

Submitted for;

Ruohua Zhou and Joshua D. Reiss, 'Music Onset Detection,' Chapter 12 in *Machine Audition: Principles, Algorithms and Systems*, Edited by W. Wang, Published by IGI Global, ISBN-13: 978-1615209194, July, 2010, p. 297-316, doi: 10.4018/978-1-61520-919-4.ch012

Music Onset Detection

Ruohua Zhou, Josh D Reiss

Center for Digital Music,

Electronic Engineering Department

Queen Mary University of London, UK

Email: zhou.ruohua@gmail.com, reiss.josh@elec.qmul.ac.uk

ABSTRACT

Music onset detection plays an essential role in music signal processing and has a wide range of applications. This chapter provides a step by step introduction to the design of music onset detection algorithms. The general scheme and commonly-used time-frequency analysis for onset detection are introduced. Many methods are reviewed, and some typical energy-based, phase-based, pitch-based and supervised learning methods are described in detail. The commonly used performance measures, onset annotation software, public database and evaluation methods are introduced. The performance difference between energy-based and pitch-based method is discussed. The future research directions for music onset detection are also described.

KEYWORDS

Music onset detection, soft onsets, energy-based, phase-based, pitch-based, RTFI, STFT, multi-band processing, supervised learning, steady time span

INTRODUCTION

The audio signal is often considered to be a succession of the discrete acoustic events. The term music onset detection refers to detection of the instant when a discrete event begins in a music signal. Music onset detection plays an essential role in music signal processing and has a wide range of applications such as automatic music transcription, beat-tracking, tempo identification and music information retrieval.

Different sound sources (instruments) have different types of onsets that are often classified as “soft” or “hard”. The human perception of the onset is usually related to the salient change in the sound’s pitch, energy or timbre. Hard onsets are characterized by sudden increases in energy, whereas soft onsets show more gradual change. 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, energy changes are usually 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 any more. 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, although many approaches use energy information, only a few pitch-based methods have been proposed in the literature.

We discuss general scheme for onset detection and how to develop an onset detection algorithm step by step. Many existing methods are described, and a few typical methods are to be described in detail. Performance evaluation and future research directions will also be discussed. The organization of this chapter is listed as follows. The ALGORITHMS section explains the general scheme for music onset

Submitted for;

Ruohua Zhou and Joshua D. Reiss, 'Music Onset Detection,' Chapter 12 in Machine Audition: Principles, Algorithms and Systems, Edited by W. Wang, Published by IGI Global, ISBN-13: 978-1615209194, July, 2010, p. 297-316, doi: 10.4018/978-1-61520-919-4.ch012

detection and typical algorithms for energy-based, phase-based, pitch-based and supervised learning approaches. In the section on PERFORMANCE EVALUATION, the evaluation of onset detection approaches is discussed, and several established evaluation results are presented. Finally, the section on FURTHER RESEARCH DIRECTIONS discusses possible research directions, inspired by human perception, which could be applied to the field of music onset detection.

ALGORITHMS

General Scheme

Many different onset detection systems have been described in the literature. As shown in Fig. 1, they typically consist of three stages; time-frequency processing, detection function generation, and peak-picking (Bello et al., 2005). At first, a music signal is transformed into different frequency bands by using a filter-bank or a spectrogram. For example, the Short Time Fourier Transform (STFT) and the Resonator Time Frequency Image (RTFI) are two useful time-frequency analysis tools for onset detection. 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.

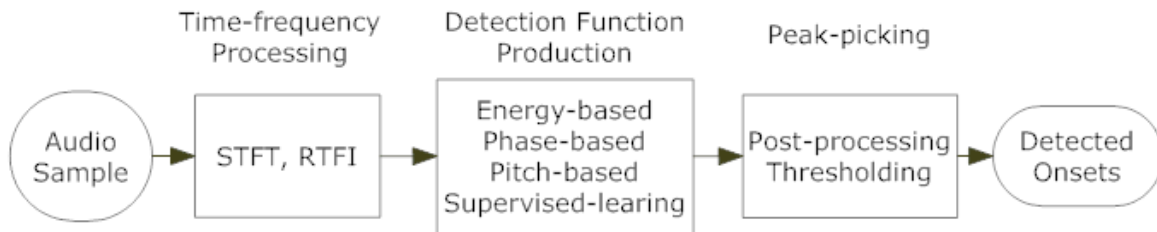


Figure 1 Three stages of music onset detection: time-frequency processing of the audio signal, producing an onset detection function, and using peak-picking to identify onsets

Time-frequency processing

Music signals are time-varying, and most of the analysis tasks require a joint time-frequency analysis. One commonly-used time-frequency analysis tool is Short Time Fourier Transform (STFT). The Fourier Transform and its inverse can transform signals between the time and frequency domains. It can make it possible to view the signal characteristics either in time or frequency domain, but not to combine both domains. In order to obtain a joint time-frequency analysis for non-stationary signals, the STFT cuts the time signal into different frames and then perform a Fourier Transform in each frame. The STFT can be defined as follows,

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

The STFT at time t is the Fourier Transform of a local signal, which is obtained by multiplication of a signal $s(t)$ and a short window function $w(\tau - t)$ centered at time t . When moving the window along the signal time axis, we can calculate the STFT at different time instants and obtain a joint time-frequency analysis. The discrete STFT of a signal $s(n)$ can be defined as follows,

$$X_m(n) = \sum_{l=-N/2}^{N/2-1} s(nh + l) w(l) e^{-2j\pi ml} \quad (2)$$

where $w(l)$ is a N -point window, h is the hop size and m denotes the different frequency bins.

Submitted for;

Ruohua Zhou and Joshua D. Reiss, 'Music Onset Detection,' Chapter 12 in Machine Audition: Principles, Algorithms and Systems, Edited by W. Wang, Published by IGI Global, ISBN-13: 978-1615209194, July, 2010, p. 297-316, doi: 10.4018/978-1-61520-919-4.ch012

Another useful time-frequency representation for music signal analysis is the Resonator Time Frequency Image (RTFI). To better explain the RTFI, first a frequency-dependent time-frequency (FDTF) analysis is defined as follows:

$$FDTF(t, \omega) = \int_{-\infty}^{\infty} s(\tau)w(\tau - t, \omega)e^{-j\omega(\tau-t)}d\tau \quad (3)$$

Unlike the STFT, the window function w of the 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, Eq. (3) can also be expressed as convolution:

$$FDTF(t, \omega) = s(t) * I(t, \omega) \quad (4)$$

where

$$I(t, \omega) = w(-t, \omega)e^{j\omega t} \quad (5)$$

Equation (3) is more suitable for expressing a transform-based implementation, whereas Eq.(4) leads to a straightforward implementation of a filter bank with impulse response functions expressed in Eq. (5). On the one hand, we can also group the terms in the STFT definition differently to obtain the filter bank interpretation for STFT. However, there are two main differences between the band-pass filter implementations of STFT and FDTF.

As illustrated in Fig. 2, one difference is that, in the implementation of the STFT, the bandwidth of the band-pass filter is kept fixed and independent of its centre frequency ω . Instead, for the FDTF, the bandwidth of the band-pass filter can be changed according to centre frequency ω . Another difference is that, in the implementation of the STFT, the output of every band-pass filter centered at ω is then demodulated to zero frequency. Such a demodulation process does not exist in the implementation of the FDTF.

