothing of spectral patterns	N.A.	No	Poly	F	Candidates F0 with best salience function, calculated by considering partial amplitude and spectral smoothness	No	No	Evaluated in MIREX 2007 (previous version) & 2008 Accuracy ≈ 62% (2008 MF0E) F-measure ≈ 25% (2008 NT)	4000 chords; random mixtures of various samples (1 to 4 voices polyphony)
[VBB08]	2008	I: N.A. O: N.A.	ERB - scale time/freq. representation (similar to STFT)	Offline	No	Poly	F	NMF methods using harmonic / inharmonic constraints on the basis spectra with fixed/adaptative tuning + spectral smoothness	Thresholding amplitude sequence of each detected pitch	Offset est. is the same as for onsets	Evaluated in MIREX 2007, improved in MIREX 2008 Accuracy ≈ 54% (2008 MF0E) F-measure ≈ 20% (2008 NT)	43 Disklavier 30 seconds excerpts
[YRR+08] (IRCAM)	2008	I: N.A. O: N.A.	Spectral analysis based on sinusoidal and noise model	N.A.	No	Poly	F	Spectral matching, spectral smoothing and synchronous amplitude evolution of single sources	F0 tracking using a high-order HMM model with two states: attack and sustain		Evaluated in MIREX 2007, improved in MIREX 2008-09 Accuracy ≈ 69% (2009 MF0E) F-measure ≈ 36% (2008 NT)	Samples from McGill, Iowa, IRCAM database
[DHP09]	2009	I: PCM audio signals fs = 44.1 kHz O: N.A.	Spectral analysis Spectrum divided into *peaks* and non-peaks regions	N.A.	No	Poly	F	Maximum Likelihood parameter estimation in the frequency domain, using also neighboring frames estimates	Build pitch trajectories by constraint clustering problem with two class: *must-link* and cannot-*link*		Evaluated in MIREX 2009 Accuracy ≈ 57% (MF0E) F-measure ≈ 22% (NT)	10 real music performances (4-parts Bach's chorales)
[ANP11]	2011	I: PCM mono or stereo 16 bits, fs = 44.1 kHz O: MIDI file; pitches, onsets & offsets list	Joint constant-Q and Bispectral (higher order spectra) analysis	Offline	No	Poly	F	Iterative 2D harmonic pattern matching (in the bispectrum domain) and subsequent cancellation	Peaks-picking in Kullback-Leibler divergence (over spectral frames)	Tracking of note events on STFT	Evaluated in MIREX 2009 1st ranked in piano NT task Accuracy ≈ 48% (MF0E) F-measure ≈ 23% (NT)	Excerpts from RWC Classical datbase;

Table 2.1 - Comparison of Automatic Music Transcription Systems (5 of 5).

ACF	Autocorrelation Function	IHC	Inner Hair Cell
AHS	Average Harmonic Structure	IIR	Infinite Impulse Response filter
DFT	Discrete Fourier Transform	MCMC	Markov Chain Monte Carlo
F0	Fundamental Frequency	MF0E	Multiple F0 Estimation MIREX task
FFT	Fast Fourier Transform	NN	Neural Network
FIR	Finite Impulse Response filter	NT	Note Tracking MIREX task
fs	Sampling Frequency	PCM	Pulse Code Modulation
HMM	Hidden Markov Models	RWC	Real World Computing database
HTC	Harmonic Temporal Clustering	STFT	Short Time Fourier Transform
HWR	Half Wave Rectification	SVM	Support Vector Machine

Table 2.2 - Definition of acronyms used in TABLE 2.1.

representation is obtained by using a bank of band-pass filters, called optimum comb filter. This has been demonstrated to be a robust but computationally expensive algorithm; the pitch estimation strategy is to search for periodicities in the input signal by minimizing the summed absolute value of its magnitude difference. The system has revealed relatively good recognition performances with synthesized strings and real guitar duets.

Piszczalski and Galler - 1977

The system by Piszczalski and Galler [PG77] operates in the Frequency domain, and the obtained spectrum is equalized with a 12 dB attenuation curve, under 500 Hz and above 3000 Hz, to enhance significative sound partials. After detecting partials with a simple peak-detection procedure, each couple of partials is analyzed in order to find the smallest harmonic number, which would correspond to a harmonic series including the two partials at issue. A weighting coefficient, related to partials amplitude, is also assigned to each processed frequency couple. This information is later used to formulate hypothesis about the candidate fundamental frequencies. Such approach makes the whole system quite robust in cases of missing fundamental frequency and inharmonic partials, as qualitatively described in the evaluation discussion. The system is evaluated against some synthetic (mainly sinusoidal) tones and real signals

(carillon bells), although detailed results are not reported.

Slaney and Lyon - 1990

Human great capabilities of perceiving pitch, even in cases of missing fundamental frequencies and partials inharmonicity, led to an increasing interest in the Auditory Scene Analysis (ASA) in the first half of 90s. One of the first and most remarkable works belonging to this area was the "Perceptual Pitch Detector" by Slaney and Lyon [SL90], based on Licklider's "Duplex Theory" of pitch perception. The system is divided into three stages:

1. A Cochlear model which approximates the behavior of the human inner ear system, particularly the response of the auditory nerve. The cochlear model consists of a multi-channel bank of second order filters modeling the propagation of sound along the Basilar Membrane (BM); an array of Half-Wave Rectifiers (HWRs), aimed at emulating the role of the inner hair cells which respond to the BM movement in only one direction; finally, a four stage Automatic Gain Control (AGC) compresses the dynamic range of the processed signal.

2. The mid-level representation is obtained by computing the short-time windowed autocorrelation of the output of each cochlear channel. Collecting such information for each channel leads to the correlogram 2D representation, which allows to find periodicities (related to the perceived pitches) of the input signal (the latter are located at horizontal positions corresponding to the correlation delay-times equal to the periods of repetition). An example of correlogram of an audio input signal is depicted in Figure

3. The pitch detector block performs a peak enhancement in the correlogram; then the value at each time-lag is summed across all the frequencies, and the obtained array show peaks in correspondence of possible periodicities in the correlogram. Each detected periodicity τ reveals the presence of a pitched sound at frequency $1/\tau$.

Maher - 1990

Maher proposed a system for duet transcription [Mah90]. Several limitations are imposed: input signals must contain only two monophonic and separate

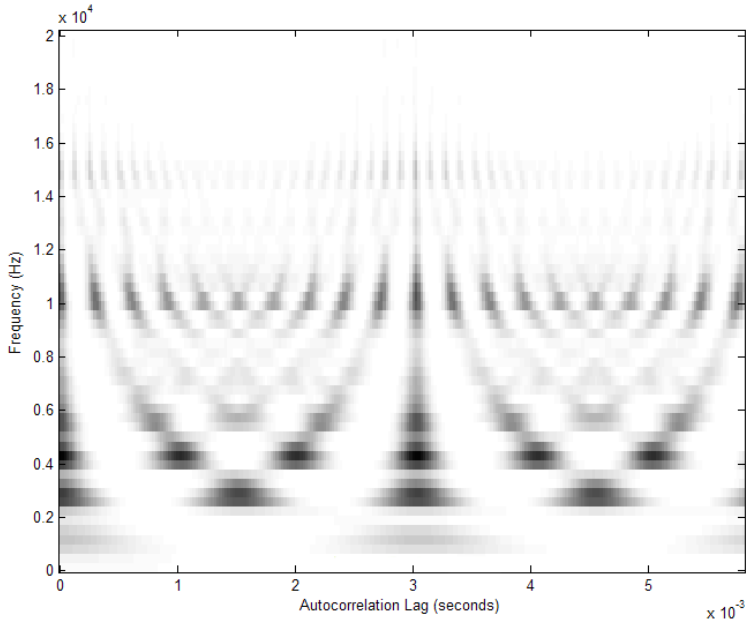


Figure 2.3: Correlogram of an upright piano E_4 at 330 Hz. ($\tau \approx 3ms$).

voices; both voices must be nearly harmonic; the frequency ranges of the fundamental frequencies belonging to the two voices should not overlap each other.

The system uses the McAulay - Quatieri sinusoidal model as a front end, which models the input signal as the sum of several time-variant sinusoidal components (similar to the Short-Time Fourier Transform). Pitch estimation task is performed by choosing the couple of frequencies which minimizes the mismatch between the predicted harmonic series of the two frequencies and the observed values. The system presents also a multi-strategy approach to resolve colliding partials: some of these techniques are based on a physical basis (analysis of beating components), other on acoustic knowledge applied (use of spectral templates). The system is not intended to work in real-time; actually the typical processing to real-time ratio exceeds 200.

Qualitative results are reported and they refer to some tests on both synthesized and real signals (clarinet/bassoon and trumpet/tuba). The initial duet assumption generally leads to the worst performance when only one voice is present (for example in solo passages). Good results are achieved when the two voices have a small number of coinciding partials; reverberation and other

ambient effects which often appear in musical recordings, represent one of the principal source of troubles.

Kashino et al. - 1995

The OPTIMA system (*Organized Processing Toward Intelligent Music scene Analysis*) proposed by Kashino, Tanaka, Nakadai and Kinoshita [KNKT95] anticipates, in some way, the introduction of the Blackboard system formalized by Martin in 1996. Actually, knowledge sources are present in this system, and they are used to find relationships among the different levels of the musical signal analysis.

The system operates in the frequency domain by extracting the frequency components of the input signal. Since simple amplitude thresholding is considered not sufficient to achieve a good estimation accuracy, the pitch detection method uses two regression planes pinching each spectral peak, in order to find temporal continuity of spectral local maxima. Rhythm information is extracted with Rosenthal's rhythm recognition method and Desain's quantization method. Onset detection is performed by combining rhythm information with beat probability, in order to determine the status (*continuous or terminated*) of the candidate fundamental frequencies. All this information is integrated in a Pearl's Bayesian network. Frequency peaks are therefore clustered according to calculated onset times, and these clusters are called processing scopes.

Six different types of external knowledge are used, in the main processing block, for information integration: chord transition dictionary, chord-note relation, chord naming rules, tone memory, timbre models, perceptual rules. With these classes of knowledge, the system aims at finding the best connectivity patterns that can explain the music played in the input signal.

Evaluation tests are organized in different levels: frequency components, notes, chords and song samples. The evaluation dataset is composed by synthesized MIDI files. Different levels are provided: frequency component level, note level, chord level, and song sample level. Detailed results are reported for note level tests only. A recognition rate from 30% to 87% is reported for two or three voices of polyphony, and an improvement is shown by using integration of knowledge, especially tone memory.

Martin - 1996

Martin proposes the Blackboard architecture for automatic music transcription [Mar96b]. This name comes from the metaphor of a group of researchers standing in front of a blackboard, working to find out the solution to a problem. This framework is a problem-solving model, which integrates knowledge from different sources and allows the interaction of different parts of the model. An expert musical knowledge, integrated with signal processing and other physical, engineering or mathematical frameworks, is considered useful to accomplish the task of automatic transcription of music.

The front end of Martin's system is an auditory model, similar to the one by Slaney and Lyon: it is a variant of the correlogram, according to Ellis' work. The filtering stage is composed by a 40 gammatone filter bank. The input signal is later half-wave rectified, and a short-time autocorrelation is made across each channel, obtaining a correlogram representation. Finally, the autocorrelations are summed across each band, and the time-lag presenting the largest peak is chosen as the pitch percept. A summary autocorrelation (*periodogram*) is obtained by averaging each frequency cell output by the zero-lag energy in the same frequency band, and then performing another average across all the frequency channels. This representation is an improvement over standard correlogram, since the periodogram presents a log-lag axis (lag, or inverse pitch, in a logarithmic scale) in addition to usual frequency channels and time axis.

The knowledge source (KS) is a set of five hypotheses (read correlogram frames, summary autocorrelation peaks, propose periodicities, note support and prune notes), which are organized into different levels of abstraction, and added to the periodogram front end, in order to improve the recognition performances. The system performs also a octave prediction test. The author reports only some qualitative examples, as evaluation, of monophonic and polyphonic transcription tests against excerpts from recorded performance of some pieces by Bach.

Tolonen and Karjalainen - 2000

Tolonen and Karjalainen proposed a variant to the *Unitary model of pitch perception* by Meddis and O'Mard [TK00]. Their system divides the input signal into two frequency regions (channels), with a cross-over frequency of 1 kHz. Then, a generalized autocorrelation of the low-channel signal, and of the

envelope of the high-channel signal, is performed. The autocorrelation functions are summed, and the summary function obtained is used for observing periodicities in the signal. Some results reported in a web page of the Helsinki University of Technology, linked from the paper, show a comparison between the two-channel system and the multi-channel algorithm by Meddis and Hewitt, thus providing similar performances. One of the main advantages of the two-channel method is the capability of operating in real-time.

Goto - 2000, 2004

Goto was one of the first who proposed a transcription system (*PreFEst*, from "Predominant $F0$ Estimation") for real-world audio signals [Got00], [Got04], characterized by complex polyphony, presence of drum or percussion, and singing voice also. To achieve such a goal, the music scene description and the signal analysis are carried out at a more specific level, focusing on the transcription of the melody and the bass line in musical fragments. Further limitations are imposed: the melody and the bass line should have the most predominant harmonic structure in the middle-high and in the low frequency regions, respectively.

The front end extracts instantaneous frequency components by using a STFT multi-rate filter bank, thus limiting the frequency regions of the spectrum with two band-pass filters. A probability density function is then assigned to each filtered frequency component; this function is a weighted combination of different harmonic-structure tone models. An Expectation-Maximization (EM) algorithm then estimates the model parameters. The frequency value that maximizes the probability function is detected as a predominant $F0$. Finally, a multi-agent architecture is used to sequentially track $F0$ peak trajectories, and to select the most stable ones; this operation is carried out by a salience detection and a dynamic thresholding procedures.

The system was evaluated against 10 excerpts from commercial classical/pop/jazz/ethnical recordings. The system reveals very good overall detection rates (88.4% for melody, 79.9% for bass). The system was realized to work in real-time.

Marolt - 2001

Marolt is the author of SONIC, a transcription system designed specifically for piano music [Mar01]. The front-end is a combination of the auditory model

and an adaptive oscillator network. The input signal is splitted into frequency bands, by using an array of 200 IIR gammatone filters with center frequencies logarithmically spaced between 70 and 6000 Hz. The output of this stage is then processed according to the Meddis' model of half-wave rectification and compression of the dynamic range of the signal.

The output of each frequency channel is send to an adaptive oscillator (a modified version of Large-Kolen oscillator), which synchronize with the input signal (assumed to be periodic) by adjusting its frequency and phase to the one of the driving signal. The observation of the synchronized frequency of each oscillator gives information about the frequency components present in the signal. The oscillators are grouped into networks, in order to track a group up to 10 harmonic related frequency components that may belong to a single tone. There are 88 oscillator networks, one for each piano key; the initial frequency of the first oscillator in each network is set to the pitch of the corresponding piano note. Finally, a set of neural networks is used to recognize single notes from the output of the oscillator network.

The evaluation dataset is composed by 120 synthesized MIDI pieces. Marolt reports an average number of correctly detected of about 90%. Octave errors and repeated note errors are the most frequent cases.

Klapuri - 2003

This method proposed by Klapuri [Kla03] is an iterative technique, consisting mainly in two procedures: a predominant $F0$ estimation, and harmonic pattern cancelation from the mixture. First, a FFT-based spectral analysis is performed, then the spectrum is processed in order to eliminate noise and to enhance sound partials' information. The obtained spectrum is divided into 18 overlapping frequency bands distributed between 50 and 6000 Hz, with a 50% overlap. In each band, a weighting factor is calculated over the frequency index. The frequency value with the highest weight is detected as the most predominant $F0$. Subsequently, its harmonic set is canceled from the mixture, and the operation is repeated for the residual, until an energy-based stop criterion is met. For evaluation, random mixed samples from McGill University, Iowa University and IRCAM audio databases are used. Results are analyzed in relation with a priori known polyphony of test data. Generally the error rate is below 10% (for polyphonies between 1 and 5 notes) and about 10% for 6-note polyphony.

Bruno and Nesi - 2005

The proposed system [BN05] processes the input audio signal through a Patterson-Meddis auditory model. A partial tracking module extracts the harmonic content, which is analyzed to estimate active pitches. Onset detection is performed by using a peak-picking algorithm on the signal envelope. Pitch tracking is carried on, for each note the onset of which has been previously estimated by a bank of neural networks. This network can be trained by a set of parameters describing several instrument models (concerning partial amplitude weights, frequency range etc.). A *Training Mode* is also available, which is needed to create automatically features and patterns for new instruments configuration.

Ryynänen and Klapuri - 2005

This system [RK05] uses a probabilistic framework, a hidden Markov Model (HMM), to track note events. The multiple $F0$ estimator front end is based on auditory model: a 70-channel bandpass filter bank splits the audio input into sub-band signals which are later compressed, half-wave rectified and low-pass filtered with a frequency response close to $1/f$. Short time Fourier Transform is then performed across the channels, and the obtained magnitude spectra are summed together into a summary spectrum. Predominant $F0$ estimation, and cancelation from the spectrum of the harmonic set of detected $F0$ is performed iteratively. Onset detection is also performed by observing positive energy variation in the amplitude of detected $F0$ values. The output of $F0$ estimator is further processed by a set of three probabilistic models: a HMM note event model tracks the likelihood for each single detected note; a silence model detects temporal intervals where no notes are played; finally, a musicological model controls the transitions between note event and silence models.

Evaluation is conducted by testing the system transcription performances on 91 recordings of several musical genres (including popular, rock, classical, jazz and world) extracted from RWC database. Notes for drums and percussions are excluded from the set of MIDI reference events. Recall of 39%, precision of 41% and mean overlap ratio of 40% are reported. The system has been also evaluated in the MIREX framework (2007 and 2008 editions), achieving an overall accuracy of 61% in the first task (MF0E - Multiple $F0$ Estimation frame by frame) and a F-measure of 34% in the second task (NT - Note Tracking).

Cemgil, Kappen and Barber - 2006

The system proposed by Cemgil and Kappen [CKB06] is based on a generative approach and a dynamical Bayesian network: note parameters to be estimated (including pitches, onsets and durations) are considered as a collection of hidden variables, while the acoustic recorded audio samples are the observed variables into a Bayesian inference problem. The modelling task must, therefore, infer the appropriate generative model which is able to reproduce better the given audio sequence. The starting model for the input signal is a space-state variant of the sinusoidal/noisy model, which is more suitable to obtain a piano-roll representation including all the unknown note parameters. The piano-roll is considered as a set of single pitch binary generators (switching Kalman filters) that can assume two possible states across time: sound and mute. Onsets are detected when the state of a generator switch from mute to sound. In this case, the actual status vector is forgotten and a new state vector is created. This aspect simplifies the high computational costs due to the nature of the problem and the use of Kalman filters. The final task of the system is to estimate the *Maximum A Posteriori* (MAP) configuration of the piano-roll which represents better the observed audio data. Main advantage of Bayesian inference is that the model can be trained, taking into account different combinations for note parameters. In addition, such approach allows to eliminate the frame by frame assumption, often used for audio signal analysis, particularly in the field of music transcription task. In this way the input musical signals can be analyzed in real time and with sample precision. For this reason this system can be considered as operating in the time-domain; actually any Fourier based (or related methods) spectral analysis is performed. The system has been evaluated against some recordings of 2, 3 voices of polyphony, reporting only qualitative results.

Kameoka, Nishimoto and Sagayama - 2007

The authors have proposed a multipitch analyzer based on harmonic temporal structured clustering (HTC) method [KNS07]. This technique aims at decomposing the energy patterns of observed power spectrum into distinct clusters, originated by separate sources. The time frequency representation of the input signal is considered as an unknown fuzzy mixture of energy components belonging to a certain number of single sources. The clustering of energy patterns is performed by introducing a spectral masking function which decomposes the

spectrum into active areas, to be associated to single sources. The decomposed spectrum is then modeled by HTC source models. The spectral masking function and the HTC model parameters are the unknown variables which have to be estimated. This is made through an Expectation-Maximization (EM) algorithm, which is composed mainly of two steps: in the first, the masking function is estimated with fixed model parameters; in the second step, the masking function is fixed and the model parameters are estimated. These operations are repeated until the unknown variables converge to a stationary value.

The system is evaluated against some excerpts of Real World Computing (RWC) database. This dataset provide PCM audio signals and corresponding reference MIDI which, however, need a careful alignment with the audio spectrogram, in order to act as a faithful ground truth reference on a frame by frame based evaluation. Accuracy rates, reported separately for eight audio tracks, vary from 61.2% to 81.2%. The system has been also evaluated in the MIREX 2007 framework, reporting an accuracy of 33.6% for the frame by frame multiple $F0$ estimation task and F-measure of 9% in the mixed set note tracking task. An improved version was submitted to MIREX 2009, achieving good results in both tasks: a frame by frame $F0$ estimation accuracy of 49% (task 1) and a F-measure of 31.9% for note tracking (task 2, 1^{st} ranked).

Vincent, Bertin and Badeau - 2008

Vincent, Bertin and Badeau have proposed a system based on Non-negative Matrix Factorization (NMF) [VBB08]. By using this technique, the observed signal spectrogram (Y) is decomposed into a weighted sum of basis spectra (contained in H) scaled by a matrix of weighting coefficients (W):

$$Y = WH.$$

Since the elements of Y are non-negative by nature, the NMF method approximates it as a product of two non-negative matrixes, W and H.

