Onset Event Decoding Exploiting the Rhythmic Structure of Polyphonic Music

Norberto Degara, *Student Member, IEEE*, Matthew E. P. Davies, Antonio Pena, *Member, IEEE*, and Mark D. Plumbley, *Member, IEEE*

***Abstract*—In this paper, we propose a rhythmically informed method for onset detection in polyphonic music. Music is highly structured in terms of the temporal regularity underlying onset occurrences and this rhythmic structure can be used to locate sound events. Using a probabilistic formulation, the method integrates information extracted from the audio signal and rhythmic knowledge derived from tempo estimates in order to exploit the temporal expectations associated with rhythm and make musically meaningful event detections. To do so, the system explicitly models note events in terms of the elapsed time between consecutive events and decodes the most likely sequence of onsets that led to the observed audio signal. In this way, the proposed method is able to identify likely time instants for onsets and to successfully exploit the temporal regularity of music. The goal of this work is to define a general framework to be used in combination with any onset detection function and tempo estimator. The method is evaluated using a dataset of music that contains multiple instruments playing at the same time, including singing and different music genres. Results show that the use of rhythmic information improves the commonly used adaptive thresholding onset detection method which only considers local information. It is also shown that the proposed probabilistic framework successfully exploits rhythmic information using different detection functions and tempo estimation algorithms.**

***Index Terms*—Music signal processing, onset detection, rhythm, tempo.**

I. Introduction

THE task of recovering the start times of events in an audio signal is known as *onset detection*. Onset detection is an important task in areas such as speech processing or audio coding. The successful extraction of onset times enables the temporal segmentation and an adaptive time–frequency representation of a signal at a meaningful time–scale [1].

Music is one of the most important sources of information on the Internet and the development of algorithms for searching, navigating, retrieving, and organizing music has become a major challenge. This field of research is known as *Music Information Retrieval* (MIR) and it has significantly gained in interest in recent years. In this domain, music onset detection forms the basis of many higher level processing tasks, including polyphonic transcription [2], beat tracking [3], and interactive musical accompaniment [4].

A. Related Work

Automatic onset detection constitutes a difficult process due to the complexity and diversity of music. Many approaches exist and several reviews can be found in the literature. For example, Bello *et al.* [5] categorize detection techniques into methods based on the use of predefined signal features and methods based on probabilistic models. Dixon [6] examines onset detection methods based on spectral features and Collins [7] places emphasis on psycho-acoustically motivated onset detection algorithms. Finally, Lacoste and Eck [8] review previous work on machine learning algorithms for onset detection. The best performing method in the Music Information Retrieval Evaluation eXchange (MIREX) 2010 was that presented by Eyben *et al.* in [9] and followed a machine learning approach. The algorithm is based on auditory spectral features and relative spectral differences processed by a bidirectional long short-term memory recurrent neural network.

The standard approach for finding onset positions is a two stage process. First, a mid-level representation, often referred to as an *onset detection function*, is extracted from the audio signal. The aim of this detection function is to exhibit peaks at likely onset locations by measuring changes in the short term properties of the audio signal such as energy, frequency content, or phase information. Once the onset detection function has been generated, the temporal locations of these peaks are recovered by picking the local maxima. Then, standard onset detection methods apply a threshold that is used to decide if a peak is likely to be an onset based on its height [5], [6]. Each peak is therefore evaluated individually and the potential temporal relations with other peaks are not considered.

Musical information is usually encoded into multiple relationships between sound events. As a result, music signals are highly structured in terms of temporal regularity, largely defining the rhythmic characteristics of a musical piece [10], [11]. The underlying periodicities can be perceived at different time levels and the most salient of these metrical levels is the *beat*, also known as foot-tapping rate or *tactus*. In general, the time between consecutive onsets corresponds to multiples and

Manuscript received September 06, 2010; revised February 14, 2011; accepted April 04, 2011. Date of publication May 23, 2011; date of current version September 16, 2011. This work was supported in part by the Galician Regional Government (2007/000023-0, 2009/062), the Spanish Government (TEC2009-14414-C03-03), the European Regional Development Fund, and the EPSRC Grants EP/G007144/1 and EP/E045235/1. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Meinard Müller.

N. Degara and A. Pena are with the Signal Theory and Communications Department, University of Vigo, Vigo 36310, Spain (e-mail: ndegara@gts.uvigo.es; apena@gts.uvigo.es).

M. E. P. Davies is with the Instituto de Engenharia de Sistemas e Computadores (INESC) do Porto, 4200–465 Porto, Portugal (e-mail: matthew.davies@eecs.qmul.ac.uk).

M. D. Plumbley is with the School of Electronic Engineering and Computer Science, Queen Mary University of London, London E1 4NS, U.K. (e-mail: mark.plumbley@eecs.qmul.ac.uk).

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

Digital Object Identifier 10.1109/JSTSP.2011.2146229

fractions of the beat period, with small deviations in timing and tempo [3]. In the following, *rhythmic structure* refers to the statistical distribution of times between onsets that largely characterizes the temporal nature of a music signal. This rhythmic structure defines a temporal context that can be used to predict the expected location of note events from signal observations.

