Music Onset Detection Based on Resonator Time Frequency Image

Ruohua Zhou, *Member, IEEE*, Marco Mattavelli, *Member, IEEE*, and Giorgio Zoia, *Member, IEEE*

***Abstract*—This paper describes a new method for music onset detection. The novelty of the approach consists mainly of two elements: the time–frequency processing and the detection stages. The resonator time frequency image (RTFI) is the basic time–frequency analysis tool. The time–frequency processing part is in charge of transforming the RTFI energy spectrum into more natural energy-change and pitch-change cues that are then used as input elements for the detection of music onsets by detection tools. Two detection algorithms have been developed: an energy-based algorithm and a pitch-based one. The energy-based detection algorithm exploits energy-change cues and performs particularly well for the detection of hard onsets. The pitch-based algorithm successfully exploits stable pitch cues for the onset detection in polyphonic music, and achieves much better performances than the energy-based algorithm when applied to the detection of soft onsets. Results for both the energy-based and pitch-based detection algorithms have been obtained on a large music dataset.**

***Index Terms*—Audio, music, onset detection.**

I. Introduction

A MUSIC signal can be considered as a succession of musical events (notes). Music onset detection aims at finding the starting time of each note. Music onset detection plays an essential role in music signal processing and has a wide range of applications such as music transcription, beat-tracking, and tempo identification. Different sound sources (instruments) have different types of onsets that are often classified as "soft" or "hard." Hard onsets are characterized by sudden increases in energy, whereas soft onsets show more gradual changes.[1]

Hard onsets can be well detected by energy-based approaches, but the detection of soft onsets remains a challenging problem. Let us suppose that a note consists of a transient, followed by a steady-state part, and the onset of the note is at the beginning of the transient. For hard onsets, usually, energy changes are significantly larger in the transients than in the steady-state parts. Conversely, when considering the case of soft onsets, energy changes in the transients and the steady-state parts are comparable, and they do not constitute reliable cues for onset detection anymore. Consequently, energy-based approaches fail to correctly detect soft onsets. Stable pitch cues enable to segment a note into a transient and a steady-state part, because the pitch of the steady-state part often remains stable. This fact can be used to develop appropriate pitch-based methods that yield better performances, for the detection of soft onsets, than energy-based methods. However, only a few pitch-based methods have been proposed in the literature, although many approaches have already used energy information.

The aim of this article is to describe a new method for music onset detection. The method consists of two stages. The first stage involves a new time–frequency analysis tool called "resonator time–frequency image" (RTFI), which transforms the analyzed signal to a time–frequency energy spectrum. Then, the specific combination of standard DSP components (e.g., low-pass filtering, use of equal loudness curves, half-wave rectification) converts the energy spectrum into more expressive representations that show pitch and energy changes more clearly. The second stage of the method employs the representations to find onsets by using two detection algorithms: an energy-based algorithm and a pitch-based one.

State-of-the-art pitch-based detection approaches often use an independent pitch estimator to track pitch changes. However, polyphonic pitch estimation remains an unsolved problem for these approaches. Differently from them, the pitch-based detection described here does not need an independent pitch estimator, but is able to use the stable pitch cues by the new approach described in Section IV. In addition, the RTFI is implemented by the lowest order filter bank so as to be computationally efficient and be able to decompose a signal into more frequency bands than the one provided by existing multiband processing approaches.

The paper is organized as follows: Section II reports a review of related work on music onset detection, Section III briefly introduces the RTFI, Section IV describes the new onset detection method, and Section V presents and discusses the experimental results. Finally, conclusions and future work are provided in Section VI.

II. Related Work

Many different onset detection systems have been described in the literature. Typically they consist of three stages: time–frequency processing, detection function generation, and peak-picking [1]. At first, a music signal is transformed into

Manuscript received January 31, 2007; revised October 14, 2007. Current version published October 17, 2008. This work was supported in part by the Swiss Commission and Innovation (CTI) under Project 6893.2 (STILE) and by European Commission Project IST-2-511299 (AXMEDIS). The associate editor coordinating the review of this manuscript and approving it for publication was Dr. George Tzanetakis.

R. Zhou and M. Mattavelli are with the Signal Processing Institute, Swiss Federal Institute of Technology, CH-1015 Lausanne, Switzerland (e-mail: ruohua.zhou@epfl.ch; marco.mattavelli@epfl.ch).

G. Zoia was with the Signal Processing Institute, Swiss Federal Institute of Technology, CH-1015 Lausanne, Switzerland. He is now with eyeP Media, 1020 Renens, Switzerland (e-mail: giorgio.zoia@eyepmedia.com).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TASL.2008.2002042

[1]As the human ear is normally sensible to events in the range of milliseconds, the terms *sudden* and *gradual* must be understood in the same scale.

different frequency bands by using a filter-bank or a spectrogram. Then, the output of the first stage is further processed to generate a detection function at a lower sampling rate. Finally, a peak-picking operation is used to find onset times within the detection function, which is often derived by inspecting the changes in energy, phase, or pitch.

A. Energy-Based Detection

In the past, differences in a signal's envelop were used to detect note onsets. However, such an approach has been proved to be inefficient. Some researchers have found it useful to separate the analyzed signal into several frequency bands and then detect onsets across the different frequency bands. This constitutes the key element of the so-called multiband processing. For example, Goto utilizes the sudden energy changes to detect onsets in seven different frequency ranges and uses these onsets to track the music beats by a multiagent architecture [2]. Klapuri divides the signal into 21 frequency bands by the nearly critical-band filter bank [3]. Then, he uses amplitude envelopes to find onsets across these frequency bands. Duxbury *et al.* introduce a hybrid multiband processing approach for onset detection [4]. In the approach, an energy-based detector is used to detect hard onsets in the upper bands, whereas a frequency based distance measure is utilized in the lower bands to improve the detection of soft onsets.

The first-order difference of energy or amplitude has been utilized to derive a detection function. However, the first-order difference is usually not able to precisely mark onset times. According to psychoacoustic principles, a perceived increase in the signal amplitude is relative to its level. The same amount of increase can be perceived more clearly in a quiet signal. Consequently, as a refinement, the relative difference can be used to better locate onset times [3].

B. Phase-Based Detection

Phase-based approaches detect onsets by using phase information [5]. The short-time Fourier transform (STFT) of the signal can be considered to be a group of sinusoid oscillators. In the steady-state parts of the signal, the frequency of each oscillator tends to remain constant. This is not the case in the transients. Therefore, the change in frequency is an indicator of a possible onset. The second difference of the phase of the oscillator is able to identify the change in its frequency. Accordingly, statistics (e.g., mean, variance, kurtosis) on the second difference of the phase can be calculated across the range of frequencies and used to derive the detection function. To detect soft onsets, phase-based approaches perform better than standard energy-based approaches. However, they are susceptible to phase distortion and to noise introduced by the phases of low-energy components. The combination of phase and energy on the complex domain can provide more robust detection [6].