The system at issue uses a family of constrained NMF models, where each basis spectrum is a sum of narrow-band spectrum (scaled by a model function of the spectral envelope) containing partials at harmonic or inharmonic frequencies. This assures that the estimated basis spectra are pitched at known fundamental frequencies; such condition is not always guaranteed if standard NMF models are applied without any of these constraints.

The input signal is first pre-processed to obtain a representation similar to the Short-time Fourier Transform, by performing an ERB-scale representation. Then, the parameters of the models are adapted by minimizing the residual loudness after applying the NMF model: the linear parameters (amplitude sequence, envelope coefficients) are multiplicatively updated, while the other nonlinear parameters (tuning and inharmonicity factors) are updated via a Newton-based optimizer. Pitches, onsets and offsets of detected notes are transcribed by simply thresholding the amplitude sequence.

Evaluation was conducted on a dataset of 43 Yamaha Disklavier piano excerpts. Standard NMF method is compared with the proposed constrained model. The former reaches an overall F-measure of about 74%, the latter reaches a maximum F-measure of 87% (harmonic-fixed method). The system has been also evaluated in the MIREX 2007 framework: the two submitted versions reached average accuracies of 46.6% and 54.3% in the task 1 (multi-$F0$ estimation over 10 ms frames) and an average F-measure of 45.3% and 52.7% in the task 2 (note tracking).

Chang, Yeh, Roebel et al. - 2008

In this method [YRR+08], instantaneous spectra are obtained by FFT analysis. A noise level estimation algorithm is applied to enhance the peaks generated by sinusoidal components (produced by an unknown number of audio sources) with respect to noise peaks. Subsequently, a matching between a set of hypothetical sources and the observed spectral peaks is made, by using a score function based on the following three assumptions: *spectral match with low inharmonicity, spectral smoothness* and *synchronous amplitude evolution*. These features are based on physical characteristics generally showed by the partials generated by a single source.

Musical notes tracking is carried out by applying a high order hidden Markov model (HMM) having two states: attack and sustain. This is a probabilistic framework aimed at describing notes evolution as a sequence of states evolving on a frame by frame basis. The goal is to estimate optimal note paths and the length of each note trajectory. The connection weights among the different states are calculated in the forward tracking stage; candidate best trajectories are estimated iteratively in the backward stage, by extracting most likely paths between recorded roots and leaves. Finally, the source streams are obtained by pruning the candidate trajectories, in order to maximize the like-

lihood of the observed polyphony.

The system has been evaluated within the MIREX 2007 framework, and improved versions were submitted to MIREX 2008 and MIREX 2009 contests. Best multiple $F0$ estimation accuracy of 69% has been achieved in 2009 running (1^{st} ranked in task 1): this is currently the highest accuracy reached in all the MIREX editions for the first task. Best performance in the note tracking task was reached in 2008 edition, with an F-measure of 35.5% (1^{st} ranked).

Pertusa and Iñesta - 2008

The algorithm proposed by the authors performs a multi $F0$ estimation, key and tempo detection on a frame by frame basis [PIn08]. Short time Fourier Transform is applied to the input signal. $F0$ candidates are extracted from each frame spectrum by amplitude thresholding. Then all the possible combinations of candidates are considered, and a salience factor is associated to each combination. The salience is computed by considering the loudness of the harmonic pattern of each $F0$ candidate, and the smoothness of the harmonics amplitude; to calculate the smoothness factor, each harmonic patter is low-pass filtered using a truncated normalized Gaussian window. The combination with the best salience (calculated as the product of loudness and smoothness, summed for each candidate) is considered the winner chord in the actual frame.

The dataset for evaluation is generated with random mixtures of different music samples, for a total of 4000 chords, and polyphony of 1, 2, 4 and 6 voices. Test results yield an overall accuracy (which corresponds, in the paper, to the standard F-measure) of 56.2%. The system was also evaluated in MIREX 2008 (and a previous version participated also in 2007 edition), reporting a maximum accuracy of 61.8% in the 2008 first task (MF0E) and a maximum F-measure of 27.7% in the 2007 second task (NT).

Yeh, Roebel and Rodet - 2010

Yeh and Roebel combined the $F0$ trajectories tracking method, described in [YRR+08], with the candidate $F0$ extraction algorithm proposed in [YRR10]. The system takes into account the sinusoids plus noise model of the musical polyphonic signal. A Rayleigh distribution is used to model the noise spectral content, and to separate from signal content before multi-pitch estimation process. The multi-$F0$ estimation is carried on by posing a set of $F0$ hypotheses on the basis of spectral and perceptual features. Candidate $F0$s are classified

into two groups: *harmonically related F0s* (HRF0s) if they are multiple of other candidate frequency values, *non- harmonically related F0s* (NHRF0s) otherwise. NHRF0s are the candidates for predominant pitch extraction and harmonic content extraction. HRF0s undergo a partial overlap treated, and also a polyphonic inference is added to estimate the number of concurrent voices. Finally, a score function merges all these combinations of hypotheses, features and estimates to extract the detected notes.

2.3 General Review and Discussion

From this review some general aspects concerning music transcription systems can be gathered. Automatic transcription of polyphonic music is to be considered as a conjunction of several tasks, which can be accomplished jointly or by using dedicated procedures. From this point of view, a modular architecture seems to be the most robust approach for a problem solution. Such construct perfectly matches with Martin's idea of a blackboard architecture [Mar96a]. Many researchers still believe that signal processing strategies are a fundamental basis, although such strategies, as widely demonstrated, can provide better results if they work jointly with other a priori knowledge sources. This statement recalls the parallel between perceptual and brain abstraction levels in human cognition process.

While human perceptual approach to music has been successfully studied and implemented through the Computational Auditory Scene Analysis (CASA), knowledge at higher levels of abstraction is more difficult to be coded into an computational framework, since it must be consistent with experience, and it often needs training to avoid misleading or ambiguous decisions. Such knowledge is commonly represented by all those models which aim at reproducing human capabilities in features extraction and grouping (e.g., harmony related models, musical key finding etc.). The experience of a well-trained musician can be understood as a greatly flexible and deep network of state-machine like hints, as well as complex matching procedures.

Review of music transcription systems in literature suggest that time-frequency representation (usually performed through short-time Fourier transform) of the signal is the most used front end, upon which pitch estimation and onset/offset detection strategies can be applied. Multi resolution spectrogram representation (obtained by using constant-Q or wavelet transform) seems to be, in our opinion, the most suitable, since it fits properly the exponential

spacing of note frequencies, and it also reduces computational load to achieve the desired time/frequency resolution. Auditory model based front ends have been largely studied and applied in the 90s; however, the interest toward this approach has decreased. Time domain techniques are becoming more and more infrequent, since they have provided poor performances in polyphonic contexts. Temporal information, however, is a relevant feature which has been used, joined with frequency analysis, to retrieve information about partials tracking [BDS06].

About pitch estimation strategies, the largely adopted class of spectral content peak-picking based algorithms has revealed to be not sufficient to achieve satisfactory transcription accuracies. Actually, amplitude thresholding in the spectrum domain, as well as simple harmonic pattern matching, leads to frequent false positive detection, if no other knowledge is applied. For this reasons, alternative thresholding methods have been investigated, for instance with variable, frequency-dependent amplitude thresholds [ES06]. A large variety of models has been proposed for spectral analysis, and it is not easy to find out if which is the best approach among the others. The most used techniques in recent literature are: Nonnegative Matrix Factorization [SB03], [Vir07], [VBB08], Hidden Markov Models [Rap02], [RK05], [YRR$^+$08], Bayesian models [KNKT95], [GD02], [GDI06], [DD07], generative harmonic models [CKB06], and the use of jointed frequency and time information.

Onset detection is often devolved upon detecting rapid spectral energy over time. Techniques such as the phase-vocoder based functions, applied to audio spectrogram, seem to be more robust with respect to peak-picking algorithms performed upon the signal envelope. Offset detection is still considered as of less perceptual importance. Statistical frameworks offer an interesting perspective in solving discontinuities in joint time-pitch information, typically yielded by lower processing levels techniques. On the contrary, other devices that usually reach a deep level of specialization, like neural networks, are more suitable for particular areas or subsets of automatic transcription; actually this kind of tools is often trained at recognizing specific notes or at inferring particular instrumental models [Mar01].

In conclusion, as a key point for future work, we can assert that model based integration seems to be an area definitely more amenable to new solutions, with respect to signal processing field. We expect that the increasing progress and improvements in computational processing will allow to build more refined systems, with a higher parallelism degree and a joint involvement of a greater

number of techniques.

Chapter 3

Constant-Q Bispectral Analysis

The bispectrum belongs to the class of Higher-Order Spectra (HOS, or polyspectra), used to represent the frequency content of a signal. An overview of the theory on HOS can be found in [Bri65], [NM93] and [NR87]. The bispectrum is defined as the third-order spectrum, being the amplitude spectrum and the power spectral density the first and second-order ones, respectively.

Previous studies on bispectral representation of audio signal have been proposed: Dubnov and associates made use of the bispectrum in order to extend the research on musical timbre, sound textures [Dub96], and instrument classification and clustering [DT95] [DTC95]. In these works it is shown that the content of bispectral analysis of a musical signal is strongly related to the harmonicity measure of concurrent sounds: all natural sustained vibration sounds contain small bandwidth random fluctuations (*jitter*) in the frequencies of their components. These fluctuations are random but coherent for all partials of a single sound. This aspect seems to be at the basis of the psycho-acoustic process according to which the human inner ear system is able to perceive separate sounds in a polyphonic mixture. Abeysekera has proposed a method for polyphonic pitch extraction based on the *frequency-lag* distribution, derived from the bispectral analysis [Abe04].

3.1 The Bispectrum

Let $x(k)$, $k = 0, 1, \ldots, K-1$, be a digital audio signal, modeled as a real, discrete and locally stationary process. The nth order moment, m_n^x, is defined [NM93] as:

$$m_n^x(\tau_1, \ldots, \tau_{n-1}) = E\{x(k)x(k+\tau_1)\ldots x(k+\tau_{n-1})\},$$

where $E\{\cdot\}$ is the statistical mean. The nth order cumulant, c_n^x, is defined [NM93] as:

$$c_n^x(\tau_1, \ldots, \tau_{n-1}) = m_n^x(\tau_1, \ldots, \tau_{n-1}) - m_n^G(\tau_1, \ldots, \tau_{n-1}),$$

where $m_n^G(\tau_1, \ldots, \tau_{n-1})$ are the nth-order moments of an equivalent Gaussian sequence having the same mean and autocorrelation sequence as $x(k)$. Under the hypothesis of a zero mean sequence $x(k)$, the relationships between cumulants and statistical moments up to the third order are:

$$c_1^x = E\{x(k)\} = 0,$$

$$c_2^x(\tau_1) = m_2^x(\tau_1) = E\big\{x(k)x(k+\tau_1)\big\},$$

$$c_3^x(\tau_1, \tau_2) = m_3^x(\tau_1, \tau_2) = E\big\{x(k)x(k+\tau_1)x(k+\tau_2)\big\}. \tag{3.1}$$

The nth-order polyspectrum, denoted as $S_n^x(f_1, f_2, \ldots, f_{n-1})$, is defined as the $(n-1)$-dimensional Fourier transform of the corresponding order cumulant, that is:

$$S_n^x(f_1, f_2, \ldots, f_{n-1}) = \sum_{\tau_1=-\infty}^{+\infty} \cdots \sum_{\tau_{n-1}=-\infty}^{+\infty} c_n^x(\tau_1, \tau_2, \ldots, \tau_{n-1}) \exp\Big(- j2\pi(f_1\tau_1 + f_2\tau_2 + \ldots + f_{n-1}\tau_{n-1})\Big).$$

The polyspectrum for $n = 3$ is also called *bispectrum*. It is also denoted as:

$$B_x(f_1, f_2) = S_3^x(f_1, f_2) = \sum_{\tau_1=-\infty}^{+\infty} \sum_{\tau_2=-\infty}^{+\infty} c_3^x(\tau_1, \tau_2) e^{-j2\pi f_1\tau_1} e^{-j2\pi f_2\tau_2}. \tag{3.2}$$

3.1.1 Relevant properties of the bispectrum

The bispectrum is a bivariate function representing some kind of signal-energy related information, as more deeply analyzed in the next section. In Figure 3.1, a contour-plot of the bispectrum of an audio signal is shown. The bispectrum presents twelve mirror symmetry regions:

$$\begin{aligned} B_x(f_1, f_2) &= B_x(f_2, f_1) = B_x^*(-f_2, -f_1) = B_x(-f_1 - f_2, f_2) = \\ &= B_x(f_1, -f_1 - f_2) = B_x(-f_1 - f_2, f_1) = B_x(f_2, -f_1 - f_2). \end{aligned}$$

Hence, the analysis can take into consideration only a single non redundant bispectral region [CE94]. Hereafter, $B_x(f_1, f_2)$ will denote the bispectrum in the triangular region $\mathcal{T}$ with vertices (0,0), (f_s/2,0) and (f_s/3,f_s/3), i.e., $\mathcal{T} = \left\{(f_1, f_2) : 0 \leq f_2 \leq f_1 \leq \frac{f_s}{2}, f_2 \leq -2f_1 + f_s\right\}$, which is depicted in Figure 3.1, where f_s is the sampling frequency.

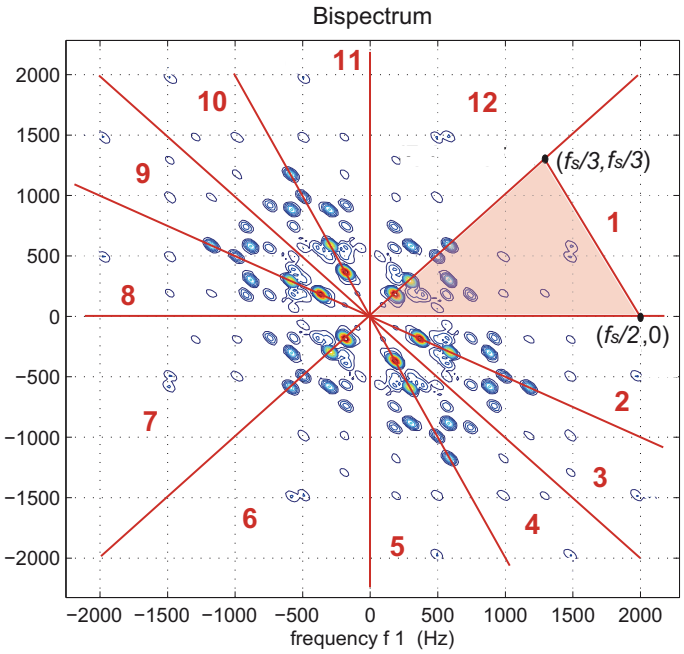


Figure 3.1: Contour plot of the magnitude bispectrum, according to Equation (3.3), of the trichord $F\sharp_3$(185 Hz), D_4(293 Hz), B_4(493 Hz) played on an acoustic upright piano and sampled at $f_s = 4$ kHz. The twelve symmetry regions are in evidence (clockwise enumerated), and the one chosen for analysis is highlighted.

It can be shown [NM93] that the bispectrum of a finite-energy signal can be expressed as:

$$B_x(f_1, f_2) = X(f_1)X(f_2)X^*(f_1 + f_2), \tag{3.3}$$

where $X(f)$ is the Fourier Transform of $x(k)$, and $X^*(f)$ is the complex conjugate of $X(f)$.

As in the case of power spectrum estimation, the estimations of the bispectrum of a finite random process are not consistent, i.e., their variance does not decrease with the observation length. Consistent estimations are obtained by averaging either in the time or in the frequency domain. Two approaches are usually considered, as described in [NM93].

The *indirect method* consists of: 1) the estimation of the third-order moments sequence, computed as temporal average on disjoint or partially overlapping segments of the signal; 2) estimation of the cumulants, computed as the average of the third-order moments over the segments; 3) computation of the estimated bispectrum as the bidimensional Fourier transform of the windowed cumulants sequence.

The *direct method* consists of: 1) computation of the Fourier transform over disjoint or partially overlapping segments of the signal; 2) estimation of the bispectrum in each segment according to (3.3) (eventually, frequency averaging can be applied); 3) computation of the estimated bispectrum as the average of the bispectrum estimates in each segment.

Finally, an interesting property involving higher spectra and gaussian processes. Consider a generic signal $y(k)$:

$$y(k) = x(k) + w(k) \qquad \text{with k } \in \mathbb{N}$$

composed by the sum of two independent processes: a non-gaussian contribute, $x(k)$, which can be considered more specifically as a deterministic signal, and a gaussian contribute $w(k)$. Under these assumptions:

$$c_n^w = 0 \qquad \forall n > 2,$$

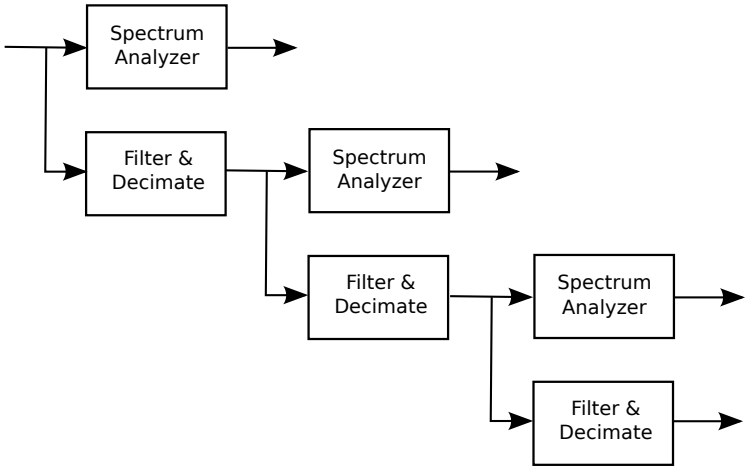
that is, all cumulant spectra of order greater than two are identically zero for gaussian additive processes. Therefore, a signal transform in the bispectrum domain suppress additive colored Gaussian noise of unknown power spectrum. For these reasons, cumulant spectra can become high signal-to-noise ratio (SNR) domains in which one may perform detection, parameters estimation, features extraction or even signal reconstruction [NM93].

3.2 Constant-Q Analysis

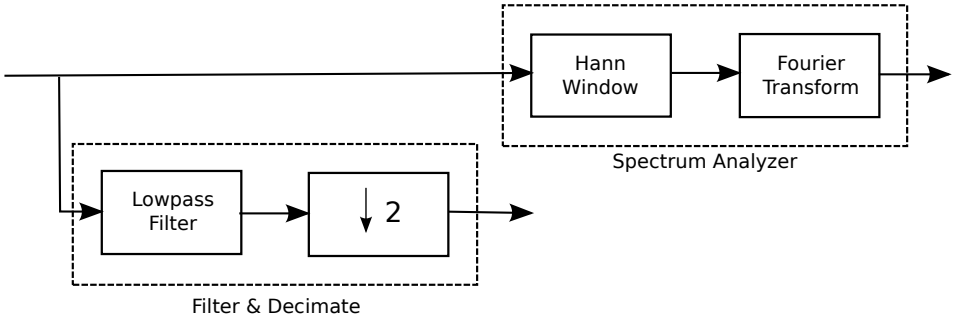
The estimation of the bispectrum according to (3.3), involves computing the spectrum $X(f)$ on each segment of the signal. In each octave, twelve semitones

need to be discriminated: since the octave spacing doubles with the octave number, the requested frequency resolution decreases when the frequency increases. For this reason, a spectral analysis with a variable frequency resolution is suitable for audio applications.

The constant-Q analysis [Bro91], [DKBN06] is a spectral representation that properly fits the exponential spacing of note frequencies. In the constant-Q analysis, the spectral content of an audio signal is analyzed in several bands. Let N be the number of bands and let



(a)



(b)

Figure 3.2: Octave Filter Bank: (a) building block of the tree, composed by a spectrum analyzer and by a filtering/downsampling block; (b) blocks combination to obtain a multi-octave analysis.

$$Q_i = \frac{f_i}{B_i},$$

where f_i is a representative frequency, e.g., the highest or the center frequency, of the ith band and B_i is its bandwidth. In a constant-Q analysis, we have $Q_i = Q$, $i = 1, 2, \ldots, N$, where Q is a constant.

A scheme that implements a constant-Q analysis is illustrated in Figure 3.2. It consists of a tree structure, shown in Figure 3.2(a), whose building block, shown in Figure 3.2(b), is composed of a spectrum analyzer block and by a filtering/downsampling block (low-pass filtering and downsampling by a factor two). The spectrum analyzer consists in windowing the input signal (Hann window with length N_H samples for each band has been used) followed by a Fourier transform that computes the spectral content at specified frequencies of interest. The low-pass filter is a zero-phase filter, implemented as a linear-phase filter followed by a temporal shift. Using zero-phase filters allows us to extract segments from each band that are aligned in time. The nominal filter cutoff frequency is at $\pi/2$. Due to the downsampling, the N_H-samples long analysis window spans a duration that doubles at each stage. Therefore, at low frequencies (i.e., at deeper stages of the decomposition tree), a higher resolution in frequency is obtained at the price of a poorer resolution in time.

