A Tutorial on Onset Detection in Music Signals

Juan Pablo Bello, Laurent Daudet, Samer Abdallah, Chris Duxbury, Mike Davies, and Mark B. Sandler, *Senior Member, IEEE*

***Abstract*—Note onset detection and localization is useful in a number of analysis and indexing techniques for musical signals. The usual way to detect onsets is to look for "transient" regions in the signal, a notion that leads to many definitions: a sudden burst of energy, a change in the short-time spectrum of the signal or in the statistical properties, etc. The goal of this paper is to review, categorize, and compare some of the most commonly used techniques for onset detection, and to present possible enhancements. We discuss methods based on the use of explicitly predefined signal features: the signal's amplitude envelope, spectral magnitudes and phases, time-frequency representations; and methods based on probabilistic signal models: model-based change point detection, surprise signals, etc. Using a choice of test cases, we provide some guidelines for choosing the appropriate method for a given application.**

***Index Terms*—Attack transcients, audio, note segmentation, novelty detection.**

I. Introduction

A. Background and Motivation

MUSIC is to a great extent an event-based phenomenon for both performer and listener. We nod our heads or tap our feet to the *rhythm* of a piece; the performer's attention is focused on each successive *note*. Even in non note-based music, there are transitions as musical timbre and tone color evolve. Without change, there can be no musical meaning.

The automatic detection of events in audio signals gives new possibilities in a number of music applications including content delivery, compression, indexing and retrieval. Accurate retrieval depends on the use of appropriate features to compare and identify pieces of music. Given the importance of musical events, it is clear that identifying and characterizing these events is an important aspect of this process. Equally, as compression standards advance and the drive for improving quality at low bit-rates continues, so does accurate event detection become a basic requirement: disjoint audio segments with homogeneous statistical properties, delimited by transitions or events, can be compressed more successfully in isolation than they can in combination with dissimilar regions. Finally, accurate segmentation allows a large number of standard audio editing algorithms and effects (e.g., time-stretching, pitch-shifting) to be more signal-adaptive.

like MIDI format

Manuscript received August 6, 2003; revised July 21, 2004. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Gerald Schuller.

J. P. Bello, S. Abdallah, M. Davies, and M. B. Sandler are with the Centre for Digital Music, Department of Electronic Engineering, Queen Mary, University of London, London E1 4NS, U.K. (e-mail: juan.bello-correa@elec.qmul.ac.uk; samer.abdallah@elec.qmul.ac.uk; mike.davies@elec.qmul.ac.uk; mark.sandler@elec.qmul.ac.uk).

L. Daudet is with the Laboratoire d'Acoustique Musicale, Université Pierre et Marie Curie (Paris 6), 75015 Paris, France (e-mail: daudet@lam.jussieu.fr).

C. Duxbury is with the Centre for Digital Music, Department of Electronic Engineering, Queen Mary, University of London, London E1 4NS, U.K., and also with WaveCrest Communications Ltd. (e-mail: christopher.duxbury@elec.qmul.ac.uk).

Digital Object Identifier 10.1109/TSA.2005.851998

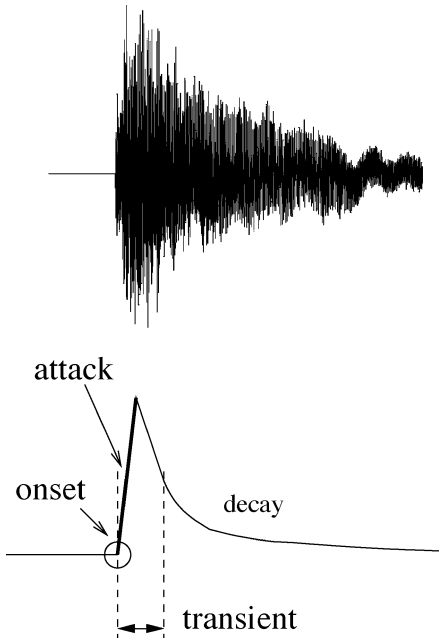


Fig. 1. "Attack," "transient," "decay," and "onset" in the ideal case of a single note.

There have been many different approaches for onset detection. The goal of this paper is to give an overview of the most commonly used techniques, with a special emphasis on the ones that have been employed in the authors' different applications. For the sake of coherence, the discussion will be focused on the more specific problem of note onset detection in musical signals, although we believe that the discussed methods can be useful for various different tasks (e.g., transient modeling or localization) and different classes of signals (e.g., environmental sounds, speech).

B. Definitions: Transients vs. Onsets vs. Attacks

A central issue here is to make a clear distinction between the related concepts of *transients*, *onsets* and *attacks*. The reason for making these distinctions clear is that different applications have different needs. The similarities and differences between these key concepts are important to consider; it is similarly important to categorize all related approaches. Fig. 1 shows, in the simple case of an isolated note, how one could differentiate these notions.

- The *attack* of the note is the time interval during which the amplitude envelope increases.

振幅包络

- The concept of *transient* is more difficult to describe precisely. As a preliminary informal definition, transients are short intervals during which the signal evolves quickly in some nontrivial or relatively unpredictable way. In the case of acoustic instruments, the transient often corresponds to the period during which the excitation (e.g., a hammer strike) is applied and then damped, leaving only the slow decay at the resonance frequencies of the body. Central to this time duration problem is the issue of the useful time resolution: we will assume that the human ear cannot distinguish between two transients less than 10 ms apart [1]. Note that the release or offset of a sustained sound can also be considered a transient period.
- The *onset* of the note is a single instant chosen to mark the temporally extended transient. In most cases, it will coincide with the start of the transient, or the earliest time at which the transient can be reliably detected.

C. General Scheme of Onset Detection Algorithms

In the more realistic case of a possibly noisy polyphonic signal, where multiple sound objects may be present at a given time, the above distinctions become less precise. It is generally not possible to detect onsets directly without first quantifying the time-varying "transientness" of the signal.

Audio signals are both *additive* (musical objects in polyphonic music superimpose and not conceal each other) and *oscillatory*. Therefore, it is not possible to look for changes simply by differentiating the original signal in the time domain; this has to be done on an intermediate signal that reflects, in a simplified form, the local structure of the original. In this paper, we refer to such a signal as a *detection function*; in the literature, the term *novelty function* is sometimes used instead [2].

Fig. 2 illustrates the procedure employed in the majority of onset detection algorithms: from the original audio signal, which can be pre-processed to improve the performance of subsequent stages, a detection function is derived at a lower sampling rate, to which a peak-picking algorithm is applied to locate the onsets. Whereas peak-picking algorithms are well documented in the literature, the diversity of existing approaches for the construction of the detection function makes the comparison between onset detection algorithms difficult for audio engineers and researchers.

D. Outline of the Paper

The outline of this paper follows the flowchart in Fig. 2. In Section II, we review a number of preprocessing techniques that can be employed to enhance the performance of some of the detection methods. Section III presents a representative cross-section of algorithms for the construction of the detection function. In Section IV, we describe some basic peak-picking algorithms; this allows the comparative study of the performance of a selection of note onset detection methods given in Section V. We finish our discussion in Section VI with a review of our findings and some thoughts on the future development of these algorithms and their applications.

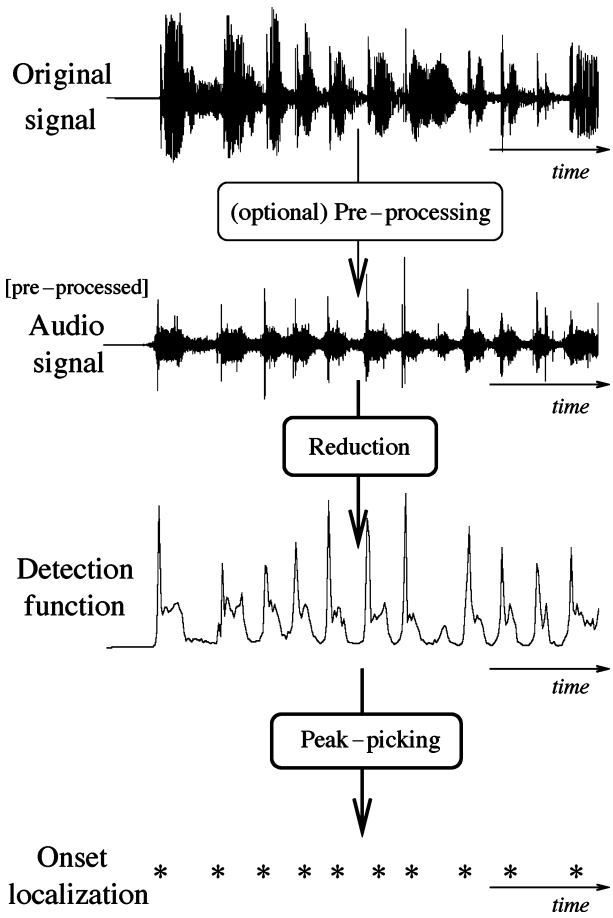


Fig. 2. Flowchart of a standard onset detection algorithm.