C. Pitch-Based Detection

The approaches that only use the information of energy and/or phase are not satisfactory for the detection of soft onsets. Pitch-based detection appears as a promising solution for the problem. Pitch-based approaches can use stable pitch cues to segment the analyzed signal into transients and steady-state parts, and then locates onsets only in the transients. Such approaches are expected to greatly reduce false positives. A pitch-based onset detection system is described in [7]. In the system, an independent constant-Q pitch detector provides pitch tracks that are used to find likely transitions between notes. For the detection of soft onsets, such system performs better than other state-of-the-art approaches. However, it is designed only for the onset detection of monophonic music. This article describes a new pitch-based approach that detects soft onsets of real polyphonic music.

Some approaches to onset detection are not compatible with the typical procedure described earlier. For example, a few methods use machine learning techniques to classify whether spectral frames are onsets or not [8], [9].

III. Introduction to RTFI

RTFI is a computationally efficient time–frequency representation for music signal analysis. Using the RTFI, different time–frequency resolutions can be selected by simply setting a few parameters.

A. Frequency-Dependent Time–Frequency Analysis

First a frequency-dependent time–frequency (FDTF) analysis is defined as follows:

$$\mathrm{FDTF}(t,\omega) = \int_{-\infty}^{\infty} s(\tau) w(\tau - t, \omega) e^{-j\omega(\tau - t)} d\tau. \quad (1)$$

Unlike STFT, the window function w of FDTF may depend on the analytical frequency ω. This means that time and frequency resolutions can be changed according to the analytical frequency. At the same time, (1) can also be expressed as

$$\mathrm{FDTF}(t,\omega) = s(t) * I(t,\omega) \quad (2)$$

where

$$I(t,\omega) = w(-t,\omega) e^{j\omega t}. \quad (3)$$

Equation (1) is more suitable for expressing a transform-based implementation, whereas (2) leads to a straightforward implementation of a filter bank with impulse response functions expressed in (3).

Computational efficiency and simplicity are the two essential criteria used to select an appropriate filter bank for implementing FDTF. The order of the filter bank needs to be as small as possible to reduce computational cost. The basic idea behind the filter-bank-based implementation of FDTF is to realize frequency-dependent frequency resolution by possibly varying the filters' bandwidths with their center frequencies. Therefore, the implementing filters must be simple so that their bandwidths can be easily controlled according to their center frequencies. A novel time–frequency representation is developed: the RTFI, which selects a first-order complex resonator filter bank to implement a frequency-dependent time–frequency analysis.

B. Resonator Time–Frequency Image

The RTFI can be expressed as follows:

$$\begin{aligned}\mathrm{RTFI}(t,\omega) &= s(t) * I_R(t,\omega) \\ &= r(\omega)\int_0^t s(\tau)e^{r(\omega)(\tau-t)}e^{-j\omega(\tau-t)}d\tau \end{aligned} \quad (4)$$

where

$$I_R(t,\omega) = r(\omega)e^{(-r(\omega)+j\omega)t}, \qquad t>0. \quad (5)$$

In these equations, I_R denotes the impulse response of the first-order complex resonator filter with oscillation frequency ω. The factor $r(\omega)$ before the integral in (4) is used to normalize the gain of the frequency response when the resonator filter's input frequency is the oscillation frequency. The decay factor r is dependent on the frequency ω and determines the exponent window length and the time resolution. At the same time, it also determines the bandwidth (i.e., the frequency resolution). The frequency resolution of time–frequency analysis implemented by the filter bank is defined as the equivalent rectangular bandwidth (ERB) of the implementing filter, according to the following equation:

$$B^{\mathrm{ERB}} = \int_0^\infty |H(f)|^2\, df \quad (6)$$

where $H(f)$ is the frequency response of a bandpass filter and the maximum value of $|H(f)|$ is normalized at 1 [10]. The ERB value of the digital filter can be expressed according to angle frequency as follows:

$$B^{\mathrm{ERB}}(\omega) = r(\omega)\left(0.5\pi + \arctan\left(\frac{\omega}{r(\omega)}\right)\right). \quad (7)$$

In most practical cases, the resonator filter exponent factor is nearly zero, so $arctan(\omega/r(\omega))$ can be approximated to 0.5π, and (7) is approximated as follows:

$$B^{\mathrm{ERB}}(\omega) \approx r(\omega)\cdot\pi. \quad (8)$$

The resolution B^{ERB} can be set through a map function between the frequency and the exponential decay factor r. For example, a frequency-dependent frequency resolution and corresponding r value can be parameterized as follows:

$$B^{\mathrm{ERB}}(\omega) = d + c\omega,\ d+c>0,\ c\geq 0,\ d\geq 0 \quad (9)$$

$$r(\omega) \approx B^{\mathrm{ERB}}(\omega)/\pi = (d+c\omega)/\pi. \quad (10)$$

The commonly used frequency resolutions for music analysis are special cases of the parameterized resolutions in (9). When $d=0$, the resolution is constant-Q; when $c=0$, the resolution is uniform; when $d = 2\pi\cdot 24.7$, $c = 0.1079$, the resolution corresponds to the widely accepted resolution of an auditory filter bank [11].

As the RTFI has a complex spectrum, it can be expressed as follows:

$$\mathrm{RTFI}(t,\omega) = A(t,\omega)e^{j\varphi(t,\omega)} \quad (11)$$

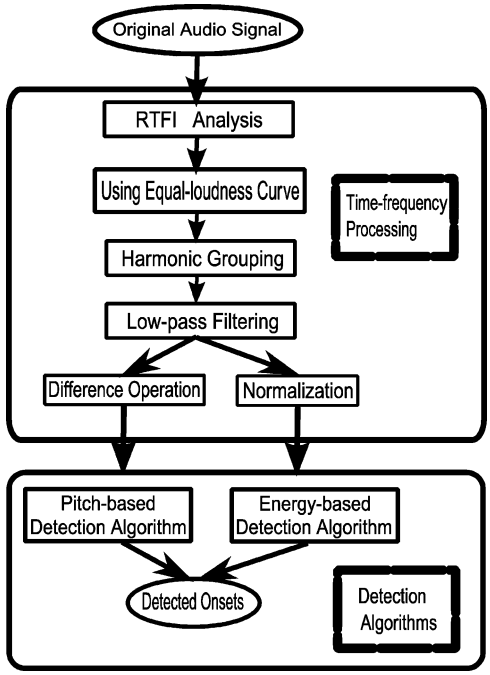

Fig. 1. Block diagram of the proposed onset detection method.

where $A(t,\omega)$ and $\varphi(t,\omega)$ are real functions

$$\mathrm{RTFI}_{\mathrm{Energy}}(t,\omega) = |A(t,\omega)|^2. \quad (12)$$

It is proposed to use a complex resonator digital filter bank for implementing a discrete RTFI. To reduce the memory usage of storing the RTFI values, the RTFI is separated into different time frames, and the average RTFI value is calculated in each time frame. The average RTFI energy spectrum can be expressed as follows:

$$A(k,\omega_m) = db\left(\frac{1}{M}\sum_{n=(k-1)M+1}^{kM} |\mathrm{RTFI}(n,\omega_m)|^2\right) \quad (13)$$

where k is the index of a frame, $db()$ converts the value to decibels, M is an integer, and the ratio of M to sampling rate is the duration time of each frame in the average process. $\mathrm{RTFI}(n,\omega_m)$ represents the value of the discrete RTFI at sampling point n and frequency ω_m.

This subsection has introduced the basic idea behind the RTFI. A detailed description of the discrete RTFI can be found in [12]. The approach to music onset detection described in this paper uses the RTFI as tool for time–frequency analysis.