3.3 Constant-Q Bispectral Analysis for Polyphonic Pitch Detection

In order to better explain the interaction of harmonics generated by a mixture of sounds, we first focus on the application of the bispectral analysis to examples of monophonic signals, and then some examples of polyphonic signals are considered.

3.3.1 Monophonic signal

Let $x(n)$ be a signal composed by a set $\mathcal{H}$ of four harmonics, namely $\mathcal{H} = \{f_1, f_2, f_3, f_4\}$, $f_k = k \cdot f_1$, $k = 2, 3, 4$, i.e.,

$$x(n) = \sum_{k=1}^{4} 2\cos(2\pi f_k n / f_s),$$

$$X(f) = \sum_{k=1}^{4} \delta(f \pm f_k),$$

where constant amplitude partials have been assumed. According to (3.3), the bispectrum of $x(n)$ is given by

$$B_x(\eta_1,\eta_2) = X(\eta_1)X(\eta_2)X^*(\eta_1+\eta_2) = \\ = \Big(\sum_{k=1}^{4}\delta(\eta_1 \pm f_k)\Big)\Big(\sum_{l=1}^{4}\delta(\eta_2 \pm f_l)\Big)\Big(\sum_{m=1}^{4}\delta(\eta_1+\eta_2 \pm f_m)\Big).$$

When the products are developed, the only terms different from zero that appear are the pulses located at (f_k, f_l), with f_k, f_l such that $f_k + f_l \in \mathcal{H}$. Hence, we have:

$$\begin{aligned} B_x(\eta_1,\eta_2) = \quad & \delta(\eta_1 \pm f_1)\delta(\eta_2 \pm f_1)\delta(\eta_1+\eta_2 \pm f_2) \\ & + \delta(\eta_1 \pm f_1)\delta(\eta_2 \pm f_2)\delta(\eta_1+\eta_2 \pm f_3) \\ & + \delta(\eta_1 \pm f_1)\delta(\eta_2 \pm f_3)\delta(\eta_1+\eta_2 \pm f_4) \\ & + \delta(\eta_1 \pm f_2)\delta(\eta_2 \pm f_1)\delta(\eta_1+\eta_2 \pm f_3) \\ & + \delta(\eta_1 \pm f_2)\delta(\eta_2 \pm f_2)\delta(\eta_1+\eta_2 \pm f_4) \\ & + \delta(\eta_1 \pm f_3)\delta(\eta_2 \pm f_1)\delta(\eta_1+\eta_2 \pm f_4). \end{aligned}$$

Note that peaks arise along the first and third quadrant bisector thanks to the fact that $f_2 = 2f_1$ and $f_4 = 2f_2$. By considering the non-redundant triangular region $\mathcal{T}$ defined in Section 3.1, the above expression can be simplified into:

$$\begin{aligned} B_x(\eta_1,\eta_2) = \quad & \delta(\eta_1 - f_1)\delta(\eta_2 - f_1)\delta(\eta_1+\eta_2 - f_2) \\ & + \delta(\eta_1 - f_2)\delta(\eta_2 - f_1)\delta(\eta_1+\eta_2 - f_3) \\ & + \delta(\eta_1 - f_3)\delta(\eta_2 - f_1)\delta(\eta_1+\eta_2 - f_4) \\ & + \delta(\eta_1 - f_2)\delta(\eta_2 - f_2)\delta(\eta_1+\eta_2 - f_4). \end{aligned} \tag{3.4}$$

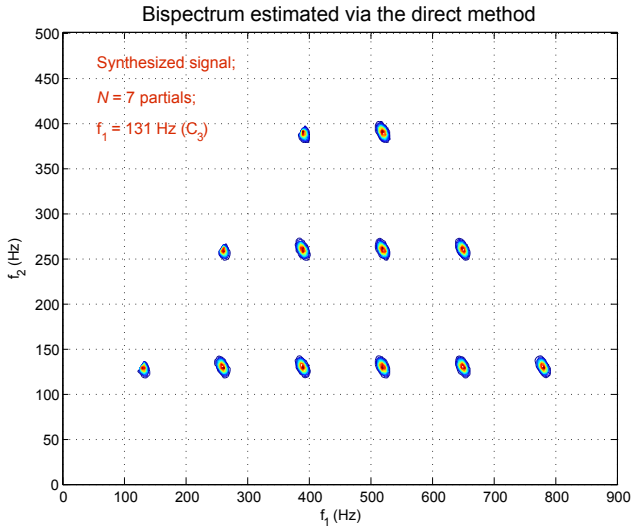
Equation (3.4) can be generalized to an arbitrary number T of harmonics as follows:

$$B_x(\eta_1,\eta_2) = \sum_{p=1}^{\lfloor T/2 \rfloor}\delta(\eta_2 - f_p)\sum_{q=p}^{T-p}\delta(\eta_1 - f_q)\delta(\eta_1+\eta_2 - f_{p+q}). \tag{3.5}$$

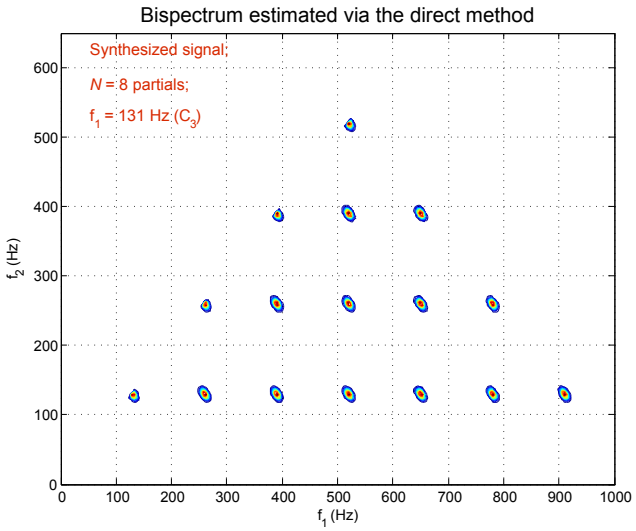
This formula shows that every monophonic signal generates a bidimensional bispectral pattern characterized by peaks positions

$$\{(f_i, f_i), (f_{i+1}, f_i), \ldots, (f_{T-i}, f_i)\}, i = 1, 2, \ldots, \lfloor \frac{T}{2} \rfloor. \tag{3.6}$$

Such a pattern is depicted in Figure 3.3 for a synthetic note at a fundamental frequency $f_1 = 131$ Hz, with $T = 7$ and $T = 8$. The energy distribution in



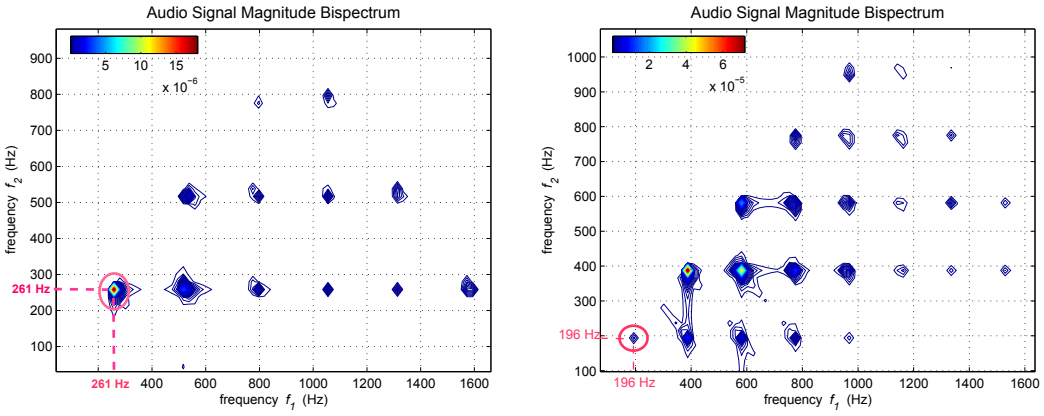
(a)



(b)

Figure 3.3: Bispectrum of monophonic signals (note C_3) synthesized with (a) $T = 7$ and (b) $T = 8$ harmonics.

the bispectrum domain is validated by the analysis of real world monophonic sounds. Figure 3.4 shows the bispectrum of a C_4 note played by an acoustic piano and a G_3 note played by a violin, both sampled at $f_s = 44100$ Hz. Even if the number of significant harmonics is not exactly known, the positions of the peaks in the bispectrum domain confirm the theoretical behavior previously shown.



(a) (b)

Figure 3.4: Bispectrum of (a) a C_4 (261 Hz) played on a upright piano, and of (b) a G_3 (196 Hz) played on a violin (bowed). Both sounds have been sampled at 44100 Hz.

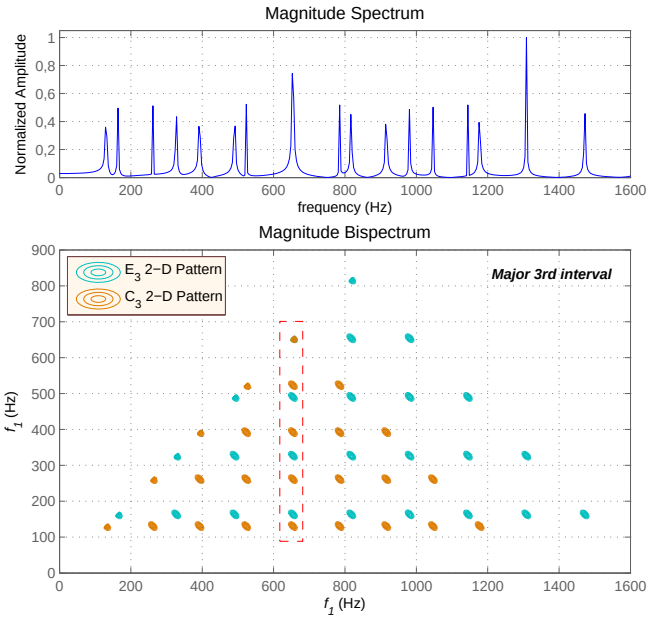
3.3.2 Polyphonic signal

Consider the simplest case of a polyphonic signal: a bichord. Accordingly with the linearity of the Fourier Transform, the spectrum of a bichord is the sum of the spectra of the component sounds. From Equation (3.3), it is clear that the bispectrum has a non-additivity nature. This means that, the bispectrum of a bichord is not equal to the sum of the bispectra of component sounds, as described in next section. In order to be more specific, two examples, in which the two notes are spaced by either a major third or a perfect fifth interval, are considered; such intervals are characterized by a significant number of overlapping harmonics. Figures 3.5-(a) and 3.5-(b) show the bispectrum of synthetic signals representing the intervals $C_3 - E_3$ and $C_3 - G_3$, respectively. For each note, ten constant-amplitude harmonics were synthesized. The top row plots in

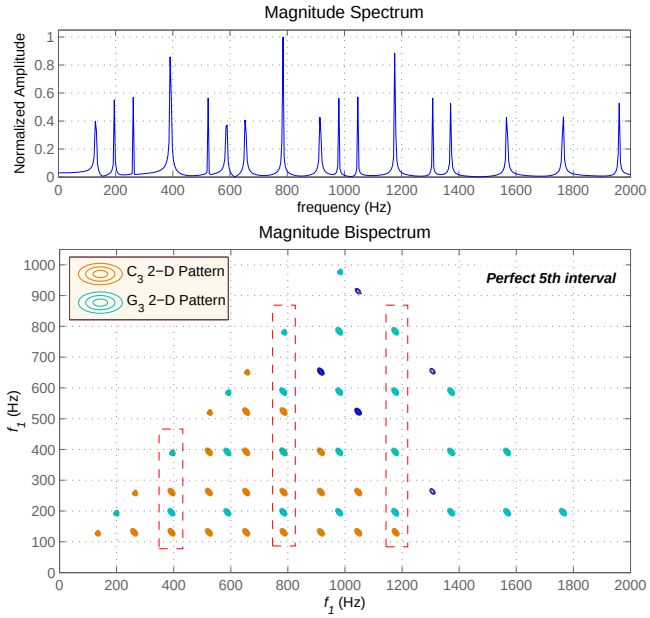
Figures 3.5-(a) and 3.5-(b) demonstrate the spectrum of the synthesized audio segments, from which the harmonics of the two notes are apparent. Overlapping harmonics, e.g., the frequencies $5i \cdot F_{0_{C_3}} = 4i \cdot F_{0_{E_3}}$ for the major third interval, with i an integer, can not be resolved. In Figure 3.6, the bispectrum of a real bichord produced by two bowed violins, playing the notes A_3 (220 Hz) and D_4 (293 Hz), is shown. The interval is a perfect fourth (characterized by a fundamental frequencies ratio equal to 4:3, corresponding to a distance of 5 semitones in the well-tempered scale), so that each third harmonic of D_4 overlaps with each fourth harmonic of A_3. Both in the synthetic and in the real sound examples, the patterns relative to each note are distinguishable, apart from a single peak on the quadrant bisector. In the next section, the bispectrum of polyphonic sound is theoretically treated, together with some examples. In particular, the cases regarding polyphonic signals with two or more sounds have been considered. In the case of bichords, one of the most interesting cases, being a perfect fifth interval, since it presents a strong partials overlap ratio. In this case, the analysis of residual coming from the difference of the real bispectrum of the bichord signal with respect to the linear composition of the single bispectra of concurrent sounds, has been performed. The formal analysis has demonstrated that the contributions of this residual are null or negligible for proposed multi-$F0$ estimation procedure. This theoretical analysis has been also confirmed by the experimental results, as shown with some examples. Moreover, the case of trichord with strong partial overlapping and a high number of harmonics per sound has confirmed the same results.

3.4 A Polyphonic Pitch Detection Case Study

In this section, the bispectrum of a bichord is theoretically treated, together with some examples. In particular, the cases regarding polyphonic signals with two or more sounds have been considered. In the case of bichords, one of the most interesting cases, being a perfect fifth interval, since it presents a strong partials overlap ratio. In this case, the analysis of residual coming from the difference of the real bispectrum of the bichord signal and the linear composition of the single bispectra of concurrent sounds, have been performed. The formal analysis has demonstrated that the contribution of this residual are null or neglectable for multi-$F0$ estimation procedure. This theoretical analysis has been also confirmed by some experimental results. Moreover, the case of trichord with strong partial overlapping and a high number of harmonics has



(a)



(b)

Figure 3.5: Spectrum and bispectrum generated by (a) a major third $C_3 - E_3$ and (b) a perfect fifth interval $C_3 - G_3$. Ten harmonics have been synthesized for each note. The regions into dotted lines in the bispectrum domain highlight that local maxima of both single monophonic sounds are clearly separated, while they overlap in the spectral representation.

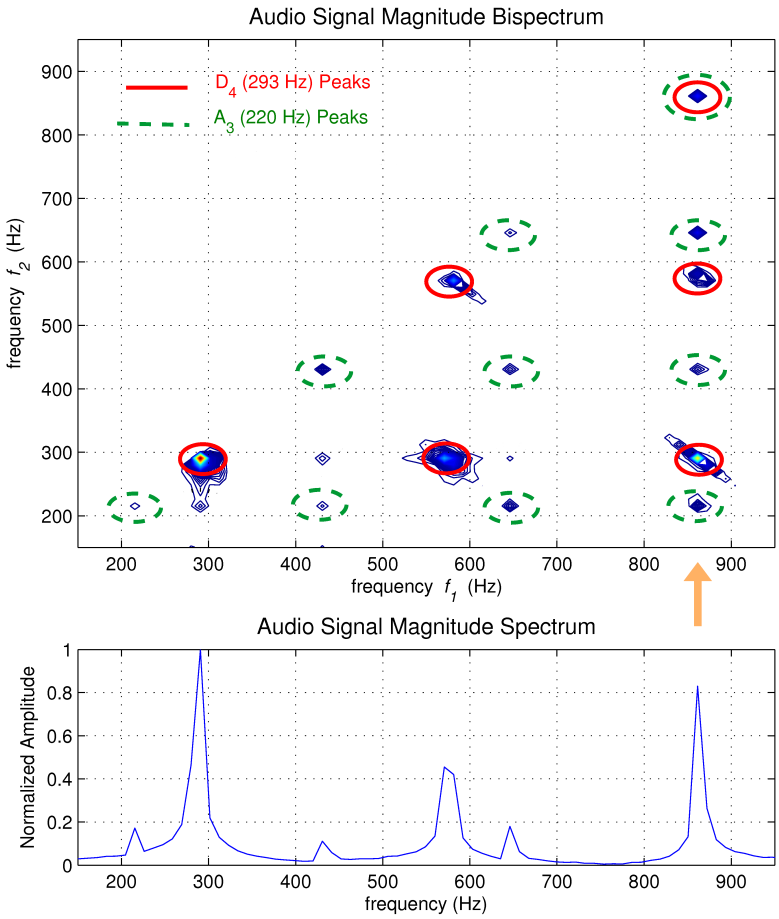


Figure 3.6: Detail (top figure) of the bispectrum of a bichord (A_3 at 220 Hz and D_4 at 293 Hz), played by two violins (bowed), sampled at 44100 Hz. The arrow highlights the frequency at 880 Hz, where the partials of the two notes overlap in the spectrum domain.

confirmed the same results.

3.4.1 Bispectrum of a polyphonic signal: a bichord

The behavior of the bispectrum for a polyphonic signal is now analyzed. Let us to recall the spectrum (positive frequencies only) of a generic monophonic sound with fundamental frequency f_0:

$$X(f) = \sum_{k=1,\dots,P} \delta(f - kf_0), \qquad kf_0 \in H,$$

where H is the set of harmonics of the sound, consisting of P partials, f_0 included: $H = \{f_0, 2f_0, 3f_0, \dots, (P-1)f_0, Pf_0\}$.

Consider now, as an example and without loss of generality, two synthesized sounds, S_1 and S_2, each one composed by five partials, so that: $H_1 = \{f_{01}, 2f_{01}, 3f_{01}, 4f_{01}, 5f_{01}\}$ and $H_2 = \{f_{02}, 2f_{02}, 3f_{02}, 4f_{02}, 5f_{02}\}$. The generated spectra are denoted as $X_1(f)$ and $X_2(f)$, respectively. Accordingly with the linearity of the Fourier Transform, let $X(f) = X_1(f) + X_2(f)$ be the spectrum of the polyphonic signal S, composed by the mixture of S_1 and S_2. Under these assumptions, the bispectrum of the polyphonic signal, computed with the direct method (defined by Equation 3.3) can be expressed as follows:

$$\begin{aligned}
B_{1,2}(f_1, f_2) &= X(f_1)X(f_2)X^*(f_1 + f_2) = \\
&\Big(X_1(f_1) + X_2(f_1)\Big)\Big(X_2(f_1) + X_2(f_2)\Big)\Big(X_1(f_1 + f_2) + X_2(f_1 + f_2)\Big)^* = \\
&X_1(f_1)X_1(f_2)X_1^*(f_1 + f_2) + X_1(f_1)X_2(f_2)X_1^*(f_1 + f_2) + \\
&X_2(f_1)X_1(f_2)X_1^*(f_1 + f_2) + X_2(f_1)X_2(f_2)X_1^*(f_1 + f_2) + \\
&X_1(f_1)X_1(f_2)X_2^*(f_1 + f_2) + X_1(f_1)X_2(f_2)X_2^*(f_1 + f_2) + \\
&X_2(f_1)X_1(f_2)X_2^*(f_1 + f_2) + X_2(f_1)X_2(f_2)X_2^*(f_1 + f_2).
\end{aligned} \tag{3.7}$$

3.4.2 Analysis of Bispectrum nonlinearity

The first and the last terms of the sum in Equation 3.7 are equal to $B_1(f_1, f_2)$ and $B_2(f_1, f_2)$, which denote the bispectra associated to signals S_1 and S_2, respectively. The bispectrum is not linear, actually:

$$B_{1,2}(f_1, f_2) \neq B_1(f_1, f_2) + B_2(f_1, f_2).$$

Let $B_{diff}(f_1, f_2)$ be the difference of the three terms $B_{1,2}(f_1, f_2)$, $B_1(f_1, f_2)$ and $B_2(f_1, f_2)$:

$$\begin{aligned} B_{diff}(f_1, f_2) =& B_{1,2}(f_1, f_2) - B_1(f_1, f_2) - B_2(f_1, f_2) = \\ & X_1(f_1)X_2(f_2)X_1^*(f_1 + f_2) + X_2(f_1)X_1(f_2)X_1^*(f_1 + f_2) + \\ & X_2(f_1)X_2(f_2)X_1^*(f_1 + f_2) + X_1(f_1)X_1(f_2)X_2^*(f_1 + f_2) + \\ & X_1(f_1)X_2(f_2)X_2^*(f_1 + f_2) + X_2(f_1)X_1(f_2)X_2^*(f_1 + f_2). \end{aligned} \tag{3.8}$$

Let us analyze each term of the sum in Equation 3.8, in order to better understand the behavior of $B_{diff}(f_1, f_2)$. The **first term** yields:

$$\begin{aligned} X_1(f_1)X_2(f_2)X_1^*(f_1 + f_2) = & \sum_{k=1,\dots,5} \delta(f_1 - kf_{01}) \sum_{l=1,\dots,5} \delta(f_2 - lf_{02}) \\ & \sum_{m=1,\dots,5} \delta(f_1 + f_2 - mf_{01}) = \prod_{\{1,2,1\}}, \end{aligned} \tag{3.9}$$

$$kf_{01} \in H_1, lf_{02} \in H_2, mf_{01} \in H_1.$$

The product $\prod_{\{1,2,1\}}$ is not null only if each term of the product itself is not null. Concerning the first two terms, this happens when $f_1 = kf_{01}$ (that is, when f_1 takes the value of any of the partials belonging to H_1) and, similarly, when $f_2 = lf_{02}$. This involves that, considering the third term, the entire product is non-zero only when it exists at least an integer value m such that $mf_{01} = kf_{01} + lf_{02}$ (where $k, l = 1, \dots, 5$ and $mf_{01} \in H_1$). To satisfy this condition, it is necessary (but not sufficient, depending on the length of H_1 and H_2) that the sounds present overlapping partials; a sufficient condition is that the two harmonic series, H_1 and H_2, share at least one frequency value.