II. Preprocessing

The concept of preprocessing implies the transformation of the original signal in order to accentuate or attenuate various aspects of the signal according to their relevance to the task in hand. It is an optional step that derives its relevance from the process or processes to be subsequently performed.

There are a number of different treatments that can be applied to a musical signal in order to facilitate the task of onset detection. However, we will focus only on two processes that are consistently mentioned in the literature, and that appear to be of particular relevance to onset detection schemes, especially when simple reduction methods are implemented: separating the signal into multiple frequency bands, and transient/steady-state separation.

A. Multiple Bands

Several onset detection studies have found it useful to independently analyze information across different frequency bands. In some cases this preprocessing is needed to satisfy the needs of specific applications that require detection in individual sub-bands to complement global estimates; in others, such an approach can be justified as a way of increasing the robustness of a given onset detection method.

As examples of the first scenario, two beat tracking systems make use of filter banks to analyze transients across frequencies.

Goto [3] slices the spectrogram into *spectrum strips* and recognizes onsets by detecting sudden changes in energy. These are used in a multiple-agent architecture to detect rhythmic patterns. Scheirer [4] implements a six-band filter bank, using sixth-order elliptic filters, and psychoacoustically inspired processing to produce onset trains. These are fed into comb-filter resonators in order to estimate the tempo of the signal.

The second case is illustrated by models such as the perceptual onset detector introduced by Klapuri [5]. In this implementation, a filter bank divides the signal into eight nonoverlapping bands. In each band, onset times and intensities are detected and finally combined. The filter-bank model is used as an approximation to the mechanics of the human cochlea.

Another example is the method proposed by Duxbury *et al.* [6], that uses a constant-Q conjugate quadrature filter bank to separate the signal into five subbands. It goes a step further by proposing a hybrid scheme that considers energy changes in high-frequency bands and spectral changes in lower bands. By implementing a multiple-band scheme, the approach effectively avoids the constraints imposed by the use of a single reduction method, while having different time resolutions for different frequency bands.

B. Transient/Steady-State Separation

The process of transient/steady-state separation is usually associated with the modeling of music signals, which is beyond the scope of this paper. However, there is a fine line between modeling and detection, and indeed, some modeling schemes directed at representing transients may hold promise for onset detection. Below, we briefly describe several methods that produce modified signals (residuals, transient signals) that can be, or have been, used for the purpose of onset detection.

Sinusoidal models, such as "additive synthesis" [7], represent an audio signal as a sum of sinusoids with slowly varying parameters. Amongst these methods, *spectral modeling synthesis* (SMS) [8] explicitly considers the residual[1] of the synthesis method as a Gaussian white noise filtered with a slowly varying low-order filter. Levine [9] calculates the residual between the original signal and a multiresolution SMS model. Significant increases in the energy of the residual show a mismatch between the model and the original, thus effectively marking onsets. An extension of SMS, *transient modeling synthesis*, is presented in [10]. Transient signals are analyzed by a sinusoidal analysis/synthesis similar to SMS *on the discrete cosine transform* of the residual, hence in a pseudo-temporal domain. In [11], the whole scheme, including tonal and transients extraction is generalized into a single matching pursuit formulation.

An alternative approach for the segregation of sinusoids from transient/noise components is proposed by Settel and Lippe [12] and later refined by Duxbury *et al.* [13]. It is based on the phase-vocoder principle of instantaneous frequency (see Section III-A.3) that allows the classification of individual frequency bins of a spectrogram according to the predictability of their phase components.

[1]The residual signal results from the subtraction of the modeled signal from the original waveform. When sinusoidal or harmonic modeling is used, then the residual is assumed to contain most of the impulse-like, noisy components of the original signal—e.g., transients.

Other schemes for the separation of tonal from nontonal components make use of lapped orthogonal transforms, such as the modified discrete cosine transform (MDCT), first introduced by Princen and Bradley [14]. These algorithms, originally designed for compression [15], [16], make use of the relative sparsity of MDCT representations of most musical signals: a few large coefficients account for most of the signal's energy. Actually, since the MDCT atoms are very tone-like (they are cosine functions slowly modulated in time by a smooth window), the part of the signal represented by the large MDCT atoms, according to a given threshold, can be interpreted as the tonal part of the signal [10], [17]. Transients and noise can be obtained by removing those large MDCT atoms.

III. Reduction

In the context of onset detection, the concept of reduction refers to the process of transforming the audio signal into a highly subsampled *detection function* which manifests the occurrence of transients in the original signal. This is the key process in a wide class of onset detection schemes and will therefore be the focus of most of our review.

We will broadly divide reduction methods in two groups: methods based on the use of explicitly predefined signal features, and methods based on probabilistic signal models.

A. Reduction Based on Signal Features

1) Temporal Features: When observing the temporal evolution of simple musical signals, it is noticeable that the occurrence of an onset is usually accompanied by an increase of the signal's amplitude. Early methods of onset detection capitalized on this by using a detection function which follows the amplitude envelope of the signal [18]. Such an "envelope follower" can be easily constructed by rectifying and smoothing (i.e., low-pass filtering) the signal

$$E_0(n) = \frac{1}{N} \sum_{m=-\frac{N}{2}}^{\frac{N}{2}-1} |x(n+m)|\, w(m) \tag{1}$$

where $w(m)$ is an N-point window or smoothing kernel, centered at $m = 0$. This yields satisfactory results for certain applications where strong percussive transients exist against a quiet background. A variation on this is to follow the local energy, rather than the amplitude, by squaring, instead of rectifying, each sample

$$E(n) = \frac{1}{N} \sum_{m=-\frac{N}{2}}^{\frac{N}{2}-1} [x(n+m)]^2\, w(m). \tag{2}$$

Despite the smoothing, this reduced signal in its raw form is not usually suitable for reliable onset detection by peak picking. A further refinement, included in a number of standard onset detection algorithms, is to work with the time derivative of the energy (or rather the first difference for discrete-time signals) so that sudden rises in energy are transformed into narrow peaks in the derivative. The energy and its derivative are commonly used in combination with preprocessing, both with filter-banks [3] and transient/steady-state separation [9], [19].

Another refinement takes its cue from psychoacoustics: empirical evidence [20] indicates that loudness is perceived *logarithmically*. This means that changes in loudness are judged relative to the overall loudness, since, for a continuous time signal, $\mathrm{d}(\log E)/\mathrm{d}t = (\mathrm{d}E/\mathrm{d}t)/E$. Hence, computing the first-difference of $\log E(n)$ roughly simulates the ear's perception of loudness. An application of this technique to multiple bands [5] showed a significant reduction in the tendency for amplitude modulation to cause the detection of spurious onsets.

2) Spectral Features: A number of techniques have been proposed that use the spectral structure of the signal to produce more reliable detection functions. While reducing the need for preprocessing (e.g., removal of the tonal part), these methods are also successful in a number of scenarios, including onset detection in polyphonic signals with multiple instruments.

Let us consider the short-time Fourier transform (STFT) of the signal $x(n)$

$$X_k(n) = \sum_{m=-\frac{N}{2}}^{\frac{N}{2}-1} x(nh+m)w(m)\mathrm{e}^{-\frac{2j\pi mk}{N}} \tag{3}$$

where $w(m)$ is again an N-point window, and h is the hop size, or time shift, between adjacent windows.

In the spectral domain, energy increases linked to transients tend to appear as a broadband event. Since the energy of the signal is usually concentrated at low frequencies, changes due to transients are more noticeable at high frequencies [21]. To emphasize this, the spectrum can be weighted preferentially toward high frequencies before summing to obtain a weighted energy measure

$$\widetilde{E}(n) = \frac{1}{N}\sum_{k=-\frac{N}{2}}^{\frac{N}{2}-1} W_k \left|X_k(n)\right|^2 \tag{4}$$

where W_k is the frequency dependent weighting. By Parseval's theorem, if $W_k = 1\ \forall k$, $\widetilde{E}(n)$ is simply equivalent to the local energy as previously defined. Note also that a choice of $W_k = k^2\ \forall k$ would give the local energy of the derivative of the signal.

Masri [22] proposes a *high frequency content* (HFC) function with $W_k = |k|$, linearly weighting each bin's contribution in proportion to its frequency. The HFC function produces sharp peaks during attack transients and is notably successful when faced with percussive onsets, where transients are well modeled as bursts of white noise.

These spectrally weighted measures are based on the instantaneous short-term spectrum of the signal, thus omitting any explicit consideration of its temporal evolution. Alternatively, a number of other approaches do consider these changes, using variations in spectral content between analysis frames in order to generate a more informative detection function.

Rodet and Jaillet [21] propose a method where the frequency bands of a sequence of STFTs are analyzed independently using a piece-wise linear approximation to the magnitude profile $X_k(l)$ for $n - M \leq l \leq n + M$, where l is a short temporal window, and M is a fixed value. The parameters of these approximations are used to generate a set of band-wise detection functions, later combined to produce final onset results. Detection results are robust for high-frequencies, showing consistency with Masri's HFC approach.

