Automatic transcription of polyphonic music based on the constant-Q bispectral analysis

Fabrizio Argenti, *Senior Member, IEEE*, Paolo Nesi, *Member, IEEE*, and Gianni Pantaleo

August 31, 2010

Abstract

In the area of music information retrieval (MIR), automatic music transcription is considered one of the most challenging tasks, to solve which many different techniques have been proposed. This paper presents a new method for polyphonic music transcription: a system that aims at estimating pitch, onset times, durations and intensity of concurrent sounds in audio recordings, played by one or more instruments. Pitch estimation is carried out by means of a front-end that jointly uses a constant-Q and a bispectral analysis of the input audio signal; subsequently, the processed signal is correlated with a fixed 2-D harmonic pattern. Onsets and durations detection procedures are based on the combination of the constant-Q bispectral analysis with information from the signal spectrogram. The detection process is agnostic and it does not need to take into account musicological and instrumental models or other *a priori* knowledge. The system has been validated against the standard RWC (Real World Computing) - Classical Audio Database. The proposed method has demonstrated good performances in the multiple $F0$ tracking task, especially for piano-only automatic transcription at MIREX 2009.

Index Terms

Music information retrieval, polyphonic music transcription, audio signals processing, constant-Q analysis, higher-order spectra, bispectrum.

I. Introduction

Automatic music transcription is 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.

An audio signal is composed of a single or a mixture of approximately periodic, locally stationary acoustic waves. According to the Fourier representation, any finite energy signal is represented as the sum of an infinite number of sinusoidal components weighted by appropriate amplitude coefficients. 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 cases, the spectral component corresponding to $F0$ can be missing), the task of music transcription, i.e., 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. 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.

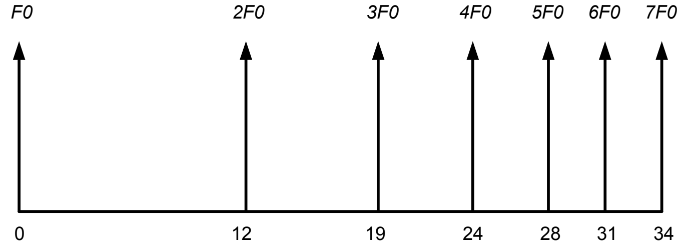


Figure 1. Fixed comb-pattern representing the harmonics set associated with every single note. Seven partials (fundamental frequency included) with the same amplitude have been considered. The distances are also expressed (bottom) as semitones.

A. Previous Work

For the monophonic transcription task, some time-domain methods were proposed based on zero-crossing detection [1], or on temporal autocorrelation [2]. Frequency-domain based approaches are better suited for multi-pitch detection of a mixture of sounds. In fact, the overlap of different period waves makes the task hard to be solved exclusively in the time-domain.

First attempts of performing polyphonic music transcription started in the late 1970s, with the pioneering work of Moorer [3] and Piszczalski and Galler [4]. During the years, 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 [5], [6] performed an iterative predominant $F0$ estimation and a subsequent cancelation of each harmonic pattern from the spectrum; Nawab [7] used an iterative pattern matching algorithm upon a constant-Q spectral representation.

In the early 1990s, other approaches, based on applied psycho-acoustic models and also known as *Computational Auditory Scene Analysis (CASA)*, from the work by Bregman [8], 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 [9], were proposed by Ellis [10], Meddis and O'Mard [11], Tolonen and Karjalainen [12] and Klapuri [13]. Marolt [14] 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 at the basis of the idea of the *Blackboard Architecture* proposed by Martin [15]. More recently, physical [16] and musicological models, like average harmonic structure (AHS) extraction in [17], as well as other *a priori* knowledge [18], and eventually temporal information [19] 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 [20], [21], [22], Bayesian networks [23], [24] or Bayesian models [25], [26]. Others, aiming at estimating the bass line [27] or the melody and bass lines [28], [29], were proposed. Currently, the approach based on non-negative matrix approximation [30] (in its different versions like non-negative matrix factorization of spectral features [31], [32], [33]) has received much attention within the music transcription community.

Higher-order spectral analysis (which includes the bispectrum as a special case) has been applied to music audio signals for source separation and instrumental modeling [34], to enhance the characterization of relevant acoustical features [35], and for polyphonic pitch detection [36].

More detailed overviews of automatic music transcription methods and related topics are contained in [37], [38].

B. Proposed Method

This paper proposes a new method for automatic transcription of real polyphonic and multi-instrumental music. Pitch estimation is here performed through a joint constant-Q and bispectral analysis of the input audio signal. The bispectrum is a bidimensional frequency representation capable of detecting nonlinear harmonic interactions.

A musical signal produces a typical 1-D pattern of local maxima in the spectrum domain and, similarly, a 2-D pattern in the bispectrum domain, as illustrated in Section III-C1. Objective of a multiple $F0$ estimation algorithm is retrieving the information relative to each single note from the polyphonic mixture. A method to perform this task, in the spectrum domain, consists in iteratively computing the cross-correlation between the audio signal and a harmonic template, and subsequently canceling/subtracting the pattern relative to the detected note. The proposed method applies this concept, opportunely adapted, in the bispectral domain.

Experimental results show that using the bispectrum analysis yields superior performances than using the spectrum domain: actually, as described in section III-C4, the local maxima distribution of the harmonic 2-D pattern generated in the bispectrum domain is more useful in gathering multiple-$F0$ information in iterative pitch estimation and harmonics extraction / cancelation methods.

A computationally efficient and relatively fast method to implement the bispectrum has been realized by using the constant-Q transform, which produces a multi-band frequency representation with variable resolution. Note duration estimation is based on a profile analysis of the audio signal spectrogram.

The goal of this research is showing the capabilities and potentialities of a constant-Q bispectrum (*CQB*) front-end applied to the automatic music transcription task. The assessment of the proposed transcription system performances has been conducted in the following way:the proposed method, based on the bispectrum front-end, and a similar system, based on a simple spectrum front-end, were compared by using audio excerpts taken from the standard RWC (Real World Computing) - Classical Audio Database [39], which is widely used in the recent literature for information music retrieval tasks; the proposed algorithm has demonstrated good performances in the multiple $F0$ tracking task, especially for piano automatic transcription at MIREX 2009 evaluation framework. The results of the comparison with the other participants are reported.

C. Paper Organization

In Section II, the bispectral analysis and the constant-Q transform are reviewed. Section III contains a detailed description of the whole architecture and the rules for pitch, onset and note duration detection. Subsequently, in section IV, experimental results, validation methods and parameters are presented. Finally, Section V is left to conclusions.

II. Theoretical preliminaries

In this section, the theoretical concepts at the basis of the proposed method are recalled.

A. Musical concepts and notation

In music, the seven notes are expressed with alphabetical letters from A to G. The octave number is indicated as a subscript. In this paper, 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.

In the well-tempered system, if f_1 and f_2 are the frequencies of two notes separated by one semitone interval, then $f_2 = f_1 \cdot 2^{1/12}$. Under these conditions (which approximates the natural tuning, or just tuning), an interval of one octave, (characterized by $f_2 = 2f_1$) it is composed of 12 semitones. Other examples of intervals between notes are the perfect fifth ($f_2 = 3/2\ f_1$, corresponding to a distance of 7 semitones in the well-tempered scale), the perfect fourth ($f_2 = 4/3\ f_1$ or 5 semitones in the well-tempered scale), and the major third ($f_2 = 5/4\ f_1$ or 4 semitones in the well-tempered scale).