As an example: consider two sounds, with harmonic sets H_1 and H_2, generate a perfect fifth interval (which presents a very strong partials overlap ratio); this implies that $2f_{02} = 3f_{01}$. Under these conditions, the contribute of $\prod_{\{1,2,1\}}$ would be non-zero only for the following couples:

$$(f_{01}, 2f_{02}); (2f_{01}, 2f_{02}),$$

with $f_{01}+2f_{02} = f_{01}+3f_{01} = 4f_{01} \in H_1$ and, similarly, $2f_{01}+2f_{02} = 5f_{01} \in H_1$. It is worthy to notice that these two couples are located in the upper triangular region of the plane (f_1, f_2), above the first quadrant bisector, and so they are outside the non-redundant region considered in the computation of the bispectrum (see Section 3.1 and Figure 3.1). For this reason, the contribute of

$\prod_{\{1,2,1\}}$ to $B_{diff}(f_1, f_2)$ is zero in this context. This analysis can be generalized for all the terms of the sum in Equation 3.8, as reported in the following.

Considering the **second term** of the sum in Equation 3.8:

$$\begin{aligned} X_2(f_1)X_1(f_2)X_1^*(f_1+f_2) = \sum_{k=1,\dots,5} \delta(f_1 - kf_{02}) \sum_{l=1,\dots,5} \delta(f_2 - lf_{01}) \\ \sum_{m=1,\dots,5} \delta(f_1 + f_2 - mf_{01}) = \prod_{\{2,1,1\}} , \end{aligned} \tag{3.10}$$

$$kf_{02} \in H_2, lf_{01} \in H_1, mf_{01} \in H_1.$$

The term $\prod_{\{2,1,1\}}$ is non-zero only if exist at least an integer values m such that $mf_{01} = kf_{02} + lf_{01}$ (where $k, l = 1, \dots, 5$ and $mf_{01} \in H_1$). Following the example of the two sounds generating a perfect fifth interval, this happens only for the couples of frequencies:

$$(2f_{02}, f_{01}); (2f_{02}, 2f_{01}).$$

As it can be noticed, this is the symmetric case of $\prod_{\{1,2,1\}}$, with respect to the first quadrant bisector, and in this circumstance these points are inside the non-redundant region considered for bispectrum computation. Therefore, $\prod_{\{2,1,1\}}$ is not null in this domain; however, $B_1(f_1, f_2)$ also generates nonnull values in correspondence of these two couples, in the equivalent form of $(3f_{01}, f_{01})$ and $(3f_{01}, 2f_{01})$ (see Equation 3.5). For this reason, $\prod_{\{2,1,1\}}$ does not generate any additional peaks in the (f_1, f_2) plane; the only effect is to add an amplitude contribute to bispectral peaks generated by $B_1(f_1, f_2)$ at the same positions in the (f_1, f_2) plane. At the end of these considerations we will show that these contributes can be considered not relevant in the computation of normalized 2-D cross-correlation, within the multi-$F0$ estimation procedure.

Consider now the **third term** in equation 3.8:

$$\begin{aligned} X_2(f_1)X_2(f_2)X_1^*(f_1+f_2) = \sum_{k=1,\dots,5} \delta(f_1 - kf_{02}) \sum_{l=1,\dots,5} \delta(f_2 - lf_{02}) \\ \sum_{m=1,\dots,5} \delta(f_1 + f_2 - mf_{01}) = \prod_{\{2,2,1\}} , \end{aligned} \tag{3.11}$$

$$kf_{02} \in H_2, lf_{02} \in H_2, mf_{01} \in H_1.$$

$\prod_{\{2,2,1\}}$ is non-zero only if exist at least an integer value m such that $mf_{01} = kf_{02} + lf_{02}$ (where $k, l = 1, \dots, 5$ and $mf_{01} \in H_1$). In our example, such a case occurs for the couple (f_{02}, f_{02}), actually:

$$f_{02} + f_{02} = 3f_{01} \in H_1.$$

This shows that $\prod_{\{2,2,1\}}$ only adds an amplitude contribute to a bispectral peak originated by $B_1(f_1, f_2)$ at the same position in the (f_1, f_2) plane, without generating any additional peaks.

Consider the **fourth term** in equation 3.8:

$$\begin{aligned} X_1(f_1)X_1(f_2)X_2^*(f_1+f_2) = \sum_{k=1,\dots,5} \delta(f_1 - kf_{01}) \sum_{l=1,\dots,5} \delta(f_2 - lf_{01}) \\ \sum_{m=1,\dots,5} \delta(f_1 + f_2 - mf_{02}) = \prod_{\{1,1,2\}}, \\ kf_{01} \in H_1, lf_{01} \in H_1, mf_{02} \in H_2. \end{aligned} \tag{3.12}$$

$\prod_{\{1,1,2\}}$ is non-zero only if exist at least an integer value m such that $mf_{02} = kf_{01}+lf_{01}$ (where $k, l = 1, \dots, 5$ and $mf_{02} \in H_2$). In our example, this happens for the following couples of frequencies:

- $(f_{01}, 2f_{01})$, actually $f_{01} + 2f_{01} = 2f_{02} \in H_2$.
- $(f_{01}, 5f_{01})$, actually $f_{01} + 5f_{01} = 4f_{02} \in H_2$.
- $(2f_{01}, 4f_{01})$, actually $2f_{01} + 4f_{01} = 4f_{02} \in H_2$.

These three couples are outside the non-redundant region considered for bispectrum computation; $\prod_{\{1,1,2\}}$ is not null only in correspondence of the following couples, which are the symmetric ones of the three ones listed above (with respect to the first quadrant bisector):

- $(2f_{01}, f_{01})$; this adds an amplitude contribute to the bispectral peak generated by $B_1(f_1, f_2)$ at the same position in the (f_1, f_2) plane;
- $(5f_{01}, f_{01})$ and $(4f_{01}, 2f_{01})$; in correspondence of these two couples, $\prod_{\{1,1,2\}}$ gives origin (in this particular case) to two additional peaks in the bispectrum: they represent an extension to the five harmonics 2-D monophonic pattern of the sound at pitch f_{01} (according to equation 3.5). The reason why $B_1(f_1, f_2)$ does not generate peaks in correspondence of these two couples is that the considered harmonic set H_1 is composed by five partials.

Consider the **fifth term** in equation 3.8:

$$\begin{aligned} X_1(f_1)X_2(f_2)X_2^*(f_1+f_2) = \sum_{k=1,\dots,5} \delta(f_1 - kf_{01}) \sum_{l=1,\dots,5} \delta(f_2 - lf_{02}) \\ \sum_{m=1,\dots,5} \delta(f_1 + f_2 - mf_{02}) = \prod_{\{1,2,2\}}, \end{aligned} \tag{3.13}$$

$$kf_{01} \in H_1, lf_{02} \in H_2, mf_{02} \in H_2.$$

$\prod_{\{1,2,2\}}$ is non-zero only if exist at least an integer value m such that $mf_{02} = kf_{01} + lf_{02}$ (where $k, l = 1, \ldots, 5$ and $mf_{02} \in H_2$). In our example, this happens for the following couples of frequencies:

- $(3f_{01}, f_{02})$ and $(3f_{01}, 2f_{02})$, in correspondence of which $\prod_{\{1,2,2\}}$ adds an amplitude contribute to the bispectral peaks generated by $B_2(f_1, f_2)$ in $(2f_{02}, f_{02})$ and $(2f_{02}, 2f_{02})$;
- $(3f_{01}, 3f_{02})$, which is outside the non-redundant region considered in the computation of the bispectrum.

Consider, finally, the **sixth term** in equation 3.8:

$$X_2(f_1)X_1(f_2)X_2^*(f_1+f_2) = \sum_{k=1,\ldots,5} \delta(f_1 - kf_{02}) \sum_{l=1,\ldots,5} \delta(f_2 - lf_{01}) \sum_{m=1,\ldots,5} \delta(f_1 + f_2 - mf_{02}) = \prod_{\{2,1,2\}}, \tag{3.14}$$

$$kf_{02} \in H_2, lf_{01} \in H_1, mf_{02} \in H_2.$$

As it can be noticed, this is the symmetric case of the previous $\prod_{\{1,2,2\}}$, with respect to the first quadrant bisector. Therefore, $\prod_{\{2,1,2\}}$ is non-zero only when exist at least an integer value m such that $mf_{02} = kf_{02} + lf_{01}$ (where $k, l = 1, \ldots, 5$ and $mf_{02} \in H_2$). In our example, this happens for the following couples of frequencies:

- $(f_{02}, 3f_{01})$, which is outside the boundaries of non-redundant region considered in the computation of the bispectrum;
- $(2f_{02}, 3f_{01})$ and $(3f_{02}, 3f_{01})$, in correspondence of which $\prod_{\{2,1,2\}}$ adds an amplitude contribute to the bispectral peaks generated by $B_2(f_1, f_2)$ in $(2f_{02}, 2f_{02})$ and $(3f_{02}, 2f_{02})$.

Eventually, let us to remember that we have illustrated an example in which the two interfering sounds present a strong partials overlap ratio. For a generic synthesized bichord, the contribute of $B_{diff}(f_1, f_2)$ gains more relevance with the increasing number of partials in the harmonic sets of the sounds, and with the increasing partials overlap ratio. In the other cases, when the two sounds don't share the value of any of their partials within their harmonic sets, the value of $B_{diff}(f_1, f_2)$ is zero.

3.4.3 An empirical example: a synthesized bichord

A graphical example could be useful to illustrate in a clearer way this argumentation. In Figure 3.7, the contour plot of the bispectrum of a synthesized 5 harmonics bichord: $C_4 - G_4$ ($C_4 : f_{01} = 261.63$ Hz, $G_4 : f_{02} = 392$ Hz), which forms a perfect fifth interval; then in Figure 3.8 the contour plot of the sum of the bispectra of C_4 and G_4 is shown. In Figure 3.7, the monophonic 2-D patterns of the two sounds are distinguishable, and also the two additional peaks generated by the contribute of the product $\prod_{\{1,1,2\}}$, located at $(5f_{01}, f_{01})$ and $(4f_{01}, 2f_{01})$, which appear to have a smaller amplitude.

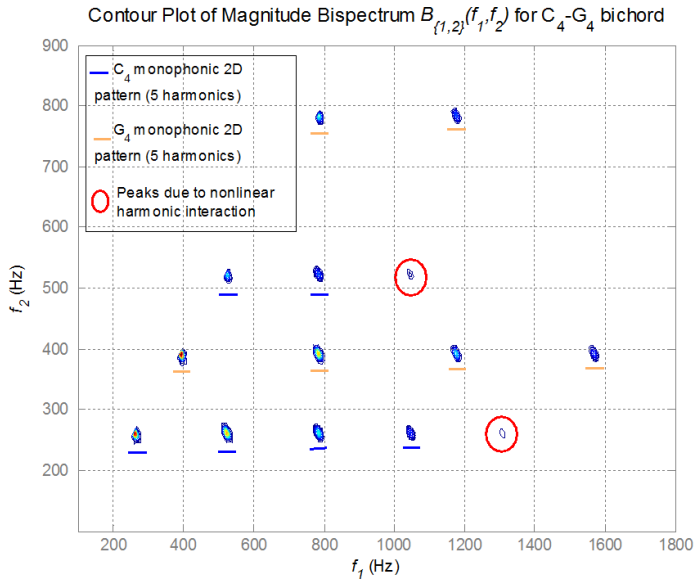


Figure 3.7: Contour plot of the bispectrum of synthesized bichord $C_4 - G_4$.

Dealing with real sounds, it is impossible to quantify the amplitude contribute given by each single term present in $B_{diff}(f_1, f_2)$, if the number of partials and their amplitude model is not known in advance for each concurrent sound. For this reason, it is difficult to perform a general qualitative analysis. On the other hand, it is possible to evaluate the normalized 2-D cross-correlations between both $B_{1,2}(f_1, f_2)$ and $B_1(f_1, f_2) + B_2(f_1, f_2)$ with a 2-D pattern, equivalent to the one used in the multi-$F0$ estimation procedure which is the core of the system described in this PhD. Thesis. The results of the two normalized 2-D cross-correlation (denoted as $\rho_{B_{1,2}}$ and $\rho_{B_1+B_2}$) and

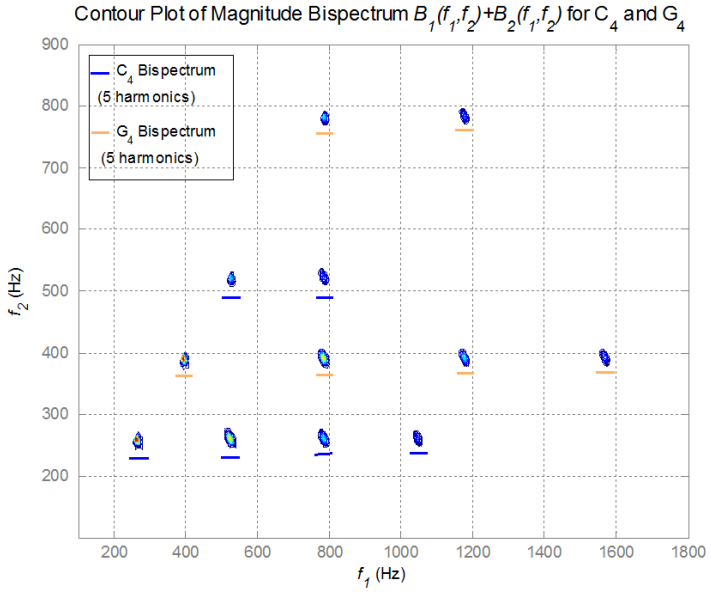


Figure 3.8: Contour plot of the bispectrum of synthesized bichord $C_4 - G_4$.

the array obtained by subtracting $\rho_{B_{1,2}}$ and $\rho_{B_1+B_2}$, are shown in Figure 3.9.

It can be noted that there are no relevant differences between the two cases (in Figure 3.9, bottom part reporting the difference, the y-axis scale has been enlarged to make difference array more readable). Moreover, the same normalized 2-D cross-correlation for other two synthesized sounds has been calculated with the same pitch by using 10 harmonics instead of 5. This operation was made in order to show that the contribute of $B_{diff}(f_1, f_2)$ would not affect significantly the values of 2-D correlation (and, therefore, the results of multi-$F0$ Estimation procedure) with increasing number of partials. The results are shown in Figure 3.10.

3.5 Comparison of multi-$F0$ estimation procedures

In this section, an example of multi-$F0$ Estimation procedure step-by-step, carried out by the transcription system presented in this work. The results are compared with those obtained by a transcription method performing an iterative 1-D pattern matching in the spectrum domain, and subsequent direct

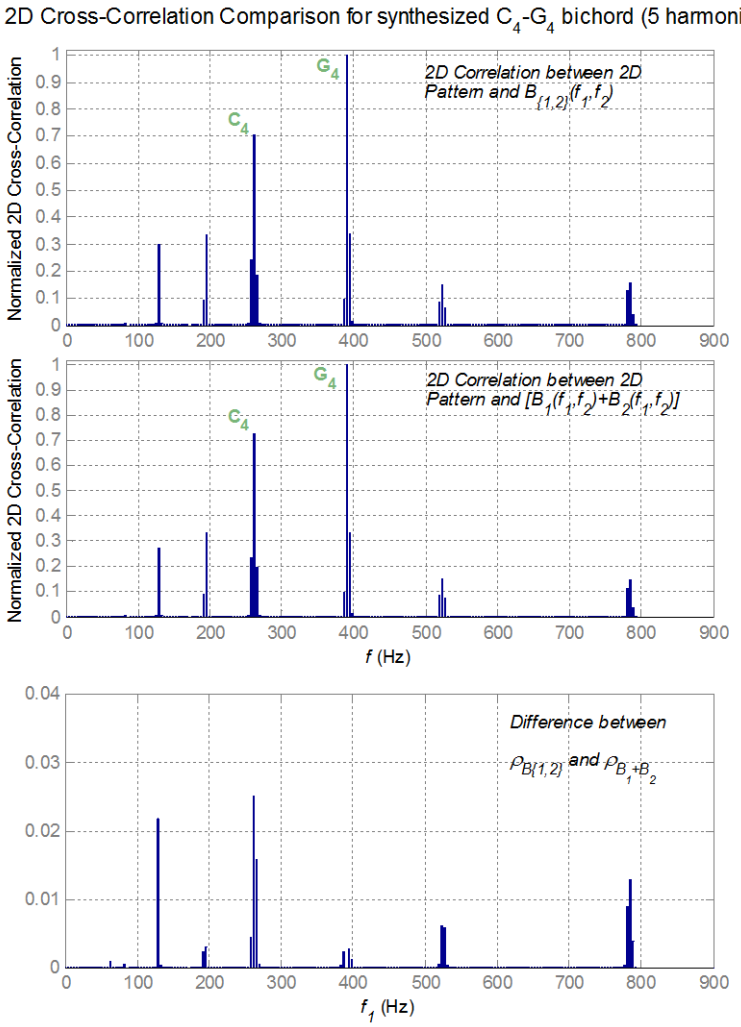


Figure 3.9: Comparison of normalized 2-D cross-correlation for 5-harmonics synthesized bichord C_4-G_4, and the difference of them (with a different scale).

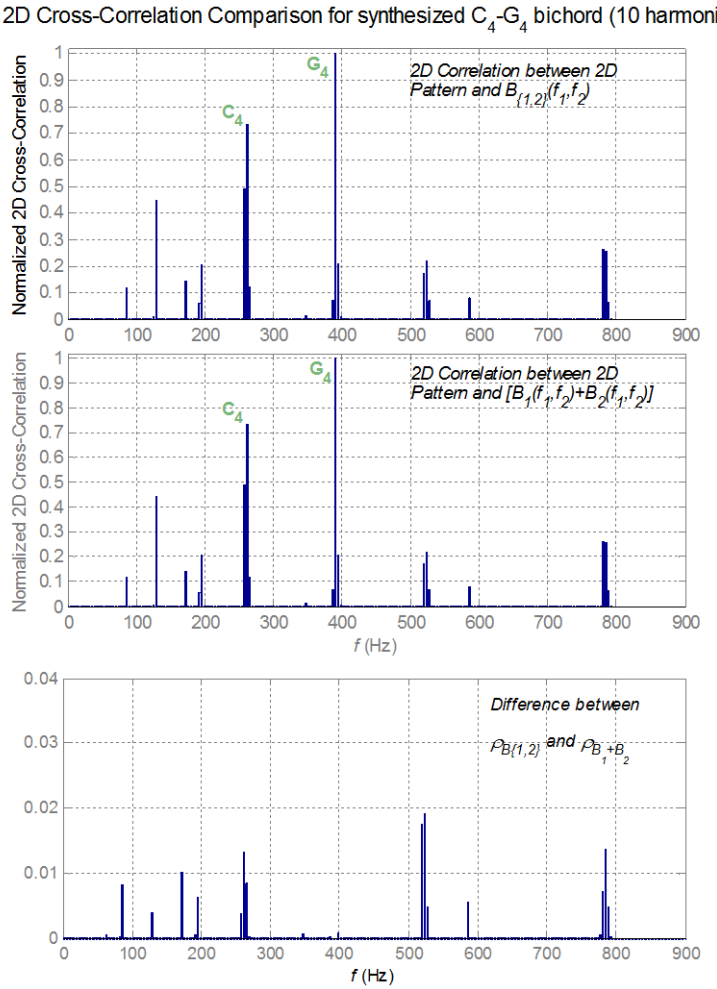


Figure 3.10: Comparison of normalized 2-D cross-correlation for 10-harmonics synthesized bichord C_4-G_4, and the difference of them (with a different scale).

cancelation of the harmonic pattern of estimated notes. The audio input source is a real signal taken from the RWC database, analyzed in a single frame for the purpose of the example. In the processed frame, notes G_2, D_4 and B_4 are playing, corresponding to MIDI notes 43, 62 and 71, respectively. These notes present a significant partials overlapping. Actually, denoting the fundamental frequencies as f_{01}, f_{02} and f_{03}, respectively, they stay in the following ratios each other:

$$f_{02} = 3f_{01}; \qquad f_{03} = 5f_{01}; \qquad f_{03} = \frac{5}{3}f_{02}.$$

These ratios are approximated, in the frequency-log scale adopted in our system (following the well-tempered scale) with distances of 19, 28 and 9 semitones.

In Figure 3.11, the amplitude spectrum and bispectrum before the $F0$ estimation process are presented.