A more general approach based on changes in the spectrum is to formulate the detection function as a "distance" between successive short-term Fourier spectra, treating them as points in an N-dimensional space. Depending on the metric chosen to calculate this distance, different *spectral difference*, or spectral flux, detection functions can be constructed: Masri [22] uses the L_1-norm of the difference between magnitude spectra, whereas Duxbury [6] uses the L_2-norm on the *rectified* difference

$$SD(n) = \sum_{k=\frac{N}{2}}^{\frac{N}{2}-1} \left\{H\left(|X_k(n)| - |X_k(n-1)|\right)\right\}^2 \tag{5}$$

where $H(x) = (x + |x|)/2$, i.e., zero for negative arguments. The rectification has the effect of counting only those frequencies where there is an *increase* in energy, and is intended to emphasize onsets rather than offsets.

A related form of spectral difference is introduced by Foote [2] to obtain a measure of "audio novelty".[2] A similarity matrix is calculated using the correlation between STFT feature vectors (power spectra). The matrix is then correlated with a "checkerboard" kernel to detect the edges between areas of high and low similarity. The resulting function shows sharp peaks at the times of these changes, and is effectively an onset detection function when kernels of small width are used.

3) Spectral Features Using Phase: All the mentioned methods have in common their use of the magnitude of the spectrum as their only source of information. However, recent approaches make also use of the phase spectra to further their analyses of the behavior of onsets. This is relevant since much of the temporal structure of a signal is encoded in the phase spectrum.

Let us define $\varphi_k(n)$ the 2π-*unwrapped* phase of a given STFT coefficient $X_k(n)$. For a steady state sinusoid, the phase $\varphi_k(n)$, as well as the phase in the previous window $\varphi_k(n-1)$, are used to calculate a value for the instantaneous frequency, an estimate of the actual frequency of the k^{th} STFT component within this window, as [23]

$$f_k(n) = \left(\frac{\varphi_k(n) - \varphi_k(n-1)}{2\pi h}\right) f_s \tag{6}$$

where h is the hop size between windows and f_s is the sampling frequency.

It is expected that, for a locally stationary sinusoid, the instantaneous frequency should be approximately constant over adjacent windows. Thus, according to (6), this is equivalent to the phase increment from window to window remaining approximately constant (cf. Fig. 3)

$$\varphi_k(n) - \varphi_k(n-1) \simeq \varphi_k(n-1) - \varphi_k(n-2). \tag{7}$$

[2]The term *novelty function* is common to the literature in machine learning and communication theory, and is widely used for video segmentation. In the context of onset detection, our notion of the *detection function* can be seen also as a novelty function, in that it tries to measure the extent to which an event is unusual given a series of observations in the past.

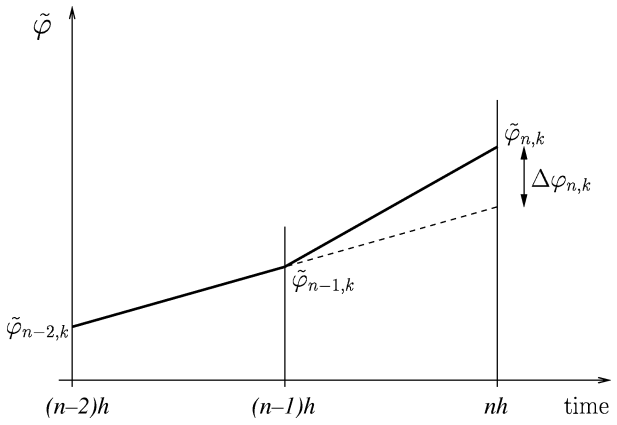


Fig. 3. Phase diagram showing instantaneous frequencies as phase derivative over adjacent frames. For a stationary sinusoid this should stay constant (dotted line).

Equivalently, the phase deviation can be defined as the second difference of the phase

$$\Delta\varphi_k(n) = \varphi_k(n) - 2\varphi_k(n-1) + \varphi_k(n-2) \simeq 0. \quad (8)$$

During a transient region, the instantaneous frequency is not usually well defined, and hence $\Delta\varphi_k(n)$ will tend to be large. This is illustrated in Fig. 3.

In [24], Bello proposes a method that analyzes the instantaneous distribution (in the sense of a probability distribution or histogram) of phase deviations across the frequency domain. During the steady-state part of a sound, deviations tend to zero, thus the distribution is strongly peaked around this value. During attack transients, $\Delta\varphi_k(n)$ values increase, widening and flattening the distribution. In [24], this behavior is quantified by calculating the inter-quartile range and the kurtosis of the distribution. In [25], a simpler measure of the spread of the distribution is calculated as

$$\zeta_p(n) = \frac{1}{N}\sum_{k=1}^{N} |\Delta\varphi_k(n)| \quad (9)$$

i.e., the mean *absolute* phase deviation. The method, although showing some improvement for complex signals, is susceptible to phase distortion and to noise introduced by the phases of components with no significant energy.

As an alternative to the sole use of magnitude or phase information, [26] introduces an approach that works with Fourier coefficients in the complex domain. The stationarity of the k^{th} spectral bin is quantified by calculating the Euclidean distance $\Gamma_k(n)$ between the observed $X_k(n)$ and that predicted by the previous frames, $\hat{X}_k(n)$

$$\Gamma_k(n) = \left\{ \left|\hat{X}_k(n)\right|^2 + |X_k(n)|^2 - 2\left|\hat{X}_k(n)\right| |X_k(n)| \cos\left(\Delta\varphi_k(n)\right) \right\}^{\frac{1}{2}}. \quad (10)$$

These distances are summed across the frequency-domain to generate an onset detection function

$$\zeta(n) = \sum_{k=1}^{N} \Gamma_k(n). \quad (11)$$

See [27] for an application of this technique to multiple bands. Other preprocessing, such as the removal of the tonal part, may introduce distortions to the phase information and thus adversely affect the performance of subsequent phase-based onset detection methods.

4) Time-Frequency and Time-Scale Analysis: An alternative to the analysis of the temporal envelope of the signal and of Fourier spectral coefficients, is the use of time-scale or time-frequency representations (TFR).

In [28] a novelty function is calculated by measuring the dissimilarity between feature vectors corresponding to a discretized Cohen's class TFR, in this case the result of convolving the Wigner-Ville TFR of the function with a Gaussian kernel. Note that the method could be also seen as a spectral difference approach, given that by choosing an appropriate kernel, the representation becomes equivalent to the spectrogram of the signal.

In [29], an approach for transient detection is described based on a simple dyadic wavelet decomposition of the residual signal. This transform, using the Haar wavelet, was chosen for its simplicity and its good time localization at small scales. The scheme takes advantage of the correlations across scales of the coefficients: large wavelet coefficients, related to transients in the signal, are not evenly spread within the dyadic plane but rather form "structures". Indeed, if a given coefficient has a large amplitude, there is a high probability that the coefficients with the same time localization at smaller scales also have large amplitudes, therefore forming dyadic trees of significant coefficients.

The significance of full-size branches of coefficients, from the largest to the smallest scale, can be quantified by a *regularity modulus*, which is a *local* measure of the regularity of the signal

$$\kappa_s[i] = \sum_{(j,k)\in\mathcal{B}[i]} 2^{js}|d_{j,k}| \quad (12)$$

where the $d_{j,k}$ are the wavelet coefficients, $\mathcal{B}[i]$ is the full branch leading to a given small-scale coefficient $d_{1,i}$ (i.e., the set of coefficients at larger scale and same time localization), and s a free parameter used to emphasize certain scales ($s = 0$ is often used in practice). Since increases of $\kappa_s[i]$ are related to the existence of large, transient-like coefficients in the branch $\mathcal{B}[i]$, the regularity modulus can effectively act as an onset detection function.

B. Reduction Based on Probability Models

Statistical methods for onset detection are based on the assumption that the signal can be described by some probability model. A system can then be constructed that makes probabilistic inferences about the likely times of abrupt changes in the signal, given the available observations. The success of this approach depends on the closeness of fit between the assumed model, i.e., the probability distribution described by the model, and the "true" distribution of the data, and may be quantified using likelihood measures or Bayesian model selection criteria.