Recently, several researchers have introduced the use of local periodicity information for onset detection. In [8], Lacoste *et al.* introduce a supervised learning algorithm for onset detection. A neural network is trained on a set of spectrogram features and a tempo trace using a large dataset for training. Hazan *et al.* [12] present a method that predicts onsets based on temporal patterns learned from past events in monophonic audio signals. Also, Grosche and Müller [13], [14] have recently proposed a new mid-level representation that captures the local periodicity of an onset detection function. At each time instant, the predominant tempo and a sinusoidal kernel that best explains the local periodic structure of the onset signal are estimated from a tempogram. Then, these kernels are accumulated along time resulting in a function that reveals tempo and beat information. When applied to onset detection, this representation shows that the number of missed detections, false negatives, is reduced at the expense of increasing the number of false onset detections, i.e., false positives. In previous work [15], we proposed a rhythmically aware onset detection algorithm. The system uses a dynamic programming algorithm to favour event locations that are rhythmically related. However, it uses an ad-hoc penalty term that is not explicitly related to any probability distribution.

A more formal probabilistic formulation could be adopted in order to exploit the temporal structure of music in onset detection. The integration of music knowledge into probabilistic models is an important field where many researchers are contributing. For example, Klapuri *et al.* [16] and Peeters [17], [18] define a probabilistic framework in the context of meter analysis and beat tracking. Raphael [19] uses a probabilistic approach to audio to score alignment. Also, examples of using probabilistic formulations for automatic transcription can be found in the work of Mauch *et al.* [20] and Ryynänen *et al.* [21]. However, little work has been done in the probabilistic integration of musical knowledge for onset detection. The inverse problem is discussed in [22], where a sequence of onsets is assigned to a corresponding rhythm and tempo process. The statistical approaches discussed in [5] and [23] assume that the signal can be described in terms of a probabilistic model; however, only local information is considered and no rhythmic information is exploited. One related work in onset detection is that of Thornburg *et al.* [24] where a Bayesian approach for joint melody extraction and onset detection for monophonic audio signals is proposed. The system models note events in terms of transient and steady-state-regions, however it does not model the temporal expectations from the rhythmic structure of the audio signal. This is proposed as a future development in a previous work of Thornburg [25], where a dynamic Bayesian network is suggested for the joint estimation of tempo, onset events, and melody.

B. Motivation

Our aim in this paper is to present a rhythmically informed approach to onset detection in polyphonic music. To incorporate musical knowledge into onset detection we propose a simple probabilistic framework for the problem. The rhythmic nature of the music signal is statistically modeled and integrated with the information extracted from the signal using a hidden Markov model (HMM). In this way, our proposed method is able to exploit the temporal predictions derived from the rhythmic properties of the music signal and decode a sequence of rhythmically meaningful onsets. This differs from standard thresholding methods for onset estimation where only individual peaks are used to find onset locations. Another innovation of the proposed model is the use of an optimality criterion to perform onset detection. Applying a maximum *a posteriori* measure, the algorithm decodes a sequence of onset events which best explain the extracted information given the underlying rhythmic structure of the audio signal. In addition, a method to weight the influence of the rhythmic information in the onset estimation is also discussed.

Previous onset detection research was mainly focused on finding new representations that detect onsets more accurately compared to other approaches, [5], [26]–[28]. On the contrary, the goal of this work is not to present a new type of onset detection function, but to propose a novel strategy for exploiting the temporal structure of music. We do this by presenting a general probabilistic framework to be used in combination with any onset detection function and tempo estimator.

C. Proposed Model

The block diagram of the proposed method is shown in Fig. 1 where the temporal expectations from the rhythmic structure of an input signal are modeled using a hidden Markov model (HMM) [29]. The audio signal is frame-wise processed and an event detection function that reflects potential onset locations is extracted. The onset detection function is then used to track the tempo of the audio signal and to compute the transition probabilities from a rhythmic model which defines the underlying temporal structure of the musical events. As in beat tracking, tempo information is estimated independently of the time events to reduce the search space [16], [18], [30]. Next, the peaks of the onset detection signal are extracted and the state-conditional probability distributions of the observed peak values are estimated. Finally, the Viterbi algorithm [31] is used to decode the most likely sequence of events by looking through all possible paths. In summary, the system integrates rhythmic information and features extracted from the signal by modeling the observations and transitions using a probabilistic framework which gives the optimal sequence of events for the onset detection function. We refer to this process as *onset rhythmic decoding* in order to make clear the difference with the traditional onset detection approach.

The remainder of this paper is structured as follows. In Section II, we describe the different elements of the onset rhythmic decoding system shown in Fig. 1. Then, Section III presents the dataset and the measures used to evaluate the performance of the proposed system. Section IV discusses the experimental results. Finally, the conclusions and future work are presented in Section V.