B. The Bispectrum

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 [40], [41] and [42]. 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.

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 [41] 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 [41] 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:

$$\begin{aligned} c_1^x &= E\{x(k)\} = 0, \\ c_2^x(\tau_1) &= m_2^x(\tau_1) = E\big\{x(k)x(k+\tau_1)\big\}, \\ c_3^x(\tau_1, \tau_2) &= m_3^x(\tau_1, \tau_2) = E\big\{x(k)x(k+\tau_1)x(k+\tau_2)\big\}. \end{aligned} \tag{1}$$

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:

$$B_x(f_1, f_2) = S_3^x(f_1, f_2) = \sum_{\tau_1=-\infty}^{+\infty} \sum_{\tau_2=-\infty}^{+\infty} c_3^x(\tau_1, \tau_2) e^{-j2\pi f_1 \tau_1} e^{-j2\pi f_2 \tau_2}. \tag{2}$$

The bispectrum is a bivariate function representing some kind of signal-energy related information, as more deeply analyzed in the next section. In Figure 2, a contour-plot of the bispectrum of an audio signal is shown. As can be noticed, the bispectrum presents twelve mirror symmetry regions:

$$B_x(f_1, f_2) = B_x(f_2, f_1) = B_x^*(-f_2, -f_1) = B_x(-f_1 - f_2, f_2) =$$
$$= B_x(f_1, -f_1 - f_2) = B_x(-f_1 - f_2, f_1) = B_x(f_2, -f_1 - f_2).$$

Hence, the analysis can take into consideration only a single non redundant bispectral region [43]. 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., $\mathcal{T} = \left\{(f_1, f_2) : 0 \le f_2 \le f_1 \le \frac{f_s}{2}, f_2 \le -2f_1 + f_s\right\}$, which is depicted in Figure 2, where f_s is the sampling frequency.

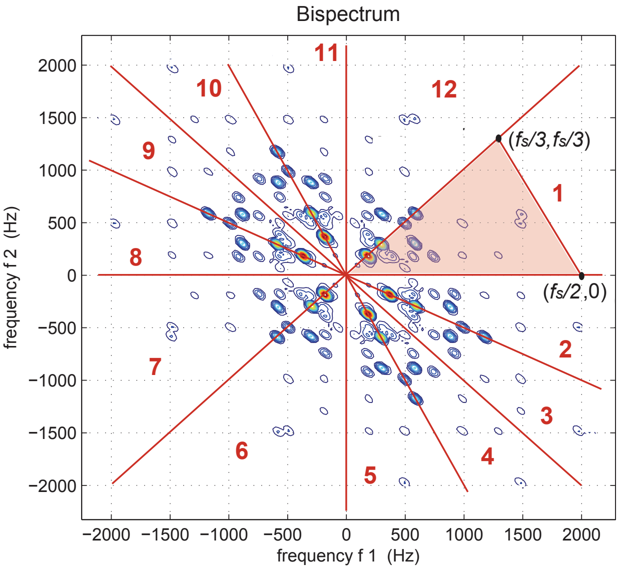


Figure 2. Contour plot of the magnitude bispectrum, according to Equation (3), of the trichord $F\sharp_3$(185 Hz), D_4(293 Hz), B_4(493 Hz) played on an acoustic upright piano and sampled at f_s = 4 kHz. The twelve symmetry regions are in evidence (clockwise enumerated), and the one chosen for analysis is highlighted.

It can be shown [41] that the bispectrum of a finite-energy signal can be expressed as:

$$B_x(f_1, f_2) = X(f_1)X(f_2)X^*(f_1 + f_2), \tag{3}$$

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 [41].

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 tansform 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) (eventually, frequency averaging can be applied); 3) computation of the estimated bispectrum as the average of the bispectrum estimates in each segment.

In this paper, in order to minimize the computational cost, the direct method has been used to estimate the bispectrum of an audio signal.

C. Constant-Q Analysis

The estimation of the bispectrum according to (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 [44], [45] 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

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

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. It consists of a tree structure, shown in Figure 3-(a), whose building block, shown in Figure 3-(b), is composed of a spectrum analyzer block and by a filtering/downsampling block (lowpass filter and downsampler 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 lowpass 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.

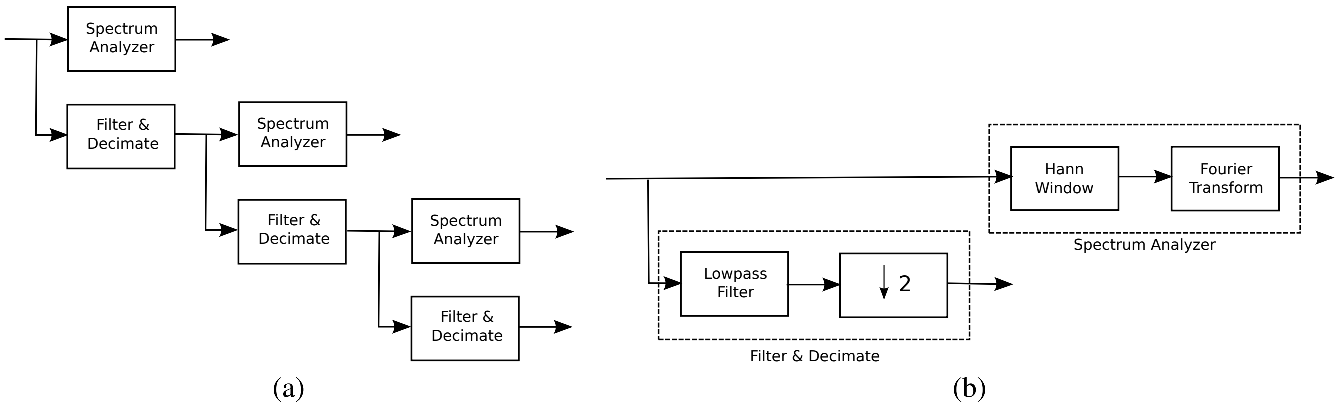


Figure 3. Octave Filter Bank: (a) building block of the tree, composed by a spectrum analyzer and by a filtering/downsampling block; (b) blocks combination to obtain a multi-octave analysis.

III. System Architecture

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

A. General Architecture

A general view of the system architecture is presented in Figure 4. 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.