1) Model-Based Change Point Detection Methods: A well-known approach is based on the *sequential probability ratio test* [30]. It presupposes that the signal samples $x(n)$ are generated

from one of two statistical models, $\mathcal{A}$ or $\mathcal{B}$. The log-likelihood ratio is defined as

$$s = \log \frac{p_{\mathcal{B}}(x)}{p_{\mathcal{A}}(x)} \tag{13}$$

where $p_{\mathcal{A}}(x)$ and $p_{\mathcal{B}}(x)$ are the probability density functions associated with the two models. The expectation of the observed log-likelihood ratio depends on which model the signal is actually following. Under model $\mathcal{A}$, the expectation is

$$E_{\mathcal{A}}(s) = -\int p_{\mathcal{A}}(x) \log \frac{p_{\mathcal{A}}(x)}{p_{\mathcal{B}}(x)} \mathrm{d}x = -D(p_{\mathcal{A}} \,||\, p_{\mathcal{B}}) < 0 \tag{14}$$

where D denotes the Kullback–Leibler divergence between the model and the observed distributions. Under model $\mathcal{B}$, the expectation is

$$E_{\mathcal{B}}(s) = \int p_{\mathcal{B}}(x) \log \frac{p_{\mathcal{B}}(x)}{p_{\mathcal{A}}(x)} \mathrm{d}x = D(p_{\mathcal{B}} \,||\, p_{\mathcal{A}}) > 0. \tag{15}$$

If we assume that the signal initially follows model $\mathcal{A}$, and switches to model $\mathcal{B}$ at some unknown time, then the short-time average of the log-likelihood ratio s will change sign. The algorithms described in [30] are concerned with detecting this change of sign. In this context, the log-likelihood ratio can be considered as a detection function, though one that produces changes in polarity, rather than localized peaks, as its detectable feature.

The method can be extended to deal with cases in which the models are unknown and must be estimated from the data. The *divergence algorithm* [31] manages this by fitting model $\mathcal{A}$ to a growing window, beginning at the last detected change point and extending to the current time. Model $\mathcal{B}$ is estimated from a sliding window of fixed size, extending back from the current time. Both Jehan [32], and Thornburg and Gouyon [33] apply variants of this method, using parametric Gaussian autoregressive models for $\mathcal{A}$ and $\mathcal{B}$.

Jehan [32] also applies Brandt's method [34], in which a fixed length window is divided at a hypothetical change point r. The two resulting segments are modeled using two separate Gaussian AR models. The model parameters and the change point r are then optimized to maximize the log-likelihood ratio between the probability of having a change at r and the probability of not having an onset at all. Change points are detected when this likelihood ratio surpasses a fixed threshold.

2) Approaches Based on 'Surprise Signals': The methods described above look for an instantaneous switch between two distinct models. An alternative is to look for *surprising moments* relative to a single global model. To this end, a detection function is defined as the moment-by-moment trace of the *negative log-probability* of the signal given its recent history, according to a global model.

The approach, introduced by Abdallah and Plumbley [35], is based on the notion of an observer which becomes "familiar" with (i.e., builds a model of) a certain class of signals, such that it is able to make predictions about the likely evolution of the signal as it unfolds in time. Such an observer will be relatively surprised at the onset of a note because of its uncertainty about when and what type of event will occur next. However, if the observer is in fact reasonably familiar with typical events (i.e., the model is accurate), that surprise will be localized to the transient region, during which the identity of the event is becoming established. Thus, a dynamically evolving measure of surprise, or novelty, can be used as a detection function.

Let us consider the signal as a multivariate random process where each vector $\mathbf{x}(n) \in \mathbb{R}^N$ is a frame of audio samples. At time n, an observer's expectations about $\mathbf{x}(n)$ will be summarized by the conditional probability according to that observer's model: $p(\mathbf{x}(n)|\mathbf{x}(n-1), \mathbf{x}(n-2), \cdots)$. When $\mathbf{x}(n)$ is actually observed, the observer will be surprised to a certain degree, which we will define as

$$S(n) \equiv S\left(\mathbf{x}(n)\right) \stackrel{\text{def}}{=} -\log p\left(\mathbf{x}(n) \middle| \left\{\mathbf{x}(j) : j < n\right\}\right). \tag{16}$$

This is closely related to the entropy rate of the random process, which is simply the expected surprise according to the "true" model.

An alternative conditional density model can be defined for an audio signal by partitioning the frame $\mathbf{x}$ into two segments $(\mathbf{x}_1, \mathbf{x}_2)$ and then expressing $p(\mathbf{x}_2|\mathbf{x}_1)$ in terms of $p(\mathbf{x}_1, \mathbf{x}_2) \equiv p(\mathbf{x})$. A detection function can then be generated from the surprise associated with $\mathbf{x}_2$

$$S(\mathbf{x}_2) = \log p(\mathbf{x}_1) - \log p(\mathbf{x}) \tag{17}$$

both terms of which may be approximated by any suitable joint density model; for example, [35] uses two separate independent component analysis (ICA) models.

In ICA, we assume that a random vector $\mathbf{x} \in \mathbb{R}^N$ is generated by linear transformation of a random vector $\mathbf{s} \in \mathbb{R}^N$ of independent non-Gaussian components; that is, $\mathbf{x} = \mathbf{A}\mathbf{s}$, where $\mathbf{A}$ is an $N \times N$ basis matrix. This model gives

$$-\log p(\mathbf{x}) = -\sum_{i=1}^{N} \log p_i(s_i) + \log \det \mathbf{A} \tag{18}$$

where $\mathbf{s}$ is obtained from $\mathbf{x}$ using $\mathbf{s} = \mathbf{A}^{-1}\mathbf{x}$, and $p_i(\cdot)$ is the assumed or estimated probability density function of the i^{th} component of $\mathbf{s}$. Estimates of $\mathbf{A}$ are relatively easy to obtain [36]. Results obtained with speech and music are given in [37].

It is worth noting that some of the detection functions described in previous sections can be derived within this probabilistic framework by making specific assumptions about the observer's probability model. For example, an observer that believes the audio samples in each frame to be independent and identically distributed according to a Laplacian (double-sided exponential) distribution, such that $p(\mathbf{x}(n)) = \prod_{i=1}^{N} (1/2) \exp -|x_i(n)|$, where $x_i(n)$ is the i^{th} component of $\mathbf{x}(n)$, would assign $S(n) = \sum_{i=1}^{N} |x_i(n)| + \text{const.}$, which is essentially an envelope follower [cf. (1)]. Similarly, the assumption of a multivariate Gaussian model for the $\mathbf{x}(n)$ would lead to a quadratic form for $S(n)$, of which the short-term energy [(2)] and weighted energy [(4)] measures are special cases. Finally, measures of spectral difference [like (5)] can be associated with specific conditional probability models of one short-term spectrum given the previous one, while the complex domain method [(10) and (11)], depending as it does on a Euclidean distance measure between predicted and observed complex spectra, is related to a time-varying Gaussian process model.

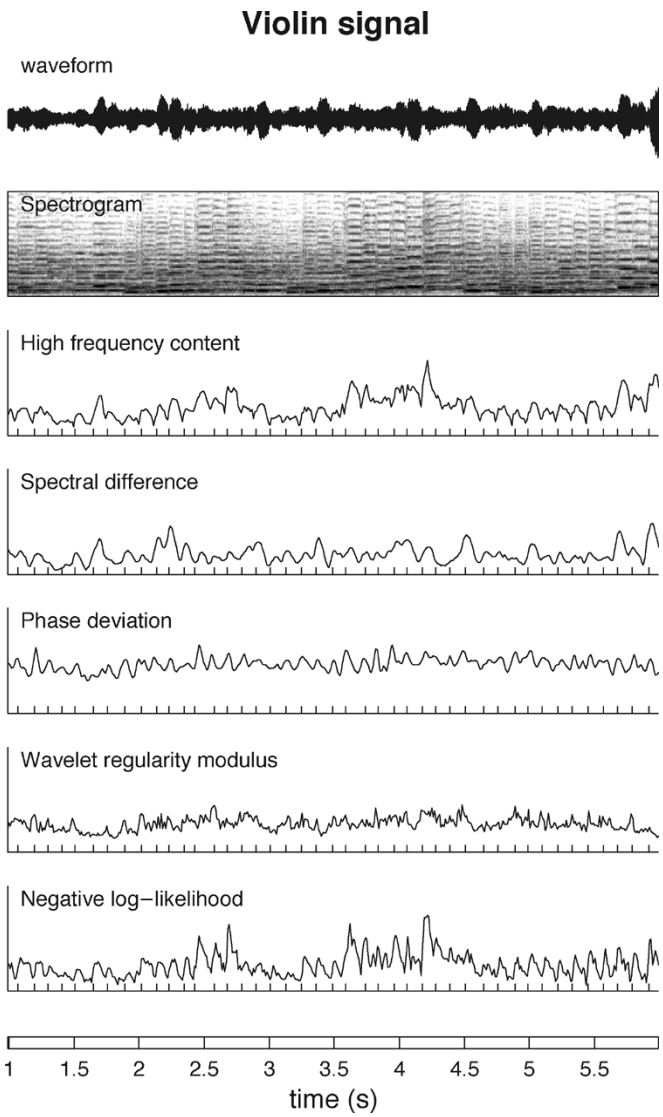


Fig. 4. Comparison of different detection functions for 5 s of a solo violin recording. From top to bottom: time-domain signal, spectrogram, high-frequency content, spectral difference, spread of the distribution of phase deviations, wavelet regularity modulus, and negative log-likelihood using an ICA model. All detection functions have been normalized to their maximum value.