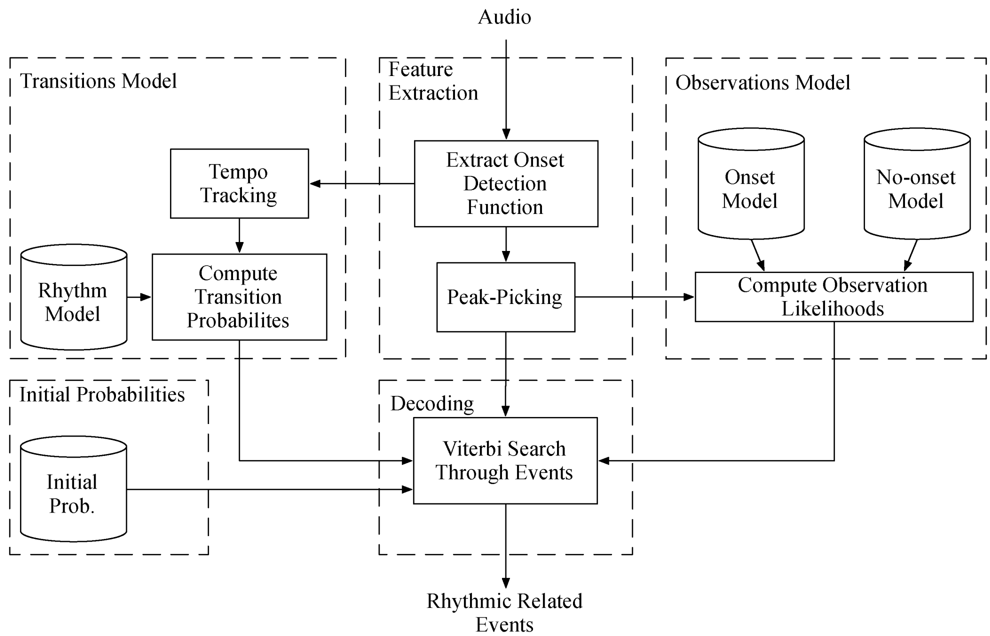

Fig. 1. Block diagram of the proposed probabilistic model for onset detection.

II. Onset Rhythmic Decoding System

This section describes the different parts of the onset rhythmic decoding system illustrated in Fig. 1. Section II-A presents the feature extraction process and the different elements of the probabilistic model are described in Section II-B.

A. Feature Extraction

The front-end to our proposed method shown in Fig. 1 is an onset detection function generated from the input audio signal that shows peaks at likely onset locations. Within our probabilistic approach we do not specify a particular onset detection function and any such function could be used, provided it has the appropriate temporal resolution.

As in [6], the temporal resolution of the onset detection function is 10 ms and the local mean is subtracted prior to peak-picking to normalize the detection signal. Let $f(t)$ denote the onset detection function at time frame t, then the normalized detection function $d(t)$ is calculated as

$$d(t) = f(t) - \frac{\sum_{k=t-mw}^{t+w} f(k)}{mw+w+1}. \tag{1}$$

A peak at time t is selected if it is a local maximum

$$d(t) \geq d(k), \quad \forall k : t-w \leq k \leq t+w \tag{2}$$

where $w = 3$ is the size of the window used to calculate the local maximum and $m = 3$ is a multiplier used to calculate the mean over a larger range before the peak, which is useful to emphasize onsets rather than offsets. This normalization accounts for changes in the local mean of the onset detection function so that small peaks in the vicinity of a large peak are not selected [6]. Note that a threshold value of 0 is implicitly used in the peak-picking process defined in (2) so every peak is considered as an onset candidate by the proposed rhythmic decoding algorithm.

As an example, Fig. 2 shows the normalized complex domain detection function, as defined in [6], of a musical excerpt. Peaks are marked with circles and ground truth onsets with vertical lines. The peaks of the detection function correspond with potential locations of onset events. However, peak time instants do not always agree with the exact location of the onset. In addition, there are spurious peaks which do not correspond to note onsets (false positives) and note onsets which are not represented by any peak in the detection function or by peaks of very low amplitude (false negatives). As we could expect, the musical events show a clear rhythmic structure since the ground truth onsets marked in Fig. 2 are regularly spaced in time.

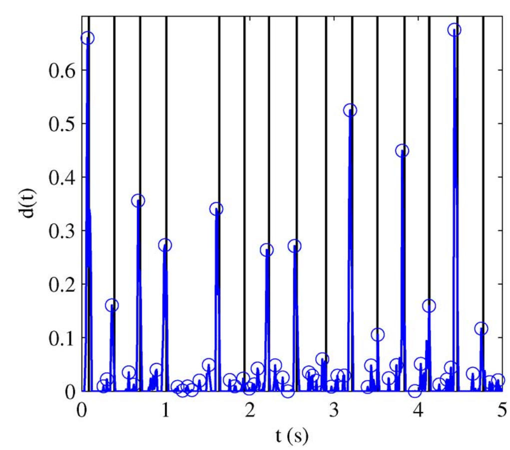

Fig. 2. Example of a normalized complex domain detection function [6] of an excerpt of the rock song "Can't Stand It" by Wilco.

B. Probabilistic Model

Hidden Markov Models are useful for representing temporal dependencies and integrating high level knowledge with signal observations. The system proposed in Fig. 1 integrates contextual information and audio features by defining an HMM where a hidden state variable τ represents the elapsed time, in temporal frames, since the last onset event. The total number of states N is determined by the maximum time between consecutive events

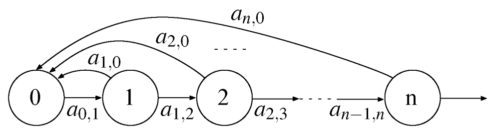

Fig. 3. Modeling temporal expectations for onset decoding in music signals using a hidden Markov model.

accepted and so the possible states for τ are $\{0, 1, \ldots, N-1\}$. Following this notation, a state $\tau = n$ implies that there have been n frames since the last onset event and a state 0 denotes an onset event. The state at any time frame t is represented as τ_t and a particular state sequence is denoted as $\tau_{1:T} = (\tau_1, \tau_2, \ldots, \tau_T)$.

The underlying temporal structure of the audio signal is then encoded in the state transition probabilities, denoted as $a_{i,j} = \mathrm{P}(\tau_t = j | \tau_{t-1} = i)$. Fig. 3 shows the HMM model where states are represented by nodes and transitions by links. The state variable τ measures the elapsed time since the last visit to state 0 and therefore the only possible transitions are from one state n to the following state $n+1$ or to the onset state event 0. This significantly reduces the search space.