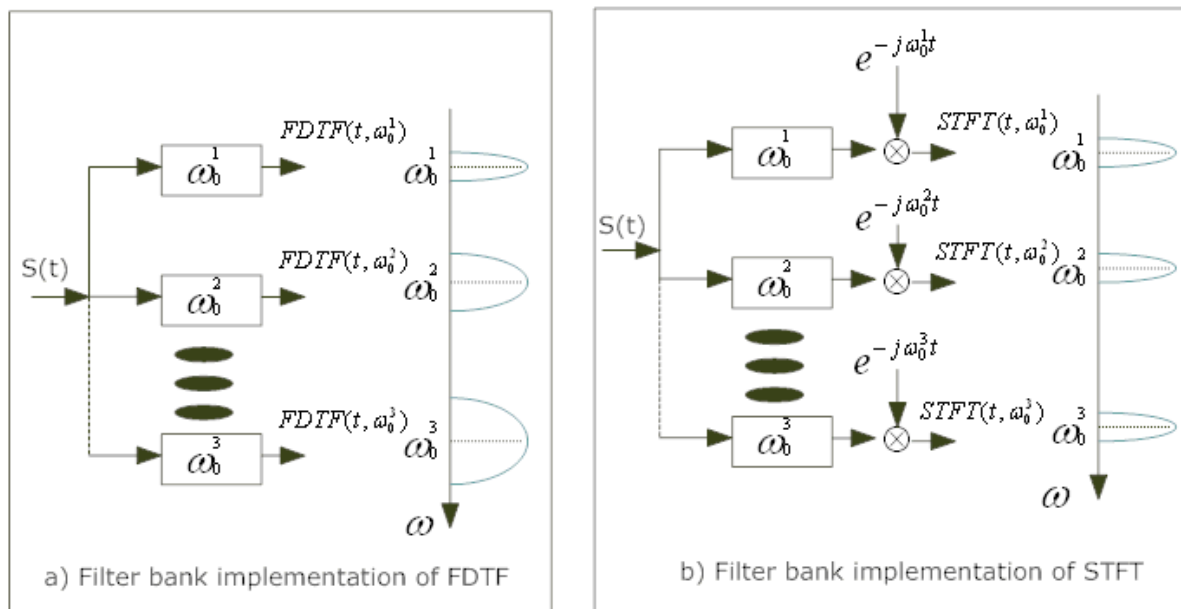


Figure 2 Filter bank implementation of STFT and FDTF

Computational efficiency and simplicity are the two essential criteria used to select an appropriate filter bank for implementing the 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

Submitted for;

Ruohua Zhou and Joshua D. Reiss, 'Music Onset Detection,' Chapter 12 in Machine Audition: Principles, Algorithms and Systems, Edited by W. Wang, Published by IGI Global, ISBN-13: 978-1615209194, July, 2010, p. 297-316, doi: 10.4018/978-1-61520-919-4.ch012

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. The RTFI selects a first-order complex resonator filter bank to implement a frequency-dependent time-frequency analysis.

The RTFI can be expressed as follows:

$$\begin{aligned} RTFI(t, \omega) &= s(t) * I_R(t, \omega) \\ &= r(\omega) \int_0^t s(\tau) e^{r(\omega)(\tau-t)} e^{-j\omega(\tau-t)} d\tau \end{aligned} \quad (6)$$

where

$$I_R(t, \omega) = r(\omega) e^{(-r(\omega)+j\omega)t}, \quad t > 0 \quad (7)$$

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 Eq. (6) 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:

$$B^{ERB} = \int_0^\infty |H(\omega)|^2 d\omega \quad (8)$$

where $H(\omega)$ is the frequency response of a band pass filter and the maximum value of $|H(\omega)|$ is normalized at 1 (Hartmann, 1997). The ERB value of the digital filter can be expressed according to angular frequency as follows:

$$B^{ERB}(\omega) = r(\omega)(0.5\pi + \arctan(\frac{\omega}{r(\omega)})) \quad (9)$$

In most practical cases, the resonator filter exponent factor is nearly zero, so $\arctan(\omega/r(\omega))$ can be approximated to 0.5π , and Eq. (9) is approximated as follows:

$$B^{ERB}(\omega) \approx r(\omega) \cdot \pi \quad (10)$$

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:

$$B^{ERB}(\omega) = d + c\omega, \quad d + c > 0, \quad c \geq 0, d \geq 0 \quad (11)$$

$$r(\omega) \approx B^{ERB}(\omega) / \pi = (d + c\omega) / \pi \quad (12)$$

The commonly used frequency resolutions for music analysis are special cases of the parameterized resolutions in Eq.(11). When $d=0$, the resolution is constant-Q; when $c=0$, the resolution is uniform; when $d = 24.7 \times 2\pi = 155.2$, $c=0.1079$, the resolution corresponds to the widely-accepted resolution of an auditory filter bank (Moore & Glasberg, 1996).

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

$$RTFI(t, \omega) = A(t, \omega) e^{j\varphi(t, \omega)} \quad (13)$$

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

$$RTFI_{Energy}(t, \omega) = |A(t, \omega)|^2 \quad (14)$$

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

Submitted for;

Ruohua Zhou and Joshua D. Reiss, 'Music Onset Detection,' Chapter 12 in *Machine Audition: Principles, Algorithms and Systems*, Edited by W. Wang, Published by IGI Global, ISBN-13: 978-1615209194, July, 2010, p. 297-316, doi: 10.4018/978-1-61520-919-4.ch012

expressed as follows:

$$A(n, \omega_m) = 20 \log_{10} \left(\frac{1}{M} \sum_{l=(n-1)M+1}^{nM} |RTFI(l, \omega_m)|^2 \right) \quad (15)$$

where n is the index of a frame, M is an integer, and the ratio of M to sampling rate is the duration time of each frame in the average process. $RTFI(l, \omega_m)$ represents the value of the discrete RTFI at sampling point l and frequency ω_m . A detailed description of the discrete RTFI can be found in the references (Zhou, 2006; Zhou & Mattavelli, 2007).

Energy-based detection

In early methods, the amplitude envelope of a music signal was used to derive the detection function. The amplitude envelope can be constructed by rectifying and smoothing the signal:

$$C(n) = \sum_{k=-N/2}^{N/2-1} |s(n+k)|w(k) \quad (16)$$

where $w(k)$ is N-point window. A variation on this is to derive the detection function from local energy, instead of amplitude.

$$E(n) = \sum_{k=-N/2}^{N/2-1} s^2(n+k)w(k) \quad (17)$$

In the simplest case, differences in a signal's amplitude or energy are used to detect note onsets. However, such an approach has been shown to give poor results. Researchers have found it useful to separate the analyzed signal into a number of bands and then detect onsets across the different frequency bands. This constitutes the key element of multi-band 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 (Goto, 2001). Klapuri divides the signal into 21 frequency bands by the nearly critical-band filter bank (Klapuri, 1999). Then, he uses amplitude envelopes to find onsets across these frequency bands. Duxbury et al. introduce a hybrid multiband processing approach for onset detection (Duxbury, Sandler and Davies, 2002). 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. Wang et al. proposes a novel approach based on music spectra magnitude (Wang et al., 2008). They first decompose the music spectra magnitude into linear temporal bases by nonnegative matrix factorization, and then use the linear temporal bases to construct detection functions.

The spectrum is used to measure the energy change in the time-frequency domain. The first-order difference of energy has been utilized to evaluate the energy change and 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 (Klapuri, 1999). Consequently, as a refinement, the relative difference can be used to better locate onset times.

For example, when the STFT is selected as the time-frequency processing tool, the spectrum D can be defined as follows,

$$D_m(n) = 20 \log_{10} (|X_m(n)|^2) - 20 \log_{10} (|X_m(n-1)|^2) \quad (18)$$

where $X_m(n)$ is the discrete STFT of the input signal.

Submitted for;

Ruohua Zhou and Joshua D. Reiss, 'Music Onset Detection,' Chapter 12 in Machine Audition: Principles, Algorithms and Systems, Edited by W. Wang, Published by IGI Global, ISBN-13: 978-1615209194, July, 2010, p. 297-316, doi: 10.4018/978-1-61520-919-4.ch012

The commonly-used energy-based detection methods can be generalized as follows,

$$M(n) = \frac{1}{N} \sum_{m=1}^N H(D_m(n)) \quad (19)$$

where $H(x) = (x + |x|)/2$ is the half-wave rectifier function, N is the total number of frequency bins in the spectrum D , and M is the detection function. 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 a threshold θ are considered as the onset candidates.