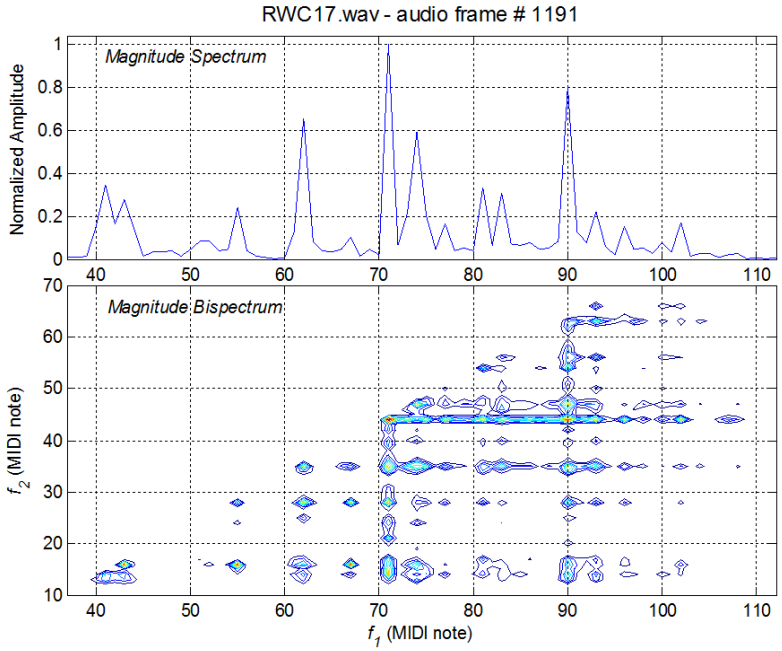


Figure 3.11: Amplitude spectrum and bispectrum of audio signal before Multi-F0 estimation.

In next Figures (Figure 3.12 and B.3) a direct comparison between both the multi-$F0$ estimation procedures is depicted, by plotting the normalized 1-D and 2-D cross-correlations for each step.

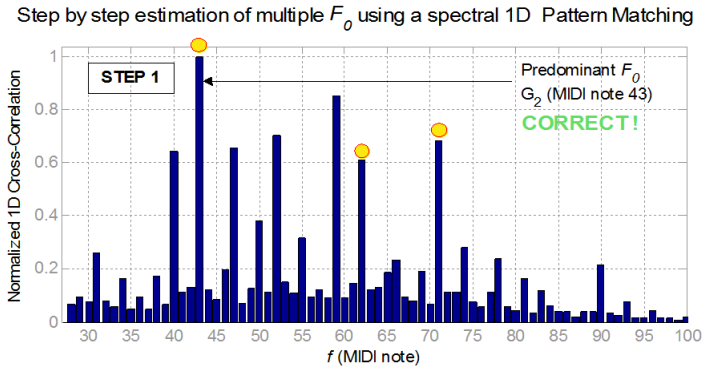
It should be noted that, the 2-D bispectral correlation is much clearer than the 1-D spectral correlation. As stated in the following (see section 4.3 in next chapter regarding the system architecture), denoting the normalized 2-D cross-correlation as $\rho(f_1, f_2)$, if a monophonic sound has a fundamental frequency corresponding to index q in the discrete log-frequency array, then the maximum of $\rho(f_1, f_2)$ is expected to be found at (q, q). For this reason, the cross-correlation is computed only for $f_1 = f_2 = q$, that is only upon the points belonging to the first quadrant bisector.

Moreover, comparing Figure 3.12 and 3.13, it can be observed that after *Step 1* (in which the lowest note G_2 is correctly identified by both the algorithms), the 2-D pattern matching method (in the bispectrum domain) succeeds in correctly estimating all the other reference pitches. On the other hand, the direct cancelation of spectral G_2 pattern (in the 1-D $F0$ estimation method) deletes some coinciding partials of the two higher sounds, including the fundamental frequencies of both D_4 and G_4, as shown in Figure 3.14.

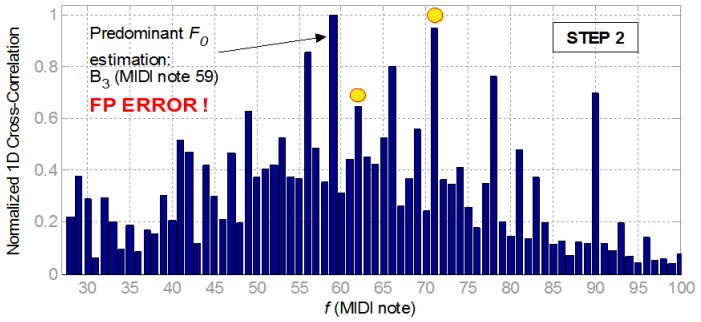
In general, the bispectral representation cannot help to resolve the underlying components of interfering partials; while it is the mechanism of extraction of the 2-D monophonic pattern of G_2 in the proposed bispectrum-based algorithm which allows keeping critical information about the peak positions of the other sounds harmonic 2-D patterns, which are:

$$(f_{02}, 2f_{02}), (2f_{02}, 2f_{02}), (f_{03}, 2f_{03}).$$

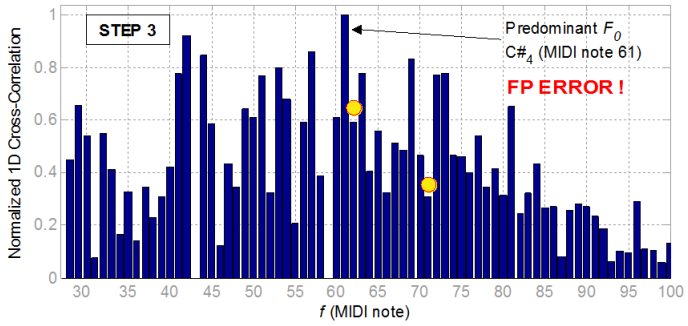
In conclusion, the system performing the iterative 2-D pattern matching and pattern extraction in the bispectrum domain successfully identifies all the three notes played in the audio source file. The system performing the iterative 1-D pattern matching and direct cancelation of the pattern in the spectrum domain identifies only the lowest note, G_2, and commits two false positive errors, due to the removal of partials of the higher sounds in the direct cancelation procedure.



(a)



(b)



(c)

Figure 3.12: Step by step multi-$F0$ estimation procedure with iterative spectral 1-D pattern matching and direct cancelation technique. The dots identify the notes played in the audio source signal.).

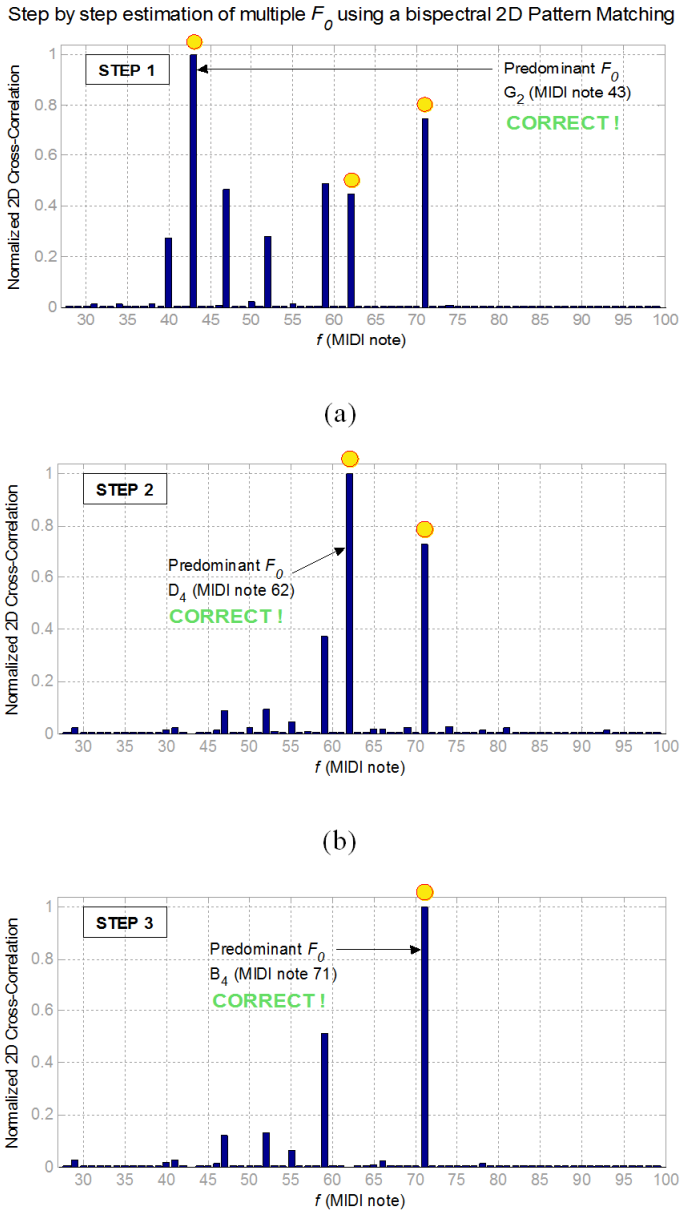


Figure 3.13: Step by step multi-$F0$ estimation procedure with iterative bispectral 2-D pattern matching and pattern extraction technique. The dots identify the notes played in the audio source signal.

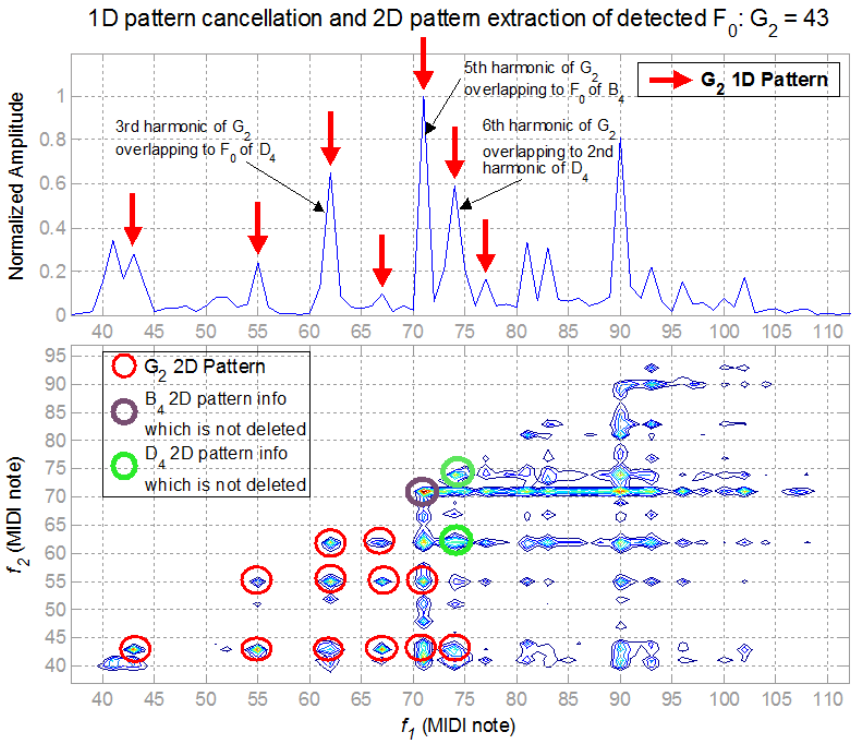


Figure 3.14: Graphical comparison between direct cancelation of 1-D pattern from the spectrum (above) and extraction of 2-D pattern from the bispectrum (below).

Chapter 4

System Architecture

In this section, a detailed description of the proposed method for music transcription, presented in [ANP11b] is given. First a general overview is given, then the main modules are discussed in detail.

4.1 General Architecture

A general view of the system architecture is presented in Figure 4.1. In the diagram, the main modules are depicted (with dashed line) as well as the blocks composing each module.

The transcriptor accepts as input a PCM Wave audio file (mono or stereo) as well as user-defined parameters related to the different procedures. The ***Pre-Processing*** module carries out the implementation of the constant-Q analysis by means of the *Octave Filter Bank* block. Then, the processed signal enters both the ***Pitch Estimation*** and ***Time Events Estimation*** modules. The ***Pitch Estimation*** module computes the bispectrum of its input, perform the 2-D correlation between the bispectrum and a harmonic-related pattern, and estimate candidate pitch values. The ***Time Events Estimation*** module is devoted to the estimation of onsets and durations of notes. The ***Post-Processing*** module discriminates notes from very short-duration events, seen as disturbances, and produces the output files: a SMF0 MIDI file (which is the transcription of the audio source) and a list of pitches, onset times and durations of all detected notes.

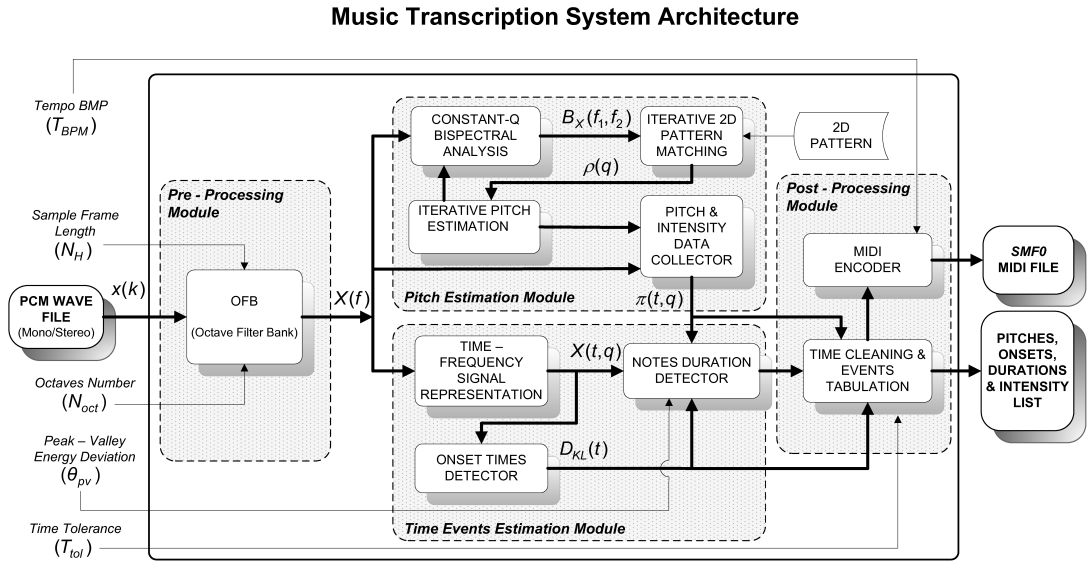


Figure 4.1: Music transcription system block architecture. The functional modules, inner blocks, input parameters and output variables and functions are illustrated.

4.2 The Pre-Processing module

The input PCM signal $x(k)$ enters the Pre-Processing module, where it is converted into a numerical array and segmented into fixed length frames by the *Signal Gather* block. The length of the time-domain analysis window W_s, in number of samples, is specified by the user. The *Signal Gather* block also calculates the energy of each frame as follows:

$$E_m = \sqrt{\sum_{i=0}^{W_s-1} x^2\big(mW_s + i\big)}, \qquad m \in [0, \ldots, M-1],$$

where $m = 0, 1, \ldots, M-1$, is the index over frames and M is the total number of frames in which the audio file is divided, given by $M = \lceil T_{tot}/(W_s/f_s) \rceil$, denoting with T_{tot} the total time length of the audio file expressed in seconds. In order to discriminate notes from silence or ground noise, E_m is compared against an energy threshold, which is set to the first frame energy, where absence of a signal is assumed. Only frames with energy higher than the threshold are passed to the *Octave Filter Bank* block.

The *Octave Filter Bank* (OFB) block performs the constant-Q analysis over a set of octaves whose number N_{oct} is provided by the user. The block produces the spectrum samples - computed by using the Fourier transform - relative to the nominal frequencies of the notes to be detected in each octave. In order to minimize detection errors due to partial inharmonicity or instrument intonation inaccuracies, two additional frequencies aside each nominal value have been considered as well. The distance between the additional and the fundamental frequencies is $\pm 2\%$ of each nominal pitch value, which is less than half a semitone spacing (assumed as approximately $\pm 3\%$); the maximum amplitude among the three spectral lines is associated with the nominal pitch frequency value. Hence, the number of spectrum samples that is passed to the successive blocks for further processing is $N_p = 12\ N_{oct}$, where 12 is the number of pitches per octave.

As an example, consider that the OFB accepts an input signal sampled at f_s=44100 Hz and consider that ideal filters, with null transition bandwidth, are used. The outputs of the first three stages of the OFB tree cover the ranges $(0, 22050)$Hz, $(0, 11025)$Hz, and $(0, 5512.5)$Hz. The spectrum analysis works only on the higher-half frequency interval of each band, whereas the lower-half frequency interval is to be analyzed in the subsequent stages. Hence, with the given sampling frequency, in the first three stages the octaves from F_9 to E_{10},

from F_8 to E_9, and from F_7 to E_8, in that order, are analyzed. In general, in the ith stage, the interval from $F_{N_{oct}+1-i}$ to $E_{N_{oct}+2-i}$, $i = 1, 2, \ldots, N_{oct}$, is analyzed.

In the case of non-ideal filters, the presence of a non-null transition band must be taken into account. Consider the branches of the building block of the OFB tree, shown in Figure 3.2-(b), the first leading to the spectral analysis sub-block, the second to filtering and downsampling sub-block. Notes, whose nominal frequency falls into the transition band of the filter, can not be resolved after downsampling and must be analyzed in the first (undecimated) branch. Useful low-pass filters are designed by choosing, in normalized frequencies, the interval $(0, \gamma\,\pi)$ as the passband, the interval $(\gamma\,\pi, \pi/2)$ as the transition band, and the interval $(\pi/2, \pi)$ as the stopband; the parameter γ ($\gamma < 0.5$) controls the transition bandwidth.

Hence, the frequency interval that must be considered into the *spectrum analysis* step at the first stage is $(\gamma f_s/2, f_s/2)$. In the second stage, the analyzed interval is $(\gamma f_s/4, \gamma f_s/2)$, and, in general, if we define $f_s^{(i)} = f_s/2^{(i-1)}$ as the sampling frequency of the input of the ith stage, the frequency interval considered by the spectrum analyzer block is (apart from the first stage) $(\gamma f_s^{(i)}/2, \gamma f_s^{(i)})$. The filter mask $H(\omega)$ and the analyzed regions are depicted in Figure 4.2.

Table 4.1 summarizes the system parameters we used to implement the OFB. With the chosen transition band, the interval from E_9 to E_{10} is analyzed in the first stage, and the interval from $E_{N_{oct}+1-i}$ to $D\sharp_{N_{oct}+2-i}$, $i = 2, \ldots, N_{oct}$, is analyzed in the ith stage. At the end of the whole process, a spectral representation from E_1 (at 41.203 Hz) to E_{10} (at 21.096 kHz), sufficient to cover the extension of almost every musical instrument, is obtained.

4.3 Pitch Estimation Module

The *Pitch Estimation* module receives as input the spectral information produced by the *Octave Filter Bank* block. This module includes the *Constant-Q Bispectral Analysis*, the *Iterative 2-D Pattern Matching*, the *Iterative Pitch Estimation* and the *Pitch & Intensity Data Collector* blocks. The first block computes the bispectrum of the input signal at the frequencies of interest. The *Iterative 2-D Pattern Matching* block is in charge of computing the 2-D correlation between the bispectral array and a fixed, bi-dimensional harmonic pattern. The objective of the *Iterative Pitch Estimation* block is detecting the

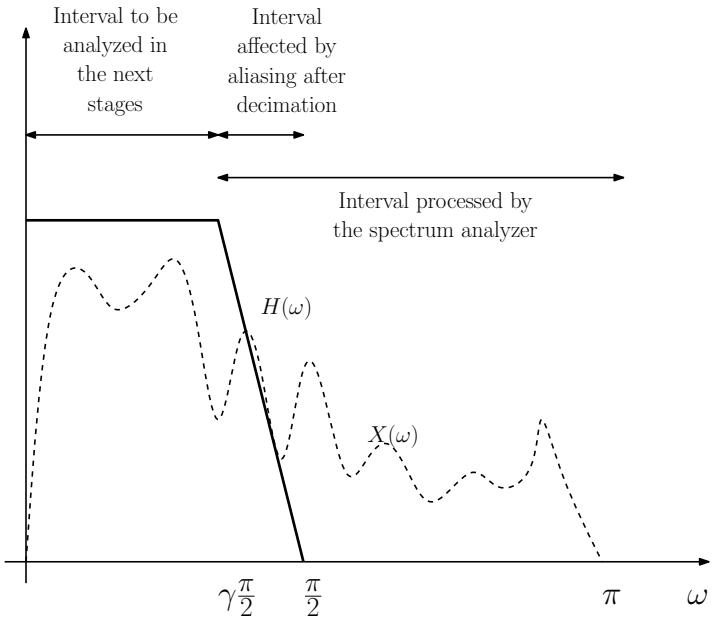


Figure 4.2: Filter mask and the analyzed regions.