At each state, the system emits an observation, o_t, which is conditioned only on the current state and the state-conditional observation probability is $P(o_t|\tau_t)$. If we represent the set of time instant frames where there is a local peak of the normalized detection function $d(t)$ as $\mathcal{T}$, the observation at this time instant o_t is

$$o_t = \begin{cases} d(t), & \text{if } t \in \mathcal{T} \\ 0, & \text{otherwise.} \end{cases} \tag{3}$$

That is, the non-null observations of o_t are the peaks of the onset detection function that have been extracted from the input audio signal. As described in [5], onset detection functions are designed to show the potential location of onset events in their peaks, and on this basis we choose these as observation for the proposed probabilistic model. It is important to note that none of the onset location candidates are discarded in (3). In addition, the selection of the peaks of the detection function as observations facilitates the estimation of the observation likelihoods as shown in Section II-B3. To make the rhythmic decoding system more robust to spurious onset detections, we could have used the detection function as observation of the proposed probabilistic model, that is $o_t = d(t) \forall t$. However, informal tests show that the global accuracy is higher when using (3).

The goal of the proposed probabilistic model is to find the sequence of states which best explains the extracted audio signal observations o_t, given the underlying rhythmic structure of the audio signal. The most likely sequence of states $\tau^*_{1:T}$ that led to the observations $o_{1:T}$ is estimated as

$$\tau^*_{1:T} = \arg\max_{\tau_{1:T}} \mathrm{P}(\tau_{1:T}|o_{1:T}). \tag{4}$$

The sequence of rhythmically related onset times is obtained by selecting the time instants that the decoded sequence of states, $\tau^*_{1:T}$, visited the onset state $\tau_t = 0$. The posterior probability of a sequence of states $\tau_{1:T}$ can be calculated as

$$\mathrm{P}(\tau_{1:T}|o_{1:T}) \propto \mathrm{P}(o_1|\tau_1)\mathrm{P}(\tau_1) \prod_{t=2}^{T} \mathrm{P}(o_t|\tau_t)\mathrm{P}(\tau_t|\tau_{t-1}) \tag{5}$$

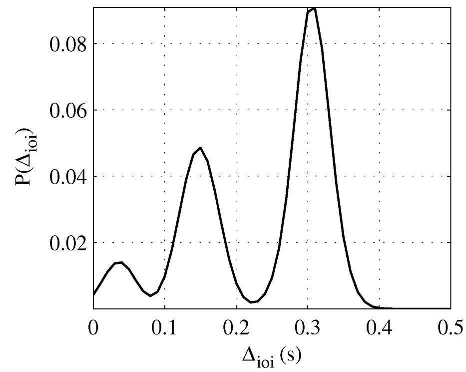

Fig. 4. Distribution estimate of the inter-onset time interval of an excerpt of the rock song "Can't Stand It" by Wilco.

where $\mathrm{P}(\tau_1)$ is the initial state distribution, $\mathrm{P}(\tau_t|\tau_{t-1})$ the transition probabilities, and $\mathrm{P}(o_t|\tau_t)$ the observation likelihoods. These probabilities together with the decoding process govern the HMM model introduced in this section and are described next. Each of these sub-systems are shown as dashed blocks in Fig. 1.

1) Initial Probabilities: The initial probabilities $\mathrm{P}(\tau_1 = n)$, for states $n = 0, \ldots, N-1$, model the time instant when the first onset is expected to happen. This is shown in the initial probabilities block in Fig. 1 and feeds into the rhythmic decoding process. While we expect there to be rhythmic dependencies between onset times, we should not place any assumptions over the location of the first event (e.g., we cannot guarantee the first onset will coincide with the start of the excerpt); therefore, we adopt a uniform distribution for the initial probabilities $\mathrm{P}(\tau_1 = n)$.

2) Transition Distribution: The rhythmic structure of a music signal defines the approximate regularity of the position of the note events in most contexts. This regularity can be used to define temporal expectations about the position of sound events as used in beat tracking systems [16] and [17]. To illustrate the regularity of musical onset events, Fig. 4 shows the distribution of the time difference between consecutive hand-labeled onsets, the inter-onset interval (IOI), of an excerpt of the rock song "Can't Stand It" by Wilco, which is part of the dataset described in Section III-A. The probability density function is estimated using a kernel density method with a normal kernel [32]. The figure reveals the predominant periodicities, in this case the beat period, 0.3 s, and half the beat period, 0.15 s. The lowest mode centered around 0.04 s is related to the inaccuracy associated with the hand labeling process of the annotations [5].

In general, onsets happen at multiples or fractions of the beat period [3], [10]. In this sense, the *rhythmic model* block in Fig. 1 encodes this periodicity information by defining the multiples and fractions of the beat period that are considered as likely locations of events. In order to account for these temporal expectations, a *tempo tracking* system that informs the onset detection task of the beat period is used. We do not specify any particular tempo estimation method since our probabilistic model defines a general framework to be used in combination with any tempo induction algorithm, provided it obtains an appropriate estimation of the beat period. Section III-B details the tempo induction methods used to analyze the performance of the proposed onset decoding algorithm.