In this paragraph, it is explained that why time-frequency decomposition can greatly improve the energy-based detection methods, and why the positive flux is necessary. In many cases, during the note transition time, the current note may decay and decrease energy, while the new note may begin and increase in energy, and the change in total energy is not noticeable. If the two successive notes have different pitch, then the energy-increasing and energy-decreasing will occur in the different frequency channels after an appropriate time-frequency decomposition. As only the energy-increasing is the useful cue for onset time, energy-decreasing during the decay of the former note should be ruled out. When summing energy change across all frequency channels to derive a detection function, the type of flux is limited to the positive flux. Accordingly, only the energy increase during the onset time of the latter note is considered.

Phase-based detection

As opposed to the standard energy-based detection, phase-based detection makes use of the spectral phase information as its source of information. The STFT can also be considered as complex band-pass filter banks with equal bandwidth, and the STFT coefficient $X_m(n)$ denotes the output of the m^{th} filter. In cases where there is only one sinusoid component passing the m^{th} band-pass filter and at the same time this sinusoid component is stable, the output of the m^{th} filter must have a nearly constant frequency. Therefore, the difference between two consecutive unwrapped phase values of $X_m(n)$ must remain nearly constant:

$$\varphi_m(n) - \varphi_m(n-1) \approx \varphi_m(n-1) - \varphi_m(n-2) \quad (20)$$

where $\varphi_m(n)$ is defined as the 2π -upwrapped phase of the STFT coefficient $X_m(n)$. The phase deviation $\Delta\varphi_m(n)$ can also be defined as:

$$\Delta\varphi_m(n) = \varphi_m(n) - 2\varphi_m(n-1) + \varphi_m(n-2) \approx 0 \quad (21)$$

During the steady-state portion of the signal, $\Delta\varphi_m(n)$ is nearly equal to zero. During the transient portion, the frequency of $X_m(n)$ is not constant, and $\Delta\varphi_m(n)$ tends to be large. The analysis can be extended to the distribution of phase deviations of all frequency bins of the STFT (Bello & Sandler, 2003). During the steady-state part of the signal, the distribution is pointed. In the transient part, the corresponding distribution is flat and wide. These observations can be quantified by calculating the inter quartile range and the kurtosis coefficient of the distribution.

Phase-based onset detection has demonstrated better performance in the detection of the soft onset than standard energy-based methods. However, it is susceptible to phase distortion and to phase noise introduced by the phases of low energy components.

Pitch-based detection

The approaches that use only energy and/or phase information are not satisfactory for the detection of soft onsets. Pitch-based detection appears as a promising solution for the problem. Pitch-based approaches can

Submitted for;

Ruohua Zhou and Joshua D. Reiss, 'Music Onset Detection,' Chapter 12 in Machine Audition: Principles, Algorithms and Systems, Edited by W. Wang, Published by IGI Global, ISBN-13: 978-1615209194, July, 2010, p. 297-316, doi: 10.4018/978-1-61520-919-4.ch012

use stable pitch cues to segment the analyzed signal into transients and steady-state parts, and then locate onsets only in the transients. A pitch-based onset detection system is described in the reference (Collins, 1999). 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 a system performs better than other state-of-the-art approaches. However, it is designed only for onset detection in monophonic music. As polyphonic pitch estimation remains a largely unsolved problem, this makes it difficult to construct a pitch-based onset detection for polyphonic music. This subsection describes a pitch-based approach that detects onsets in real polyphonic music (Zhou, Mattavelli and Zoia, 2008).

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 Eq. (11) 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 10ms.

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 (Jensen & Andersen, 2004), 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 (Collins, 2005a). 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:

$$Y(k, \omega_m) = A(k, \omega_m) - Eq(\omega_m) \quad (22)$$

where ω_m represents the angular frequency of the m^{th} frequency bin. $A(k, \omega_m)$ denotes the average RTFI energy spectrum that is defined in Eq. (15).

The music signal is structured according to notes. It is interesting to observe that an energy spectrum is organized according to note pitches rather 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:

$$R(k, \omega_m) = \frac{1}{5} \sum_{i=1}^5 Y(k, i \cdot \omega_m) \quad (23)$$

In practical cases, instead of using Eq. (23), the spectrum R can be easily calculated on the logarithmic scale by the following approximation:

$$R(k, \omega_m) \approx \frac{1}{5} \sum_{i=1}^5 Y(k, \omega_{m+A[i]}) \quad (24)$$

$$A[5] = [0, 120, 190, 240, 279] \quad (25)$$

In Eq. (24) and (25), $\omega_m = 2\pi \cdot 26 \cdot 2^{m/120}$, m is from 1 to 680 and the corresponding pitch range is 26Hz to 1.32kHz. To reduce noise, a 5×5 mean filter is used for the low-pass filtering of the spectrum R according to the expression:

Submitted for;

Ruohua Zhou and Joshua D. Reiss, ‘Music Onset Detection,’ Chapter 12 in *Machine Audition: Principles, Algorithms and Systems*, Edited by W. Wang, Published by IGI Global, ISBN-13: 978-1615209194, July, 2010, p. 297-316, doi: 10.4018/978-1-61520-919-4.ch012

$$S(k, \omega_m) = \frac{1}{25} \sum_{i=-2}^2 \sum_{j=-2}^2 R(k+i, \omega_{m+j}) \quad (26)$$

To show energy changes more clearly, the spectrum D is calculated by the n^{th} -order difference of spectrum S .

$$D(k, \omega_m) = S(k, \omega_m) - S(k-n, \omega_m) \quad (27)$$

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

$$F(k, \omega_m) = S(k, \omega_m) - \max((S(k, \omega_m))_{m=1:N}) \quad (28)$$

where N is the total number of frequency bins, and spectrum F is the relative measure of the maximum of S .

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

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 in the first several harmonic components. Let us suppose that all the energy of a note is distributed in the first 10 harmonic components. For a monophonic note with fundamental frequency ω , usually its spectrum Y (Eq.22) can have peaks $P(\omega, A_1)$, $P(2\omega, A_2) \dots P(10\omega, A_{10})$ 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 (Eq. 23) presents the strongest spectral peak $P(\omega, (A_1+ A_2+ A_3+ A_4+ A_5)/5)$ at the fundamental frequency of the note. Accordingly, the fundamental frequency of a monophonic note can be estimated by searching for the maximum peak in the note’s spectrum R . For a polyphonic note, the predominant pitches can be estimated by searching the spectral peaks that have values approaching or equal to the maximum in spectrum R . These peaks are near the fundamental frequencies of the note’s predominant pitches; hence, the peaks are named “predominant peaks”. The spectrum F (Eq. 28) is the relative measure of the maximum of R . Consequently, in F , the predominant peaks have values approximately equal to 0 dB. To know how a pitch changes in a music signal, F can be calculated in each short time frame in units of 10ms 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 the time-frequency spectrum are very useful to validate algorithm development by visual inspection. Several different music signals and their spectra 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 after 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 occurring pitch rightly has an approximate fundamental frequency ω_{m1} . In each

Submitted for;

Ruohua Zhou and Joshua D. Reiss, 'Music Onset Detection,' Chapter 12 in Machine Audition: Principles, Algorithms and Systems, Edited by W. Wang, Published by IGI Global, ISBN-13: 978-1615209194, July, 2010, p. 297-316, doi: 10.4018/978-1-61520-919-4.ch012

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 10ms) is considered to be steady if it meets the following three conditions:

$$(F(k, \omega_m))_{m=m1, k=k1:k2} > \alpha_1 \quad (29)$$

$$\max((F(k, \omega_m))_{m=m1, k=k1:k2}) > \alpha_2 \quad (30)$$

where α_1 and α_2 are two thresholds.