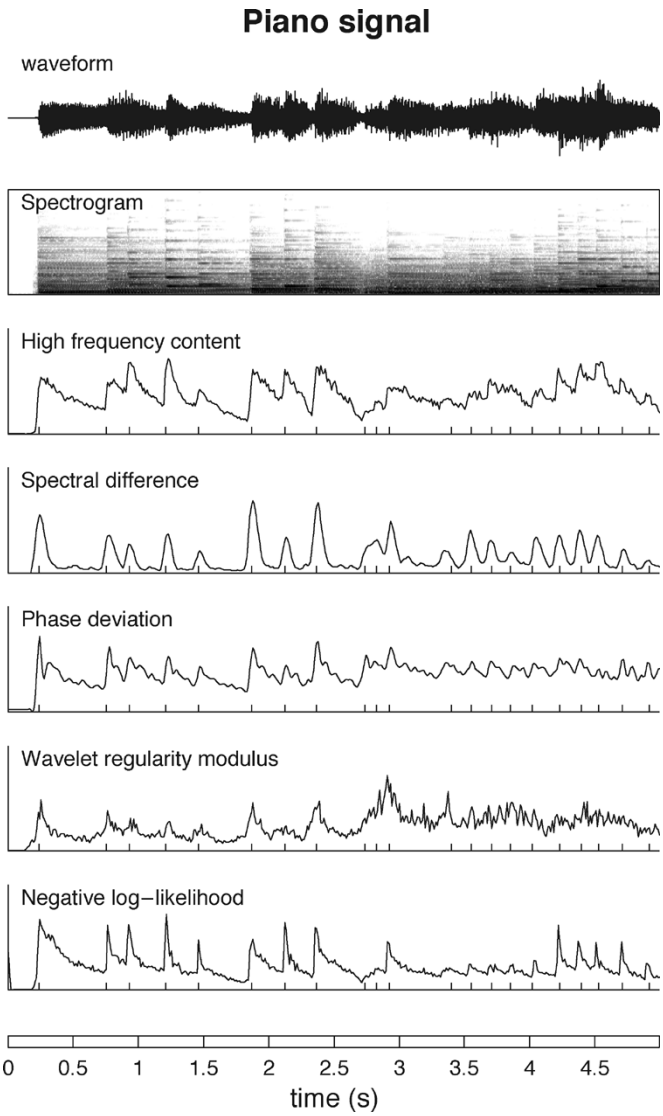


Fig. 5. Comparison of different detection functions for 5 s of a solo piano recording. From top to bottom: time-domain signal, spectrogram, high-frequency content, spectral difference, spread of the distribution of phase deviations, wavelet regularity modulus, and negative log-likelihood using an ICA model. All detection functions have been normalized to their maximum value.

C. Comparison of Detection Functions

All the approaches described above provide a solution to the problem of onset detection in musical signals. However, every method presents shortcomings depending both on its definition and on the nature of the signals to be analyzed. What follows is a discussion of the merits of different reduction approaches, with an emphasis on the ones that have been employed in the various applications developed by the authors. Figs. 4–6 are included to support the discussion. They correspond, respectively, to a pitched nonpercussive sound (solo violin), a pitched percussive sound (solo piano), and a complex mixture (pop music). The figures show the waveforms, spectrograms, and a number of different detection functions for comparison. The hand-labeled onsets for each signal are marked with ticks in the time-axis of the detection functions.

Temporal methods are simple and computationally efficient. Their functioning depends on the existence of clearly identifiable amplitude increases in the analysis signal, which is the case only for highly percussive events in simple sounds. The robustness of amplitude-based onset detection decreases when facing amplitude modulations (i.e., vibrato, tremolo) or the overlapping of energy produced by simultaneous sounds. This is true even after dividing the signal into multiple bands or after extracting the transient signal. For nontrivial sounds, onset detection schemes benefit from using richer representations of the signal (e.g., a time-frequency representation).

The commonly used HFC [22, eq. (4)] is an example of a spectral weighting method. It is successful at emphasizing the percussiveness of the signal [cf. Figs. 5 and 6], but less robust at detecting the onsets of low-pitched and nonpercussive events [cf. Fig. 4], where energy changes are at low frequencies and hence de-emphasized by the weighting. In some signals, even broadband onsets are susceptible to masking by continuous high-frequency content such as that due to open cymbals in a pop recording. This problem can be overcome by using temporal difference methods such as the L_2-norm of the rectified

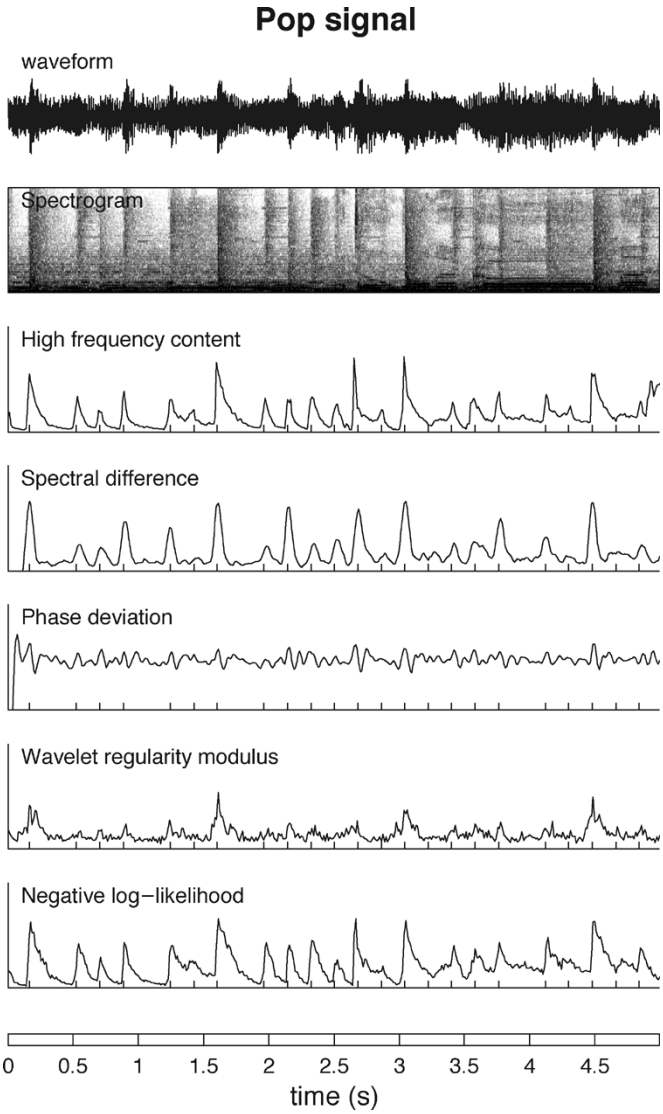


Fig. 6. Comparison of different detection functions for 5 s of a pop song. From top to bottom: time-domain signal, spectrogram, high-frequency content, spectral difference, spread of the distribution of phase deviations, wavelet regularity modulus, and negative log-likelihood using an ICA model. All detection functions have been normalized to their maximum value.

spectral difference [[6, eq. (5)], as these can respond to changes in the *distribution* of spectral energy, as well as the total, in any part of the spectrum. However, the difference calculation relies solely on magnitude information, thus neglecting the detection of events without a strong energy increase: e.g., low notes, transitions between harmonically related notes or onsets played by bowed instruments (cf. Fig. 4).

Phase-based methods, such as the *spread of the distribution of phase deviations* in (9) (see [25]), are designed to compensate for such shortcomings. They are successful at detecting low and high-frequency tonal changes regardless of their intensity. The approach suffers from variations introduced by the phases of noisy low-energy components, and from phase distortions common to complex commercial music recordings (e.g., audio effects, post-production treatments—cf. Fig. 6).

The *wavelet regularity modulus* [29] in (12), is an example of an approach using an alternative time-scale representation that can be used to precisely localize events down to a theoretical resolution of as little as two samples of the original signal, which for typical audio sampling rates is considerably better than the ear's resolution in time. The price of this is a much less smooth detection function (cf. all figures), therefore emphasizing the need for post-processing to remove spurious peaks. The method provides an interesting alternative to other feature-based methods, but with an increase in algorithmic complexity.

Approaches based on probabilistic models provide a more general theoretical view of the analysis of onsets. As shown in Section III-B.2, previous reduction methods can be explained within the context of measuring surprise relative to a probabilistic model, while new methods can be proposed and evaluated by studying refinements or alternatives to existing models. An example is the surprise-based method using ICA to model the conditional probability of a short segment of the signal, calculated in (17) as the difference between two *negative log-likelihoods* [35]. If the model is adequate (i.e., the assumptions behind the model are accurate and the parameters well-fitted), then robust detection functions for a wide range of signals can be produced. Examples are at the bottom of Figs. 4–6. However, for adaptive statistical models such as ICA, these advantages accrue only after a potentially expensive and time-consuming training process during which the parameters of the model are fitted to a given training set.