As shown in the *transitions model* of Fig. 1, the transition probabilities $\mathrm{P}(\tau_t|\tau_{t-1})$ are estimated using this tempo and rhythmic model information. Let T_b be the estimated beat period and $\{m_k\}$ with $k = 1, \ldots, K_i$ the set of beat period ratios that define a rhythmic template $M_i = \{m_k\}$, where i denotes a specific template. The probability distribution function of the inter-onset time, $\mathrm{P}(\Delta_{ioi})$, for the input audio signal is estimated as a mixture of independent and equally likely Gaussians

$$\mathrm{P}(\Delta_{ioi}) = \frac{1}{K_i} \sum_{k=1}^{K_i} \frac{1}{\sqrt{2\pi\sigma_k^2}} \exp\left(-\frac{(\Delta_{ioi} - \mu_k)^2}{2\sigma_k^2}\right) \quad (6)$$

where μ_k and σ_k are the mean and the standard deviation of each component k. The Gaussians are centered at a value related to the beat period as

$$\mu_k = m_k T_b \quad (7)$$

and the value of σ_k decreases proportionally to μ_k

$$\sigma_k = \frac{\mu_k}{\beta}. \quad (8)$$

The parameter m_k is a positive rational number to allow Gaussians to be centered at, for example, 1/2 or 2/3 the beat period T_b. To avoid overlap between Gaussians the parameter $\beta > 1$ defines the rate of change in the value of σ_k. The lower the mean value μ_k, the smaller the standard deviation σ_k. Through informal testing a value of $\beta = 18$ was found to be appropriate. The width of the Gaussians, set by σ_k, allows for departures from strict metrical timing that occur in musical performances [11] and for timing deviations from the exact location of onsets caused by the detection function [33].

Once the distribution of the inter-onset time has been estimated, the temporal structure derived from the detection function is encoded in the state transition probabilities. As shown in Fig. 5, if there are n frames between two consecutive onset events, the $a_{i,j}$'s and the distribution of Δ_{ioi}, in frames, can be related as

$$\mathrm{P}(\Delta_{ioi} = n) = a_{n-1,0} \prod_{k=0}^{n-2} a_{k,k+1} \quad (9)$$

where $a_{i,j}$ denotes the state transition probabilities $\mathrm{P}(\tau_t = j|\tau_{t-1} = i)$. The state transition probabilities, for $n = 1, \ldots, N$, can be iteratively calculated as follows:

$$a_{n-1,0} = \frac{\mathrm{P}(\Delta_{ioi} = n)}{\prod_{k=0}^{n-2} a_{k,k+1}} \quad (10)$$

$$a_{n-1,n} = 1 - a_{n-1,0}. \quad (11)$$

Note that (11) reflects the fact that the only possible transitions defined in our model are the transitions from state n to the following state $n+1$ or to the onset event 0.

The transition probability depends on the rhythmic template, M_i, assumed for the signal to be analyzed. In order to include several musical styles such as pop, rock, techno, classic, and jazz music, a set of multiple rhythmic templates, $\{M_i\}$, have to be considered. Similar to [34], the following rhythmic templates are defined: a *single template*, $M_{\text{single}} = \{1\}$, which assumes potential locations of events at the beat period; a *half template*, $M_{\text{half}} = \{1, 1/2\}$, which takes events at the beat period and half the beat period; a *double template*, $M_{\text{double}} = \{1, 2\}$, which considers potential locations of events at the beat period and double the beat period; and, finally, an *odd template*, $M_{\text{odd}} = \{1, 2/3, 1/3\}$, which assumes events at the beat period, 2/3 and 1/3 the beat period. Informal tests show that including additional templates such as $\{1,\ 1/2,\ 2\}$ does not increase the performance of the system. With these set of templates we want to account for odd divisions on the time between events and for ambiguities associated to the metrical level at which the tempo estimation occurs [35]. The most common situations are the half and double tapping rate where the estimated tempo is half or twice the annotated tempo. Therefore, the proposed *half* and *double templates* are intended to deal with these ambiguities in the beat period estimation process. Also, depending on the class of music to be analyzed, a different set of templates could be defined. In fact, if the study was restricted to pop music, a *single* and a *half template* would probably be enough since beat structures tend to be simple in pop music.

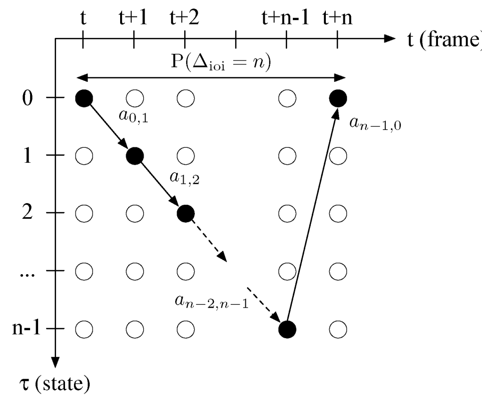

Fig. 5. State transition probabilities calculation.

To decode the sequence of rhythmic related onsets we must select a rhythmic template. Using Bayes' formula, the probability of the rhythmic model M_i given the observation sequence is

$$\mathrm{P}(M_i|o_{1:T}) = \frac{\mathrm{P}(o_{1:T}|M_i)\mathrm{P}(M_i)}{\mathrm{P}(o_{1:T})}. \quad (12)$$

This allows us to choose the template with the largest probability. Assuming that the initial probability of the rhythmic model $\mathrm{P}(M_i)$ is uniformly distributed, the model that best explain the observations can be chosen as

$$M^* = \underset{\{M_i\}}{\arg\max} \left[\mathrm{P}(o_{1:T}|M_i)\right]. \quad (13)$$

As long as the input audio is equally distributed along the set of rhythmic templates this assumption is valid. Otherwise, a contextual music model would be required to characterize the model probabilities $\mathrm{P}(M_i)$ of the templates. This is left as a topic for future work.

3) Observation Likelihood: At every time frame t, the system is in a state τ_t and it emits the observation o_t. The observation likelihood $\mathrm{P}(o_t|\tau_t)$ represents the probability of observing o_t given that the system is in state τ_t.