IV. New Onset Detection Method

A. System Overview

The new onset detection method, reported in Fig. 1, consists of two main stages: time–frequency processing and detection algorithms.

B. Time–Frequency Processing

The selection of time–frequency resolution has an important effect on the performance of a music analysis system. The following explains how it may be reasonable to select a nearly con-

stant-Q resolution for general-purpose music signal analysis. In case of the common western music (CWM), the fundamental frequency and corresponding partials of a music note can be described as

$$f^0_{k'} = 440 \cdot \left(2^{\frac{k'-69}{12}}\right) \text{ and } f^m_{k'} = m \cdot f^0_{k'}, \quad k' \geq 1 \qquad (14)$$

using the music instrument digital interface (MIDI) note number for note k'. Supposing that the energy of every music note mainly distributes over the first 10 partials, and Energy$(f^m_{k'}) \approx 0$ for $m \geq 11$, the frequency ratio between the partials of one note and the fundamental frequency of other notes is as follows:

$$\begin{aligned} 2f^0_{k'} &= f^0_{k'+12},\ 3f^0_{k'}/f^0_{k'+19} = 0.9989 \\ 4f^0_{k'} &= f^0_{k'+24},\ 5f^0_{k'}/f^0_{k'+28} = 1.0079 \\ 6f^0_{k'}/f^0_{k'+31} &= 0.9989,\ 7f^0_{k'}/f^0_{k'+34} = 1.018 \\ 8f^0_{k'} &= f^0_{k'+36},\ 9f^0_{k'}/f^0_{k'+38} = 0.9977 \\ 10f^0_{k'}/f^0_{k'+40} &= 1.0079. \end{aligned}$$

This means that the first ten partials always either completely or in part overlap with another fundamental frequency. Since the fundamental frequencies follow an exponential law (14), most of the energy is concentrated in frequency bins, which are exponentially spaced and then equally spaced according to a logarithmic axis. This is the reason why the required resolution is constant-Q.

The monaural music signal is used as the input signal at a sampling rate of 44.1 kHz. The system applies the RTFI as the time–frequency analysis. The center frequencies of the discrete RTFI are set according to a logarithmic scale. The resolution parameters in (9) are set as $d = 0$ and $c = 0.0058$. The frequency resolution is constant-Q and equal to 0.1 semitones. Ten filters are used to cover the frequency band of one semitone. A total of 960 filters are necessary to cover the analyzed frequency range that extends from 26 Hz to 6.6 kHz. The RTFI energy spectrum is averaged to produce the RTFI average energy spectrum in units of 10 ms.

It is well known that the human auditory system reacts with different sensitivities in the different frequency bands. This fact is often described by tracing equal-loudness contours. Jensen suggests a detection function called the perceptual spectral flux [13], in which he weighs the difference frequency bands by the equal-loudness contours. Collins uses the equal-loudness contours to weight the different ERB scale bands and derive another detection function [14]. Considering these works, in the method described here, the average RTFI energy spectrum is transformed following the Robinson and Dadson equal-loudness contours, which have been standardized in the international standard ISO-226. To simplify the transformation, only an equal-loudness contour corresponding to 70 dB is used to adjust the average RTFI energy spectrum. The standard provides equal-loudness contours limited to 29 frequency bins. Then, this contour is used to get the equal-loudness contours of 960 frequency bins by cubic spline interpolation in the logarithmic frequency scale. Let us identify this equal-loudness contour as $Eq(\omega_m)$ in dB. Then, the spectrum Y can be calculated as follows:

$$Y(k, \omega_m) = A(k, \omega_m) - Eq(\omega_m) \qquad (15)$$

where ω_m represents the angle frequency of the mth frequency bin.

The music signal is structured according to notes. It is more interesting to observe that an energy spectrum is organized according to note pitches than to a single frequency component. Then, the spectrum Y is further recombined to yield the spectrum R according to a simple harmonic grouping principle:

$$R(k, \omega_m) = \frac{1}{5} \sum_{i=1}^{5} Y(k, i \cdot \omega_m). \qquad (16)$$

In practical cases, instead of using (16), the spectrum R can be easily calculated in the logarithm scale by the following approximation:

$$\begin{aligned} R(k, \omega_m) &\approx \frac{1}{5} \sum_{i=1}^{5} Y\left(k, \omega_{m+A[i]}\right) \\ A[5] &= [0, 120, 190, 240, 279]. \end{aligned} \qquad (17)$$

As shown in Table I, the deviation between the approximate and ideal values is negligible for the purposes of the spectral analysis.

In (16) and (17), $\omega_m = 2\pi \cdot 26 \cdot 2^{m/120}$, m is from 1 to 680 and the corresponding pitch range is 26 Hz to 1.32 kHz.

To reduce noise, a 5×5 mean filter is used for the low-pass filtering of the spectrum R according to the expression

$$S(k, \omega_m) = \frac{1}{25} \sum_{i=-2}^{2} \sum_{j=-2}^{2} R(k+i, \omega_{m+j}). \qquad (18)$$

To show energy changes more clearly, the spectrum D is calculated by the n-order difference of spectrum S

$$D(k, \omega_m) = S(k, \omega_m) - S(k-n, \omega_m) \qquad (19)$$

where the difference order n is set as 3 in a heuristic way

$$F(k, \omega_m) = S(k, \omega_m) - \max\left(\left(S(k, \omega_m)\right)_{m=1:N}\right) \qquad (20)$$

where N is the total number of frequency bins.

Finally, the spectra D and F together are considered as the input for the second stage of the onset detection algorithms.

TABLE I
DEVIATION BETWEEN APPROXIMATION AND IDEAL VALUES

i	1	2	3	4	5
$\frac{\omega_{m+A[i]}}{i \cdot \omega_m}$	0%	0%	-0.1%	0%	0.2%

C. Energy-Based Detection Algorithm

The energy-based detection algorithm can be described by the following expression:

$$L(k, \omega_m) = H\left(D(k, \omega_m) - \theta_1\right), \qquad \theta_1 > 0 \qquad (21)$$

where $H(x) = x + |x|/2$ is the half-wave rectifier function, followed by the detection function

$$M(k) = \frac{1}{N}\sum_{m=1}^{N} L(k, \omega_m) \tag{22}$$

where N is the total number of frequency bins in the spectrum D (19).

As shown in (21), D is subtracted by a threshold θ_1 and then half-wave rectified to produce L, which is considered to be a possible transient cue. Then, L is averaged across all frequency bins to generate the detection function M. The detection function is further smoothed by a moving average filter and a simple peak-picking operation is used to find the note onsets. In the peak-picking operation, only those peaks having values greater than threshold θ_2 are considered as the onset candidates.

Fig. 2 reports the results of the energy-based detection algorithms for a popular music example with duration time of 4 s. The vertical line in the image denotes the time labels of the true onsets. The first image is the spectrum Y according to (15). And the second image is the limited spectrum D with a threshold $\theta_1 = 3$ dB according to (21). In this example, it is obvious that most of the main energy variations only exist in the onset times. L is averaged across all the frequency channels to generate the detection function as expressed in (22); this detection function is further smoothed. The smoothed detection function is shown in the third subimage, and the blue lines in this image represent the positions of the true note onsets. Finally, a simple peak-picking operation is used with the second threshold $\theta_2 = 0.02$ dB. In addition, if there exist two successive onset candidates and the position difference between them is smaller or equal to 50 ms, only the onset candidate with the larger value is kept.