The third condition is that $Sum(\omega_m)$ has a spectral peak at the frequency ω_{m1} ,

$$Sum(\omega_m) = \sum_{k=k1}^{k2} F(k, \omega_m) \quad (31)$$

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} .

$$F_t(k) = (F(k, \omega_m))_{m=m1} \quad (32)$$

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

$$P(k) = \begin{cases} 1, & F_t(k) \geq \alpha_1 \\ 0, & F_t(k) < \alpha_1 \end{cases} \quad (33)$$

$$G(k) = P(k) - P(k-1) \quad (34)$$

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 (Eq. 27). 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:

$$Onset_C(k, \omega_m) \geq 0, m=1, 2, 3, \dots, N, \quad (35)$$

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

If $Onset_C(k, \omega_m)=0$, no onset exists in the k^{th} time frame of the m^{th} frequency channel. If $Onset_C(k, \omega_m)>0$, there is an onset candidate in the k^{th} time frame of the m^{th} frequency channel, and the value of $Onset_C(k, \omega_m)$ 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,

$$DF(k) = \frac{1}{N} \sum_{m=1}^N Onset_C(k, \omega_m) \quad (36)$$

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

Submitted for;

Ruohua Zhou and Joshua D. Reiss, 'Music Onset Detection,' Chapter 12 in *Machine Audition: Principles, Algorithms and Systems*, Edited by W. Wang, Published by IGI Global, ISBN-13: 978-1615209194, July, 2010, p. 297-316, doi: 10.4018/978-1-61520-919-4.ch012

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, it is seen that there are many spurious peaks that are, in fact, not related to the true note onsets. Consequently, the energy-based detection algorithm shows very poor performance in this example.

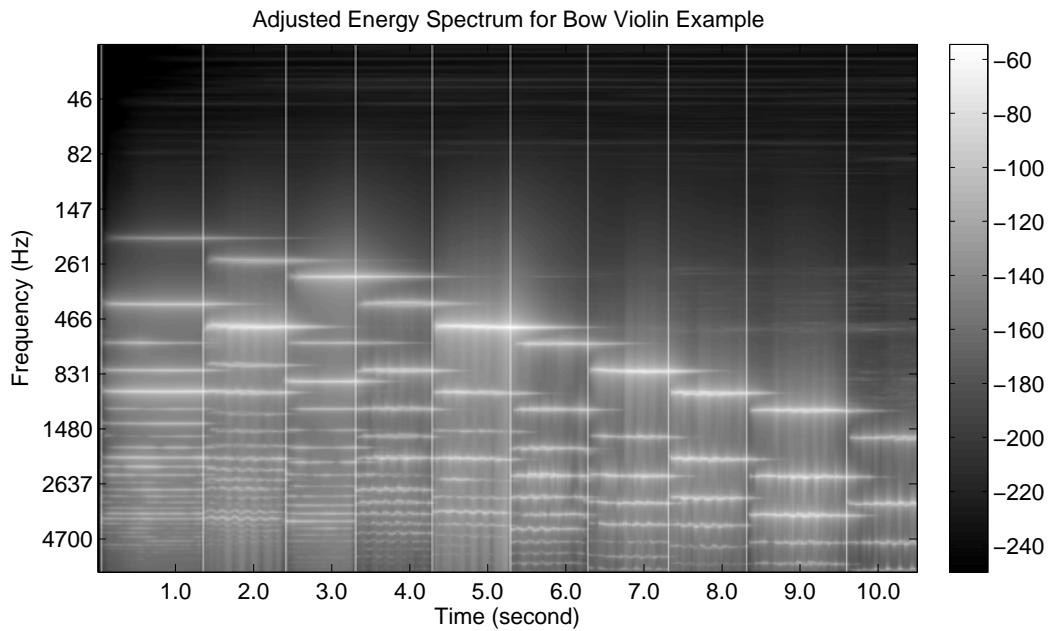


Figure 3 Bow Violin example: adjusted energy spectrum (spectrum Y).

Submitted for;

Ruohua Zhou and Joshua D. Reiss, 'Music Onset Detection,' Chapter 12 in Machine Audition: Principles, Algorithms and Systems, Edited by W. Wang, Published by IGI Global, ISBN-13: 978-1615209194, July, 2010, p. 297-316, doi: 10.4018/978-1-61520-919-4.ch012

Normal Pitch Energy Spectrum for Bow Violin Example

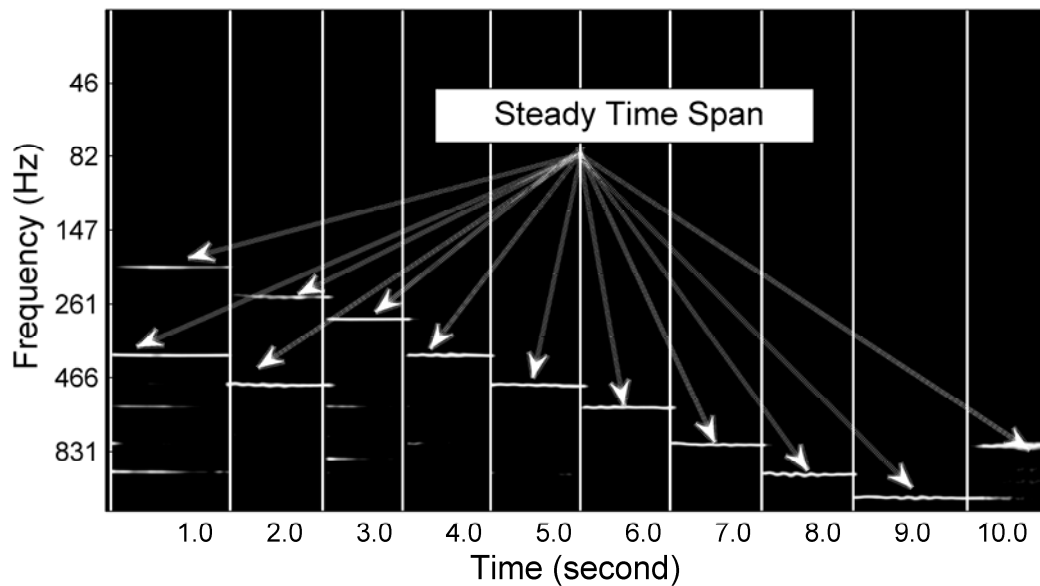


Figure 4 Bow Violin example: normal pitch energy spectrum (spectrum F).
The vertical lines in the image denote the positions of the true onsets.
The Figure is adapted from (Zhou, Mattavelli and Zoia, 2008)

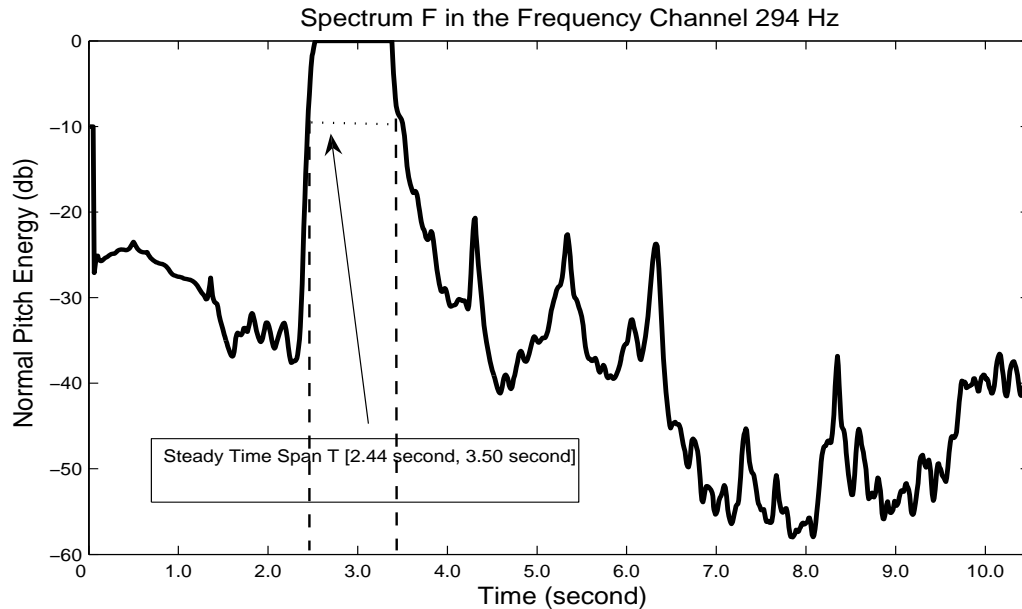


Figure 5 Bow Violin example: search of steady time spans in one frequency channel.
The Figure is adapted from (Zhou, Mattavelli and Zoia, 2008)