Table 4.1: OFB specifications

Sampling frequency (f_s)	44.1 kHz
Number of octaves (N_{oct})	9
Frequency range	[40 Hz , 20 kHz]
Hann's window length (N_H)	256 samples
FIR passband	$(0, 0.46\ \pi)$
FIR stopband	$(\pi/2, \pi)$
FIR ripples ($\delta_1 = \delta_2$)	10^{-3}
Filter length	187 samples

presence of the pitches, and subsequently extracting the 2-D harmonic pattern of detected notes from the bispectrum of the actual signal frame. Finally, the *Pitch & Intensity Data Collector* block associates energy information to corresponding pitch values in order to collect the intensity information.

4.3.1 Harmonic pattern correlation

Consider a 2-D harmonic pattern as dictated by the distribution of the bispectral local maxima of a monophonic musical signal expressed in semitone intervals. The chosen pattern, shown in Figure 4.3, has been validated and refined by studying the actual bispectrum computed on several real monophonic audio signals. The pattern is a sparse matrix with all non-zero values (denoted as dark dots) set to one. The *Iterative 2-D Pattern Matching* block computes

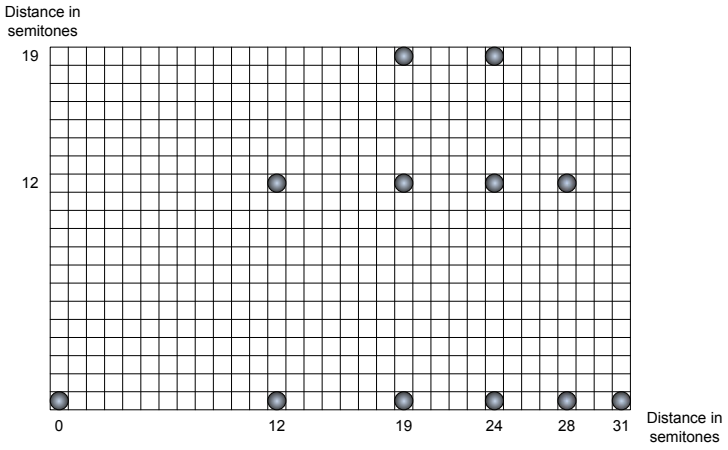


Figure 4.3: Fixed 2-D harmonic pattern used in the validation tests of the proposed music transcriptor. It represents the theoretical set of bispectral local maxima for a monophonic 7-partials sound all weights are set equal to unity.

the similarity between the actual bispectrum (produced by the *Constant-Q Bispectral Analysis* by using the spectrum samples given by the *Octave Filter Bank* block) of the analyzed signal and the chosen 2-D harmonic pattern. Since only $12N_{oct}$ spectrum samples (at the fundamental frequencies of each note) are of interest, the bispectrum results to be a $12N_{oct} \times 12N_{oct}$ array. The

cross-correlation between the bispectrum and the pattern is given by:

$$\rho(k_1, k_2) = \sum_{m_1=0}^{C_P-1} \sum_{m_2=0}^{R_P-1} P(m_1, m_2) \left| B_x(k_1 + m_1, k_2 + m_2) \right|, \qquad (4.1)$$

where $1 \leq k_1, k_2 \leq 12N_{oct}$ are the frequency indexes (spaced by semitone intervals), and P denotes the sparse $R_P \times C_P$ 2-D harmonic pattern array. The ρ coefficient is assumed to take a maximum value when the template array P exactly matches the distribution of the peaks of the played notes. If a monophonic sound has a fundamental frequency corresponding to index q, then the maximum of $\rho(k_1, k_2)$ is expected to be positioned at (q, q), upon the first quadrant bisector. For this reason, $\rho(k_1, k_2)$ is computed only for $k_1 = k_2 = q$ and denoted in the following as $\rho(q)$. The 2-D cross-correlation computed in this way is far less noisy than the 1-D cross-correlation calculated on the spectrum (as illustrated in the example in Appendix B). Finally, the ρ array is normalized to the maximum value over each temporal frame.

The *Iterative 2-D Pattern Matching* block output is used by the *Iterative Pitch Estimation* block, whose task is ascertaining the presence of multiple pitches in an audio signal.

4.3.2 Pitch Detection

(4a) - Recall on Spectrum Domain. Several methods based on pattern matching in the spectrum domain were proposed for multiple-pitch estimation [Kla03, Kla05, NAW01, BLW07]. In these methods, an iterative approach is used. First, a single $F0$ is estimated by using different criteria (e.g., maximum amplitude, or lowest peak-frequency); then, the set of harmonics related to the estimated pitch is directly canceled from the spectrum and the residual is further analyzed until its energy is less than a given threshold. In order not to excessively degrade the original information, a partial cancelation (subtraction) can be performed based on perceptual criteria, spectral smoothness, etc. The performance of direct/partial cancelation techniques, on the spectrum domain, significantly degrades when the number of simultaneous voices increases.

(4b) - Proposed Method. The method proposed in the present work and described also in [ANP11b]uses an *iterative procedure for multiple $F0$ estimation based on successive 2-D pattern extraction in the bispectrum domain.* Consider two concurrent sounds, with fundamental frequencies F_l and F_h ($F_l < F_h$), such that $F_h : F_l = m : n$. Let $F_{ov} = nF_h = mF_l$ be the frequency value of

the first overlapping partial. Consider now the bispectrum generated by the mixture of the two notes (as an example, see Figure 3.5). A set of peaks is located at the same abscissa F_{ov}, that is at the co-ordinates $(F_{ov}, k_l F_l)$ and $(F_{ov}, k_h F_h)$, where $k_l = 1, 2, \ldots, m-1$, $k_h = 1, 2, \ldots, n-1$. Hence, the peaks have the same abscissa but are separated along the y-axis. If, for example, F_l is detected as the first $F0$ candidate, extracting its 2-D pattern from the bispectrum does not completely eliminate the information carried by the harmonic F_{ov} related to F_h, that is the peaks at $(F_{ov}, k_h F_h)$ are not removed. On the contrary, if F_h is detected as the first $F0$ candidate, in a similar way the peaks at $(F_{ov}, k_l F_l)$ are not removed. This is strongly different than in methods based on direct harmonic cancelation in the spectrum, where the cancelation of the 1-D harmonic pattern, after the detection of a note, implies a complete loss of information about the overlapping harmonics of concurrent notes.

The proposed procedure can be summarized as follows:

1. Compute the 2-D correlation $\rho(q)$ between the bispectrum and the chosen template, only upon the first quadrant bisector:

$$\rho(q) = \sum_{m_1=0}^{C_P-1} \sum_{m_2=0}^{R_P-1} P(m_1, m_2) \left| B_x(q+m_1, q+m_2) \right|, \tag{4.2}$$

   derived directly from Equation (4.1);

2. Select the frequency value q_0 yielding the highest peak of $\rho(q)$ as the index of a candidate $F0$;

3. Cancel the entries of the bispectrum array that correspond to the harmonic pattern having q_0 as fundamental frequency;

4. Repeat steps 1-3 until the energy of the residual bispectrum is higher than $\theta_E E_B$, where θ_E, $0 < \theta_E < 1$ is a given threshold and E_B is the initial bispectrum energy.

Once multiple $F0$ candidates have been detected, the corresponding energy values in the signal spectrum are taken by the *Pitch & Intensity Data Collector* block, in order to collect also the intensity information. The output of this block is the array $\pi(t, q)$, computed over the whole musical signal, where q is the pitch index and t is the discrete time variable over the frames: $\pi(t, q)$ contains either zero values (denoting the absence of a note) or the energy of the detected note. This array is used later in the *Time Events Estimation* module

to estimate note durations, as explained in the next section. In Appendix B, an example of multiple $F0$ estimation procedure, carried out by using the proposed method is illustrated step by step. Results are compared with those obtained by a transcription method performing a 1-D direct cancelation of the harmonic pattern in the spectrum domain. The test file is a real audio signal, taken from RWC music database [GHNO02], analyzed in a single frame.

In conclusion, the component of the spectrum at the frequency F_{ov} is due to the combination of two harmonics related to the notes F_l and F_h. According to eq. (3.3), the spectrum amplitude at F_{ov} affects all the peaks in the bispectrum located at $(F_{ov}, k_l F_l)$ and $(F_{ov}, k_h F_h)$. Interference of the two notes occurring at these peaks is not resolved; nevertheless, we deem that the geometry of the bispectral local maxima is a relevant information that is an added value of the bispectral analysis with respect to the spectral analysis, as experimental results confirm.

4.3.3 Time Events Estimation

The aim of this module is the estimation of the temporal parameters of a note, i.e., *onset* and *duration* times. The module is composed of three blocks, namely the *Time-Frequency Representation* block, the *Onset Times Detector* block, and the *Notes Duration Detector* block.

The *Time-Frequency Representation* block collects the spectral information $X(f)$ of each frame, used also to compute the bispectrum, in order to represent the signal in the time-frequency domain. The output of this block is the array $X(t, q)$, where t is the index over the frames, and q is the index over pitches, $1 \leq q \leq 12N_{oct}$.

The *Onset Times Detector* block uses the variable $X(t, q)$ to detect the onset time of the estimated notes, which is related to the *attack* stage of a sound. Mechanical instruments produce sounds with rapid volume variations over time. Four different phases have been defined to describe the envelope of a sound, that is *Attack*, *Decay*, *Sustain* and *Release* (*ADSR envelope* model). The ADSR envelope can be extracted in the time domain - without using spectral information - for monophonic audio signals, whereas this approach results less efficient in a polyphonic context. Several techniques [BN05], [Moo78], [Dol01] have been proposed for onset detection in the time-frequency domain. The methods based on the phase-vocoder functions [Moo78], [Dol01] try to detect rapid spectral-energy variations over time: this goal can be achieved either

by simply calculating the amplitude difference between consecutive frames of the signal spectrogram or by applying more sophisticated functions. The proposed solution uses the *Modified Kullback-Liebler Divergence* function, which achieved the best performance in [Bro06]. This function aims at evaluating the distance between two consecutive spectral vectors, highlighting large positive energy variations and inhibiting small ones. The modified Kullback-Liebler divergence $D_{KL}(t)$ is defined by:

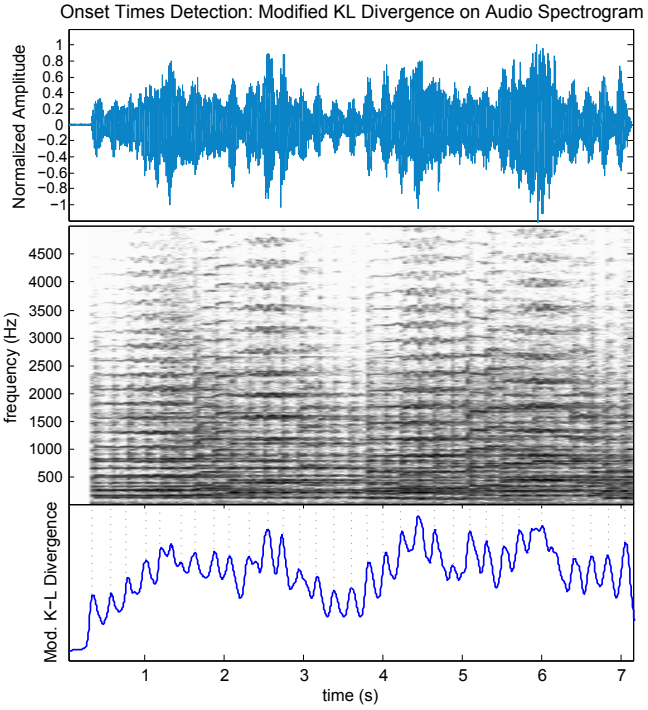
$$D_{KL}(t) = \sum_{q=1}^{12N_{oct}} \log\left(1 + \frac{|X(t,q)|}{|X(t-1,q)| + \varepsilon}\right),$$

where $t \in [2, \ldots, M]$, with M the total number of frames of the signal; ε is a constant, typically $\varepsilon \in [10^{-6}, 10^{-3}]$, which is introduced to avoid large variations when very low energy levels are encountered, thus preventing $D_{KL}(t)$ to diverge in proximity of the release stage of sounds. $D_{KL}(t)$ is an $(M-1)$-element array, whose local maxima are associated with the detected onset times. Some example plots of $D_{KL}(t)$ are shown in Figure 4.4.

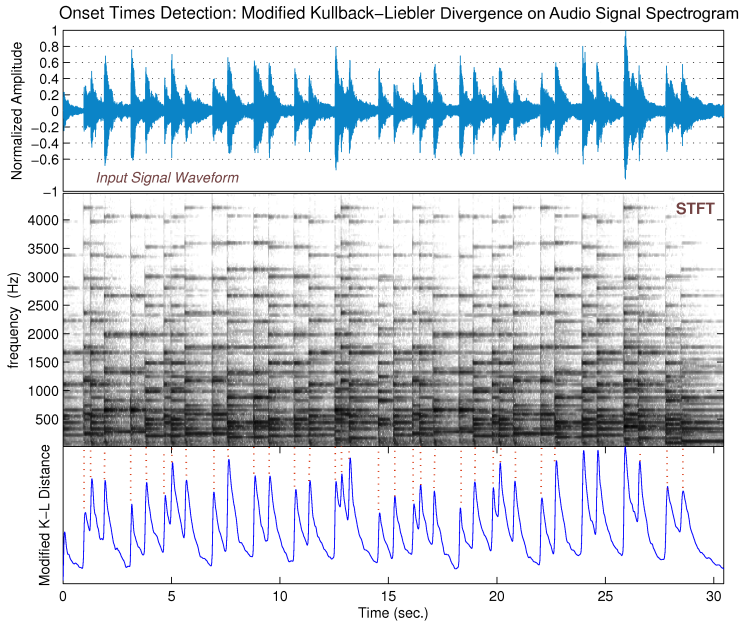
The *Notes Duration Detector* block carries out the estimation of notes duration. The beginning of a note relies on the $D_{KL}(t)$ onset locations. The end of a note is assumed to coincide with the release phase of the ADSR model and is based on the time-frequency representation. A combination of the information coming from both the functions $X(t,q)$ and $\pi(t,q)$ (the latter computed in the *Pitch Estimation* module, see 4.3.2) is used, as described below. The rationale for using this approach stems from the observation of the experimental results: $\pi(t,q)$ supplies a robust but time-discontinuous representation of the detected notes, whereas $X(t,q)$ contains more robust information about notes duration. The algorithm is the following:

For each $\bar{q}$ such that $\exists \pi(t,\bar{q}) \neq 0$ for some t, do:

1. Execute a smoothing (simple averaging) of array $X(t,\bar{q})$ along the t-axis;
2. Identify the local maxima (peaks) and minima (valley) of the smoothed $X(t,\bar{q})$;
3. Select from consecutive peak-valley points the couples whose amplitude difference exceed a given threshold θ_{pv};
4. Let (V_1, P_1) and (P_2, V_2) be two consecutive valley-peak and peak-valley couples that satisfy the previous criterion: the extremals (V_1, V_2) identify a "possible note" event;



(a)



(b)

Figure 4.4: Example of onset detection procedure: (a) 7 seconds extracted from Mozart's *String Quartet n. 19, K465*; (b) first 30 seconds of Mozart's *Sonata for piano K331*.

5. For each "possible note" event, do:

   (a) Estimate $(\bar{V}_1, \bar{V}_2) \subset (V_1, V_2)$ such that $(\bar{V}_1, \bar{V}_2)$ contains a given percentage of the energy in (V_1, V_2);

   (b) Set the onset time ON_T of the note equal to the maximum of the $D_{KL}(t)$ array nearest to $\bar{V}_1$;

   (c) Set the offset time OFF_T of the note equal to $\bar{V}_2$;

   (d) If $\pi(t, \bar{q})$, with $t \in (ON_T, OFF_T)$ contains non-zero entries, then a note at the pitch value $\bar{q}$, beginning at ON_T and with duration OFF_T - ON_T is detected.

4.4 System Output Data

The ***Post-Processing*** module tasks are the following. First, a cleaning operation in the time-domain is made in order to delete events having a duration shorter than a user defined time tolerance parameter T_{TOL}. Then, all the information concerning the estimated note is tabulated into an output list file. These data are eventually sent to a *MIDI Encoder* (taken from the Matlab® MIDI Toolbox in [ET04]), which generates the output MIDI SMF0 file, provided that the user defines a tempo value T_{BPM}, expressed in beats per minute.

Chapter 5

Experimental Results and Validation

In this section, the experimental tests that have been set up to assess the performances of the proposed method are described. First, the evaluation parameters are defined. Then, some results obtained by using excerpts from the standard RWC-C database are shown, in order to highlight the advantages of the bispectrum approach with respect to spectrum methods based on direct pattern cancelation. Finally, the results of the comparison of the proposed method with others participating at the MIREX 2009 contest are presented.

5.1 Evaluation parameters

In order to assess the performances of the proposed method, the evaluation criteria that have been proposed in MIREX 2009, specifically those related to the multiple $F0$ estimation (frame level and $F0$ tracking), were chosen.

The evaluation parameters are the following [PE07]:

- *Precision*: the ratio of correctly transcribed pitches to all transcribed pitches for each frame, i.e.,

$$\text{Prec} = \frac{TP}{TP + FP},$$

  where TP is the number of the true positives (correctly transcribed voiced frames) and FP is the number of false positives (unvoiced note-frames transcribed as voiced).

- *Recall*: the ratio of correctly transcribed pitches to all ground truth reference pitches for each frame, i.e.,

$$\text{Rec} = \frac{TP}{TP + FN},$$

where FN is the number of false negatives (voiced note-frames transcribed as unvoiced).

- *Accuracy*: an overall measure of the transcription system performance, given by

$$\text{Acc} = \frac{TP}{TP + FN + FP}.$$

- *F-measure*: a measure yielding information about the balance between FP and FN, that is

$$\text{F-measure} = 2 \times \frac{\text{Prec} \times \text{Rec}}{\text{Prec} + \text{Rec}}.$$

5.2 Validation of the proposed method

5.2.1 Experimental data set: RWC database

The performances of the proposed transcription system have been evaluated by testing it on some audio fragments taken from the standard RWC - Classical Music database. The sample frequency is 44.1 kHz and a frame length of 256 samples (which is approximately 5.8 ms) have been chosen.

For each audio file, segments containing one or more complete musical phrases have been taken, so that the excerpts have different time lengths. In Table 5.1, the main features of the used test audio files are reported. The set includes about 100000 one-frame-long voiced events.

The musical pieces were selected with the aim of creating an heterogeneous dataset: the list includes piano solo, piano plus soloist, strings quartet and strings plus soloist recordings. Several metronomic tempo values were chosen.

The proposed transcription system has been realized and tested in Matlab® environment installed on a dual core 64-bit processor 2.6 GHz with 3 GB of RAM. With this equipment, the system performs the transcription in a period which is approximately fifteen times the input audio file duration.

# Data	Author	Title	Catalog Number RWC-MDB	Instruments
(1)	J.S. Bach	Ricercare a 6, *BWV 1079*	C-2001 n. 12	2 Vns, Vc
(2)	W. A. Mozart	String Quartet n. 19, *K 465*	C-2001 n. 13	Vn, Vla, Vc, Cb
(3)	J. Brahms	Clarinet Quintet, *op. 115*	C-2001 n. 17	Cl, Vla, Vc
(4)	M. Ravel	Ma Mère l'Oye, Petit Poucet	C-2001 n. 23B	Piano
(5)	W. A. Mozart	Sonata *K 331*, 1st mov.	C-2001 n. 26	Piano
(6)	C. Saint - Saëns	Le Cygne	C-2001- n. 42	Piano and Violin
(7)	G. Fauré	Sicilienne, *op. 78*	C-2001 n. 43	Piano and Flute

Table 5.1: Test data set from RWC - Classical database. Vn(s): Violin(s); Vla: Viola; Vc: Cello; Cb: Contrabass; Cl: Clarinet

5.2.2 Comparison of bispectrum and spectrum based approaches

In this section, the performances of bispectrum and spectrum based methods for multiple $F0$ estimation are compared. The comparison is made on a frame-by-frame basis, that is every frame of the transcribed output is matched with every corresponding frame of the ground truth reference of each audio sample, and the mismatches are counted.

The proposed bispectrum based algorithm, referred to as BISP in the following, has been described in Section 4.3. A spectrum-based method, referred to as SP1 in the following, is obtained in a way similar to the proposed method by making the following changes: 1) the bispectrum front-end is substituted by a spectrum front-end; 2) the 2-D correlation in the bispectrum domain, using the 2-D pattern in Figure 4.3, is substituted by a 1-D correlation in the spectrum domain, using the 1-D pattern in Figure 1.1. Both bispectrum and spectrum based algorithms are iterative and perform subsequent 2-D harmonic pattern extraction and 1-D direct pattern cancelation, after an $F0$ has been detected. The same pre-processing (constant-Q analysis), onset and duration, and post-processing modules have been used for both algorithms. A second spectrum-based method, referred to as SP2 in the following, in which $F0$ estimation is performed by simply thresholding the 1-D correlation output without direct cancelation, has been also considered.

The frame-by-frame evaluation method requires a careful alignment between the ground truth reference and the input audio. The ground truth reference data have been obtained from the MIDI files associated to each audio sample. The RWC-C database reference MIDI files, even though quite

faithful, do not supply an exact time correspondence with the real audio executions. Hence, time alignment between MIDI files and the signal spectrogram has been carefully checked. An example of the results of the MIDI-spectrogram alignment process is illustrated in Figure 5.1.