D. Pitch-Based Detection Algorithm

The energy-based detection algorithm does not perform well for detecting soft onsets. Consequently, a pitch-based algorithm has been developed to improve detection accuracy of soft onsets. A music signal can be separated into transients and steady-state parts. The basic idea behind the algorithm is to find the steady-state parts by using stable pitch cues and then look backward to locate onset times in the transients by inspecting energy changes.

In most cases, a note has a spectral structure where dominant frequency components are approximately equally spaced. The energy of a note is mainly distributed on the first several harmonic components. Let us suppose that all energies of a note are distributed in the first ten harmonic components; for a monophonic note with fundamental frequency ω, usually its spectrum Y [(15)] can have peaks $P(\omega, A_1), P(2\omega, A_2), \ldots, P(10\omega, A_{10})$ at the harmonic frequencies. $P(\omega, A)$ denotes the spectral peak that has value A at frequency ω. In most cases, the corresponding spectrum R [(16)] can present the strongest spectral peak $P(\omega, (A_1 + A_2 + A_3 + A_4 + A_5)/5)$ rightly at the fundamental frequency of the note. Accordingly, the fundamental frequency of a monophonic note can be estimated by searching the maximum peak at the note's spectrum R. For a polyphonic note, the predominant pitches can be estimated by searching the spectral

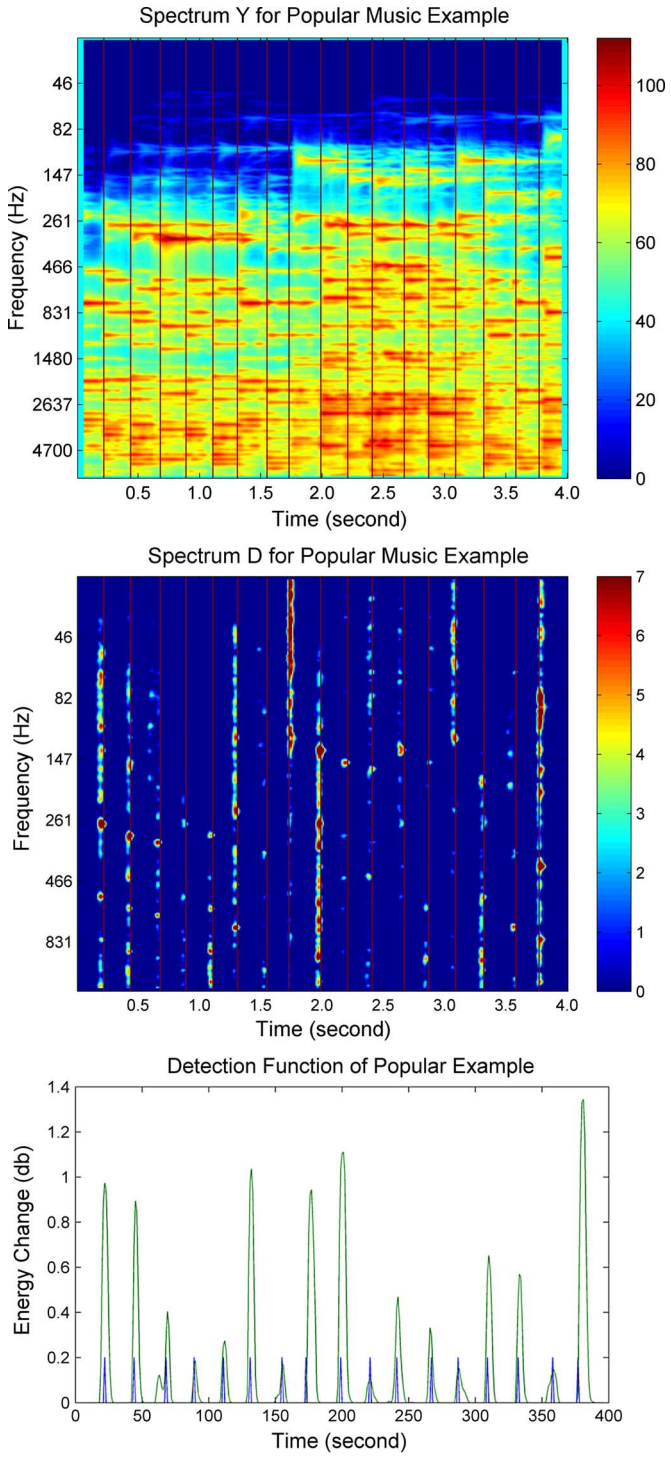

Fig. 2. Energy-based detection of a popular music example. The first image is the energy spectrum adjusted according to (15). And the second image is the limited energy spectrum with a threshold $\theta_1 = 3$ dB according to (21).

peaks that have values approaching or equal to the maximum in spectrum R. These peaks are nearly around the fundamental frequencies of the note's predominant pitches; hence, the peaks are named "predominant peaks." The spectrum F [(20)] is the relative measure of the maximum of R. Consequently, in spectrum F, the predominant peaks have values approximate or equal to 0 dB. To know how a pitch changes in a music signal, the spectrum F can be calculated in each short time frame

in units of 10 ms to get a two-dimensional time–frequency spectrum. Given the time–frequency spectrum F of a signal, if there is always a predominant peak around a frequency ω_{m1} in every time frame of a time span, this means that there is a stable pitch in the time span, and it can be assumed that the time span corresponds to a steady-state part. The time span can be called "steady time span." The images of time–frequency spectrum are very useful to validate algorithm development by visual inspection. Several different music signals and their spectrum F have been analyzed during the experimental work. It can be commonly observed that, during the steady-state part of a note, there are always one or more steady time spans, which are located just behind the note's onset. Consequently, the steady-state parts of a signal can be found by searching steady time spans in the signal's spectrum F.

The pitch-based algorithm described here consists of two steps:

1) searching possible note onsets in every frequency channel;
2) combining the detected onset candidates across all the frequency channels.

In the first step, the algorithm searches for possible pitch onsets in every frequency channel. When searching in a certain frequency channel with frequency ω_{m1}, the detection algorithm tries to find only the onset where the newly occurred pitch rightly has an approximate fundamental frequency ω_{m1}. In each frequency channel with frequency ω_{m1}, the algorithm searches the steady time spans, each of which corresponds to the steady-state part of a note having a predominant pitch with fundamental frequency ω_{m1}. Given a time–frequency spectrum $F(k,\omega_m)$, a time span $T[k1,k2]$ (in units of 10 ms) is considered to be steady if it meets the following three conditions:

$$(F(k,\omega_m))_{m=m1,k=k1:k2} > \alpha_1 \tag{23}$$

$$\max\left((F(k,\omega_m))_{m=m1,k=k1:k2}\right) > \alpha_2 \tag{24}$$

$\mathrm{Sum}(\omega_m)$ has a spectral peak at the frequency ω_{m1}

$$\mathrm{Sum}(\omega_m) = \sum_{k=k_1}^{k_2} F(k,\omega_m). \tag{25}$$

The boundary ($k1$ and $k2$) of a time span can be easily determined as follows. $F_t(k)$ is the time–frequency spectrum F in the frequency channel with frequency ω_{m1}

$$F_t(k) = (F(k,\omega_m))_{m=m1}. \tag{26}$$

Then, a two-value function $P(k)$ is defined as

$$P(k) = \begin{cases} 1, & F_t(k) \geq \alpha_1 \\ 0, & F_t(k) < \alpha_1 \end{cases} \tag{27}$$

$$G(k) = P(k) - P(k-1) \tag{28}$$

where $G(k)$ is the first-order difference of P(k). The beginning of a time span corresponds to the time at which $G(k)$ assumes the value 1 and the end of the time span is the first instant, when $G(k)$ assumes the value -1.