Submitted for;
Ruohua Zhou and Joshua D. Reiss, 'Music Onset Detection,' Chapter 12 in Machine Audition: Principles, Algorithms and Systems, Edited by W. Wang, Published by IGI Global, ISBN-13: 978-1615209194, July, 2010, p. 297-316, doi: 10.4018/978-1-61520-919-4.ch012

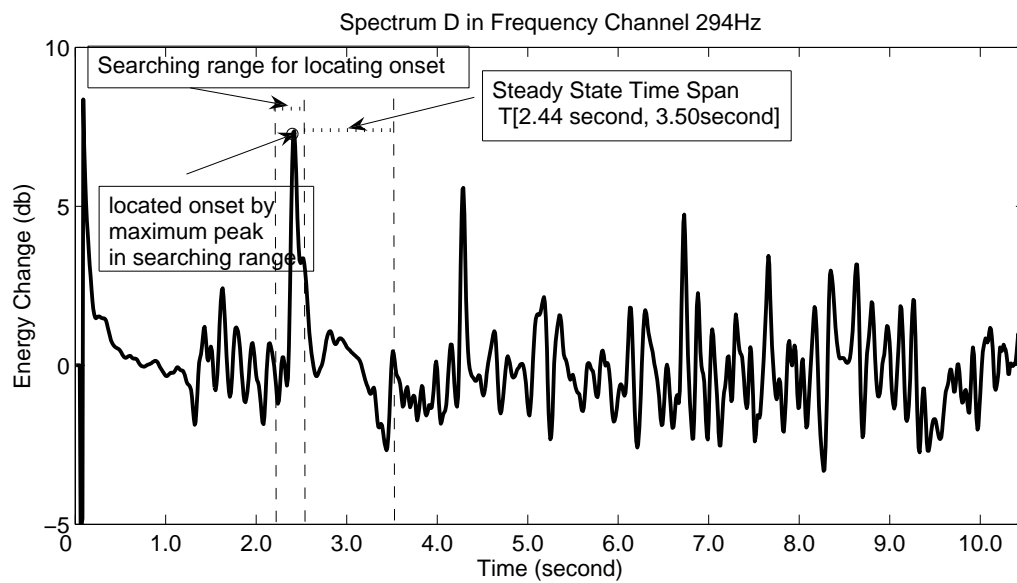


Figure 6 Bow Violin example: location of the onset position backward from steady time span.
The Figure is adapted from (Zhou, Mattavelli and Zoia, 2008)

Submitted for;

Ruohua Zhou and Joshua D. Reiss, 'Music Onset Detection,' Chapter 12 in Machine Audition: Principles, Algorithms and Systems, Edited by W. Wang, Published by IGI Global, ISBN-13: 978-1615209194, July, 2010, p. 297-316, doi: 10.4018/978-1-61520-919-4.ch012

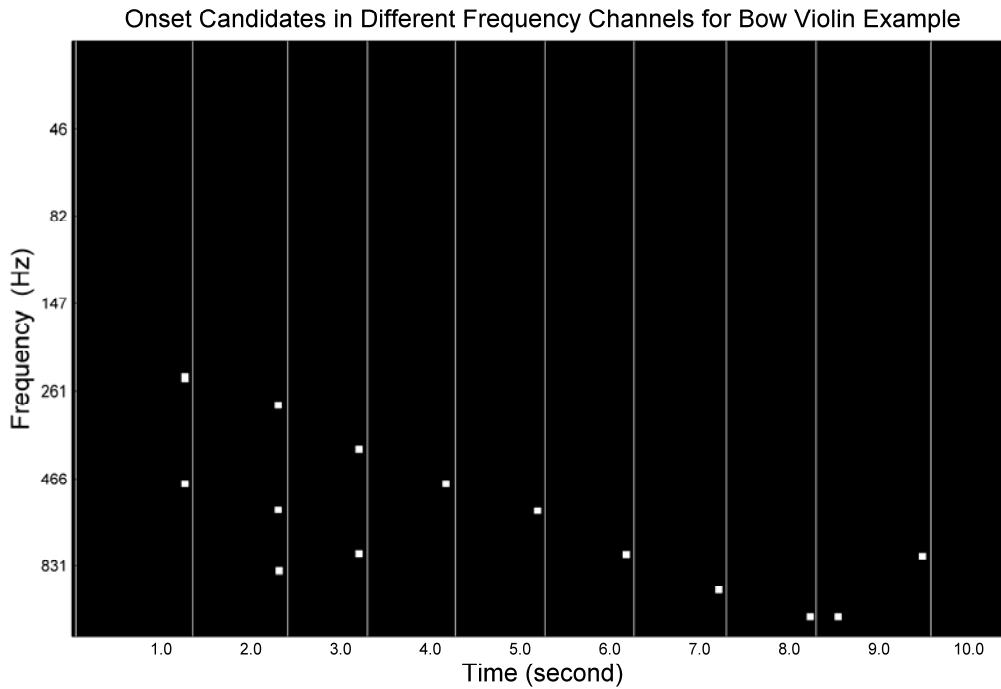


Figure 7 Bow Violin example: onset candidates in all the frequency channels. The dots denote the detected onset candidates, the vertical lines are true onsets.

Fig. 4 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. 5, 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 $\alpha_2 = -10dB$, and presents its maximum up to 0 dB. There is also a peak rightly at a frequency of 294Hz in the $Sum^T(\omega_m)$, which is obtained by the following expression:

$$Sum^T(\omega_m) = \sum_{k=244}^{320} Fv(k, \omega_m) \quad (37)$$

$Fv(k, \omega_m)$ 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 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. 6, in the spectrum D of the frequency channel 294Hz, a peak with a value larger than the threshold α_3 is positioned near 2.42 s. 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

Submitted for;

Ruohua Zhou and Joshua D. Reiss, 'Music Onset Detection,' Chapter 12 in Machine Audition: Principles, Algorithms and Systems, Edited by W. Wang, Published by IGI Global, ISBN-13: 978-1615209194, July, 2010, p. 297-316, doi: 10.4018/978-1-61520-919-4.ch012

dots in Fig. 7 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.

Supervised learning

Some approaches to onset detection are not compatible with the typical procedure described earlier, and they are based on machine learning (Marolt, et al., 1999; Chuan & Elaine, 2008; Davy & Godsill, 2002). Generally speaking, these methods first perform a time-frequency analysis, such as STFT or constant-Q analysis for the music signal. Then, corresponding energy and phase information is used to classify every frame as being onset or non-onset by supervised learning algorithms.

A typical supervised learning method is described as follows (Lacoste & Eck, 2007).



Figure 8 A music onset detection method based on supervised learning

As shown in Fig. 8, the method first performs a STFT analysis to produce a spectrogram, which is then used as the input variables to a feed-forward neural network (FNN). The FNN is trained to evaluate how much every spectrogram frame can be classified as being onset. The outputs of the network are then treated with a simple peak-picking algorithm based on a moving average. The FNN has two hidden layers and a single neuron in the output layer. The tanh activation function is used for hidden layers, and the logistic sigmoid activation function is selected for the output layer. The conjugate gradient descent is employed as learning function.