B. The Pre-Processing module

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 $\pm 2\%$

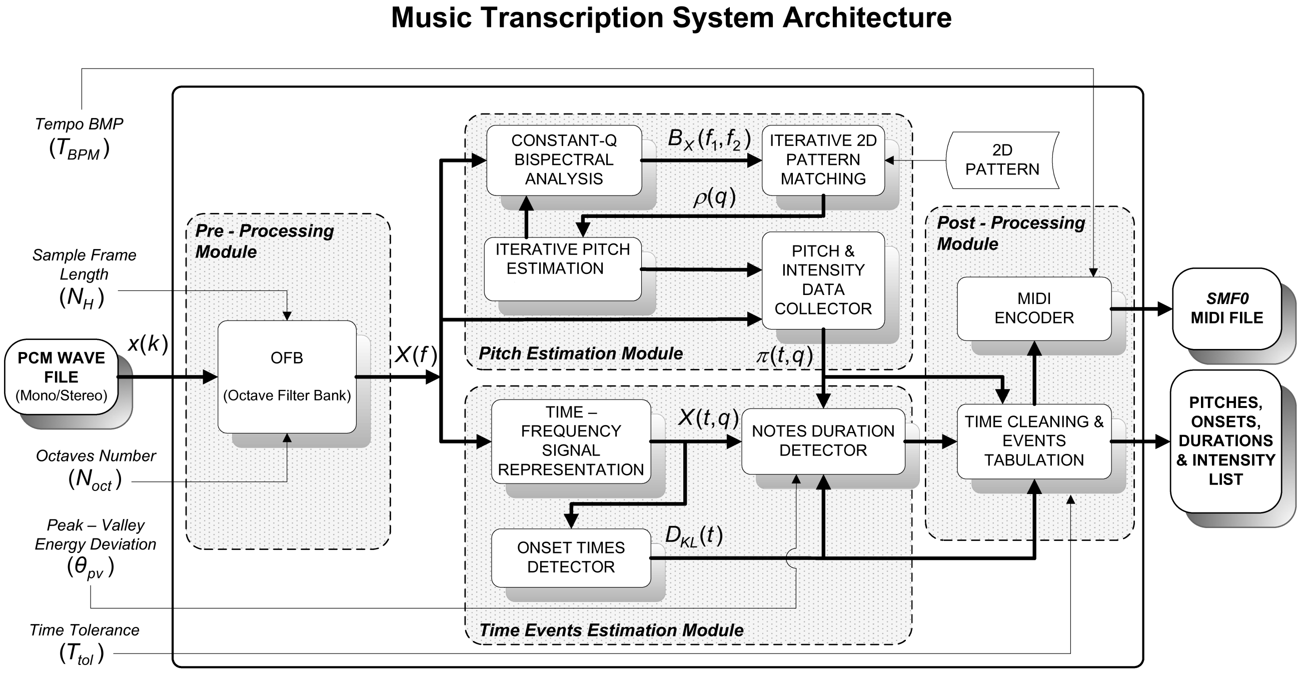


Figure 4. Music transcription system block architecture. The functional modules, inner blocks, input parameters and output variables and functions are illustrated.

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 $(0, 22050)$, $(0, 11025)$, and $(0, 5512.5)$. 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-(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 lowpass 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 5.

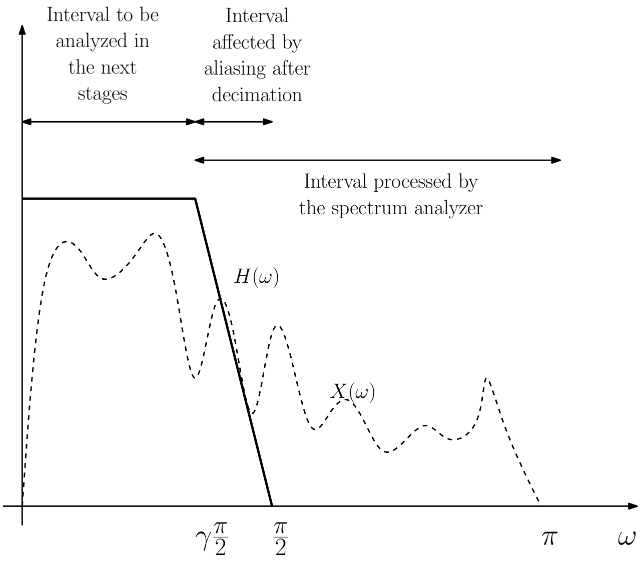

Interval to be analyzed in the next stages
Interval affected by aliasing after decimation
Interval processed by the spectrum analyzer
$H(\omega)$
$X(\omega)$
$\gamma\frac{\pi}{2}$
$\frac{\pi}{2}$
π
ω

Figure 5. Filter mask and the analyzed regions.

Table I 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.

Table I
OFB CHARACTERISTICS

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

C. 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 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.

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.

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), the bispectrum of $x(n)$ is given by

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

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

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

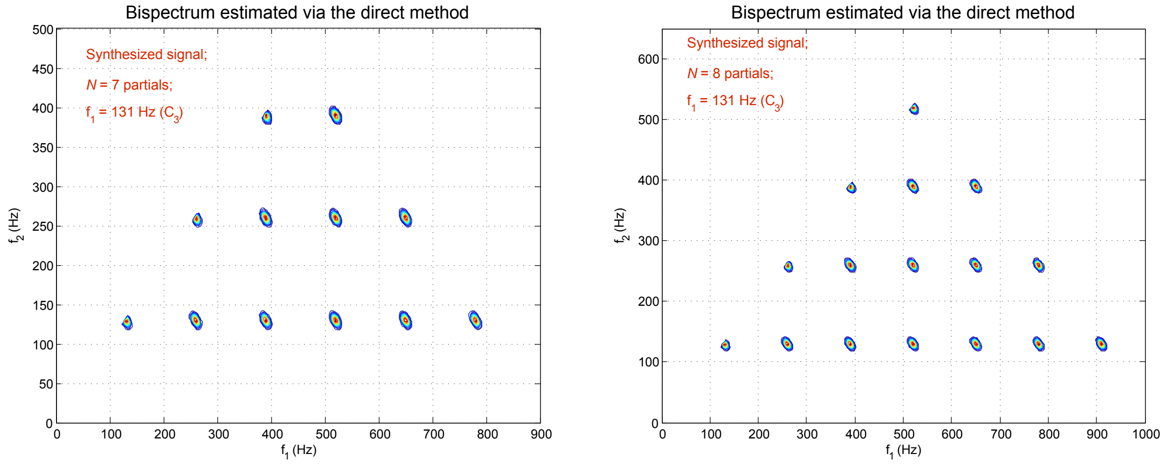
Note that peaks arise along the first and third quadrant bisector thanks to the fact that $f_2 = 2f_1$ and $f_4 = 2f_2$. By considering the non-redundant triangular region $\mathcal{T}$ defined in Section II-B, the above expression can be simplified into

$$\begin{aligned} B_x(\eta_1, \eta_2) =& \delta(\eta_1 - f_1)\delta(\eta_2 - f_1)\delta(\eta_1+\eta_2 - f_2) + \delta(\eta_1 - f_2)\delta(\eta_2 - f_1)\delta(\eta_1+\eta_2 - f_3) \\ &+ \delta(\eta_1 - f_3)\delta(\eta_2 - f_1)\delta(\eta_1+\eta_2 - f_4) + \delta(\eta_1 - f_2)\delta(\eta_2 - f_2)\delta(\eta_1+\eta_2 - f_4). \end{aligned} \tag{4}$$

Equation (4) can be generalized to an arbitrary number T of harmonics as follows:

$$B_x(\eta_1, \eta_2) = \sum_{p=1}^{\lfloor T/2 \rfloor} \delta(\eta_2 - f_p) \sum_{q=p}^{T-p} \delta(\eta_1 - f_q)\delta(\eta_1 + \eta_2 - f_{p+q}). \tag{5}$$

This formula shows that every monophonic signal generates a bidimensional bispectral pattern characterized by peaks positions $\{(f_i, f_i), (f_{i+1}, f_i), \ldots, (f_{T-i}, f_i)\}$, $i = 1, 2, \ldots, \lfloor \frac{T}{2} \rfloor$. Such a pattern is depicted in Figure 6 for a synthetic note at a fundamental frequency $f_1 = 131$ Hz, with $T = 7$ and $T = 8$.



(a)

(b)

Figure 6. Bispectrum of monophonic signals (note C_3) synthesized with (a) $T = 7$ and (b) $T = 8$ harmonics.