After all the steady time spans have been determined, the algorithm looks backward to locate onsets from the beginning of each steady time span using the spectrum D (19). For a steady time span $T[k1,k2]$, the detection algorithm locates the onset time by searching for most noticeable energy-change peak larger than the threshold α_3 in spectrum $(D(k,\omega_m))_{m=m1,k=(k1-30):k1}$. The search is done backward from the beginning of a steady time span, and the searching range is limited inside the 0.3-s window before the steady time span. The time position of this energy-change peak of the spectrum D is considered as a candidate pitch onset.

After all frequency channels have been searched, the pitch onset candidates are found and can be expressed as follows:

$$\text{Onset\_}C(k,\omega_m) \geq 0, \quad m = 1,2,3,\ldots N \tag{29}$$

where k is the index of time frame and N is the total number of the frequency channels.

If $\text{Onset\_}C(k,\omega_m) = 0$, no onset exists in the kth time frame of the mth frequency channel. If $\text{Onset\_}C(k,\omega_m) > 0$, there is an onset candidate in the kth time frame of the mth frequency channel, and the value of $\text{Onset\_}C(k,\omega_m)$ is set to the value of $D(k,\omega_m)$.

In the second step, the detection algorithm combines the pitch onset candidates across all the frequency channels to generate the detection function as follows:

$$DF(k) = \frac{1}{N}\sum_{m=1}^{N} \text{Onset\_}C(k,\omega_m). \tag{30}$$

The detection function is low-pass filtered by a moving average filter. Then, a peak-picking operation is used to find the onset times. If two onset candidates are neighbors in a 0.05-s time window, then only the onset candidate with the larger value is kept.

A bow violin excerpt is provided to exemplify the specific usage and advantage of the pitch-based algorithm. The example is a slow-attacking violin sound. Very strong vibrations can be observed from its spectrum Y reported in Fig. 3. Because of the vibrations, noticeable energy changes also exist in the steady-state parts of the signal. Therefore, the energy changes are not reliable for onset detection in this case. In the energy-based detection function [Fig. 4], it is seen that there are many spurious peaks that are, in fact, not related to the true note onsets (the dotted lines represent the positions of the true onsets). Consequently, the energy-based detection algorithm shows very poor performance in this example.

Fig. 5 illustrates the spectrum F of the example, and the vertical lines in the image denote the positions of the true onsets. It can be clearly observed that there is always at least one steady time span (white spectral line) just behind an onset position. The algorithm searches every frequency channel to find steady time spans, each of which is assumed to correspond to a steady-state part.

For example, steady time spans are searched in frequency channel 294 Hz. As shown in Fig. 6, in the spectrum F of this frequency channel, there is a time span $T[244,320]$ (in units of 10 ms). T has values larger than the threshold $\alpha_2 = -10$ dB,

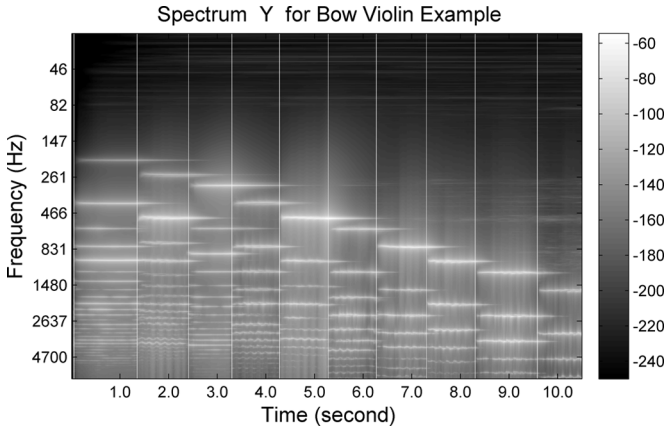

Fig. 3. Bow violin example: adjusted energy spectrum (spectrum Y).

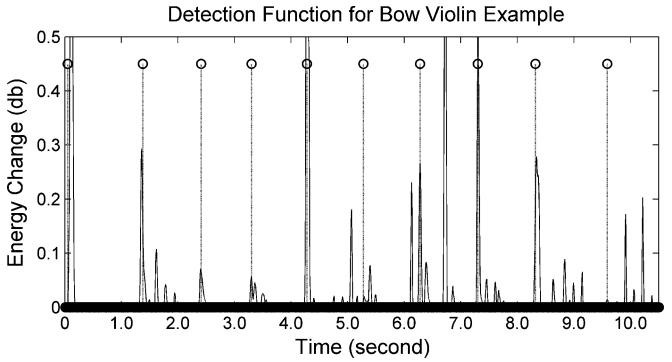

Fig. 4. Bow violin example: energy-based detection function. The dotted lines represent the positions of the true onsets.

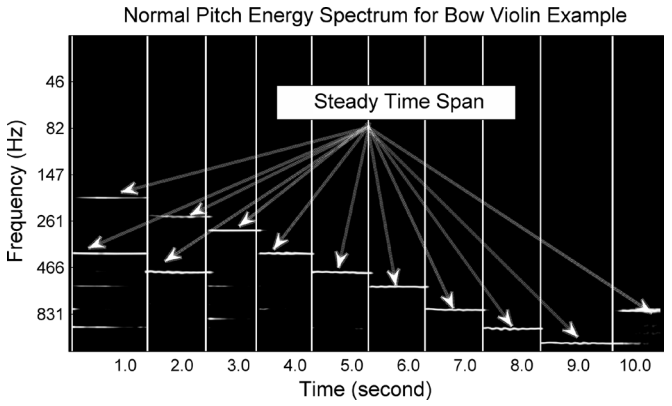

Fig. 5. Bow violin example: normal pitch energy spectrum (spectrum F). The vertical lines in the image denote the positions of the true onsets.

and presents its maximum up to 0 dB. There is also a peak rightly at a frequency of 294 Hz in the $\mathrm{Sum}^T(\omega_m)$, which is obtained by the following expression:

$$\mathrm{Sum}^T(\omega_m) = \sum_{k=244}^{320} Fv(k, \omega_m). \tag{31}$$

$Fv(k, \omega_m)$ is the time–frequency spectrum F of the bow violin example. T is considered to be a steady time span because it meets the three conditions, which were introduced earlier and

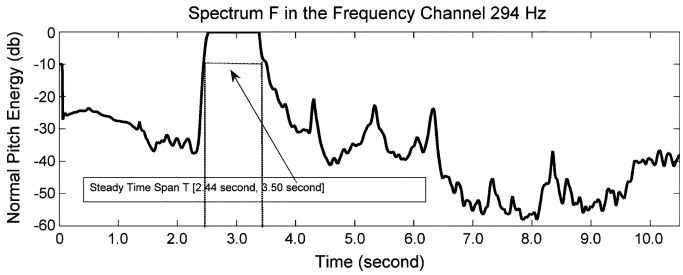

Fig. 6. Bow violin example: search of steady time spans in one frequency channel.