Although every spectrogram frame could also be classified as a simple 0 or 1 (onset/nononset), it is more reasonable to use smooth changes to model the transition between onset and nononset. Accordingly, the target trace for every frame can be a combination of the peaked Gaussians, which is centred on the labelled onset time.

$$T(t) = \sum_i \exp\left(-\frac{(t - \tau_i)^2}{\sigma^2}\right) \quad (38)$$

Where $T(t)$ represents the target trace value of the spectrogram frame at the time position of t , τ_i is the i^{th} labelled onset time and σ is the width of the peak, chosen to be 10 milliseconds. In the training phase, the network produces a predicted value for every time step. As shown in the following equation, the error function is the sum of the squared error over all input patterns.

$$E = \sum_j (T(t_j) - O(t_j))^2 \quad (39)$$

where $O(t_j)$ represents the output of the neural network for pattern j .

The goal of the neural network is to predict an onset trace at the time of every time step. The input variables to the neural network are selected from the spectrogram of the input signal. The number of input variables should be as small as possible, because the large size of input variables makes it very difficult to train the neural network. Accordingly, when predicting an onset trace for a certain frame, it is reasonable that only the spectrogram points near the frame are selected as the input variables. In the method, the

Submitted for;

Ruohua Zhou and Joshua D. Reiss, 'Music Onset Detection,' Chapter 12 in Machine Audition: Principles, Algorithms and Systems, Edited by W. Wang, Published by IGI Global, ISBN-13: 978-1615209194, July, 2010, p. 297-316, doi: 10.4018/978-1-61520-919-4.ch012

input variables are some spectrogram points within a time-frequency window centred on the predicted frame. To reduce the number of input variables, the spectrum points in the input window are randomly sampled both on the time and frequency axis. Uniform sampling is used along frequency axis, and the values along the time axis are sampled according to a normal distribution. More than 200 input variables are sufficient to yield a good result.

The input window has a width of 200 milliseconds. The window height is 90% of the height of the spectrogram. Thus, when moving the input window across the frequency axis, there are multiple similar input windows for the same time step and this will yield a more robust model. Accordingly, the network has multiple predicted values for a single frame, and the multiple values are merged to generate a detection function

Finally, a simple peak picking algorithm is used to find onset times from the detection function. The peak picking algorithm can be expressed as follows.

$$p(t) = d(t) - u(t) \quad (40)$$

where

$$u(t) = g * d(t) \quad (41)$$

where g is the Gaussian filter, d denotes the detection function, and p is the peak trace.

In the peak picking algorithm, a high-pass filter is used to isolate the interesting fast change information from the slow change, which is considered to be not related to onsets. The high pass filter is implemented by subtracting the Gaussian-filtered signal from the original signal. In the final step, in the peak trace p , each peak with a value larger than threshold K is considered to represent an onset time. The position of the onset is calculated as the centre of mass of all points inside the peak. The optimal threshold K can be selected from training samples.

PERFORMANCE EVALUATION

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 $[t-50ms, t+50ms]$, 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:

$$P = \frac{N_{CD}}{N_{CD} + N_{FP}} \quad (42)$$

$$R = \frac{N_{CD}}{N_{CD} + N_{FN}} \quad (43)$$

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:

$$F = \frac{2PR}{P + R} \quad (44)$$

Onset Annotation

It is important to construct a reference database with reliable annotated onset times. Onset detection is a subjective task, usually there is no perfect reference for onset times. In most cases, the onset times need to be hand labeled. For percussive onsets, the reference times can be precisely labeled by visualizing the

Submitted for;

Ruohua Zhou and Joshua D. Reiss, 'Music Onset Detection,' Chapter 12 in *Machine Audition: Principles, Algorithms and Systems*, Edited by W. Wang, Published by IGI Global, ISBN-13: 978-1615209194, July, 2010, p. 297-316, doi: 10.4018/978-1-61520-919-4.ch012

waveform in audio display software. In the spectrogram of the analyzed signal, the concurrent fast energy-increasing of different harmonic components of a music note can be often visualized at the onset time. Accordingly, the spectrogram is also a very useful tool to help annotate the onset times, although it is not very precise. The most precise and flexible method for onset annotation is to listen to signal slice with the support of visualizations.

The Sound Onset Labelizer (SOL) is a free onset annotation software for research purpose (Leveau, Daudet and Richard, 2004). This software provides a user-friendly interface to annotators. Both the waveform and the spectrogram of the analyzed signal can be seen in the screen of the software's GUI. The spectrogram and waveform parts have the same time axis. Using the tool, onset labels can be first annotated in the spectrogram by visual inspection, and then they can be more precisely adjusted by aural feedbacks. More reliable reference onset times can be achieved by cross-validation among the different expert annotators.

As onset annotation is a very time consuming task, it is more efficient to share some public database among different researchers. Leveau et al. provide a reliable public distributed database for onset detection, which contains 17 short music sequences in different music instruments and genres (Leveau, Daudet and Richard, 2004). The validated onset labels for more than 700 corresponding onsets are also freely distributed.

Performance Comparison and MIREX Evaluation

It is quite difficult to make performance comparison among the different methods for onset detection, because there is a shortage of public database available, and different researchers use different test databases. There is a direct comparison between an energy-based method and a pitch-based one in the literature (Zhou, Mattavelli and Zoia, 2008). Both methods are based on the RTFI time-frequency analysis. The pitch-based method has been described in the above section. The same test dataset 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-minutes of time duration. The music files are classified into the following classes: piano, plucked string, sustained string, brass, winds, complex mixes. Here the piano is considered as a single class because most of the piano music contains many hard onsets.

The total test results on the test dataset are summarized in Table I. 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 string, in which note onsets are considered to be softer. For the sustained string, the pitch-based algorithm gains

Table 1 Comparison between the energy-based and pitch-based methods

	Piano	Complex Mixes	Plucked String (Guitar, Violin, Cello)	Brass (Trumpet, Horn)	Winds (Clarinet, Flute, Oboe)	Sustained String (Quartet, Violin, Viola)
Average F-Measure (Pitch-based)	92.7%	82.6%	87.6%	93.2%	88.4%	87.0%
Average F-Measure (Energy-based)	97.7%	91.0%	83.6%	87.8%	80.8%	44.1%

Submitted for;

Ruohua Zhou and Joshua D. Reiss, 'Music Onset Detection,' Chapter 12 in *Machine Audition: Principles, Algorithms and Systems*, Edited by W. Wang, Published by IGI Global, ISBN-13: 978-1615209194, July, 2010, p. 297-316, doi: 10.4018/978-1-61520-919-4.ch012

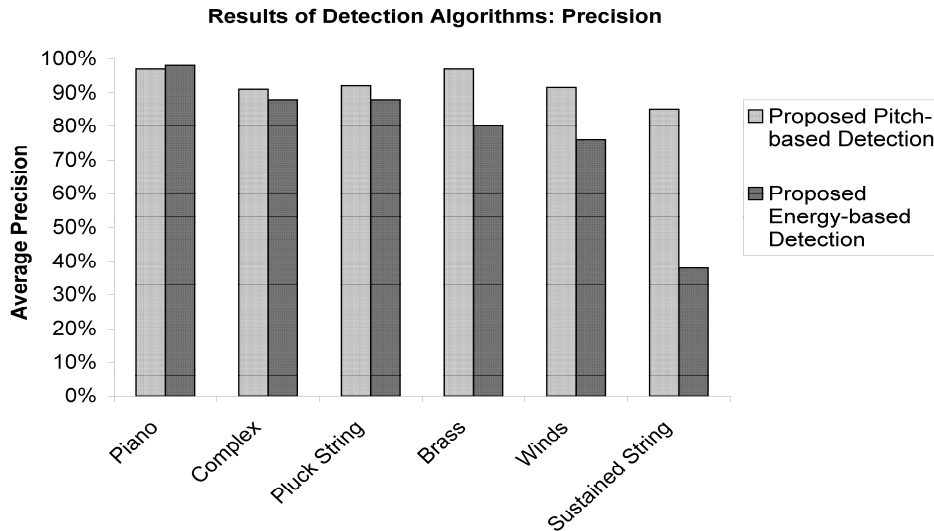


Figure 9 Precision comparison of energy-based and pitch-based onset detections.