The energy distribution in the bispectrum domain is validated by the analysis of real world monophonic sounds. Figure 7 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 behaviour previously shown.

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), 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 Appendix A. 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 8-(a) and 8-(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 8-(a) and 8-(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 9, the bispectrum of a real bichord produced by two bowed violins,

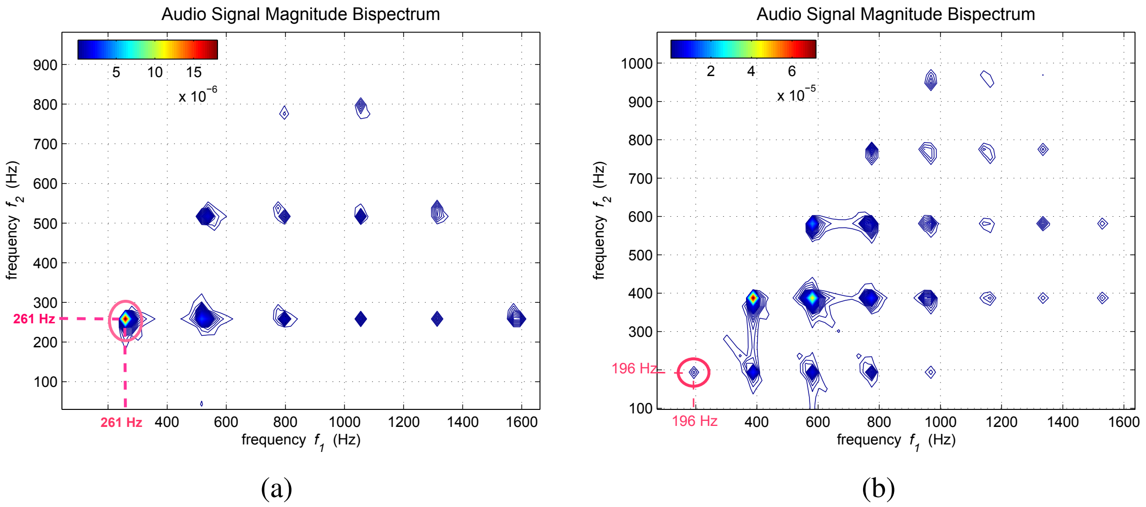


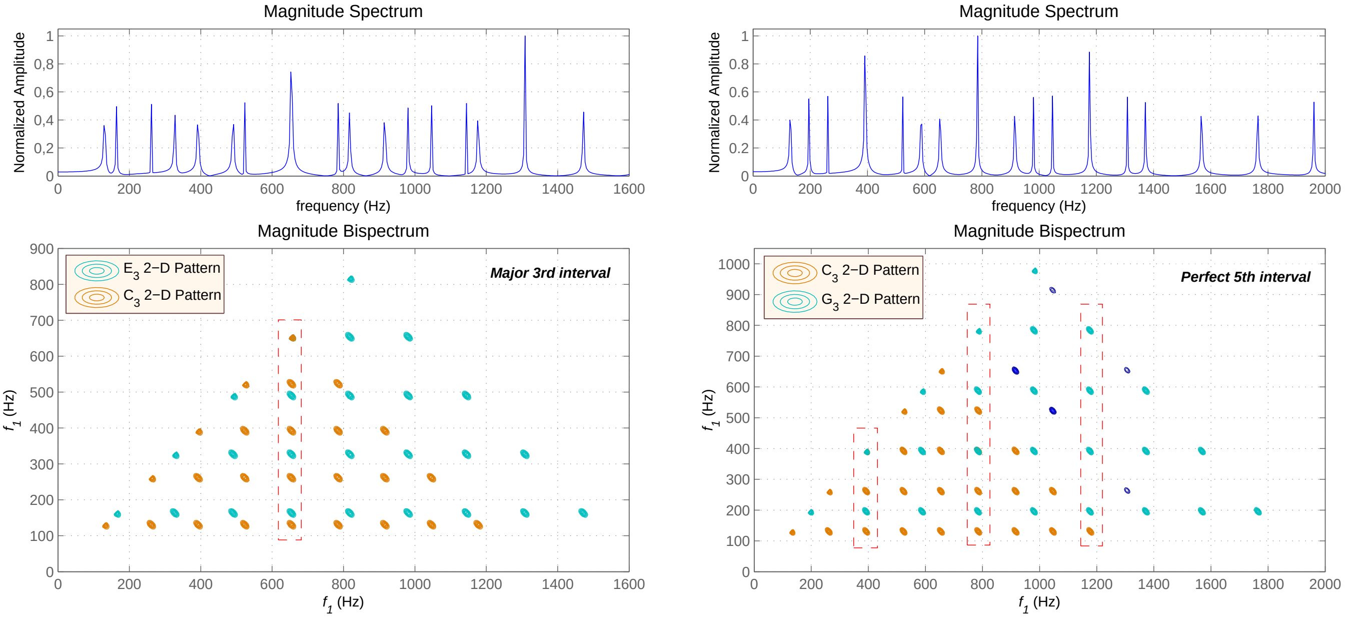
Figure 7. Bispectrum of (a) a C$_4$ (261 Hz) played on a upright piano, and of (b) a G$_3$ (196 Hz) played on a violin (bowed). Both sounds have been sampled at 44100 Hz.

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 Appendix A, 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 tri-chord with strong partial overlapping and a high number of harmonics per sound has confirmed the same results.

3) 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 10, 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 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



(a) (b)

Figure 8. 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.

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:

$$\rho(k_1, k_2) = \sum_{m_1=0}^{C_P-1} \sum_{m_2=0}^{R_P-1} P(m_1, m_2) \left| B_x(k_1 + m_1, k_2 + m_2) \right|, \tag{6}$$

where $1 \leqslant k_1, k_2 \leqslant 12N_{oct}$ 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) Pitch Detection: (4a) - Recall on Spectrum Domain. Several methods based on pattern matching in the spectrum domain were proposed for multiple-pitch estimation [5], [6], [7], [46]. In these methods, an iterative approach is used. First, a single $F0$ is estimated by using different criteria (e.g., maximum amplitude, or lowest

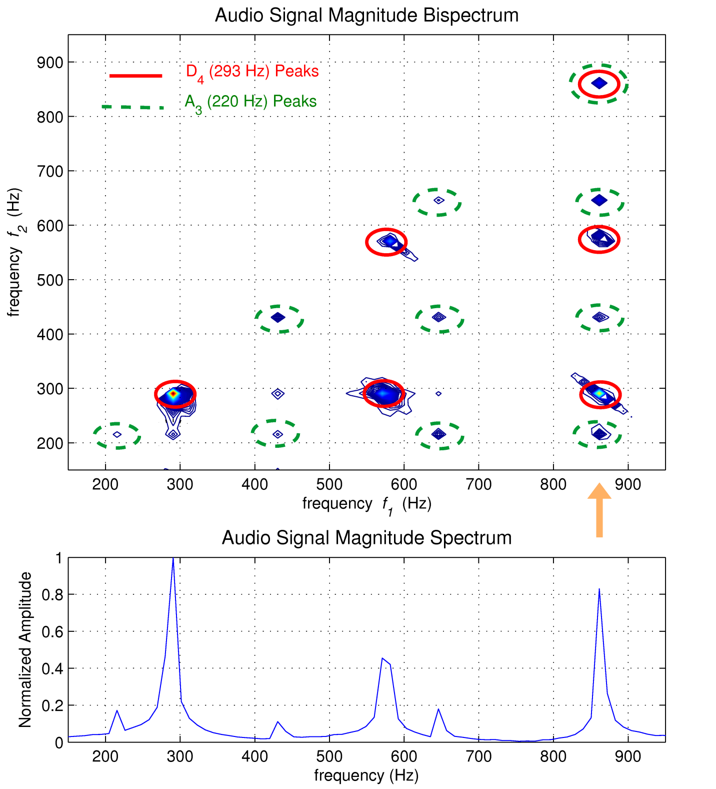


Figure 9. Detail (top figure) of the bispectrum of a bichord (A_3 at 220 Hz and D_4 at 293 Hz), played by two violins (bowed), sampled at 44100 Hz. The arrow highlights the frequency at 880 Hz, where the partials of the two notes overlap in the spectrum domain.