IV. Peak-Picking

If the detection function has been suitably designed, then onsets or other abrupt events will give rise to well-localized identifiable features in the detection function. Commonly, these features are local maxima (i.e., peaks), generally subject to some level of variability in size and shape, and masked by 'noise', either due to actual noise in the signal, or other aspects of the signal not specifically to do with onsets, such as vibrato. Therefore a robust peak-picking algorithm is needed to estimate the onset times of events within the analysis signal.[3]

We will divide the process of peak-picking a detection function in three steps: post-processing, thresholding, and a final decision process.

A. Post-Processing

Like preprocessing, post-processing is an optional step that depends on the reduction method used to generate the detection function. The purpose of post-processing is to facilitate the tasks of thresholding and peak-picking by increasing the uniformity and consistency of event-related features in the detection function, ideally transforming them into isolated, easily detectable local maxima. Into this category fall processes intended to reduce the effects of noise (e.g., smoothing) and processes needed for the successful selection of thresholding parameters for a wide range of signals (e.g., normalization and DC removal).

B. Thresholding

For each type of detection function, and even after post-processing, there will be a number of peaks which are not related to

[3]It is worth noting that identifiable features are not necessarily peaks, they could be steep rising edges or some other characteristic shape. An algorithm able to identify characteristic shapes in detection functions is presented in [38].

onsets. Hence, it is necessary to define a threshold which effectively separates event-related and nonevent-related peaks. There are two main approaches to defining this threshold: fixed thresholding and adaptive thresholding.

Fixed thresholding methods define onsets as peaks where the detection function exceeds the threshold: $d(n) \geq \delta$, where δ is a positive constant and $d(n)$ is the detection function. Although this approach can be successful with signals with little dynamics, music generally exhibits significant loudness changes over the course of a piece. In such situations, a fixed threshold will tend to miss onsets in the most quiet passages, while over-detecting during the loud ones.

For this reason, some adaptation of the threshold is usually required. Generally, an adaptive threshold $\tilde{\delta}[n]$ is computed as a smoothed version of the detection function. This smoothing can be linear, for instance using a low-pass FIR-filter

$$\tilde{\delta}(n) = \delta + \sum_{i=0}^{M} a_i \, d(n-i) \tag{19}$$

with $a_0 = 1$. Alternatively, this smoothing can be nonlinear, using for instance the square of the detection function

$$\tilde{\delta}(n) = \delta + \lambda \sum_{i=-M}^{M} w_i d^2(n+i) \tag{20}$$

where λ is a positive constant and $\{w_i\}_{i=-M \ldots M}$ is a (smooth) window. However, a threshold computed in this way can exhibit very large fluctuations when there are large peaks in the detection function, tending to mask smaller adjacent peaks. Methods based on percentiles (such as the local median) are less affected by such outliers

$$\tilde{\delta}(n) = \delta + \lambda \operatorname{median}\left\{|d(n-M)|, \ldots, |d(n+M)|\right\}. \tag{21}$$

C. Peak-Picking

After post-processing and thresholding the detection function, peak-picking is reduced to identifying local maxima above the defined threshold. For a review of a number of peak-picking algorithms for audio signals, see [39].

For our experiments the detection functions were first normalized by subtracting the mean and dividing by the maximum absolute deviation, and then low-pass filtered. An adaptive threshold, calculated using a moving-median filter [(21)], was then subtracted from the normalized detection function. Finally, every local maximum above zero was counted as an onset. Both the filter and the thresholding parameters (cutoff frequency, M, λ and δ) were hand-tuned based on experimenting, thus resulting in a separate parameter set for each detection function. Values for the cutoff frequency are selected according to the inherent characteristics of each detection method, as discussed in Section III-C; M is set to the longest time interval on which the global dynamics are not expected to evolve (around 100 ms); while λ is set to 1, as it is not critical for the detection. However, experiments show sensitivity to variations of δ, such that error rates can be minimized by changing it between different types of music signals (e.g., pitched percussive, nonpercussive, etc). The signal dependency of the onset detection process is further discussed in Section V-C.

V. Results

A. About the Experiments

This section presents experimental results comparing some of the onset detection approaches described in Section III-C: the high frequency content, the spectral difference, the spread of the distribution of phase deviations, the wavelet regularity modulus and the negative log-likelihood of the signal according to a conditional ICA model. Peak-picking was accomplished using the moving-median adaptive threshold method described in Section IV.

The experiments were performed on a database of commercial and noncommercial recordings covering a variety of musical styles and instrumentations. All signals were processed as monaural signals sampled at 44.1 kHz.

The recordings are broadly divided into four groups according to the characteristics of their onsets: pitched nonpercussive (e.g., bowed strings), pitched percussive (e.g., piano), nonpitched percussive (e.g., drums) and complex mixtures (e.g., pop music). The number of onsets per category is given in Table I; there are 1065 onsets in total.

Onset labeling was done mostly by hand, which is a lengthy and inaccurate process, especially for complex recordings such as pop music: typically including voice, multiple instruments and post-production effects. A small subsection of the database corresponds to acoustic recordings of MIDI-generated piano music which removes the error introduced by hand-labeling. Correct matches imply that target and detected onsets are within a 50-ms window. This relatively large window is to allow for the inaccuracy of the hand labeling process.

B. Discussion: Comparison of Performance

Fig. 7 depicts a graphical comparison of the performance of the different detection functions described in this paper. For each method, it displays the relationship between the percentage of true positives (i.e., correct onset detections relative to the total number of existing onsets) and percentage of false positives (i.e., erroneous detections relative to the number of detected onsets). All peak-picking parameters (e.g., filter's cutoff frequency, λ) were held constant, except for the threshold δ which was varied to trace out the performance curve. Better performance is indicated by a shift of the curve upwards and to the left. The optimal point on a particular curve can be defined as the closest point to the top-left corner of the axes, where the error is at its minimum.

By reading the different optimal points we can retrieve the best set of results for each onset detection method. For the complete database, the negative log-likelihood (90.6%, 4.7%) performs the best, followed by the HFC (90%, 7%), spectral difference (83.0%, 4.1%), phase deviation (81.8%, 5.6%), and the wavelet regularity modulus (79.9%, 8.3%).

However, optimal points are just part of the story. The shape of each curve is also important to analyze, as it contains useful information about the properties of each method that may be

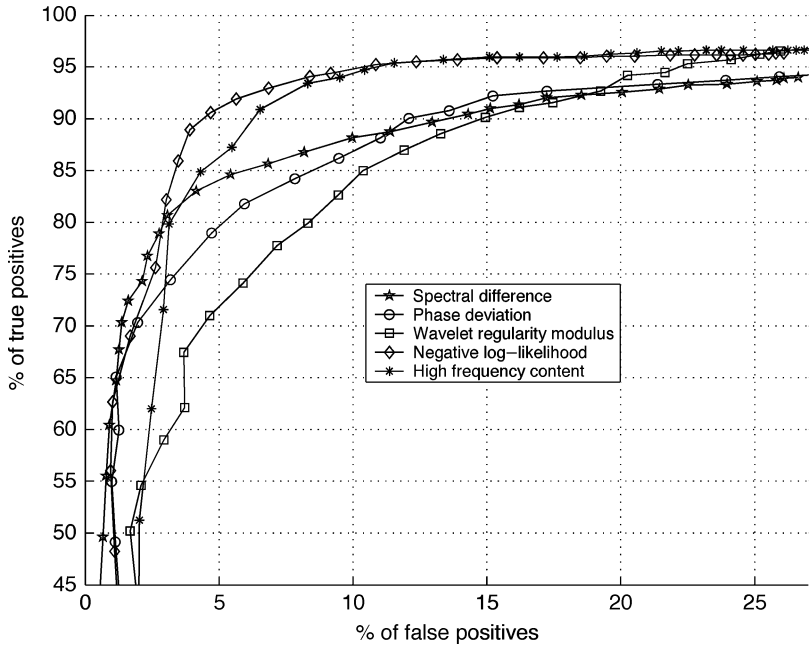


Fig. 7. Comparison of onset detection algorithms: spectral difference, phase deviation, wavelet regularity modulus, negative log-likelihood, and high-frequency content.

relevant to a number of applications. For example, certain applications (e.g., tempo estimation) may require high confidence in the events actually detected even at the expense of under-detecting, while other applications (e.g., time-stretching) require a maximum percentage of detected onsets, regardless of an increase in false detections.

In this context, the negative log-likelihood shows appeal for a number of applications by remaining close to the top-left corner of the axes ([100% TP, 0% FP] point). The method successfully characterizes all types of onsets while producing little unrelated noise.

The HFC is able to retrieve a large proportion of the existing onsets for relatively few false positives, reaching 95% true positives for 10% false positives. However, there is a drop in the number of correctly detected onsets as the rate of false positives is brought below 5%. This is similar to the performance of the wavelet regularity modulus, although the corresponding performance curve rises more slowly as the percentage of false positives increases. Both measures generate sharply defined peaks in their detection functions, and are therefore well-suited for precise time-localization of onsets. This also means that both methods are susceptible to producing identifiable peaks even when no onsets are present.