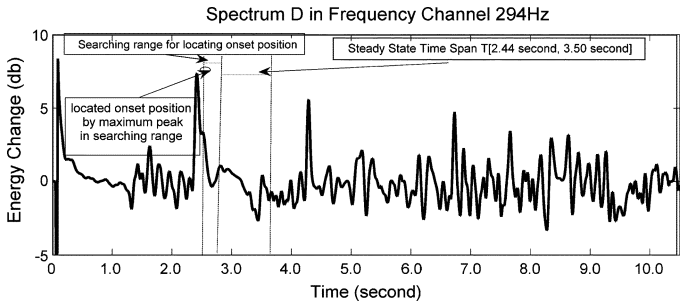

Fig. 7. Bow violin example: location of the onset position backward from steady time span.

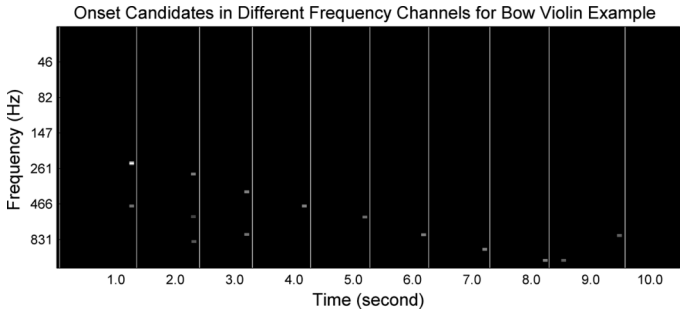

Fig. 8. Bow violin example: onset candidates in all the frequency channels. The dots denote the detected onset candidates, the vertical lines are true onsets.

used to judge if the time span is steady. Then, the detection algorithm locates the onset position by searching for a noticeable energy change peak larger than the threshold α_3 (in this example, $\alpha_3 = 2$) in the spectrum D of the frequency channel. The searching window is limited inside the 0.3-s window before the steady time span T. As shown in Fig. 7, in the spectrum D of the frequency channel 294 Hz, a peak with a value larger than the threshold α_3 is positioned nearly at the 2.42 s instant. The time position is considered as a candidate onset time.

Here, the pitch-based algorithm uses stable pitch cues to separate the signal into the transients and the steady-state parts, and searches the onset candidates by energy changes only in the transients. So, the energy changes caused by the vibrations in steady-steady parts are not considered as detection cues. The dots in Fig. 8 denote the detected onset candidates in the different frequency channels by the pitch-based detection algorithm. It can be observed that the onset candidates are nearly around the true onset positions. Finally, the detection algorithm combines the pitch onset candidates across all the frequency channels to get the final result.

TABLE II
TRAINING DATABASE

#	Content	Reference	Duration	Onset Num
1	Solo Piano	RWC-M06 TR1,2:28-2:48	20 s	110
2	Folk music	RWC-M07 TR5, 0:00-0:11	11 s	36
3	Solo Guitar, Classics	RWC-M06 TR4, 0:00-0:30	15 s	64
4	Female Singing	RWC-M07 TR11, 0:00-0:15	15 s	35
5	Country music	RWC-M07TR10, 0:00-0:15	16 s	37
6	Piano, Classical Baroque	RWC-M06TR2, 0:00-0:15	15 s	76
7	String Quartet	RWC-M06TR2, 3:04-3:28	20 s	40
8	Solo Trumpet	Commercial CD	21 s	43
9	Solo Clarinet	Commercial CD	30 s	31
10	Solo Bow Violin	Commercial CD	46 s	43
Total			3 min 29 s	515

V. EXPERIMENTS AND RESULTS

A. Performance Measures

To evaluate the detection method, the detected onset times must be compared with the reference ones. For a given reference onset at time t, if there is a detection within a tolerance time-window $[t-50\text{ ms}, t+50\text{ ms}]$, it is considered to be a correct detection (CD). If not, there is a false negative (FN). The detections outside all the tolerance windows are counted as false positives (FP). The F-measure, Recall, and Precision measures are used to summarize the results. The Precision and Recall can be expressed as

$$P = \frac{N_{\mathrm{CD}}}{N_{\mathrm{CD}} + N_{\mathrm{FP}}} \tag{32}$$

$$R = \frac{N_{\mathrm{CD}}}{N_{\mathrm{CD}} + N_{\mathrm{FN}}} \tag{33}$$

where N_{CD} is the number of correct detections, N_{FP} is the number of false positives, and N_{FN} is the number of false negatives. These two measures can be summarized by the F-measure defined as

$$F = \frac{2PR}{P+R}. \tag{34}$$

B. Datasets

Input data used for experiments are separated into two data sets: one training data set and one test data set. The training data set is used to set the optimal parameter values for the detection method.

The training data set contains ten different music files belonging to different genres. The detailed information of the data set is reported in Table II. Among them, seven files were taken from the RWC music database [15]. The positions of these files in the RWC database are reported in the Reference column of Table II. The other three files were selected from commercial CDs.

One test data set was used for the evaluation. The test database contains 30 music sequences of different genres and instruments. In total there are 2543 onsets and more than 15-min. of time duration. The reference [11] contains the detailed information about each file of the dataset, such as duration time, instruments or genres, and the number of labeled onsets. In the test data set, some files were selected from two public databases: the RWC music database and *Leveau* database [16]. The other files were collected from commercial music CDs. Similar to the MIREX 2005 [17], the music files are classified into the following classes: plucked string, sustained string, brass, winds, complex mixes. There are some differences between this data set and the MIREX data set. In MIREX, only monophonic music is contained in the classes: plucked string, sustained string, brass, and winds. Conversely, this test data set also contains polyphonic music for these classes. In addition, here the piano is considered as a single class because most of the piano music contains many hard onsets.

The onsets of the training and test data sets were labeled by an annotation tool: *Sound Onset Labellizer* [16]. Using the tool, onset labels were first annotated in the spectrogram by visual inspection, and then they were more precisely adjusted by aural feedbacks.

C. Setting Parameters

Given a test data set, better results could be achieved by setting ad-hoc parameters. Consequently, performances may be overestimated because parameters have been optimally selected to fit the testing data set. To avoid overestimation, optimal parameter values have been selected by using the training data set. The parameter values that yielded the best average F-measure on the training data set were assumed optimal.

Consequently, the energy-based algorithm selected the parameter thresholds: $\theta_1 = 3$; $\theta_2 = 0.02$ with the best average F-measure at 77.8% on the training data set, while the pitch-based algorithm selected the parameter thresholds: $\alpha_1 = -10$; $\alpha_2 = -3$; $\alpha_3 = 2$ with a best average F-measure at 92.0%. With these fixed parameter values, the detection algorithms were evaluated on the test data sets.

D. Results Comparison Between the Energy-Based and Pitch-Based Detection Algorithms

The total test results on the test data set are summarized in Table III. More detailed test results on each file can be found in [12].