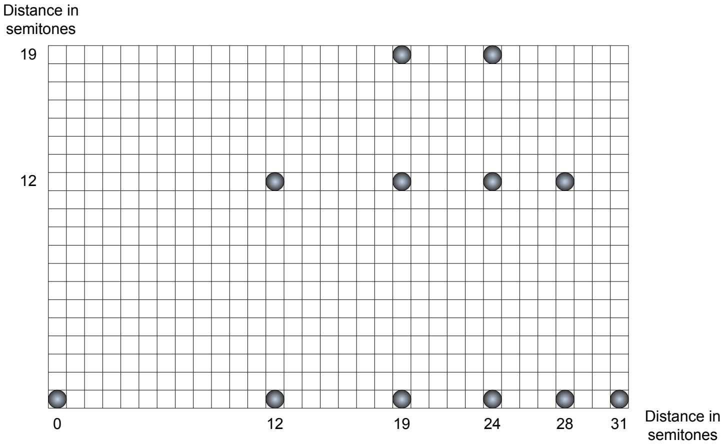


Figure 10. Fixed 2-D harmonic pattern used in the validation tests of the proposed music transcriptor. It represents the theoretical set of bispectral local maxima for a monophonic 7-partials sound all weights are set equal to unity.

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 this paper 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 8). 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:

$$\rho(q) = \sum_{m_1=0}^{C_P-1} \sum_{m_2=0}^{R_P-1} P(m_1, m_2) \left| B_x(q+m_1, q+m_2) \right|, \tag{7}$$

derived directly from Equation (6)

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 [39], 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), 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.

D. 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.

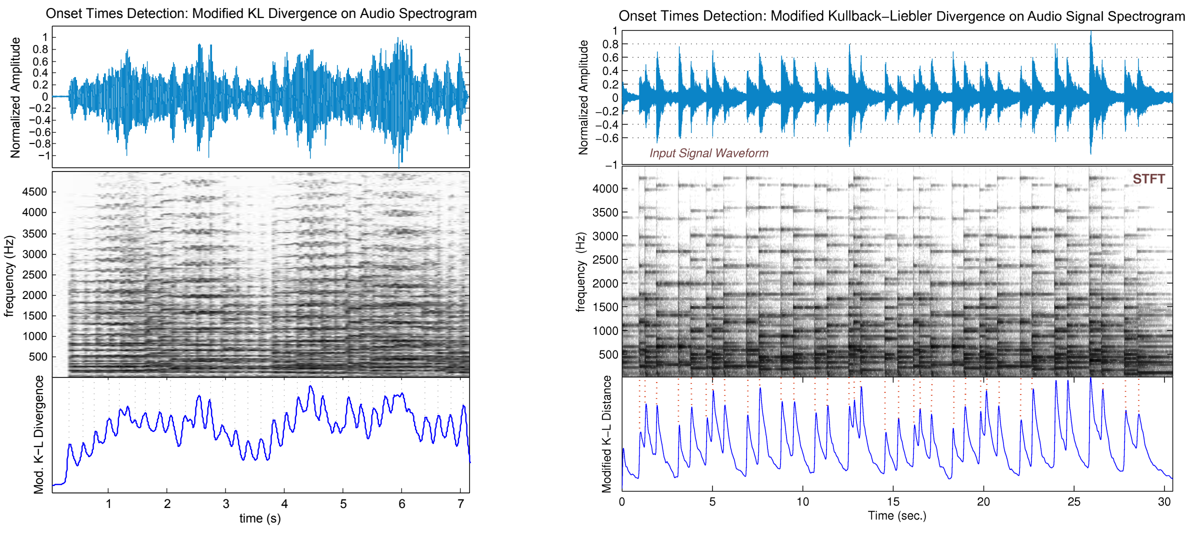
The *Time-Frequency Representation* block collects the spectral information $X(f)$ of each frame, used also to compute the bispectrum, in order to represent the signal in the time-frequency domain. The output of this block is the array $X(t, q)$, where t is the index over the frames, and q is the index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

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 [47], [48], [49] have been proposed for onset detection in the time-frequency domain. The methods based on the phase-vocoder functions [48], [49] 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 method proposed in this paper uses the *Modified Kullback-Liebler Divergence* function, which achieved the best performance in [50]. This function aims at evaluating the distance between two consecutive spectral vectors, highlighting large positive energy variations and inhibiting small ones. The modified Kullbak-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 11.

The *Notes Duration Detector* block carries out the estimation of notes duration. The beginning of a note relies on



(a) (b)

Figure 11. Results of onset detection procedure obtained applying the *Modified Kullback-Liebler Divergence* over audio spectrogram for two fragments from RWC - Classical Database: (a) 7 seconds extracted from Mozart's *String Quartet n. 19, K465*; (b) the first 30 seconds of Mozart's first movement of *Sonata for piano in A major K331*.

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 III-C4) 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;
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.

E. 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 [51]), which generates the output MIDI SMF0 file, provided that the user defines a tempo value T_{BPM}, expressed in beats per minute.

IV. 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 cancellation. Finally, the results of the comparison of the proposed method with others participating at the MIREX 2009 contest are presented.

A. 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 [52]:

- *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

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

B. Validation of the proposed method by using the RWC-C database

1) Experimental data set: 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 II, the main features of the used test audio files are reported. The set includes about 100000 one-frame-long voiced events.

Table II
TEST DATA SET FROM RWC - CLASSICAL DATABASE. VN(S): VIOLIN(S); VLA: VIOLA; VC: CELLO; CB: CONTRABASS; CL: CLARINET

# 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 Mï£¡re 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. Faurï£¡	Sicilienne, *op. 78*	C-2001 n. 43	Piano and Flute

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.

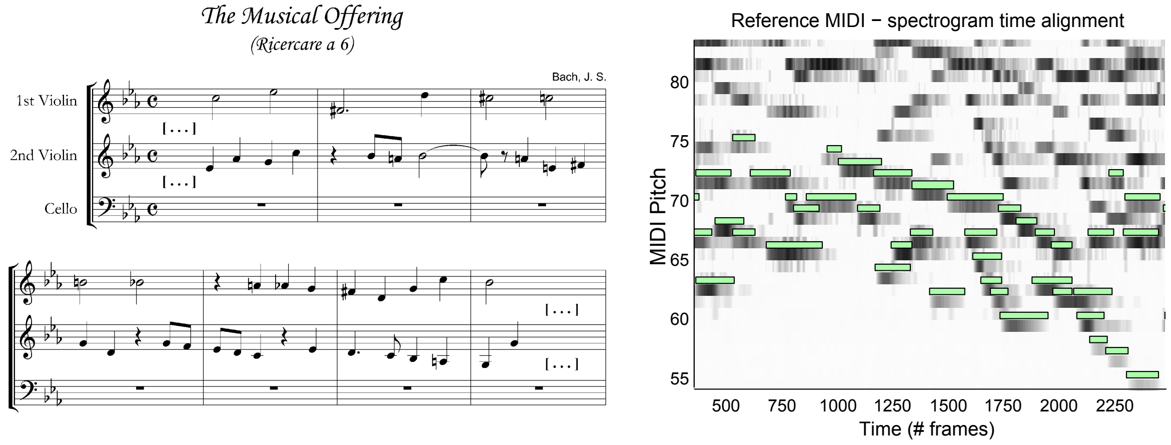
The proposed transcription system has been realized and tested in Matlab® environment installed on a dual core 64-bit processor 2.6 GHz with 3 GB of RAM. With this equipment, the system performs the transcription in a period which is approximately fifteen times the input audio file duration.

