AUTOMATIC TRANSCRIPTION OF POLYPHONIC MUSIC BASED ON CONSTANT-Q BISPECTRAL ANALYSIS FOR MIREX 2009

Paolo Nesi and Gianni Pantaleo
Dep. of Systems and Informatics,
University of Florence, Italy
nesi@dsi.unifi.it
pantaleo@dsi.unifi.it

Fabrizio Argenti
Dep. of Electronics and Telecommunication,
University of Florence, Italy
fabrizio.argenti@unifi.it

ABSTRACT

This paper describes a new method submitted for MIREX 2009 task: Multiple Fundamental Frequency Estimation & Tracking. Specifically, it is submitted for the first two subtasks: frame by frame evaluation and event contour detection and tracking. In this framework [1], multiple-F0 estimation is carried out by means of a front-end that jointly uses a constant-Q and a bispectral (higher-order spectra) analysis of the input audio signal; subsequently, the processed signal is correlated with a fixed two-dimensional harmonic pattern. Onset and duration detection procedures are based on the combination of the constant-Q bispectral analysis with information from the signal spectrogram.

1. 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 *loudness*.

In our system, multiple-F0 estimation has been performed, on a frame by frame basis, 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, which are typically present in musical audio signals. Furthermore, in the presence of interfering sound partials, the bispectrum ideally enables to resolve the contribution of each single harmonic by performing a bidimensional cross-correlation procedure with a fixed 2D-harmonic pattern.

A computationally efficient and relatively fast method to implement the bispectrum has been realized using the constant-Q transform, which produces a multi-band frequency representation with variable resolution.

F0-tracking has been obtained by detecting note onsets and durations through the analysis of the signal spectrogram.

2. 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 of HOS can be found in [2] and [3]. The bispectrum is defined to be the third-order spectrum, being the amplitude spectrum and the power spectral density, respectively, the first and second-order ones.

Comparing higher-order spectra with simple amplitude spectral analysis, the former result to be more useful for music transcription tasks. Actually, first and second order spectra are not always able to resolve partials' overlapping due to sounds interference.

Let $x(k)$ be a real, discrete and locally stationary process of K elements; a digitalized audio signal belongs to this class of signals, apart from the fact that the locally stationary assumption is satisfied only in its sustain stage.

We used the *direct method* [2] to compute the bispectrum B_x of the signal, which is given by:

$$B_x(f_1, f_2) = X(f_1)X(f_2)X^*(f_1 + f_2), \qquad (1)$$

where $X(f)$ is the Fourier Transform of $x(k)$. In order to obtain a consistent estimation of B_x, the sequence $x(k)$ should be segmented and the estimations of the different bispectral 2D-arrays, obtained from each segment, should be averaged. In our system we performed an averaging over five consecutive frames. In Figure 1, a contour-plot of the bispectrum of a real audio signal is shown. It can be noticed that the bispectrum presents twelve mirror symmetry regions, and it can be shown that the analysis can take into consideration only a single non redundant region: hereafter, $B_x(f_1, f_2)$ will denote the bispectrum in the triangular region with vertices (0,0), (f_s/2,0) and (f_s/3,f_s/3), being f_s the sampling frequency.

The bispectrum of a synthetic monophonic audio signal composed by T harmonics, $x(n) = \sum_{k=1}^{T} 2\cos(2\pi f_k n/f_s)$, according to equation 1, is given by:

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

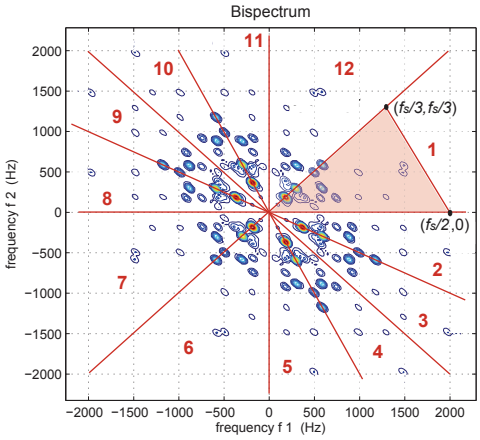

Figure 1. Contour plot of the magnitude bispectrum of a real audio signal, sampled at f_s = 4 kHz. The twelve symmetry regions are highlighted, and the one chosen for analysis is stressed.