The performances of algorithms BISP, SP1 and SP2 applied to the audio data set described in section 5.2.1 are shown in Tables 5.2, 5.3 and 5.4. The Tables show the overall accuracy and the F-measure evaluation metrics, as well as the TP, FP and FN for each audio sample. A comparison of the results is presented in Figure 5.2, and a graphical comparison between the output of BISP and SP1 is shown in Figure 5.4. In Figure 5.3, a graphical view of the matching between the ground truth reference and the system piano-roll output representations is illustrated. The results show that the proposed BISP algorithm outperforms spectrum based methods. BISP shows an overall accuracy of 57.6%, and an F-measure of 72.1%. Since pitch detection is performed in the same way, such results highlight the advantages of the bispectrum representation with respect to spectrum one. The results are encouraging considering also the complex polyphony and the multi-instrumental environment of the test audio fragments.

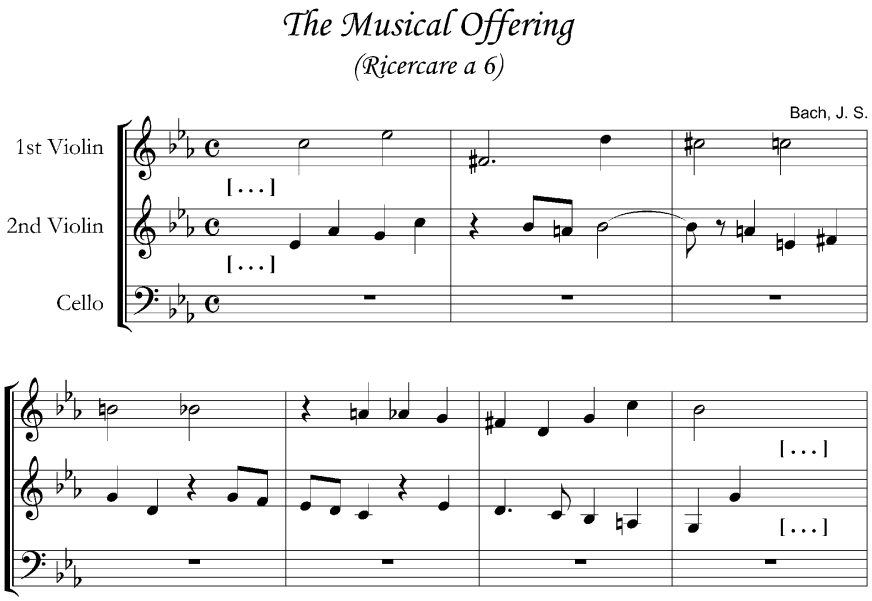
The comparison with other automatic transcription methods is described in the next section, where the results of the MIREX 2009 evaluation framework are reported.

# Data	Reference events	TP	FP	FN	Accuracy%	F-measure%
(1)	16063	11025	2482	5038	59.4	74.6
(2)	6584	4401	2158	2223	50.1	66.8
(3)	12652	8865	2079	3787	60.2	75.1
(4)	12424	10663	2655	1761	70.8	82.8
(5)	6054	4120	1294	1934	56.1	71.8
(6)	20032	15122	6746	4910	56.5	72.2
(7)	21653	16563	9933	5090	52.4	68.8
TOTAL	**95412**	**70759**	**27347**	**24743**	**57.6%**	**72.1%**

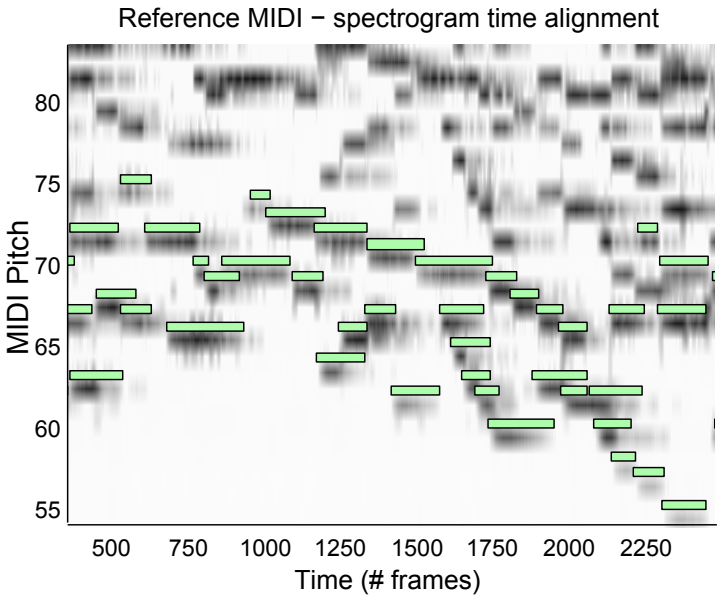
Table 5.2: BISP: transcription results obtained with the test data set listed in Table 5.1.

5.2.3 Results from MIREX 2009

The Music Information Retrieval Evaluation eXchange (MIREX) is the community-based framework for the formal evaluation of Music Information Retrieval



(a)



(b)

Figure 5.1: Graphical view of the alignment between reference MIDI file data (represented as rectangular objects) and the spectrogram of the corresponding PCM Wave audio file (b). The detail shown here is taken from a fragment of Bach's *Ricercare a 6, The Musical Offering, BWV 1079* (a), which belongs to the test data set.

# Data	Reference events	TP	FP	FN	Accuracy%	F-measure%
(1)	16063	10348	6327	5715	46.4	63.2
(2)	6584	3216	2021	3318	38.0	54.6
(3)	12652	6026	8187	6626	29.0	44.9
(4)	12424	10363	3920	2061	63.8	77.6
(5)	6054	4412	4542	1642	42.0	58.8
(6)	20032	9952	7558	10080	36.2	53.0
(7)	21653	11727	9813	9926	37.4	54.3
TOTAL	**95412**	**56044**	**42368**	**39368**	**40.7%**	**57.8%**

Table 5.3: SP1: transcription results obtained with the test data set listed in Table 5.1.

# Data	Reference events	TP	FP	FN	Accuracy%	F-measure%
(1)	16063	10234	7857	5829	42.8	59.9
(2)	6584	2765	2243	3769	31.5	47.9
(3)	12652	6206	9590	6446	27.9	43.6
(4)	12424	9471	3469	2953	59.6	74.7
(5)	6054	3642	3844	2412	36.8	53.8
(6)	20032	7769	6692	12263	29.1	45.0
(7)	21653	10399	8023	11254	35.0	51.9
TOTAL	**95412**	**50486**	**41718**	**44926**	**36.8%**	**53.8%**

Table 5.4: SP2: transcription results obtained with the test data set listed in Table 5.1.

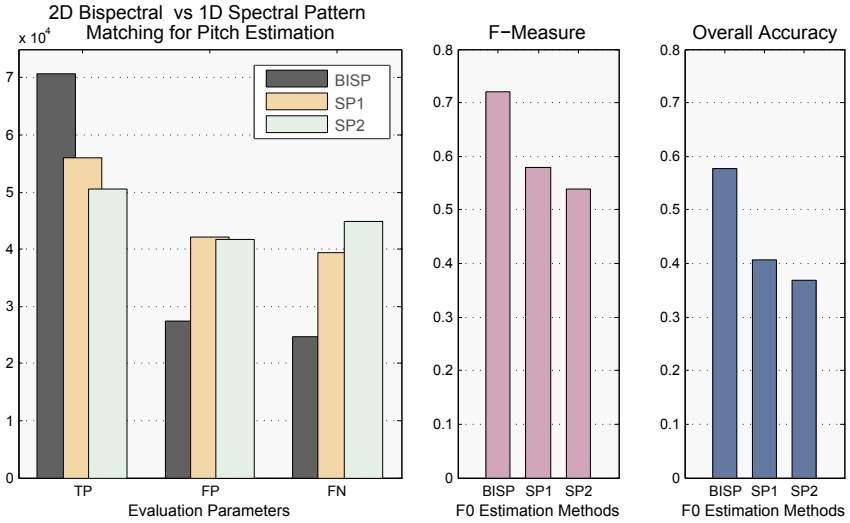
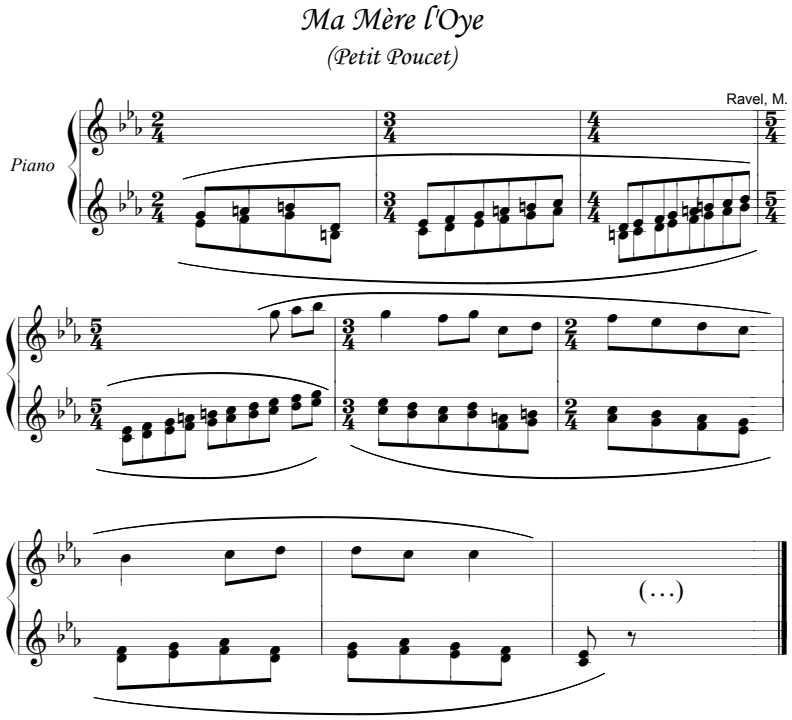
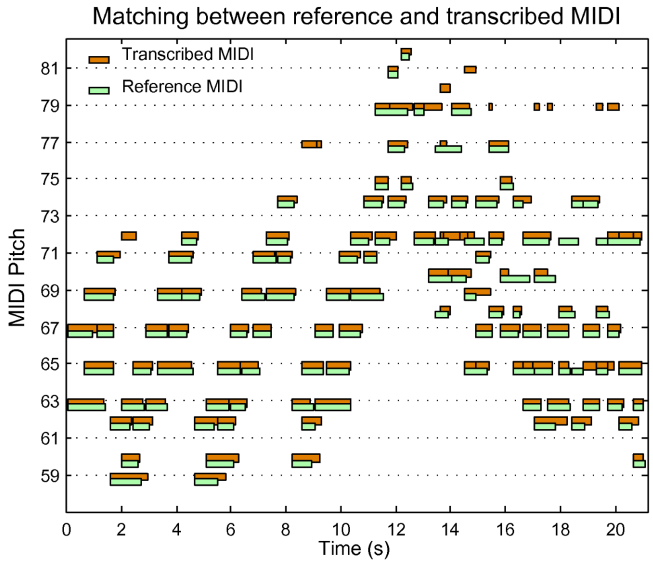


Figure 5.2: Results of comparison between bispectrum based (BISP) and spectrum based (SP1 and SP2) multi-$F0$ estimation methods. SP1 performs iterative pitch estimation and harmonic pattern subtraction; SP2 performs simple thresholding of cross-correlation measure.

(MIR) systems and algorithms [Dow08]. In 2009, MIREX has reached its fifth running. The proposed BISP method has been submitted for an evaluation and a comparison with the other participants in the field of *Multiple Fundamental Frequency Estimation & Tracking*, which is divided into the following tasks: 1) Multiple Fundamental Frequency Estimation (MF0E); 2A) Mixed Set Note Tracking (NT); and 2B) Piano Only Note Tracking. Task 1 is a frame level evaluation (similar to that described in section 5.2.2) of the submitted methods. Task 2 considers as events to be detected notes characterized by pitches, onset and offset times. For a specific definition of tasks and evaluation criteria, the reader should refer to [MIRb]. Two different versions of the proposed system have been submitted to MIREX: they are referred to as *NPA1* and *NPA2* as team-ID. The differences between the two versions regard mainly the use of the *Time Events Estimation* module: *NPA1* simply performs a multiple-$F0$ estimation without onset and duration times detection, whereas *NPA2* uses the procedures described in Section 4.3.3. As a result, *NPA2* has reported better results than *NPA1* in all the three tasks considered. A detailed overview of the overall performance results is available at [MIRa], see section



(a)



(b)

Figure 5.3: Graphical (piano-roll) view of event matching between the ground truth reference and transcribed MIDI (b), related to Ravel's *Ma Mère l'Oye - Petit Poucet* (a), present in the test data set.

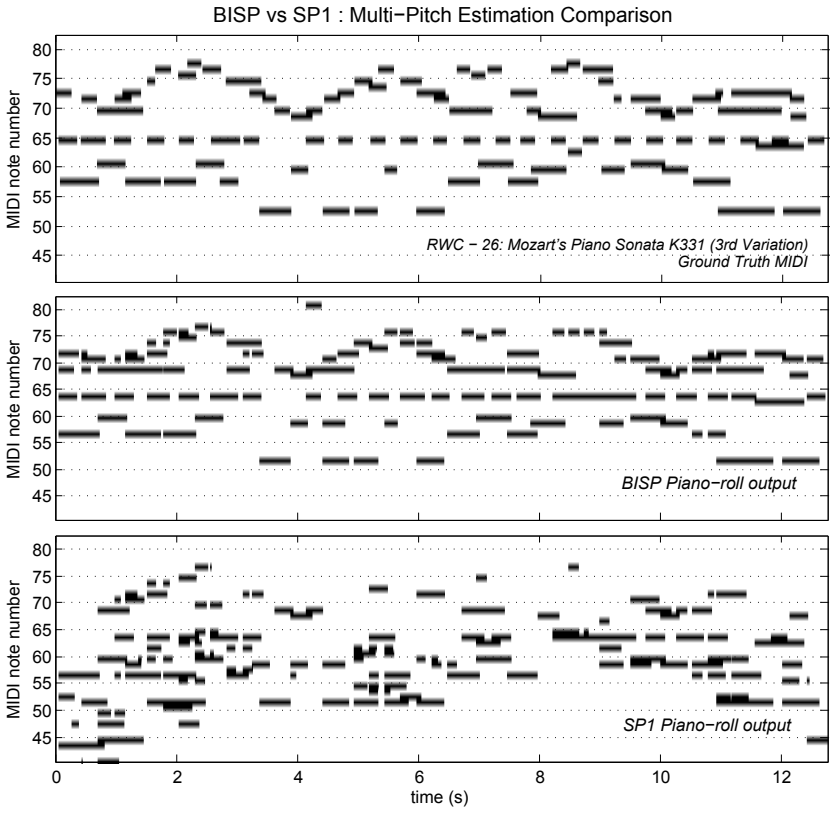
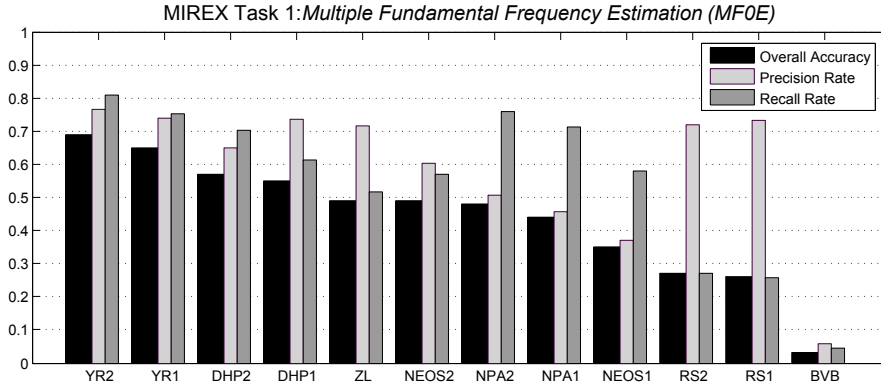


Figure 5.4: Graphical comparison between piano-roll output of BISP and SP1, and the reference ground truth data. The test audio example is a fragment of the 3*rd* variation of Mozart's Piano Sonata K 331.

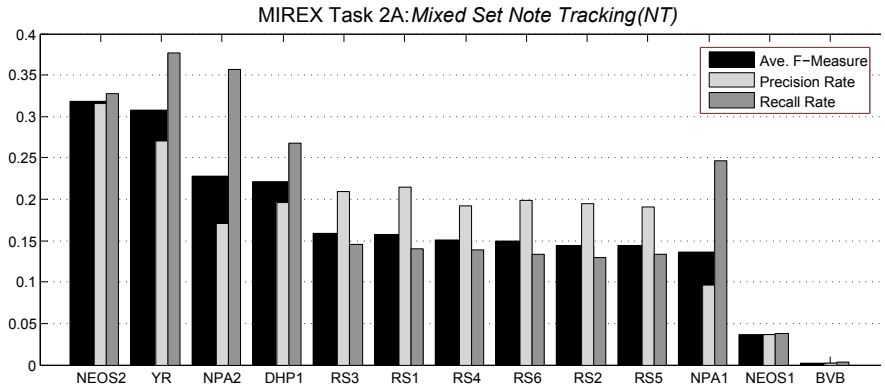
Multiple Fundamental Frequency Estimation and Tracking Results.

For Task 1 (MF0E), accuracy has been chosen as a key performance indicator. The proposed system *NPA2* is mid-level ranked, with an accuracy of 48%; anyway, it presents the second highest recall rate (76%); this demonstrates that the proposed system has a good capability in detecting ground truth reference notes, showing a tendency in detecting more false positives than false negatives. For Task 2A (Mixed Set NT) and Task 2B (Piano Only NT), F-measure has been chosen as the overall performance indicator. In Task 2A, the proposed system *NPA2* has achieved the third highest F-measure rate and the second highest recall rate; again the precision rate show a quite high false positive detection rate. In Task 2B, the proposed system *NPA2* is top-ranked, outperforming all the other competitors' systems.

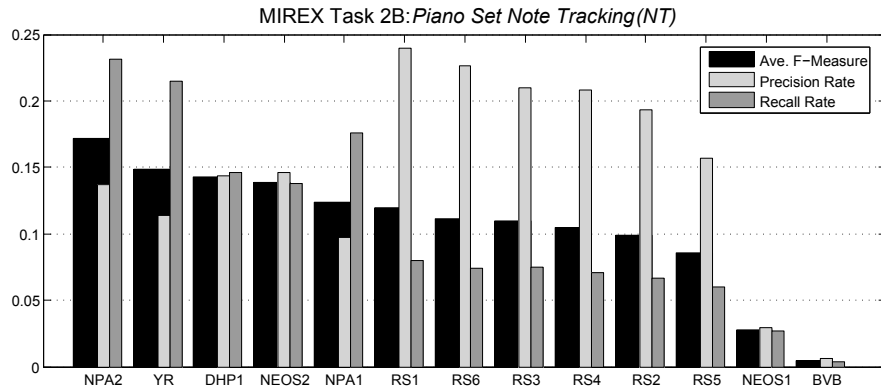
Results of MIREX 2009 are summarized in Figure 5.5.



(a)



(b)



(c)

Figure 5.5: Results of MIREX 2009 evaluation framework. The system proposed in this work has been submitted in two different versions, referred to as *NPA1* and *NPA2*, from the name of the authors; (a) task 1: *multi-F*0 *estimation*; (b) task 2A: *Mixed-set note tracking (NT)*; (c) task 2B: *Piano-only note tracking (NT)*.

Chapter 6

Conclusions and Future Work

The automatic music transcription system described in this Thesis implements a novel front-end, presented in [ANP11b], obtained by a constant-Q bispectral analysis of the input audio signal, which offers advantages with respect to lower dimensional spectral analysis in polyphonic pitch estimation. In every frame, pitch estimation is performed by means of a 2-D correlation between signal bispectrum and a fixed bi-dimensional harmonic pattern, while information about intensity of detected pitches is taken directly from the magnitude spectrum. Onset times are detected by a procedure that highlights large energy variations between consecutive frames of the time-frequency signal representation. Such a representation is also the basis for note durations estimation: a pitch against time representation of detected notes is compared with the audio spectrogram; the duration of each detected note event in the former is adjusted to the duration of corresponding event in the latter. All these data concerning pitches, onset times, durations and volumes are tabulated and output as a numerical list and a standard MIDI file is produced.

The capabilities and the performance of the proposed transcription system have been compared with a spectrum based transcription system. The evaluation data set has been extracted from the standard RWC - Classical database; for this purpose the whole architecture has been left the most general as possible, without introducing any *a priori* knowledge. Standard parameters have been used for validation. Our system successfully identified over 57% of voiced events, with an overall F-measure of 72.1%. Finally, a comparison with other methods have been made within the MIREX 2009 evaluation framework, in which the proposed system has achieved good rankings: in particular, it has been top ranked in the piano-only tracking task. The MIREX results show

a very good overall recall rate in all the three tasks the proposed system was submitted to.

The weakest aspect seems to be a still quite high false positive rate, which affects the precision rate. This could be further improved with the introduction of physical / musicological / statistical models, or any other knowledge that may be useful to solve the challenging task of music transcription. The added values of the proposed solution, with respect to the methods based on multi-$F0$ estimation via direct cancelation on the spectrum domain, are the less leakage of information in presence of partial overlapping, and the computation of a clearer 2-D cross-correlation which leads to stronger decision capabilities.