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 III-C. 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 10, is substituted by a 1-D correlation in the spectrum domain, using the 1-D pattern in Figure 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 12.



(a)

(b)

Figure 12. 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.

The performances of algorithms BISP, SP1 and SP2 applied to the audio data set described in section IV-B1 are shown in Tables III, IV and V. 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 13, and a graphical comparison between the output of BISP and SP1 is shown in Figure 15. In Figure 14, 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 demanded to the next section, where the results of the MIREX 2009 evaluation framework are reported.

Table III
BISP: TRANSCRIPTION RESULTS OBTAINED WITH THE TEST DATA SET LISTED IN TABLE II.

# 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 IV
SP1: TRANSCRIPTION RESULTS OBTAINED WITH THE TEST DATA SET LISTED IN TABLE II.

# 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%**

C. *Results from MIREX 2009*

The Music Information Retrieval Evaluation eXchange (MIREX) is the community-based framework for the formal evaluation of Music Information Retrieval (MIR) systems and algorithms [53]. 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 IV-B2) of the

Table V
SP2: TRANSCRIPTION RESULTS OBTAINED WITH THE TEST DATA SET LISTED IN TABLE II.

# 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%**

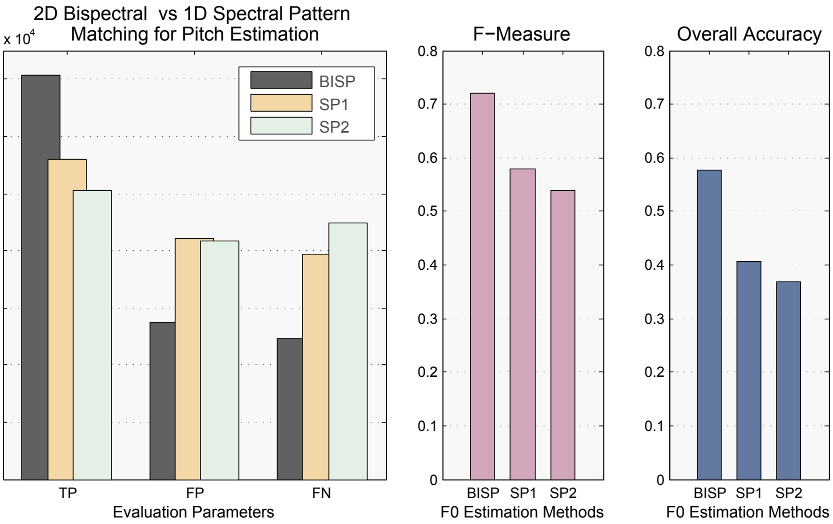
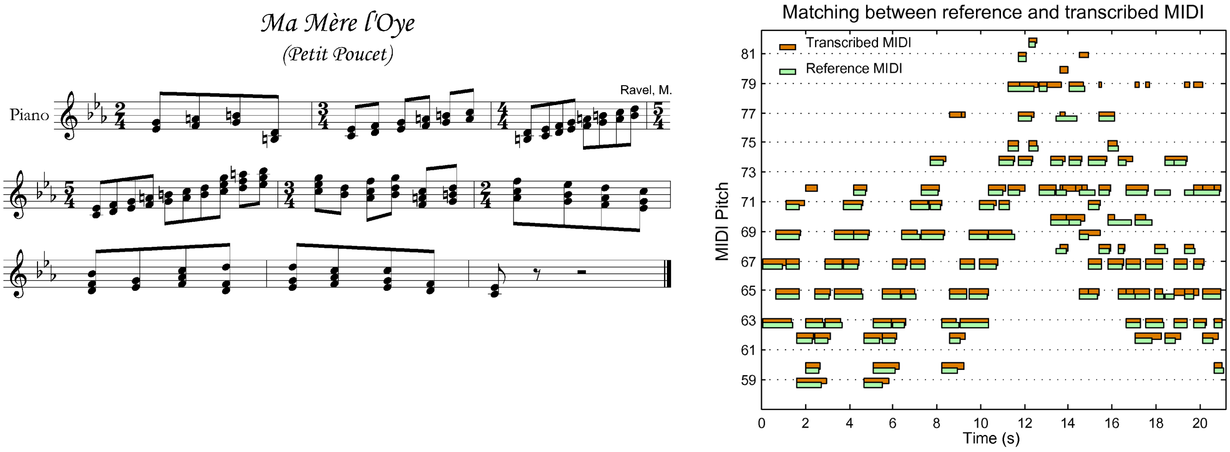


Figure 13. Results of comparison between bispectrum based (BISP) and spectrum based (SP1 and SP2) multi-$F0$ estimation methods. SP1 performs iterative pitch estimation and harmonic pattern subtraction; SP2 performs simple thresholding of cross-correlation measure.



(a)

(b)

Figure 14. Graphical (piano-roll) view of event matching between the ground truth reference and transcribed MIDI (b), related to Ravel's *Ma Mï£¡re l'Oye - Petit Poucet* (a), present in the test data set.

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 [54]. 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 III-D. 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 [55], see section *Multiple Fundamental*

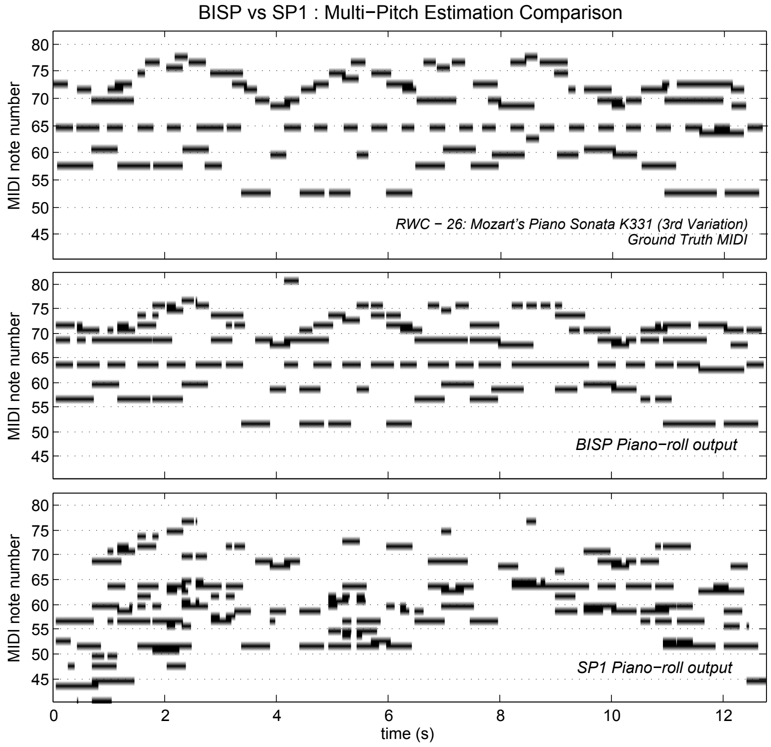


Figure 15. Graphical comparison between piano-roll output of BISP and SP1, and the reference ground truth data. The test audio example is a fragment of the 3*rd* variation of Mozart's Piano Sonata K 331.