This formula shows that every monophonic signal having fundamental frequency f_0, generates a bidimensional bispectral pattern, like the one depicted in Figure 2, characterized by peaks at the positions:

$$\{(f_i, f_i), (f_{i+1}, f_i), \ldots, (f_{T-2i-1}, f_i)\},$$

for $i = 0, 1, \ldots, \lfloor \frac{T}{2} \rfloor - 1$. We have experimentally verified

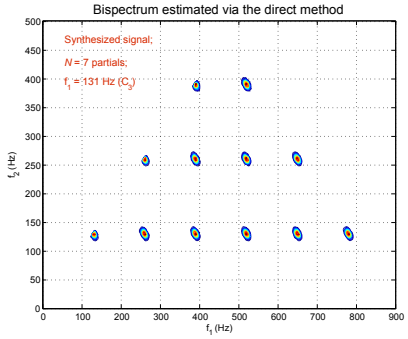

Figure 2. Bispectrum of monophonic signals (note C_3) synthesized with $T = 7$ harmonics.

that the pattern is actually generated by real audio signals.

The richness of bispectral analysis is more evident in a polyphonic context. Although the amplitude of the peaks belonging to each sound's bispectral pattern is still affected by the overlap of partials of different sounds, what we consider more interesting is the geometry of bispectral patterns local maxima. Actually, in the presence of simultaneously played notes, while in the spectral domain we may have overlap of the sounds' harmonics, in the bispectral domain each note generates almost non overlapping patterns apart from few peaks on the bisector of the quadrant (as shown in Figure 3), which can be more easily detected by a 2D correlation with a bidimensional harmonic pattern.

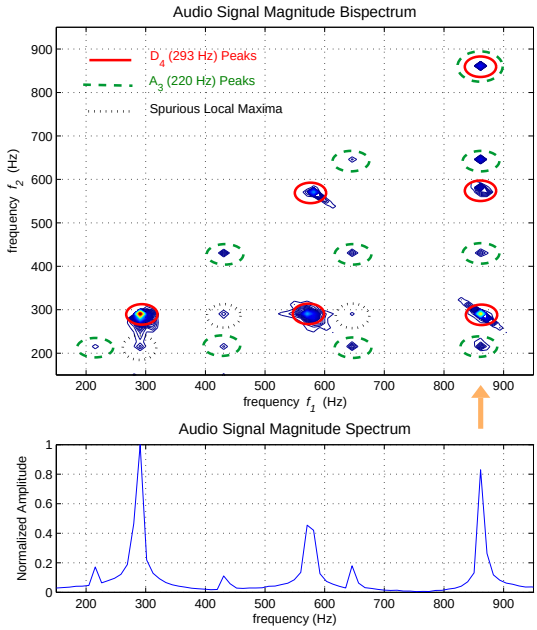

Figure 3. Detail (top figure) of the bispectrum of the bichord A_3 - D_4, played by two violins (bowed), sampled at 44 100 Hz. The arrow highlights the frequency at 880 Hz, where the two notes partials overlap. Peaks in the bispectrum are separated, which does not occur in the spectrum (shown in the bottom figure).

3. SYSTEM DESCRIPTION

3.1 Constant-Q Analysis

Figure 4 shows a block diagram of the proposed system. A constant-Q transform has been implemented for the computation of $X(f)$, which is used to calculate the bispectrum according to relation (1), since it allows a variable frequency resolution to be achieved. Such a spectral representation properly fits the exponential spacing of note frequencies. The *Octave Filter Bank* (OFB) block carries out this operation.

Constant-Q analysis [4] provides a multi-band spectral representation. Let N be the number of bands and let

$$Q_i = \frac{f^i_{max}}{B_i}, \qquad \forall i \in [1, \ldots, N],$$

where f^i_{max} is the highest (or center) frequency of the ith band and B_i is its bandwidth. A constant-Q analysis is such that $Q_i = Q$, for $i = 1, 2, \ldots, N$, where Q is a constant.

In this paper, the constant-Q analysis is implemented by a multistage low-pass filtering (a 189-tap FIR linear phase equiripple filter has been used) and subsampling by a factor 2 applied to the output of each stage. In this way, the spectrum is divided into octave-wide bands, which are later reassembled to produce the whole spectrum.