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 methods is shown in Fig. 9. The comparison clearly suggests that the pitch-based method has a much better precision than the energy-based method.

The pitch-based method overperforms the energy-based algorithm for the detection of soft onsets. The reason for improved 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 can not 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.

Similar to pitch-based onset detection, phase-based method has a good performance for the detection of soft onsets. But phase-based detection is very susceptible to phase distortion and to phase noise introduced by the phases of low energy components. Compared to pitch-based detection, phase-based detection shows very low performance for the onset detection in real polyphonic music signal.

Submitted for;

Ruohua Zhou and Joshua D. Reiss, 'Music Onset Detection,' Chapter 12 in *Machine Audition: Principles, Algorithms and Systems*, Edited by W. Wang, Published by IGI Global, ISBN-13: 978-1615209194, July, 2010, p. 297-316, doi: 10.4018/978-1-61520-919-4.ch012

Music Information Retrieval Evaluation Exchange (MIREX) provides a formal evaluation framework (MIREX, 2005). From MIREX 2005 to MIREX 2007, several onset detection methods have been evaluated on the same dataset in the audio onset detection task. The test dataset contains 85 music sequences of different genres and instruments. In total there are 14-minutes of time duration. The dataset contains four different classes such as solo drum, solo polyphonic pitched instruments, solo monophonic pitched instruments and complex mixtures. In addition, the monophonic pitched instruments class is subdivided into 5 different subclasses: brass, wind, sustained strings, plucked strings, bars and bells, singing voice.

In the MIREX 2005~2007 onset detection tasks, many methods have been evaluated. Almost all the methods fail to detect onsets in singing voice. This result suggests that it is quite difficult to develop a general onset detection method, which can work well for both singing voice and the other music. Most of the submitted methods are energy-based and show a poor performance on the detection of the classes: solo brass, solo wind, solo sustained string. These classes usually contain a large number of soft onsets. The significant energy changes in the steady-state parts can mislead energy-based approaches and cause many false positives. As explained before, the pitch-based detection can clearly outperform the energy-based detection for the detection of soft onsets. In addition, the phase information is also useful to improve the detection for soft onsets. The Zhou and Reiss method combines the energy-based and pitch-based detection so that the method has much better performance on the solo brass class and solo wind class than other methods (Zhou and Reiss, 2007). The Lee method combines energy and phase information and achieves the best performance on the sustained string class (Lee, Shiu and Kuo, 2007). According to the average F-measure, the overall performance difference between the first four best methods [Zhou and Reiss, 81%; Lee 80%; Lacoste and Eck, 80%; Robel, 80%] is minor (Zhou and Reiss, 2007; Lee, Shiu and Kuo, 2007; Lacoste & Eck, 2005; Robel, 2007).

FUTURE RESEARCH DIRECTIONS

Since the nature of music signals is quite varied, no single method is suitable for all music signals. different detection methods could be used for different types of sound events to achieve better performances (Collins, 2005b; Ricards, 2005). Inspired by the human perceptual system, further improvements may arise by combining multiple simpler methods using a voting mechanism. These improvements could be achieved by developing more efficient classification algorithms capable of assisting music onset detection. The classification algorithms would automatically estimate the dominant onset type for the music signal being analyzed. Then, the adaptive combination of different methods is expected to improve the overall performance. In addition, as the human ear performs much better for onset detection than an automatic detection method, computational auditory models such as a loudness model could also play an important role in the further research.

REFERENCES

Bello, J.P. & Sandler, M. (2003). Phase-based note onset detection for music signals. *Proc. IEEE Int. Conf. Acoustics, Speech, and Signal Processing (ICASSP-03)*, Hong Kong, pp.49-52.

Bello, J.P., Daudet, L., Abadia, S., Duxbury, C., Davies, M. & Sandler, M.B. (2005). A tutorial on onset detection in music signals, *IEEE Trans. Speech and Audio Signal Processing*, vol. 13, pp. 1035–1047.

Submitted for;

Ruohua Zhou and Joshua D. Reiss, 'Music Onset Detection,' Chapter 12 in *Machine Audition: Principles, Algorithms and Systems*, Edited by W. Wang, Published by IGI Global, ISBN-13: 978-1615209194, July, 2010, p. 297-316, doi: 10.4018/978-1-61520-919-4.ch012

Chuan, C. & Elaine C. (2008). *Audio Onset Detection Using Machine Learning Techniques: The Effect and Applicability of Key and Tempo Information*. University of Southern California Computer Science Department Technical Report No. 08-895.

Collins, N.(1999). Using a pitch detector as an onset detector. *Proc. International Conf. On Music Information Retrieval*.

Collins, N. (2005a). A comparison of sound onset detection algorithms with emphasis on psychoacoustically motivated detection functions. *AES Convention 118*, Barcelona.

Collins, N. (2005b). A change discrimination onset detector with peak scoring peak picker and time domain correction. . *MIREX 2005 audio onset detection contest*: from <http://www.music-ir.org/evaluation/mirex-results/articles/onset/collins.pdf>

Davy, M. & Godsill S. (2002). Detection of abrupt spectral changes using support vector machines an application to audio signal segmentation," in *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP'02)*, vol. 2, pp. 1313-1316, Orlando, Fla, USA.

Duxbury, C., Sandler, M., & Davies, M. (2002). A hybrid approach to musical note onset detection. *Proc. 5th International Conf. Digital Audio Effects (DAFX-02)*, Hamburg, Germany.

Goto, M. (2001). An audio-based real-time beat tracking system for music with or without drum-sounds. *Journal of New Music Research*, vol. 30, No. 2, pp.159–171.

Hartmann, W.M. (1997). *Signals Sound and Sensation*, American Institute of Physics Press.

Jensen, K. & Andersen, T.H. (2004). Causal rhythm grouping. *Proc. 2nd International Symposium on Computer Music Modeling and Retrieval*, Esbjerg, Denmark.

Klapuri, A. (1999). Sound onset detection by applying psychoacoustic knowledge. *Proc. IEEE International Conf. Acoustics, Speech, and Signal Processing (ICASSP-99)*, pp. 3089–3092.

Lacoste, A. & Eck, D. (2005) Onset detection with artificial neural network for MIREX 2005. *MIREX 2005 audio onset detection contest*: from <http://www.music-ir.org/evaluation/mirex-results/articles/onset/lacoste.pdf>

Lacoste, A. & Eck, D. (2007). A supervised classification algorithm for Note Onset Detection. *EURASIP Journal on Advances in Signal Processing*, vol. 2007, Article ID 43745, 13 pages.

Lee, W., Shiu, Y. and Kuo, C. (2007) Musical onset detection with linear predication and joint features. *MIREX 2007 audio onset detection contest*: from http://www.music-ir.org/mirex/2007/abs/OD_lee.pdf

Leveau, P., Daudet, L. & Richard, G. (2004). Methodology and Tools for the evaluation of automatic onset detection algorithms in music. *Proc. 5th International Conf. On Music Information Retrieval*, Barcelona.

Submitted for;

Ruohua Zhou and Joshua D. Reiss, 'Music Onset Detection,' Chapter 12 in *Machine Audition: Principles, Algorithms and Systems*, Edited by W. Wang, Published by IGI Global, ISBN-13: 978-1615209194, July, 2010, p. 297-316, doi: 10.4018/978-1-61520-919-4.ch012

Marolt, M., Kavcic, A. & Privosnik, M. (1999). Neural networks for note onset detection in piano music. *Proc. International Conf. On Computer Music*.