On the other hand, methods that take information from a number of temporal frames into consideration (e.g., spectral difference, phase deviation) present a smoother detection function profile, minimizing the amount of spurious detections. The cost of this is a reduced ability to resolve all onsets as distinct detectable features. This is reflected in a performance curve that manages relatively high correct onset detection rates for low numbers of false positives, while obtaining comparatively fewer good detections for high rates of false positives (more than 25%). These methods are also less precise in their time localization.

C. Discussion: Dependency on the Type of Onsets

The above analysis emphasizes the dependency of the results on the characteristics of each tested method. In Table I, results are categorized according to the different types of onsets in the database. The idea is to illustrate the dependency of the results on the type of analysis signals. The results in the table correspond to the methods' optimal points for each subset of the database.

The selection of a particular method depends on the type and the quality of the input signal. For example, the phase deviation performs successfully for pitched sounds (both percussive and nonpercussive) where tonal information is key to the detection of onsets, while returning poor results for purely percussive sounds and complex mixtures (where it is affected by phase distortions and the artifacts introduced by speech utterances). On the other hand, the HFC performs better for highly percussive sounds and complex mixtures (with drums) than for music with softer onsets. The spectral difference sits in the middle, slightly below phase deviation for pitched sounds and just under-performing HFC for more percussive and complex sounds.

The wavelet regularity modulus performance is at its best when dealing with simple percussive sounds, otherwise performing poorly with respect to the other methods. Notably, the negative log-likelihood performs relatively well for almost all types of music. This shows the method's effectiveness when fitted with an appropriate model.

These results, while depicting a general trend in the behavior of these approaches, are not absolute. As confirmed by the results in Table I, detection results are strongly signal-dependent, and therefore the plots in Fig. 7 might have been significantly different had a different database been used. In addition, the hand-labeling of onsets is in some rare cases (e.g., in the pop signal) ambiguous and subjective. Finally, for the sake of a fair comparison between the detection functions, we opted to use

TABLE I
ONSET DETECTION RESULTS. COLUMNS SHOW THE PERCENTAGE OF TRUE POSITIVES (TP%) AND PERCENTAGE OF FALSE POSITIVES (FP%) FOR EACH METHOD

PITCHED NON-PERCUSSIVE - 93 ONSETS

METHOD	TP %	FP %
High frequency content	81.7	14.7
Spectral difference	87.1	8.6
Phase deviation	95.7	4.3
Wavelet reg. modulus	92.5	10.1
Neg. log-likelihood	96.8	3.2

PITCHED PERCUSSIVE - 489 ONSETS

METHOD	TP %	FP %
High frequency content	94.1	5.4
Spectral difference	94.9	1.6
Phase deviation	95.5	0.3
Wavelet reg. modulus	92.7	5.1
Neg. log-likelihood	92.4	3.1

NON-PITCHED PERCUSSIVE - 212 ONSETS

METHOD	TP %	FP %
High frequency content	96.7	0.0
Spectral difference	81.6	5.5
Phase deviation	80.7	5.5
Wavelet reg. modulus	88.7	2.2
Neg. log-likelihood	92.9	1.7

COMPLEX MIX - 271 ONSETS

METHOD	TP %	FP %
High frequency content	84.5	10.8
Spectral difference	80.4	10.4
Phase deviation	80.1	24.7
Wavelet reg. modulus	81.9	27.7
Neg. log-likelihood	86.0	10.8

a common post-processing and peak-picking technique. However, performance can be improved for each detection function by fine tuning the peak-picking algorithm for specific tasks.

VI. CONCLUSIONS

In this paper, we have described and compared a variety of commonly used techniques and emerging methods for the detection of note onsets in audio signals. Given the scope of the paper, we have not mentioned methods that are not explicitly devised for this task but that may nevertheless hold some relevance (e.g., matching pursuits and time-frequency adaptive tiling).

Direct comparisons of performance such as those in Section V have to be carefully considered with respect to the different requirements that a given application may have and the type of used audio signals. Generally speaking, a set of guidelines can be drawn to help find the appropriate method for a specific task.

A. Guidelines for Choosing the Right Detection Function

The general rule of thumb is that one should choose the method with minimal complexity that satisfies the requirements of the application. More precisely, good practice usually requires a balance of complexity between preprocessing, construction of the detection function, and peak-picking.

- If the signal is very percussive (e.g., drums), then time-domain methods are usually adequate.
- On the other hand, spectral methods such as those based on phase distributions and spectral difference perform relatively well on strongly pitched transients.
- The complex-domain spectral difference seems to be a good choice in general, at the cost of a slight increase in computational complexity.
- If very precise time localization is required, then wavelet methods can be useful, possibly in combination with another method.
- If a high computational load is acceptable, and a suitable training set is available, then statistical methods give the best overall results, and are less dependent on a particular choice of parameters.

A more detailed description of relative merits can be found in Section III-C and Section V.

B. Perspectives

In this paper, we have only covered the basic principles of each large class of methods. Each one of these methods needs a precise fine-tuning, as described in the relevant papers (referenced in Section III). However, it is not expected that a single method will ever be able to perform perfectly well for all audio signals, due to the intrinsically variable nature of the beginning of sound events, especially between percussive (when transients are related to short bursts of energy) and sustained-note instruments (when transients are related to changes in the spectral content, possibly on a longer time-scale). In fact, we believe that the most promising developments for onset detection schemes lie in the combination of cues from different detection functions [6], [26], which is most likely the way human perception works [40]. More generally, there is a need for the development of analysis tools specifically designed for strongly nonstationary signals, which are now recognized to play an important part in the perceived timbre of most musical instruments [41].

ACKNOWLEDGMENT

The authors wish to thank G. Monti and M. Plumbley for fruitful discussions and help; the S2M team at the Laboratoire de Mécanique et d'Acoustique, Marseille, France, for kindly letting us use their Yamaha Disklavier; and the two anonymous reviewers for their great help in improving the structure of this article.

REFERENCES

[1] B. C. J. Moore, *An Introduction to the Psychology of Hearing*, 5th ed. New York: Academic, 1997.
[2] J. Foote, "Automatic audio segmentation using a measure of audio novelty," in *Proc. IEEE Int. Conf. Multimedia and Expo (ICME2000)*, vol. I, New York, Jul. 2000, pp. 452–455.
[3] M. Goto and Y. Muraoka, "Beat tracking based on multiple-agent architecture—a real-time beat tracking system for audio signals," in *Proc. 2nd Int. Conf. Multiagent Systems*, Dec. 1996, pp. 103–110.
[4] E. D. Scheirer, "Tempo and beat analysis of acoustic musical signals," *J. Acoust. Soc. Amer.*, vol. 103, no. 1, pp. 588–601, Jan. 1998.
[5] A. Klapuri, "Sound onset detection by applying psychoacoustic knowledge," in *Proc. IEEE Int. Conf. Acoustics, Speech and Signal Processing (ICASSP-99)*, Phoenix, AZ, 1999, pp. 115–118.