In this evaluation, average F-Measure is used to evaluate detection performance. The energy-based algorithm performs better than does the pitch-based algorithm on the piano and complex music, which contains several hard onsets. The energy-based detection gains 5.0% for piano music and 8.4% for the complex music. Conversely, the pitch-based detection algorithm performs better in the brass, winds and sustained

TABLE III
RESULTS OF THE TWO PROPOSED ONSET DETECTION ALGORITHMS

Class	File Num	Onset Num	Duration	Average F-Measure (Pitch-based)	Average F-Measure (Energy-based)
Piano	2	449	2 min	92.7%	97.7%
Complex Mixes	9	690	3 min 28 s	82.6%	91.0%
Plucked String (*Guitar, Violin, Cello*)	5	421	5 min 15 s	87.6%	83.6%
Brass (*Trumpet, Horn*)	3	230	1 min 17 s	93.2%	87.8%
Winds (*Clarinet, Flute, Oboe*)	5	375	2 min 32 s	88.4%	80.8%
Sustained String (Quartet, Violin, Viola)	6	378	4 min 12 s	87.0%	44.1%

TABLE IV
RESULTS OF THE TWO DETECTION ALGORITHMS FOR PUBLICLY AVAILABLE DATABASE

Description	Onset Num	Duration	Average F-Measure (Pitch-based)	Average F-Measure (Energy-based)
Solo piano	18	15 s	94.4%	94.4%
Solo cello	61	14 s	74.1%	68.6%
Solo violin	72	15 s	85.7%	81.9%
Solo distorted guitar	19	6 s	82.1%	82.3%
Solo electric guitar	36	15 s	87.5%	83.9%
Solo steel guitar	55	15 s	96.3%	98.2%
Techno	48	6 s	73.7%	84.3%
Rock	56	15 s	82.3%	80.4%
Jazz (octet)	44	14 s	74.8%	74.7%
Jazz (contrabass)	52	11 s	66.7%	83.0%
Classic I	37	20 s	67.5%	64.5%
Classic II	11	14 s	74.0%	20.8%
Popular	27	15 s	81.3%	73.2%
Solo clarinet	22	30 s	93.2%	76.0%
Solo sax	10	10 s	80.0%	50.0%
Solo synthetic bass	22	7 s	76.9%	87.0%
Solo trumpet	52	14 s	90.0%	93.7%

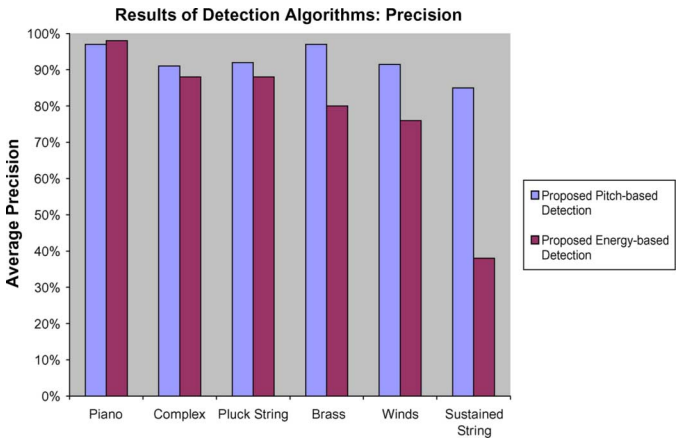

Fig. 9. Precision comparison of energy-based and pitch-based onset detections.

string, in which note onsets are considered to be softer. For the sustained string, the pitch-based algorithm gains 42.9% and greatly improves the performance from 44.1% to 87.0%. In addition, the pitch-based algorithm gains 5.4%, 7.6% for brass and winds, respectively.

A comparison between the precisions of the pitch-based and energy-based algorithms is shown in Fig. 9. The comparison clearly suggests that the pitch-based algorithm has a much better precision than the energy-based algorithm.

The pitch-based algorithm over-performs the energy-based algorithm for the detection of soft onsets. The reason of such better performance can be explained as follows. Energy-based approaches are based on the assumption that there are relatively more salient energy changes at the onset times than in the steady-state parts. In case of soft onsets, the assumption cannot stand. The significant energy changes in the steady-state parts can mislead energy-based approaches and cause many false positives. Conversely, the proposed pitch-based algorithm can first utilize stable pitch cues to separate the music signal into the transients and the steady-state parts, and then find note onsets only in the transients. The pitch-based algorithm reduces the false positives that are caused by the salient energy changes in the steady-state parts, and greatly improves the onset detection performance of the music signal with many soft onsets. Because of the reduction of false positives, it also gets a better precision.

The detailed test results of the public distributed database [16] are reported in Table IV. This makes it possible for other researchers to compare their methods with ours if they will use the same public database.

TABLE V
RESULTS OF THE TWO PROPOSED ONSET DETECTION ALGORITHMS FOR DIFFERENT TOLERANCE WINDOW

	Window 10ms	Window 20ms	Window 30ms	Window 40ms	Window 50 ms
Pitch-based detection	41.3%	70.6%	80.9%	86.0%	87.6%
Energy-based detection	79.5 %	87.0%	89.6%	90.6%	91.0%

The localization performances of the two algorithms have also been compared. To evaluate the localization capabilities, the size of tolerance window has been changed. Several music files were collected for this comparison. Both the algorithms perform well on these files when a 50-ms tolerance window is considered. Average F-measures with the different tolerance window sizes are reported in Table V. It can be observed that, when reducing the size of the tolerance window, the pitch-based algorithm has more decrease in performance than the energy-based algorithm. This suggests that the energy-based algorithm yields better localization performance than the pitch-based algorithm.

E. MIREX 2007 Results

With the combination of the energy-based and pitch-based algorithms, the method described in this paper has been evaluated in the MIREX 2007 audio onset detection task [18].

According to the overall performance, the method outperforms all other techniques which were evaluated in this task. In particular, the method performed best on the overall average F-measure, which was the primary criterion for evaluation. Different methods can perform significantly better for different classes. The method also yields the best performances for the classes: solo drum, solo brass, and solo wind. For the solo brass and solo wind, the method outperforms the second best methods by about 8% and 9%, respectively. Such performances can be contributed to the combination of the pitch-based detection.

VI. CONCLUSION AND FUTURE WORK

In this paper, a new method for onset detection in polyphonic music is described. The proposed method includes two detection algorithms classified as "energy-based" and "pitch-based." The energy-based detection algorithm yields better performance than the pitch-based algorithm for music signals with hard onsets. In addition, the energy-based algorithm also has better localization performance. However, for music signals presenting several soft onsets, energy changes are not reliable for onset detection. In such case, the energy changes in the steady-state parts can mislead an energy-based detection and produce many false positives. The pitch-based algorithm utilizes stable pitch cues and greatly reduces false positives so that higher precisions and better performances are achieved for the detection of soft onsets.

As discussed in [19] and [20], different detection methods could be used for different types of sound events to achieve better performances. Further improvements from the approach could be achieved by developing more efficient classification algorithms capable of assisting music onset detections. The classification algorithms could automatically estimate the dominant onset type for the music signal being analyzed. In such an approach, an energy-based detection algorithm should be selected when the dominant onset type has been estimated as hard; conversely, the pitch-based detection should be selected. Therefore, the adaptive combination of energy-based and pitch-based detection is expected to improve the overall performance.