These observation likelihoods need to be estimated for $n = 0, \ldots, N-1$ possible states. However, a very large number of training samples would be required to estimate the parameters of the N probability distribution functions $\mathrm{P}(o_t|\tau_t = n)$. To overcome this situation the distribution of all non-onset states $n = 1, \ldots, N-1$ are assumed to be the same and tied together, equivalent to the data model simplification used in [19]. In other words, the output distribution is assumed to depend only on an onset state $n = 0$ and a non-onset state $n \neq 0$ and therefore only two output distributions have to be estimated, one for the onset state $\mathrm{P}(o_t|\tau_t = 0)$ and another one for all the non-onset states $\mathrm{P}(o_t|\tau_t \neq 0)$. It could be argued that the distribution of the observations for states next to an onset state, for example $n = 1$, would resemble the observation onset distribution, $\mathrm{P}(o_t|\tau_t = 0)$, instead of the non-onset distribution, $\mathrm{P}(o_t|\tau_t \neq 0)$, since we expect large values of the detection function in the neighborhood of an onset. However, we know from (3) that the observation sequence o_t is made up of the local peaks of the detection function $d(t)$ and the observations o_t in a neighborhood of an onset are set to 0. Then this state-tying simplification agrees with the observation vector o_t.

On the one hand no-onset observations will typically be zero and the probability will go down as the value of the observation increases. Accordingly, an exponential distribution is chosen for the *no-onset model* shown in Fig. 1, with parameter λ for the no-onset states $n \neq 0$:

$$\mathrm{P}(o_t|\tau_t \neq 0) = \lambda \exp(-\lambda o_t). \tag{14}$$

On the other hand, onset observations will be generated by different sources of events and its score will distribute around a mean peak value. Therefore, a Gaussian distribution, corresponding to the *onset model* in Fig. 1, with parameters μ and σ is chosen for the onset state $n = 0$:

$$\mathrm{P}(o_t|\tau_t = 0) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(o_t - \mu)^2}{2\sigma^2}\right). \tag{15}$$

The parameters of these models can be either fitted or trained using actual data. In our case, the observation likelihoods are directly estimated by fitting the distributions detailed above using the peak data observations obtained from the detection function $d(t)$.

4) Decoding: The estimates for the initial probabilities $\mathrm{P}(\tau_1)$, the transition probabilities $\mathrm{P}(\tau_t|\tau_{t-1})$, and the observation likelihoods $\mathrm{P}(o_t|\tau_t)$ define the underlying model for the input audio signal. Given this model and (4) and (5), we can determine the sequence of states $\tau^*_{1:T}$ which best explains the extracted audio signal observations o_t as

$$\tau^*_{1:T} = \underset{\tau_{1:T}}{\arg\max} \left[\mathrm{P}(o_1|\tau_1)\mathrm{P}(\tau_1) \prod_{t=2}^{T} \mathrm{P}(o_t|\tau_t)\mathrm{P}(\tau_t|\tau_{t-1})\right]. \tag{16}$$

Reformulating the optimization by taking the negative logarithm in (16) yields to

$$\tau^*_{1:T} = \underset{\tau_{1:T}}{\arg\min} \left[-\ln\left(\mathrm{P}(o_1|\tau_1)\mathrm{P}(\tau_1)\right) - \sum_{t=2}^{T} \ln\left(\mathrm{P}(o_t|\tau_t)\right) - \sum_{t=2}^{T} \ln\left(\mathrm{P}(\tau_t|\tau_{t-1})\right)\right]. \tag{17}$$

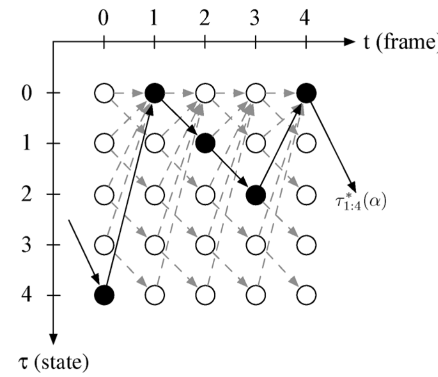

Fig. 6. Illustration of the probability evaluation for onset decoding.

Analogous to the weighted dynamic program proposed for beat tracking in [36], an additional parameter α is introduced to evaluate the influence of the rhythmic structure in the onset decoding performance:

$$\tau^*_{1:T}(\alpha) = \underset{\tau_{1:T}}{\arg\min} \left[-\alpha \ln\left(\mathrm{P}(o_1|\tau_1)\mathrm{P}(\tau_1)\right) - (1-\alpha)\sum_{t=2}^{T} \ln\left(\mathrm{P}(o_t|\tau_t)\right) - \alpha \sum_{t=2}^{T} \ln\left(\mathrm{P}(\tau_t|\tau_{t-1})\right)\right] \tag{18}$$

where α can vary between 0 and 1. This parameter balances the importance of the rhythmic structure encoded in $\mathrm{P}(\tau_t|\tau_{t-1})$ and the observation likelihoods $\mathrm{P}(o_t|\tau_t)$. As α approaches 0 the optimization becomes equivalent to a simple thresholding method since the distribution of the observations is the only term to be considered. In this case, we can expect to minimise the number of false negative events. As α approaches 1 the optimization becomes similar to a beat tracker and the number of false positive onset events is minimized. Setting $\alpha = 0.5$ leads us to the optimization objective defined in (17).