[6] C. Duxbury, M. Sandler, and M. Davies, "A hybrid approach to musical note onset detection," in *Proc. Digital Audio Effects Conf. (DAFX,'02)*, Hamburg, Germany, 2002, pp. 33–38.
[7] R. J. McAulay and T. F. Quatieri, "Speech analysis/synthesis based on a sinusoidal representation," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. ASSP-34, pp. 744–754, 1986.
[8] X. Serra and J. O. Smith, "Spectral modeling synthesis: a sound analysis/synthesis system based on a deterministic plus stochastic decomposition," *Comput. Music J.*, vol. 14, no. 4, pp. 12–24, winter 1990.
[9] S. Levine, "Audio Representations for Data Compression and Compressed Domain Processing," Ph.D. dissertation, Stanford Univ., Stanford, CA, 1998.
[10] T. Verma, S. Levine, and T. Meng, "Transient modeling synthesis: A flexible analysis/synthesis tool for transient signals," in *Proc. Int. Computer Music Conf.*, Thessaloniki, Greece, 1997, pp. 164–167.
[11] T. Verma and T. Meng, "Sinusoidal modeling using frame-based perceptually weighted matching pursuits," in *Proc. Int. Conf. Acoustics, Speech, and Signal Processing*, vol. 2, Phoenix, AZ, 1999, pp. 981–998.
[12] Z. Settel and C. Lippe, "Real-time musical applications using the FFT-based resynthesis," in *Proc. Int. Computer Music Conf. (ICMC94)*, Aarhus, Denmark, 1994, pp. 338–343.
[13] C. Duxbury, M. Davies, and M. Sandler, "Extraction of transient content in musical audio using multiresolution analysis techniques," in *Proc. Digital Audio Effects Conf. (DAFX '01)*, Limerick, Ireland, 2001, pp. 1–4.
[14] J. Princen and A. Bradley, "Analysis/synthesis filter bank design based on time domain aliasing cancellation," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. ASSP-34, no. 5, pp. 1153–1161, Oct. 1986.
[15] S. Shlien, "The modulated lapped transform, its time-varying forms, and its applications to audio coding standards," *IEEE Trans. Speech Audio Process.*, vol. 5, no. 4, pp. 359–366, 1997.
[16] M. Purat and P. Noll, "Audio coding with a dynamic wavelet packet decomposition based on frequency-varying modulated lapped transforms," in *Proc. ICASSP*, Atlanta, GA, 1996, pp. 1021–1024.
[17] L. Daudet and B. Torrésani, "Hybrid representations for audiophonic signal encoding," *Signal Process.*, vol. 82, no. 11, pp. 1595–1617, 2002.
[18] A. W. Schloss, "On the Automatic Transcription of Percussive Music—From Acoustic Signal to High-Level Analysis," Ph.D. dissertation, Tech. Rep. STAN-M-27, Dept. Hearing and Speech, Stanford Univ., Stanford, CA, 1985.
[19] C. Duxbury, M. Davies, and M. Sandler, "Improved time-scaling of musical audio using phase locking at transients," in *Proc. AES 112th Conv.*, Munich, Germany, 2002, p. 5530.
[20] B. Moore, B. Glasberg, and T. Bear, "A model for the prediction of thresholds, loudness and partial loudness," *J. Audio Eng. Soc.*, vol. 45, no. 4, pp. 224–239, 1997.
[21] X. Rodet and F. Jaillet, "Detection and modeling of fast attack transients," in *Proc. Int. Computer Music Conf.*, Havana, Cuba, 2001, pp. 30–33.
[22] P. Masri, "Computer Modeling of Sound for Transformation and Synthesis of Musical Signal," Ph.D. dissertation, Univ. of Bristol, Bristol, U.K., 1996.
[23] M. Dolson, "The phase vocoder: a tutorial," *Comput. Music J.*, vol. 10, no. 4, 1986.
[24] J. P. Bello and M. Sandler, "Phase-based note onset detection for music signals," in *Proc. IEEE Int. Conf. Acoustics, Speech, and Signal Processing (ICASSP-03)*, Hong Kong, 2003, pp. 49–52.
[25] C. Duxbury, J. P. Bello, M. Davies, and M. Sandler, "A combined phase and amplitude based approach to onset detection for audio segmentation," in *Proc. 4th European Workshop on Image Analysis for Multimedia Interactive Services (WIAMIS-03)*, London, U.K., Apr. 2003, pp. 275–280.
[26] 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 Proces. Lett.*, vol. 11, no. 6, pp. 553–556, Jun. 2004.
[27] C. Duxbury, "Signal Models for Polyphonic Music," Ph.D. dissertation, Dept. Electron. Eng., Queen Mary, Univ. of London, London, U.K., 2004.
[28] M. Davy and S. Godsill, "Detection of abrupt spectral changes using support vector machines. An application to audio signal segmentation," in *Proc. IEEE Int. Conf. Acoustics, Speech and Signal Processing (ICASSP-02)*, Orlando, FL, 2002, pp. 1313–1316.
[29] L. Daudet, "Transients modeling by pruned wavelet trees," in *Proc. Int. Computer Music Conf. (ICMC'01)*, Havana, Cuba, 2001, pp. 18–21.
[30] M. Basseville and I. V. Nikiforov, *Detection of Abrupt Changes—Theory and Application*. Englewood Cliffs, NJ: Prentice-Hall, 1993.
[31] M. Basseville and A. Benveniste, "Sequential detection of abrupt changes in spectral changes of digital signals," *IEEE Trans. Inform. Theory*, vol. 29, pp. 709–724, 1983.
[32] T. Jehan, "Musical Signal Parameter Estimation," M.S. thesis, Univ. of California, Berkeley, CA, 1997.
[33] H. Thornburg and F. Gouyon, "A flexible analysis-synthesis method for transients," in *Proc. Int. Computer Music Conf. (ICMC-2000)*, Berlin, 2000, pp. 400–403.
[34] A. von Brandt, "Detecting and estimating parameter jumps using ladder algorithms and likelihood ratio test," in *Proc. ICASSP*, Boston, MA, 1983, pp. 1017–1020.
[35] S. A. Abdallah and M. D. Plumbley, "Probability as metadata: event detection in music using ICA as a conditional density model," in *Proc. 4th Int. Symp. Independent Component Analysis and Signal Separation (ICA2003)*, Nara, Japan, 2003, pp. 233–238.
[36] J.-F. Cardoso, "Infomax and maximum likelihood for blind source separation," *IEEE Signal Process. Lett.*, vol. 4, no. 4, pp. 112–114, Apr. 1997.
[37] S. A. Abdallah and M. D. Plumbley, "If edges are the independent components of natural scenes, what are the independent components of natural sounds?," in *Proc. 3rd Int. Conf. Independent Component Analysis and Signal Separation (ICA2001)*, San Diego, CA, 2001, pp. 534–539.
[38] ——, "Unsupervised onset detection: a probabilistic approach using ICA and a hidden Markov classifier," in *Cambridge Music Processing Colloq.*, Cambridge, U.K., 2003, [Online] Available: http://www-sigproc.eng.cam.ac.uk/music_proc/2003/contributors.html.
[39] I. Kauppinen, "Methods for detecting impulsive noise in speech and audio signals," in *Proc. 14th Int. Conf. Digit. Signal Process. (DSP2002)*, vol. 2, Santorini, Greece, Jul. 2002, pp. 967–970.
[40] A. S. Bregman, *Auditory Scene Analysis: The Perceptual Organization of Sound*. Cambridge, MA: MIT Press, 1990.
[41] M. Castellengo, "Acoustical analysis of initial transients in flute-like instruments," *Acta Acustica*, vol. 85, no. 3, pp. 387–400, 1999.

Juan Pablo Bello received the engineering degree in electronics from the Universidad Simon Bolivar, Caracas, Venezuela, in 1998 and the Ph.D. degree from Queen Mary, University of London, London, U.K., in 2003.

After a brief period working in industry, he received awards from institutions in Venezuela, the U.S., and the U.K. to pursue Ph.D. studies. He is currently a postdoctoral Researcher for the Centre for Digital Music, Queen Mary, University of London. His research is mainly focused on the semantic analysis of musical signals and its applications to music information retrieval and live electronics.

Laurent Daudet received the degree in statistical and nonlinear physics from the Ecole Normale Supérieure, Paris, in 1997 and the Ph.D. degree in mathematical modeling from the Université de Provence, Marseilles, France, in 2000 on audio coding and physical modeling of piano strings.

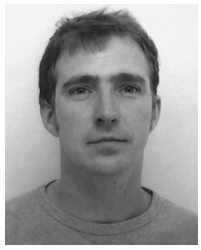
In 2001 and 2002, he was a Marie Curie post-doctoral fellow at the Department of Electronic Engineering, Queen Mary, University of London, London, U.K. Since 2002, he has been a Lecturer at the Université Pierre et Marie Curie (Paris 6), where he joined the Laboratoire d'Acoustique Musicale. His research interests include audio coding, time-frequency and time-scale transforms, sparse representations of audio, and music signal analysis.

Samer Abdallah was born in Cairo, Egypt, in 1972. He received the B.A. degree in natural sciences from Cambridge University, Cambridge, U.K., in 1994, and the M.Sc. and Ph.D. degrees from King's College London, London, U.K., in 1998 and 2003, respectively.

He spent three years working in industry. He is now a postdoctoral Researcher at the Centre for Digital Music, Queen Mary, University of London.

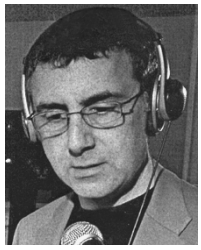
Chris Duxbury received the B.Eng. degree in computer systems from King's College London, London, U.K., in 2000. In 2005, he received the Ph.D. degree from the Centre for Digital Music, Queen Mary, University of London.

He is now a developer for WaveCrest Communications Ltd.



Mike Davies has worked in signal processing and nonlinear modeling for 15 years and held a Royal Society Research Fellowship at the University College of London, London, U.K., and Cambridge University, Cambridge, U.K., from 1993 to 1998. In 2001, he co-founded the DSP Research Group at Queen Mary, University of London, where he is a Reader in digital signal processing. He specializes in nonlinear and non-Gaussian signal processing with particular application to audio signals. His interests include non-Gaussian statistics, independent component analysis, sparse signal representations and machine learning in DSP.

Mr. Davies is currently an Associate Editor for the IEEE TRANSACTIONS ON SPEECH AND AUDIO PROCESSING.



Mark B. Sandler (M'87–SM'95) was born in London, U.K., in 1955. He received the B.Sc. and Ph.D. degrees from the University of Essex, U.K., in 1978 and 1984, respectively.

He is Professor of signal processing at Queen Mary, University of London, where he moved in 2001, after 19 years at King's College London. He was founder and CEO of Insonify Ltd., an Internet Audio Streaming startup for 18 months. He has published over 250 papers in journals and conferences.

Dr. Sandler is a Fellow of the IEE and of the Audio Engineering Society. He is a two-time recipient of the IEE A. H. Reeves Premium Prize.