As the pitch-based detection algorithm requires high-frequency resolutions so that the number of frequency channels is quite large (up to 960), the main computational cost is due to the RTFI processing. In the current implementation it requires 1.6 times of music real-time when running on a common desktop computer. The faster RTFI filter implementations could be realized by means of specific software optimizations.

REFERENCES

[1] J. P. Bello, L. Daudet, S. Abadia, C. Duxbury, M. Davies, and M. B. Sandler, "A tutorial on onset detection in music signals," *IEEE Trans. Speech Audio Process.*, vol. 13, no. 5, pp. 1035–1047, Sep. 2005.

[2] M. Goto, "An audio-based real-time beat tracking system for music with or without drum-sounds," *J. New Music Res.*, vol. 30, no. 2, pp. 159–171, 2001.

[3] A. Klapuri, "Sound onset detection by applying psychoacoustic knowledge," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP'99)*, Mar. 1999, pp. 3089–3092.

[4] C. Duxbury, M. Sandler, and M. Davies, "A hybrid approach to musical note onset detection," in *Proc. 5th Int. Conf. Digital Audio Effects (DAFX-02)*, Hamburg, Germany, 2002, pp. 33–38.

[5] J. P. Bello and M. Sandler, "Phase-based note onset detection for music signals," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Proces. (ICASSP'03)*, Hong Kong, China, 2003, pp. 49–52.

[6] J. P. Bello, C. Duxbury, M. Davies, and M. Sandler, "On the use of phase and energy for musical onset detection in the complex domain," *IEEE Signal Process. Lett.*, vol. 11, no. 6, pp. 553–556, Jun. 2004.

[7] N. Collins, "Using a pitch detector as an onset detector," in *Proc. Int. Conf. Music Inf. Retrieval*, London, U.K., Sep. 1999, pp. 100–106.

[8] M. Marolt, A. Kavcic, and M. Privosnik, "On detecting note onsets in piano music," in *Proc. IEEE Int. Conf. Mediterranean Electrotech.*, Cairo, Egypt, May. 2002, pp. 385–389.

[9] A. Lacoste and D. Eck, "A supervised classification algorithm for note onset detection," *EURASIP J. Adv. Signal Process.*, vol. 2007, 2007, article ID 43745.

[10] W. M. Hartmann, *Signals Sound and Sensation*. College Park, MD: AIP, 1997.

[11] B. C. J. Moore and B. R. Glasberg, "A revision of Zwicker's loudness model," *ACTA Acust.*, vol. 82, pp. 335–345, 1996.

[12] R. Zhou, "Feature extraction of musical content for automatic music transcription" Ph.D. dissertation, Swiss Federal Inst. of Technol., Lausanne, Oct. 2006 [Online]. Available: http://www.library.epfl.ch/en/theses/?nr=3638

[13] K. Jensen and T. H. Andersen, "Causal rhythm grouping," in *Proc. 2nd Int. Symp. Comput. Music Modeling and Retrieval*, Esbjerg, Denmark, May 2004, pp. 83–95.

[14] N. Collins, "A comparison of sound onset detection algorithms with emphasis on psychoacoustically motivated detection functions," in *Proc. AES Convention 118*, Barcelona, Spain, May 2005, paper 6363.

[15] M. Goto, H. Hashiguchi, T. Nishimura, and R. Oka, "RWC music database: Music genre database and musical instrument sound database," in *Proc. Int. Conf. Music Inf. Retrieval*, Washington, DC, Oct. 2003, pp. 229–230.

[16] P. Leveau, L. Daudet, and G. Richard, "Methodology and tools for the evaluation of automatic onset detection algorithms in music," in *Proc. 5th Int. Conf. Music Inf. Retrieval*, Barcelona, Spain, Oct. 2004, pp. 72–75.

[17] in *Proc. 1st Annu. Music Inf. Retrieval Evaluation eXchange (MIREX)*, 2005 [Online]. Available: http://www.music-ir.org/mirex2005/index.php/Audio_Onset_Detection

[18] R. Zhou and J. D. Reiss, "Music onset detection combining energy-based and pitch-based approaches," in *Proc. MIREX Audio Onset Detection Contest*, 2007 [Online]. Available: http://www.music-ir.org/mirex2007/abs/OD_zhou.pdf

[19] N. Collins, "A change discrimination onset detector with peak scoring peak picker and time domain correction," in *Proc. 1st Annu. Music Inf. Retrieval Evaluation eXchange (MIREX)*, 2005 [Online]. Available: http://www.music-ir.org/evaluation/ mirex-results/articles/onset/collins.pdf.

[20] J. Ricard, "An implementation of multi-band onset detection," in *Proc. 1st Annu. Music Inf. Retrieval Evaluation eXchange (MIREX)*, 2005 [Online]. Available: http://www.music-ir.org /evaluation/mirex-results/ articles/onset/ricard.pdf

Ruohua Zhou received the B.S. degree from the Electronics Engineering Department, Beijing Institute of Technology, Beijing, China, in 1994, the M.S. degree of engineering in microelectronics and semiconductor devices from Microelectronics R&D Center, Chinese Academy of Sciences, Beijing, in 1997, and the Ph.D. degree from the Signal Processing Laboratory (LTS), Swiss Federal Institute of Technology (EPFL), Lausanne, Switzerland, in 2006 for the thesis: "Feature extraction of musical content for automatic music transcription."

In 2001, he joined the Signal Processing Laboratory (LTS), EPFL. His research focuses on the music signal processing and music information retrieval. He is currently an Assistant Researcher in the Signal Processing Institute, EPFL.

Marco Mattavelli received the Diploma degree in electrical engineering from the Politecnico di Milano, Milan, Italy, in 1987 and the Ph.D. degree from the Signal Processing Laboratory (LTS), Swiss Federal Institute of Technology (EPFL), Lausanne, Switzerland, in 1996 for the thesis: "Motion analysis and estimation: From ill-posed discrete inverse linear problems to MPEG-2 coding."

In 1995, he was Visiting Researcher at the Center of Operational Research and Applied Mathematics, Cornell University, Ithaca, NY. He has been involved in several collaborations with industries and in the ISO/IEC JTC1/SC29/WG11 standardization activities (better known as MPEG), for which he is currently Chairman of the Implementation Study Group (ISG). His major research activities and interests include architectures and systems for audio/video coding, real-time multimedia systems, high-speed image acquisition and audio/video processing, motion analysis and estimation, neural networks for image and signal processing, and applications of combinatorial optimization to signal processing. He is the author or coauthor of more than 80 research papers and one book.

Dr. Mattavelli received the ISO/IEC Award in 1998 and in 2001 for his work and contributions on the standardization of MPEG-4.

Giorgio Zoia received the Laurea degree in Ingegneria Elettronica from Politecnico di Milano, Milan, Italy, and the Ph.D. degree in technical sciences from Ecole Polytechnique Fédérale de Lausanne (EPFL), Lausanne, Switzerland.

He is Senior Software Engineer at eyeP Media SA, Renens, Switzerland. Fields of experience include audio–visual synthesis and coding, 3-D spatialization, analysis, representations and description of sound, interaction, and intelligent user interfaces for media control. Other research interests include compilers, virtual architectures, and fast execution engines for digital audio processing.