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 Figures 16-18

V. Conclusions

In this paper a new technique for automatic transcription of real, polyphonic and multi-instrumental music has been presented. The system implements a novel front-end, 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

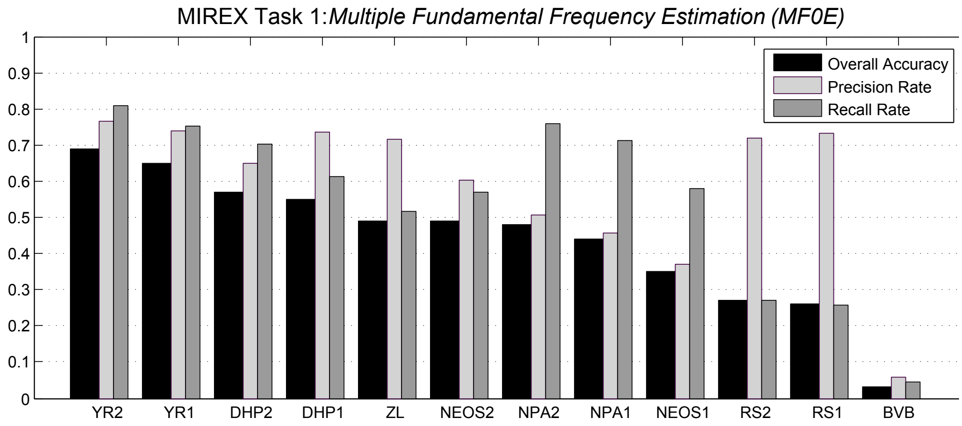


Figure 16. Results of MIREX 2009 evaluation task 1: *Multiple $F0$ estimation on a frame by frame level (MF0E)*. The system proposed in this paper has been submitted in two different versions, referred to as *NPA1* and *NPA2*, from the name of the authors.

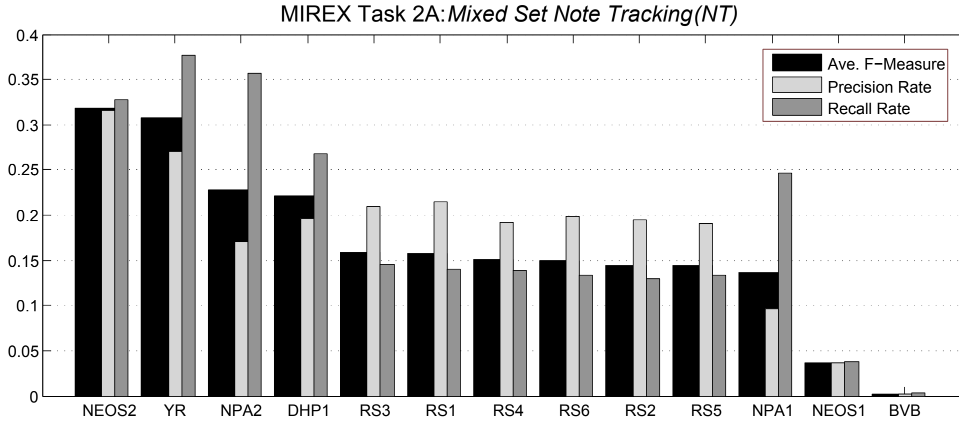


Figure 17. Results of MIREX 2009 evaluation task 2A: *Mixed-set note tracking (NT).*

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

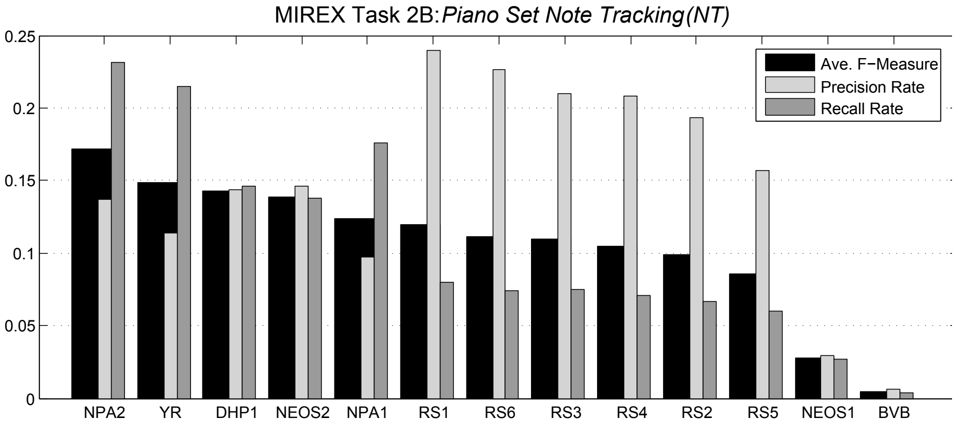


Figure 18. Results of MIREX 2009 evaluation task 2B: *Piano-only note tracking (NT)*.

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 cancellation 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.

VI. Acknowledgment

The constructive remarks of the anonymous reviewers, which have notably improved the quality of presentation of this paper, are gratefully appreciated.

References

[1] A. Noll, "Pitch determination of human speech by the harmonic product spectrum, the harmonic sum spectrum, and a maximum likelihood estimate," *Proc. of the Symposium on Computer Processing Communications*, vol. 19, pp. 779 – 797, 1969.

[2] L. Rabiner, A. Rosenberg, and C. McGonegal, "A comparative performance study of several pitch detecting algorithms," *IEEE Transactions on Acoustic, Speech and Signal Processing*, vol. 24, no. 5, pp. 399 – 418, Oct. 1976.

[3] J. Moorer, "On the transcription of musical sound by computer," *Computer Music Journal*, vol. 1, no. 4, pp. 32 – 38, 1977.

[4] M. Piszczalski and B. Galler, "Automatic music transcription," *Computer Music Journal*, vol. 1, no. 4, pp. 24 – 31, 1977.

[5] A. Klapuri, "Multiple fundamental frequency estimation based on harmonicity and spectral smoothness," *IEEE Transactions on Speech and Audio Processing*, vol. 11, no. 6, pp. 804 – 816, Nov. 2003.

[6] A. Klapuri, "A Perceptually Motivated Multiple-F0 Estimation Method," *Proc. on IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, pp. 291 – 294, Oct. 2005.

[7] 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*, vol. 5, pp. 3373 – 3376, 2001.

[8] A. Bregman, *Auditory Scene Analysis*. The MIT Press, 1990.

[9] M. Slaney and R. Lyon, "A Perceptual Pitch Detector," *IEEE International Conference on Acoustics, Speech and Signal Processing - ICASSP '90*, pp. 357 – 360, Apr. 1990.

[10] D. Ellis, *Prediction-Driven Computational Auditory Scene Analysis*. PhD thesis, Department of Electrical Engineering and Computer Science, Massachusetts Institute of Technology (MIT), USA, 1996.