The Viterbi algorithm [31] is used to determine the most likely sequence of states $\tau^*_{1:T}(\alpha)$ that led to the observations $o_{1:T}$ given the underlying rhythmic model. This is illustrated in Fig. 6, where the black-arrowed line shows the best target sequence $\tau^*_{1:T}(\alpha)$ decoded from (18). Note that the proposed system looks through all the possible combinations of state path sequences represented by gray-arrowed lines and therefore every peak extracted from the detection function, which defines the system observations o_t, is considered as an onset candidate. The model takes into account the temporal dependencies between successive peak candidates and performs a noncausal decoding of the onsets. The search space is larger than the one considered in traditional thresholding, where each peak candidate is individually considered. Also, the computational load is larger when using rhythmic information.

Finally, the most likely set of rhythmically related onset times $\mathcal{T}^*_{\text{onset}}(\alpha)$ is obtained by selecting the time instants that the decoded sequence of states $\tau^*_{1:T}(\alpha)$ visited the onset state $n = 0$:

$$\mathcal{T}^*_{\text{onset}}(\alpha) = \left\{t : \tau^*_t(\alpha) = 0\right\}. \tag{19}$$

TABLE I
DATASET OF COMPLEX MIXTURES

No.	Reference	File	Genre	Duration	Onsets
1	Bello *et. al.* [5]	Jaillet 65	Pop	3s	15
2		Jaillet 70	Pop	4s	19
3		Dido	Pop	12s	56
4		Fiona	Pop	8s	40
5		Jaxx	Techno	6s	45
6		Wilco	Rock	15s	63
7		Metheny	Jazz	6s	33
8	Daudet *et. al.* [33]	11	Techno	6s	56
9		12	Rock	15s	59
10		13	Jazz	14s	47
11		14	Jazz	11s	53
12		15	Classic	20s	38
13		17	Pop	15s	27
Total onsets					551

III. EXPERIMENTAL SETUP

This section describes the database and the performance measures used to evaluate the proposed onset rhythmic decoding system. In addition, we detail the detection functions and the tempo estimation methods used to estimate the peak candidates and the beat period of the input audio signal. The reference system that defines the baseline performance used for comparison is also described.

A. Dataset and Evaluation

To evaluate our onset detection model we use audio examples from existing databases. Since traditional onset detection algorithms and, more recently, score-level fusion approaches have shown good performance on single-instrument audio signals, [27] and [37], we focus on the case of complex mixtures using the excerpts from [5] and [33]. The selected audio signals are mixtures of multiple instruments and singing, including different music genres. The ground truth onsets were hand annotated by musical experts. Details about the annotation process can be found in [5] and [33]. The dataset is comprised of 13 polyphonic audio files with a total 551 annotated onsets. The size of the dataset is comparable to the MIREX complex mixture class for onset detection [38]. This dataset was also recently evaluated in [27] and it has been shown that fusion of multiple onset detection functions does not provide any benefit on complex mixtures over traditional thresholding. Table I shows the genre, duration and the number of onsets for each file in the dataset.

Although the signals of the database are short, the excerpts are long enough to be able to obtain a reliable estimate of the beat period. Tempo estimation methods generally use an analysis window of 4 s to 8 s and, as indicated in [13] and [30], this length is long enough to provide a robust estimate of the periodicity of the detection function. The signals of the database are between 4 s and 20 s except for the first test signal whose duration is 3 s. This is a pop excerpt with a clear rhythmic structure where the tempo estimation algorithms included in the system obtain a reasonable estimate of the beat period.

For the quantitative evaluation of our onset detection system we follow the approach from [6] using the following quantities: precision, P, recall, R, and F-measure, F;

$$P = \frac{n_{\mathrm{cd}}}{n_{\mathrm{cd}} + n_{\mathrm{fp}}} \tag{20}$$

$$R = \frac{n_{\mathrm{cd}}}{n_{\mathrm{cd}} + n_{\mathrm{fn}}} \tag{21}$$

$$F = \frac{2PR}{P + R} \tag{22}$$

where n_{cd} is the number of correctly detected onsets, n_{fp} is the number of false positives (detection of an onset when no ground truth onset exists), and n_{fn} is the number of false negatives (missed detections). A correct detection is defined as one occurring within a $+/-50$ ms tolerance window of each ground truth onset [38]. Total performance results are calculated by accumulating the number of false positives, false negatives and correct detection across all the files of the dataset.

B. Reference Systems

The proposed method defines a general probabilistic framework to be used in combination with any onset detection function and tempo induction algorithm. This differs from current onset detection research which focuses on the definition of a feature that works for multiple signals of different nature. Therefore, to analyze the performance of our rhythmic decoding system under different conditions, three state-of-the-art onset detection functions will be used: the Complex Domain detection function [5], the Spectral Flux detection function [5] and the Resonator Time Frequency Image detection function [26]. Also, the effect of tempo estimation in the accuracy of the decoded onsets will be analyzed by comparing the proposed probabilistic model when using two different tempo estimators: a HMM version of the tempo estimation method presented in [30] by Davies *et al.* and the algorithm introduced by Ellis in [36].

To evaluate the benefit of using rhythmic information, the performance of the proposed model is compared with the commonly used adaptive thresholding algorithm described in [6] and [5]. In adaptive thresholding the detection function is first normalized using its local mean as in (1), then the peaks are extracted as in (2) and finally a fixed threshold is used to decide if a peak is an onset or not. As discussed in Section II-A, this normalization process accounts for changes in the local dynamics of the onset detection function so that small peaks in the vicinity of a large peak are not selected as onset candidates. Also, note that a threshold value of 0 is implicitly used in (2), therefore every peak of the normalized detection function is initially considered as an onset candidate in our rhythmic decoding method. Then, from the set of all peak candidates, our rhythmic decoding approach uses the temporal structure of the music signal to filter out those peak candidates that are not rhythmically related. The adaptive thresholding approach simply applies a fixed threshold to the set of all peak candidates.