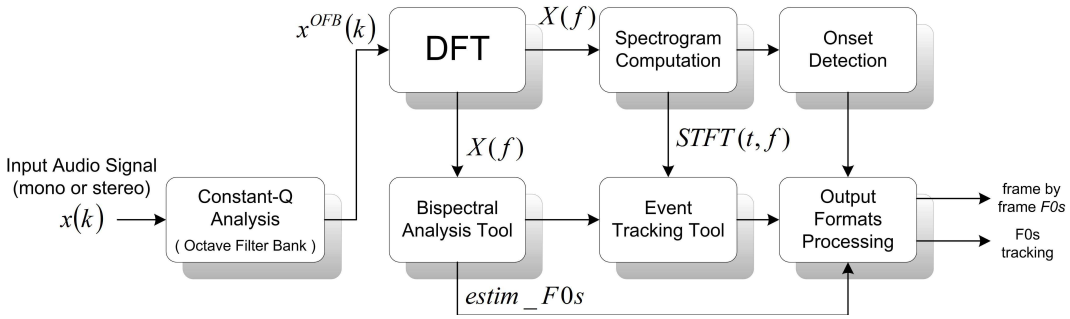

Figure 4. Block diagram of the submitted system.

3.2 Multiple-F0 Estimation

The multiple-F0 estimation is performed frame by frame by means of 2D correlation between the bispectral image of the signal and a bidimensional harmonic pattern. The geometry of the pattern is fixed, since the frequency analysis of the signal is based on a constant-Q framework, and it is dictated by the distribution of the bispectral local maxima of monophonic synthetic musical signals expressed in semitone intervals. It has been refined by studying the actual bispectrum computed on several real monophonic audio signals. The chosen pattern is the one which gave the best results.

3.3 Multiple-F0 tracking

For this subtask, information about event onsets and durations are extracted from the signal spectrogram, obtained by assembling all the spectral arrays collected for each frame. Onsets are estimated by detecting rapid spectral-energy variations over time. We used the *Modified Kullback-Liebler Distance* function that has been proved to yield the best results in [5]. Let $X_t(f)$ be the spectral value computed in the t-th frame and at frequency f, in the range of analysis $F0_{inf} \leq f \leq F0_{sup}$. The modified Kullbak-Liebler distance D_{mkl} is defined by the following relation:

$$D_{mkl}(t) = \sum_{q=F0_{inf}}^{F0_{sup}} \log\left(1 + \frac{|X_t(f)|}{|X_{t-1}(f)| + \varepsilon}\right).$$

It is assumed that $t \in [2, \ldots, N_{fr}]$, where N_{fr} is the total number of frames of the signal; ε is constant , typically $\varepsilon \in [10^{-6}, 10^{-3}]$ introduced to avoid large variations when very low energy levels are encountered, thus preventing D_{mkl} to assume large values in proximity of the release phases of sounds. D_{mkl} results to be an $(N_{fr}-1)$-element array, whose local maxima are the detected onset times.

Offsets are estimated, for each candidate F0, through a profile analysis of the events on the spectrogram row corresponding to the F0 value over time, by applying adaptative energy thresholds.

Finally, the transcribed notes are converted into the required format for MIREX tasks.

4. RESULTS

Before the submission to MIREX 2009 contest, some previous results that show the capabilities of the system have been reported [1]. Tests have been made on audio fragments extracted from the standard RWC (Real World Computing) - Classical Audio Database [6]. The evaluation has been made on a frame basis, like the MIREX subtask 1, over a data set of 119238 pitched frames, with an overall precision rate of 74.7%. The precision rate is intended as the one defined in the MIREX 2009 Wiki, i.e., the portion of correct retrieved pitches for all pitches retrieved for each frame.

5. REFERENCES

[1] Nesi, P., Argenti, F. and Pantaleo, G.: "Automatic Transcription of Real, Polyphonic and Multi-Instrumental Music Based on Constant-Q Bispectral Analysis," *Tech. Report by Dep. of Systems and Informatics, University of Florence, Italy*, Feb. 2009.

[2] Nikias, C. L. and Mendel, J. M.: "Signal Processing with Higher-Order Spectra," *IEEE Signal Processing Magazine*, Vol. 10, No. 3, pp. 10-37, 1993.

[3] Nikias, C. L. and Raghuveer, M. R.: "Bispectrum Estimation: A Digital Signal Processing Framework," *Proceedings of the IEEE*, Vol. 75, No. 7, pp. 869-891, 1987.

[4] Brown, J. C.: "Calculation of a Constant-Q Spectral Transform," *Journal of the Acoustical Society of America*, Vol. 89, No. 1, pp. 425 - 434, 1991.

[5] Brossier, P. M.: "Automatic Annotation of Musical Audio for Interactive Applications," *PHD Thesis, Centre for Digital Music, Queen Mary, University of London*, 2006.

[6] Goto, M., Hashiguchi, H., Nishimura, T. and Oka, R.: "RWC Music Database: Popular, Classical, and Jazz Music Database," *Proc. on ISMIR*, pp. 287 - 288, 2002.