6.1 Guidelines for Future Work

The solution adopted and described in the present work is mainly based on a signal representation technique which is quite novel for music transcription systems, while it has been already applied for sound source separation, instrument timbre modelling, classification and clustering. The Constant-Q bispectral front end has revealed to be a robust front end for multi-pitch estimation, outperforming traditional spectrum-based techniques.

In Section 3.4.2, an analysis of bispectrum nonlinearity behavior, with respect to harmonic interactions retrieval, has been conducted. We believe that further investigation in this direction could reveal the real advantages and additional information hidden in the bispectral signal representation: additional bispectral peaks, generated by nonlinear harmonic interactions, could be used to map a richer tone patterns and instrumental models, maybe weighted by opportune coefficients or statistically treated in a modified model, at the expense of a higher computational cost.

Sound source separation is an interesting field of research, which could be used jointly with traditional pitch estimation and note tracking techniques. Some experiments have been conducted [Oli09] on the music transcription system described in this work. Results of this experimentation have shown good performance, for multi-$F0$ estimation, of the constant-Q bispectral analysis jointly applied to a source separation algorithm; with higher degrees of polyphony (4 voices or more), this method even improve estimation accuracy. However, the source separation is not blind: the user has to specify the number of known sound sources, and this affects the black-box condition of the proposed system. For the future, agnostic source separation techniques should

be implemented to evaluate performance improvement.

References

[Abe04] S. S. Abeysekera, *Multiple pitch estimation of polyphonic audio signals in a frequency-lag domain using the bispectrum*, Proc. on the 2004 International Symposium on Circuits and Systems - ISCAS '04 **3** (2004), 469 – 472. (Cited on pp. 13 and 39.)

[ANP11a] F. Argenti, P. Nesi, and G. Pantaleo, *Automatic music transcription: from monophonic to polyphonic*, Musical Robots and Interactive Multimodal Systems (2011), 27–46. (Cited on pp. 13.)

[ANP11b] ———, *Automatic transcription of polyphonic music based on the constant-Q bispectral analysis*, IEEE Transactions on Audio, Speech and Language Processing **19** (2011), no. 6, 1610–1630. (Cited on pp. 65, 71 and 88.)

[BCN11] P. Bellini, A. Cappuccio, and P. Nesi, *Collaborative and assisted SKOS generation and management*, Proc. Of the 17th International Conference on Distributed Multimedia Systems. (DMS 2011) (2011), 28–33. (Cited on pp. x.)

[BDS06] J. P. Bello, L. Daudet, and M. B. Sandler, *Automatic piano transcription using frequency and time-domain information*, IEEE Transactions on Audio, Speech, and Language Processing **14** (2006), no. 6, 2242 – 2251. (Cited on pp. 13 and 37.)

[Bel03] J. P. Bello, *Towards the automated analysis of simple polyphonic music: A knowledge-based approach* , Ph.D. thesis, Jan. 2003. (Cited on pp. 7 and 8.)

[BLW07] J. G. A. Barbedo, A. Lopes, and P. J. Wolfe, *High time resolution estimation of multiple fundamental frequencies*, Proc.

of the 8^{th} International Conference on Music Information Retrieval (ISMIR) (2007), 399 – 402. (Cited on pp. 71.)

[BN05] I. Bruno and P. Nesi, *Automatic music transcription supporting different instruments*, Journal of New Music Research **34** (2005), no. 2, 139 – 149. (Cited on pp. 31 and 73.)

[BNP11] P. Bellini, P. Nesi, and G. Pantaleo, *Palamede: a Multi-Press Open Journal System*, Proc. Of the 17th International Conference on Distributed Multimedia Systems. (DMS 2011) (2011), 58–63. (Cited on pp. ix.)

[Bre90] A. S. Bregman, *Auditory scene analysis*, The MIT Press, 1990. (Cited on pp. 12.)

[Bri65] D. R. Brillinger, *An introduction to polyspectra*, The Annals of Mathematical Statistics **36** (1965), no. 5, 1351 – 1374. (Cited on pp. 39.)

[Bro91] J. C. Brown, *Calculation of a Constant-Q spectral transform*, Journal of the Acoustical Society of America **89** (1991), no. 1, 425 – 434. (Cited on pp. 43.)

[Bro06] P. M. Brossier, *Automatic annotation of musical audio for interactive applications*, Ph.D. thesis, Centre for Digital Music, Queen Mary, University of London, August 2006. (Cited on pp. 7, 8 and 74.)

[BS04] R. E. Berg and D. G. Stork, *The physics of sound*, Addison Wesley, 2004. (Cited on pp. 2.)

[CE94] V. Chandran and S. Elgar, *A general procedure for the derivation of principal domains of higher-order spectra*, IEEE Transactions on Signal Processing **42** (1994), no. 1, 229 – 233. (Cited on pp. 41.)

[CJ86] C. Chafe and D. Jaffe, *Source separation and note identification in polyphonic music*, Proc. on IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP '86) **11** (1986), 1289 – 1292. (Cited on Table 2.1.)

[CKB06] A. T. Cemgil, B. Kappen, and D. Barber, *A generative model for music transcription*, IEEE Transactions on Audio, Speech, and Language Processing **14** (2006), no. 2, 679 – 694. (Cited on pp. 10, 13, 32 and 37.)

[dCK02] A. de Cheveigné and H. Kawahara, *YIN, a fundamental frequency estimator for speech and music*, The Journal of the Acoustical Society of America **111** (2002), no. 4, 1917 – 1930. (Cited on Table 2.1.)

[DD07] C. Dubois and M. Davy, *Joint detection and tracking of time-varying harmonic components: a general bayesian framework*, IEEE Transactions on Audio, Speech and Language Processing **15** (2007), no. 4, 1283 – 1295. (Cited on pp. 13 and 37.)

[DHP09] Z. Duan, J. Han, and B. Pardo, *Harmonically Informed Multi-pitch Tracking*, Proc. on 10th International Society for Music Information Retrieval Conference (ISMIR '09) (2009). (Cited on Table 2.1.)

[DKBN06] F. C. C. Diniz, I. Kothe, L. W. P. Biscainho, and S. L. Netto, *A bounded-Q fast filter bank for audio signal analysis*, Proc. of the IEEE Telecommunications Symposium, 2006 International (2006), 1015 – 1019. (Cited on pp. 43.)

[Dol01] M. B. Dolson, *The phase vocoder: s tutorial*, Computer Music Journal **26** (2001), no. 4, 14 – 27. (Cited on pp. 73.)

[Dow08] J. S. Downie, *The music information retrieval evaluation exchange (2005–2007): s window into music information retrieval research*, Acoustical Science and Technology **29** (2008), no. 4, 247 – 255. (Cited on pp. 83.)

[DT95] S. Dubnov and N. Tishby, *Clustering of musical sounds using polyspectral distance measures*, Proceedings of the International Computer Music Conference, Banff (1995). (Cited on pp. 39.)

[DTC95] S. Dubnov, N. Tishby, and D. Cohen, *Hearing beyond the spectrum*, Journal of New Music Research **24** (1995), no. 4. (Cited on pp. 39.)

[Dub96] S. Dubnov, *Polyspectral analysis of musical timbre*, Ph.D. thesis, Institute for Computer Science and Center for Neural Computation, Hebrew University, 1996. (Cited on pp. 39.)

[DZZS08] Z. Duan, Y. Zhang, C. Zhang, and Z. Shi, *Unsupervised single-channel music source separation by average harmonic structure modeling*, IEEE Transactions on Audio, Speech and Language Processing **16** (2008), no. 4, 766 – 778. (Cited on pp. 13.)

[Ell96] D. Ellis, *Prediction-driven computational auditory scene analysis*, Ph.D. thesis, Department of Electrical Engineering and Computer Science, Massachusetts Institute of Technology (MIT), USA, 1996. (Cited on pp. 9 and 12.)

[ES06] M. R. Every and J. E. Szymanski, *Separation of synchronous pitched notes by spectral filtering of harmonics*, IEEE Transactions on Audio, Speech and Language Processing **14** (2006), no. 5, 1845 – 1856. (Cited on pp. 37.)

[ET04] T. Eerola and P. Toiviainen, *Midi toolbox: Matlab tools for music research*, University of Jyväskylä, Jyväskylä, Finland, 2004. (Cited on pp. 76.)

[FCCQ98] P. Fernández-Cid and F. J. Casajús-Quirós, *Multi-pitch estimation for Polyphonic Musical Signals*, Proc. of IEEE Int. Conf. on Acoustics Speech and Signal Processing (ICASSP '98) **6** (1998), 3565 – 3568. (Cited on Table 2.1.)

[Fri79] D. H. Friedman, *Multichannel Zero-Crossing-Interval pitch estimation*, Proc. on IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP '79) **4** (1979), 764 – 767. (Cited on Table 2.1.)

[GD02] S. Godsill and M. Davy, *Bayesian harmonic models for musical pitch estimation and analysis*, Proc. on IEEE International Conference on Acoustics, Speech and Signal Processing - ICASSP '02 (2002), 1769 – 1772. (Cited on pp. 37.)

[GDI06] S. Godsill, M. Davy, and J. Idier, *Bayesian analysis of polyphonic western tonal music*, Journal of the Acoustical Society

of America **119** (2006), no. 4, 2498 – 2517. (Cited on pp. 13 and 37.)

[Ghe03] D. Gherard, *Pitch extraction and fundamental frequency: history and current techniques*, Tech. report, Dep. of Computer Science, University or Regina, Canada, November 2003. (Cited on pp. 8.)

[GHNO02] M. Goto, H. Hashiguchi, T. Nishimura, and R. Oka, *RWC Music Database: popular, classical, and jazz music database*, Proc. on the 3^{th}'International Conference on Information Music Retrieval (ISMIR) (2002), 287 – 288. (Cited on pp. 73.)

[Got00] M. Goto, *A robust predominant-F0 estimation method for real-time detection of melody and bass lines in CD recordings*, Proc. on IEEE International Conference on Acoustics, Speech and Signal Processing ICASSP 2000 **2** (2000), 757 – 760, Istanbul, Turkey. (Cited on pp. 13 and 29.)

[Got04] ———, *A real-time music-scene-description system: predominant-F0 estimation for detecting melody and bass lines in real-world audio signals*, Speech Communication - ISCA Journal **43** (2004), no. 4, 311 – 329. (Cited on pp. 13 and 29.)

[GR77] B. Gold and L. R. Rabiner, *On the use of autocorrelation analysis for pitch detection*, IEEE Transactions on Acoustics, Speech and Signal Processing **25** (1977), no. 1, 24 – 33. (Cited on pp. 12.)

[Hal91] D. E. Hall, *Musical acoustics*, Brooks/Cole Pub Co, 1991. (Cited on pp. 2.)

[KI89] H. Katayose and S. Inokuchi, *The KANSEI Music System*, Computer Music Journal **13** (1989), no. 4, 72 – 77. (Cited on Table 2.1.)

[Kla98] A. Klapuri, *Automatic transcription of music*, Ph.D. thesis, Master Thesis, Tampere University of Technology (Dep. of Information Technology), 1998. (Cited on pp. 3.)

[Kla03] ———, *Multiple fundamental frequency estimation based on harmonicity and spectral smoothness*, IEEE Transactions on Speech and Audio Processing **11** (2003), no. 6, 804 – 816. (Cited on pp. 12, 30 and 71.)

[Kla04a] ———, *Automatic music transcription as we know it today*, Journal of New Music Research **33** (2004), no. 3, 269 – 282. (Cited on pp. 5, 7 and 8.)

[Kla04b] ———, *Signal processing methods for the automatic transcription of music*, Ph.D. thesis, Tampere University of Technology, March 2004. (Cited on pp. 7, 8 and 14.)

[Kla05] ———, *A Perceptually motivated multiple-F0 estimation method*, IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (2005), 291 – 294. (Cited on pp. 12 and 71.)

[Kla08] ———, *multipitch analysis of polyphonic music and speech signals using an auditory model*, IEEE Transactions on Audio, Speech, and Language Processing **16** (2008), no. 2, 255 – 266. (Cited on pp. 13.)

[KNKT95] K. Kashino, K. Nakadai, T. Kinoshita, and H. Tanaka, *Application of bayesian probability network to music scene analysis*, Proc. on IJCAI Workshop on Computational Auditory Scene Analysis (CASA) (1995). (Cited on pp. 10, 13, 27 and 37.)

[KNS07] H. Kameoka, T. Nishimoto, and S. Sagayama, *A multipitch analyzer based on harmonic temporal structured clustering*, IEEE Transactions on Audio, Speech, and Language Processing **15** (2007), no. 3, 982 – 994. (Cited on pp. 32.)

[KR07] A. Klapuri and M. P. Ryynänen, *Automatic bass line transcription from streaming polyphonic audio*, IEEE International Conference on Acoustics, Speech and Singal Processing - ICASSP '07 **4** (2007), 1437 – 1440. (Cited on pp. 13.)

[KT92] K. Kashino and H. Tanaka, *A Sound Source Separation System Using Spectral Features Integrated by Dempster's Law of Com-*

bination, Annual Report of the Engineering Research Institute **52** (1992). (Cited on Table 2.1.)

[KT93] ———, *A Sound Source Separation System with the Ability of Automatic Tone Modeling*, Proc. of International Computer Music Conference (ICMC '93) (1993), 248 – 255. (Cited on Table 2.1.)

[Mah90] R. C. Maher, *Evaluation for a method for separating digitized duet signals*, Journal of Acoustic Engineering Society **38** (1990), no. 12, 956 – 979. (Cited on pp. 25.)

[Mar72] J. D. Markel, *The SIFT algorithm for fundamental frequency estimation*, IEEE Transactions on Audio and Electroacoustics **20** (1972), no. 5, 366 – 367. (Cited on pp. 12.)

[Mar96a] K. D. Martin, *A blackboard system for automatic transcription of simple polyphonic music*, Perceptual Computing Technical Report 385, MIT Media Lab (1996). (Cited on pp. 10, 13 and 36.)

[Mar96b] ———, *Automatic transcription of simple polyphonic music: robust eront end processing*, Tech. Report #399, MIT Media Lab, Perceptual Computing Section, July 1996. (Cited on pp. 28.)

[Mar01] M. Marolt, *SONIC: transcription of polyphonic piano music with neural networks*, Audiovisual Institute, Pompeu Fabra University, 2001, pp. 217 – 224. (Cited on pp. 13, 29 and 37.)

[Mar04] ———, *Networks of adaptive oscillators for partial tracking and transcription of music recordings*, Journal of New Music Research **33** (2004), no. 1, 49 – 59. (Cited on pp. 13.)

[Mil75] N. J. Miller, *Pitch detection by data reduction*, IEEE Transactions on Acoustic, Speech and Signal Processing **23** (1975), no. 1, 72 –79. (Cited on pp. 12.)

[MIRa] *MIREX 2009 results, Web URL*, http://www.music-ir.org/mirex/wiki/2009:MIREX2009_Results. (Cited on pp. 83.)

[MIRb] *MIREX 2009, Web home page URL*, http://www.music-ir.org/mirex/wiki/2009:Main_Page. (Cited on pp. 10, 16 and 83.)

[MO97] R. Meddis and L. O'Mard, *A unitary model of pitch perception*, The Journal of the Acoustical Society of America **102** (1997), no. 3, 1811 – 1820. (Cited on pp. 9 and 12.)

[Moo77] J. A. Moorer, *On the transcription of musical sound by computer*, Computer Music Journal **1** (1977), no. 4, 32 – 38. (Cited on pp. 12 and 18.)

[Moo78] ———, *The use of the phase vocoder in computer music applications*, Journal of the Audio Engineering Society **10** (1978), no. 1/2, 42 – 45. (Cited on pp. 73.)

[NAW01] S. H. Nawab, S. A. Ayyash, and R. Wotiz, *Identification of musical chords using constant-Q spectra*, IEEE Proc. on Acoustic, Speech and Signal Processing - ICASSP '01 **5** (2001), 3373 – 3376. (Cited on pp. 12 and 71.)

[NM93] C. L. Nikias and J. M. Mendel, *Signal processing with higher-order Spectra*, IEEE Signal Processing Magazine **10** (1993), no. 3, 10 – 37, ISSN: 1053-5888. (Cited on pp. 39, 40, 41 and 42.)

[NR87] C. L. Nikias and M. R. Raghuveer, *Bispectrum estimation: a digital signal processing framework*, Proceedings of the IEEE **75** (1987), no. 7, 869 – 891. (Cited on pp. 39.)

[OBCQTG05] L. I. Ortiz-Berenguer, F. J. Casajús-Quirós, and S. Torres-Guijarro, *Multiple piano note identification using a spectral matching method with derived patterns*, Journal of Audio Engineering Society **53** (2005), no. 1/2. (Cited on pp. 13.)

[Oli09] F. Olivieri, *Teoria e metodi per la separazione di componenti parziali nelle sorgenti audio*, Master's thesis, Department of Systems and Informatics, University of Florence, Italy, 2009, (Italian language). (Cited on pp. 89.)

[PC03] I. Peretz and M. Coltheart, *Modularity of music processing*, Nature Neuroscience **6** (2003), no. 7, 688 – 691. (Cited on pp. 8.)

[PE07] G. E. Poliner and D. P. W. Ellis, *A discriminative model for polyphonic piano transcription*, EURASIP Journal on Advances on Signal Processing (2007). (Cited on pp. 77.)

[PG77] M. Piszczalski and B. A. Galler, *Automatic music transcription*, Computer Music Journal **1** (1977), no. 4, 24 – 31. (Cited on pp. 12 and 24.)

[PIn08] A. Pertusa and J. M. Iñesta, *Multiple fundamental frequency estimation using gaussian smoothness*, Proc. of the IEEE Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP 2008) (2008), 105 – 108, Las Vegas (USA). (Cited on pp. 35.)

[Rab77] L. R. Rabiner, *On the use of autocorrelation analysis for pitch detection*, IEEE Transactions on Acoustics, Speech and Signal Processing **25** (1977), no. 1, 24 – 33. (Cited on pp. 7.)

[Rap02] C. Raphael, *Automatic transcription of piano music*, Proc. on 3rd International Conference on Music Information Retrieval (2002), 15 – 19. (Cited on pp. 13 and 37.)

[RK05] M. P. Ryynänen and A. Klapuri, *Polyphonic music transcription using note event modeling*, IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (2005), 319 – 322. (Cited on pp. 10, 13, 31 and 37.)

[ROS07] S. Raczynksi, N. Ono, and Sagayama, *Multipitch analysis with harmonic nonnegative matrix approximation*, Proc. of the 8^{th} International Conference on Music Information Retrieval (ISMIR) (2007), 381 – 386. (Cited on pp. 13.)

[RRM76] L. R. Rabiner, A. E. Rosenberg, and C. A. McGonegal, *A comparative performance study of several pitch detecting algorithms*, IEEE Transactions on Acoustic, Speech and Signal Processing **24** (1976), no. 5, 399 – 418. (Cited on pp. 12.)

[SB03] P. Smaragdis and J. C. Brown, *Non-negative matrix factorization for polyphonic music transcription*, IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (2003), 177 – 180, New Paltz (NY). (Cited on pp. 13 and 37.)

[SL90] M. Slaney and R. F. Lyon, *A Perceptual pitch detector*, Proc. of IEEE International Conference on Acoustics, Speech and Signal Processing - ICASSP '90 (1990), 357 – 360. (Cited on pp. 9, 12 and 25.)

[Sor95] R. Sorce, *Music theory for the music professional*, Ardsley House Publishers, Inc., 1995. (Cited on pp. 5.)

[TK00] T. Tolonen and M. Karjalainen, *A computationally efficient multipitch analysis model*, IEEE Transactions on Speech and Audio Processing **8** (2000), no. 6, 708 – 716. (Cited on pp. 12 and 28.)

[VBB08] E. Vincent, N. Bertin, and R. Badeau, *Harmonic and inharmonic nonnegative matrix factorization for polyphonic pitch transcription*, IEEE International Conference on In Acoustics, Speech and Signal Processing (ICASSP '08) (2008), 109 – 112. (Cited on pp. 10, 13, 33 and 37.)

[Vir07] T. Virtanen, *Monaural sound source separation by nonnegative matrix factorization with remporal continuity and sparseness criteria*, IEEE International Conference on Computational Intelligence for Measurement Systems and Applications **15** (2007), no. 3, 1066 – 1074. (Cited on pp. 10, 13 and 37.)

[Yeh08] C. Yeh, *Multiple fundamental frequency estimation of polyphonic rercordings*, Ph.D. thesis, Ecole Doctorale Edite, University of Paris VI, 2008. (Cited on pp. 7, 8 and 9.)

[YRR+08] C. Yeh, A. Roebel, X. Rodet, W. C. Chang, and A. W. Y. Su, *Multiple F0 tracking based on a high-order HMM model*, Proc. of the 11^{th} Conference on Digital Audio Effects (DAFx-08), Espoo, Finland (2008). (Cited on pp. 10, 13, 34, 35 and 37.)

[YRR10] C. Yeh, A. Roebel, and X. Rodet, *Multiple fundamental frequency estimation and polyphony inference of polyphonic music signals*, IEEE Transactions on Audio, Speech and Language Processing **18** (2010), no. 6. (Cited on pp. 35.)