MIREX (2005), First Annual Music Information Retrieval Evaluation eXchange (MIREX), 2005. from http://www.music-ir.org/mirex2005/index.php/Audio_Onset_Detection .

Moore, B.C.J. & Glasberg, B.R. (1996). A revision of Zwicker's loudness model. *ACTA Acustica*, vol. 82, pp. 335-345.

Ricard, J. (2005). An implementation of multi-band onset detection. *MIREX 2005 audio onset detection contest*: from <http://www.music-ir.org/evaluation/mirex-results/articles/onset/ricard.pdf>

Robel, A. (2007) Onset detection in polyphonic signals by means of transient peak classification. *MIREX 2007 audio onset detection contest*: from http://www.music-ir.org/mirex/2007/abs/OD_roebel.pdf

Wang W., Luo Y., Chambers J.A. & Sanei S.(2008). Note Onset Detection via Nonnegative Factorization of Magnitude Spectrum. *EURASIP Journal on Advances in Signal Processing*, vol. 2008, Article ID 231367, 15 pages, 2008. doi:10.1155/2008/231367.

Zhou, R. (2006). Feature Extraction of Musical Content for Automatic Music Transcription. Ph.D. dissertation, Swiss Federal Institute of Technology, Lausanne, from <http://library.epfl.ch/en/theses/?nr=3638>.

Zhou, R. & Mattavelli (2007). A new time-frequency representation for music signal analysis. *Signal Processing and Its Applications, 2007. ISSPA 2007. 9th International Symposium on*, pp. 1-4.

Zhou, R. & Reiss, J.D. (2007) Music onset detection combining energy-based and pitch-based approaches. *MIREX 2007 audio onset detection contest*: from http://www.music-ir.org/mirex2007/abs/OD_zhou.pdf

Zhou, R., Mattavelli, M. & Zoia, G. (2008). Music onset detection based on Resonator Time-frequency Image. *IEEE Trans. Audio, Speech and Language Processing*, vol. 16, pp. 1685–1695.

ADDITIONAL READING SECTION

Bello, J.P. & Sandler, M. (2003). Phase-based note onset detection for music signals. *Proc. IEEE Int. Conf. Acoustics, Speech, and Signal Processing (ICASSP-03)*, Hong Kong, pp.49-52.

Bello, J.P., Duxbury, C., Davies, M., & Sandler, M.(2004). On the use of phase and energy for musical onset detection in the complex domain. *IEEE Signal Processing Letter*, vol. 11, no. 6, pp. 553-556.

Bello, J.P., Daudet, L., Abadia, S., Duxbury, C., Davies, M. & Sandler, M.B. (2005). A tutorial on onset detection in music signals, *IEEE Trans. Speech and Audio Signal Processing*, vol. 13, pp. 1035–1047.

Chuan, C., & Elaine C. (2008). Audio Onset Detection Using Machine Learning Techniques: The Effect and Applicability of Key and Tempo Information. University of Southern California Computer Science Department Technical Report No. 08-895.

Submitted for;

Ruohua Zhou and Joshua D. Reiss, 'Music Onset Detection,' Chapter 12 in *Machine Audition: Principles, Algorithms and Systems*, Edited by W. Wang, Published by IGI Global, ISBN-13: 978-1615209194, July, 2010, p. 297-316, doi: 10.4018/978-1-61520-919-4.ch012

Chuan, C., & Elaine C. (2008). *Audio Onset Detection Using Machine Learning Techniques: The Effect and Applicability of Key and Tempo Information*. University of Southern California Computer Science Department Technical Report No. 08-895.

Collins, N. (1999). Using a pitch detector as an onset detector. *Proc. International Conf. On Music Information Retrieval* Sep..

Collins, N. (2005a). A comparison of sound onset detection algorithms with emphasis on psychoacoustically motivated detection functions. *AES Convention 118*, Barcelona.

Collins, N. (2005b). A change discrimination onset detector with peak scoring peak picker and time domain correction. First Annual Music Information Retrieval Evaluation eXchange (MIREX), 2005.

Dixon, S. (2006). Onset detection revisited in *Proceedings of the 9th International Conference on Digital Audio Effects*, pp. 133–137.

Duxbury, C., Sandler, M., & Davies, M. (2002). A hybrid approach to musical note onset detection. *Proc. 5th International Conf. Digital Audio Effects (DAFX-02)*, Hamburg, Germany.

Goto, M. (2001). An audio-based real-time beat tracking system for music with or without drum-sounds. *Journal of New Music Research*, vol. 30, No. 2, pp.159–171.

Hartmann, W.M. (1997). *Signals Sound and Sensation*, American Institute of Physics Press.

Jensen, K. & Andersen, T.H. (2004). Causal rhythm grouping. *Proc. 2nd International Symposium on Computer Music Modeling and Retrieval*, Esbjerg, Denmark.

Klapuri, A. (1999). Sound onset detection by applying psychoacoustic knowledge. *Proc. IEEE International Conf. Acoustics, Speech, and Signal Processing (ICASSP-99)*, pp. 3089–3092.

Lacoste, A. & Eck, D. (2007). A supervised classification algorithm for Note Onset Detection. *EURASIP Journal on Advances in Signal Processing*, vol. 2007, Article ID 43745, 13 pages.

Leveau, P., Daudet, L. & Richard, G. (2004). Methodology and Tools for the evaluation of automatic onset detection algorithms in music. *Proc. 5th International Conf. On Music Information Retrieval*, Barcelona.

Marolt, M., Kavcic, A. & Privosnik, M. (1999). Neural networks for note onset detection in piano music. *Proc. International Conf. On Computer Music*.

MIREX (2005), First Annual Music Information Retrieval Evaluation eXchange (MIREX), 2005. from http://www.music-ir.org/mirex2005/index.php/Audio_Onset_Detection

Moore, B.C.J. & Glasberg, B.R. (1996). A revision of Zwicker's loudness model. *ACTA Acustica*, vol. 82, pp. 335-345.

Submitted for;

Ruohua Zhou and Joshua D. Reiss, 'Music Onset Detection,' Chapter 12 in *Machine Audition: Principles, Algorithms and Systems*, Edited by W. Wang, Published by IGI Global, ISBN-13: 978-1615209194, July, 2010, p. 297-316, doi: 10.4018/978-1-61520-919-4.ch012

Ricard, J. (2005). An implementation of multi-band onset detection. First Annual Music Information Retrieval Evaluation eXchange (MIREX).

Wang W., Luo Y., Chambers J.A. & Sanei S.(2008). Note Onset Detection via Nonnegative Factorization of Magnitude Spectrum. *EURASIP Journal on Advances in Signal Processing*, vol. 2008, Article ID 231367, 15 pages, 2008. doi:10.1155/2008/231367.

Zhou, R. (2006). Feature Extraction of Musical Content for Automatic Music Transcription. Ph.D. dissertation, Swiss Federal Institute of Technology, Lausanne, from <http://library.epfl.ch/en/theses/?nr=3638>.

Zhou, R. & Mattavelli (2007). A new time-frequency representation for music signal analysis. *Signal Processing and Its Applications, 2007. ISSPA 2007. 9th International Symposium on*, pp. 1-4.

Zhou, R. & Reiss, J.D. (2007) Music onset detection combining energy-based and pitch-based approaches. *MIREX 2007 audio onset detection contest*: from http://www.music-ir.org/mirex2007/abs/OD_zhou.pdf

Zhou, R., Mattavelli, M. & Zoia, G. (2008). Music onset detection based on Resonator Time-frequency Image. *IEEE Trans. Audio, Speech and Language Processing*, vol. 16, pp. 1685–1695.