[11] R. Meddis and L. O'Mard, "A Unitary Model of Pitch Perception," *The Journal of the Acoustical Society of America*, vol. 102, no. 3, pp. 1811 – 1820, 1997.

[12] T. Tolonen and M. Karjalainen, "A computationally efficient multipitch analysis model," *IEEE Transactions on Speech and Audio Processing*, vol. 8, no. 6, pp. 708 – 716, Nov. 2000.

[13] A. Klapuri, "Multipitch Analysis of Polyphonic Music and Speech Signals Using an Auditory Model," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 16, no. 2, pp. 255 – 266, 2008.

[14] M. Marolt, "Networks of adaptive oscillators for partial tracking and transcription of music recordings," *Journal of New Music Research*, vol. 33, no. 1, pp. 49 – 59, 2004.

[15] K. D. Martin, "A Blackboard System for Automatic Transcription of Simple Polyphonic Music," *Perceptual Computing Technical Report 385, MIT Media Lab*, 1996.

[16] 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*, vol. 53, no. 1/2, 2005.

[17] 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*, vol. 16, no. 4, pp. 766 – 778, 2008.

[18] H. Kameoka, T. Nishimoto, and S. Sagayama, "A Multipitch Analyzer Based on Harmonic Temporal Structured Clustering," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 3, pp. 982 – 994, 2007.

[19] J. Bello, L. Daudet, and M. B. Sandler, "Automatic Piano Transcription Using Frequency and Time-Domain Information," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 14, no. 6, pp. 2242 – 2251, 2006.

[20] C. Raphael, "Automatic transcription of piano music," *Proc. on 3rd International Conference on Music Information Retrieval*, pp. 15 – 19, 2002.

[21] M. Ryynänen and A. Klapuri, "Polyphonic Music Transcription Using Note Event Modeling," *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, pp. 319 – 322, Oct. 2005.

[22] C. Yeh, A. Roebel, X. Rodet, W. Chang, and A. Su, "Multiple F0-tracking based on a high-order HMM model," *Proc. of the* 11^{th} *Conference on Digital Audio Effects, Espoo, Finland*, 2008.

[23] 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.

[24] A. Cemgil, B. Kappen, and D. Barber, "A generative model for music transcription," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 14, no. 2, pp. 679 – 694, 2006.

[25] S. Godsill, M. Davy, and J. Idier, "Bayesian analysis of polyphonic western tonal music," *Journal of the Acoustical Society of America*, vol. 119, no. 4, pp. 2498 – 2517, 2006.

[26] 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*, vol. 15, no. 4, pp. 1283 – 1295, 2007.

[27] A. Klapuri and M. Ryynänen, "Automatic Bass Line Transcription from Streaming Polyphonic Audio," *IEEE International Conference on Acoustics, Speech and Singal Processing - ICASSP '07*, vol. 4, pp. 1437 – 1440, 2007.

[28] 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, vol. 2, pp. 757 – 760, 2000. Istanbul, Turkey.

[29] M. Goto, "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*, vol. 43, no. 4, pp. 311 – 329, 2004.

[30] 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)*, pp. 381 – 386, 2007.

[31] 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*, pp. 177 – 180, 2003. New Paltz (NY).

[32] T. Virtanen, "Monaural Sound Source Separation by Nonnegative Matrix Factorization With Temporal Continuity and Sparseness Criteria," *IEEE International Conference on Computational Intelligence for Measurement Systems and Applications*, vol. 15, no. 3, pp. 1066 – 1074, 2007.

[33] 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)*, pp. 109 – 112, 2008.

[34] S. Dubnov, N. Tishby, and D. Cohen, "Hearing Beyond The Spectrum," *Journal of New Music Research*, vol. 24, no. 4, 1995.

[35] J. J. G. De La Rosa, I. Lloret, J. E. Ruzzante, R. Piotrkowski, M. Armeite, and M. I. López Pumarega, "Higher Order Analysis of Acoustic Emission Signals," *IEEE International Conference on Computational Intelligence for Measurement Systems and Applications*, pp. 296–300, 2005.

[36] S. Abeysekera, "Multiple Pitch Estimation Of Poly-phonic Audio Signals in a frequency-lag domain using the bispectrum," *Proc. on the 2004 International Symposium on Circuits and Systems - ISCAS '04*, vol. 3, pp. 469 – 472, May 2004.

[37] A. Klapuri, *Signal Processing Methods for the Automatic Transcription of Music*. PhD thesis, Tampere University of Technology, March 2004.

[38] C. Yeh, *Multiple Fundamental Frequency Estimation of Polyphonic Rercordings*. PhD thesis, Ecole Doctorale Edite, University of Paris VI, 2008.

[39] 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)*, pp. 287 – 288, 2002.

[40] D. R. Brillinger, "An Introduction to Polyspectra," *The Annals of Mathematical Statistics*, vol. 36, no. 5, pp. 1351 – 1374, 1965.

[41] C. L. Nikias and J. M. Mendel, "Signal Processing with Higher-Order Spectra," *IEEE Signal Processing Magazine*, vol. 10, no. 3, pp. 10 – 37, 1993. ISSN: 1053-5888.

[42] C. L. Nikias and M. R. Raghuveer, "Bispectrum Estimation: A Digital Signal Processing Framework," *Proceedings of the IEEE*, vol. 75, no. 7, pp. 869 – 891, 1987.

[43] V. Chandran and S. Elgar, "A General Procedure for the Derivation of Principal Domains of Higher-Order Spectra," *IEEE Transactions on Signal Processing*, vol. 42, no. 1, pp. 229 – 233, 1994.

[44] J. C. Brown, "Calculation of a Constant Q Spectral Transform," *Journal of the Acoustical Society of America*, vol. 89, no. 1, pp. 425 – 434, 1991.

[45] 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*, pp. 1015–1019, 2006.

[46] 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)*, pp. 399 – 402, 2007.

[47] I. Bruno and P. Nesi, "Automatic Music Transcription Supporting Different Instruments," *Journal of New Music Research*, vol. 34, no. 2, pp. 139 – 149, 2005. Guest Editor: Kia Ng.

[48] J. A. Moorer, “The use of the phase vocoder in computer music applications,” *Journal of the Audio Engineering Society*, vol. 10, no. 1/2, pp. 42 – 45, 1978.

[49] M. B. Dolson, “The Phase Vocoder: A tutorial,” *Computer Music Journal*, vol. 26, no. 4, pp. 14 – 27, 2001.

[50] P. M. Brossier, *Automatic Annotation of Musical Audio for Interactive Applications*. PhD thesis, Centre for Digital Music, Queen Mary, University of London, 2006.

[51] T. Eerola and P. Toiviainen, *MIDI Toolbox: MATLAB Tools for Music Research*. Jyväskylä, Finland: University of Jyväskylä, 2004.

[52] G. E. Poliner and D. P. W. Ellis, “A Discriminative Model for Polyphonic Piano Transcription,” *EURASIP Journal on Applied Signal Processing*, vol. 2007, no. 1, 2006.

[53] J. S. Downie, “The music information retrieval evaluation exchange (2005–2007): A window into music information retrieval research,” *Acoustical Science and Technology*, vol. 29, no. 4, pp. 247 – 255, 2008.

[54] *MIREX 2009, Web Home Page URL*. http://www.music-ir.org/mirex/2009/index.php/Main_Page.

[55] *MIREX 2009 Results, Web URL*. http://www.music-ir.org/mirex/2009/index.php/MIREX2009_Results.