IV. RESULTS AND DISCUSSION

In this section, the advantages of using rhythmic information are presented in Section IV-A. Then, Sections IV-B and

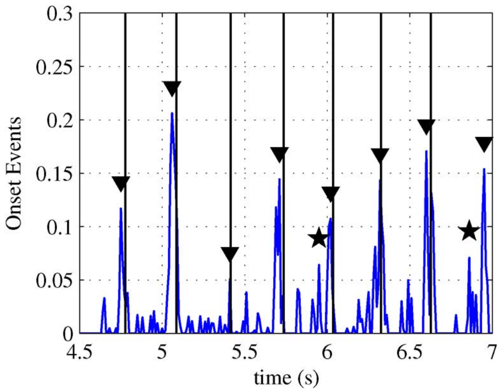

Fig. 7. Decoding results for an excerpt of Wilco's audio file [5]. The solid line corresponds to the normalized Complex Domain detection function, ground truth annotations are marked with vertical lines and onset estimates of the proposed probabilistic method are labeled with down triangles symbols. Peaks marked with star symbols represent potential false positives when using an adaptive thresholding algorithm.

IV-C discuss the effect of using different detection functions and tempo estimation algorithms in the proposed probabilistic framework. Finally, Section IV-D introduces a detailed analysis of the results discussing possible limitations of the system.

A. *Is Rhythmic Information Helpful?*

To illustrate the potential advantage of using rhythmic information over the standard adaptive thresholding approach described in [5] and [6] we refer to Fig. 7. This figure shows the onsets decoded using rhythmic information in an excerpt of the rock song "Can't Stand It" by Wilco with a rhythmic weighting parameter $\alpha = 0.5$. The solid line corresponds to the Complex Domain detection function [6] normalized according to the adaptive-thresholding paradigm, ground truth annotations are marked with vertical lines and the onset event estimates of the proposed probabilistic method are labeled with down triangles symbols. The starred time instants, at 5.9 s and 6.8 s, respectively, are examples of large peaks in the Complex Domain detection function which do not correspond to annotated onsets. Comparing the height of these peaks with the detected onset at 5.4 s we can observe an intrinsic problem with traditional thresholding methods. We cannot easily define a threshold that will allow the low peak at 5.4 s but disregard the peaks at 5.9 s and 6.8 s. This results in the following trade off: either we use a threshold low enough to catch the peak at 5.4 s and incur false positives for the higher non-onset peaks, or we use a higher threshold and incur a false negative for the 5.4 s peak. Therefore, to accurately detect all the onsets in this example we must use an alternative method than standard thresholding. In this sense, the probabilistic integration of rhythmic knowledge allows our system to exploit the temporal expectations associated with rhythmic hypothesis and make musically meaningful event decisions.

We now compare onset detection based on adaptive thresholding and the proposed decoding system by plotting the precision (P) versus recall (R) measures defined in (21) and (22). As described in Section II-A, the local mean is subtracted prior to the peak picking to normalize the detection function. Therefore, to trace out the performance curve of the original detection function using adaptive thresholding, a fixed threshold δ that is varied between 0 and 1 is applied to the peak observations o_t obtained from the normalized detection function. The performance curve of the proposed method is calculated by varying the rhythmic weighting parameter α between 0 and 1 and evaluating the precision and recall of the onset event decoding sequence $\mathcal{T}^*_{\text{onset}}(\alpha)$ given by (19).

Fig. 8(a) presents the total performance curve for the dataset described in Section III using the Complex Domain detection function [6]. The tempo induction system that is used to inform the onset detection task of the beat period estimation is Davies *et al.* method [30]. Adaptive thresholding results are shown with a dashed line and the proposed method with a solid line. Better performance is indicated by a shift of a curve to the top-right corner of the axes which corresponds to a 100% rate of recall and precision. Contours of equal F measure are plotted with dotted lines and their corresponding values are also indicated. As the threshold δ increases, the precision tends to increase and the recall decreases. Similarly, as the value of α is larger, the precision increases and the recall decreases. The reason is that the only peaks that are chosen as onset events are those that are strongly related in time and obviously the number of onset states that the decoded path visits is smaller. For $\alpha = 0.5$, shown with an x-mark, both precision and recall tend to be balanced. As can be seen, the performance curve of the rhythmic decoding approach lies above adaptive thresholding. In fact, the maximum F measure is 83.5% for rhythmic decoding and 76.5% for adaptive thresholding, therefore a gain of 7.0% can be obtained by integrating rhythmic information. In this case, the proposed system successfully exploits the temporal expectations associated with the rhythmic hypothesis. Note that this increase in performance is obtained at the expense of a noncausal decoding and a larger computational load. Therefore, the improvement is valid if the application is not constrained by computational or real-time restrictions.

As expected, the value of the threshold parameter δ has a large impact on the results. Fig. 8(d) shows the total F measure versus the weighting parameter α and the adaptive threshold value δ. In adaptive thresholding, the maximum value of F is obtained for $\delta^* = 0.1$ and the performance decreases significantly as we move away from this optimal threshold value δ^*. Therefore, as already discussed in [6], it seems very difficult to select an appropriate value of threshold automatically since small changes in δ can have a large impact on performance. However, rhythmic decoding is robust to the selection of the weighting parameter α. The F measure stays around 83.0% for values of α ranging from 0.2 to 0.8 and within this range is always above the performance curve of the adaptive thresholding approach.

B. *Onset Detection Function Dependence Analysis*

We turn now to evaluate the performance of our rhythmic decoding system using different onset detection functions to extract the peak candidates. Fig. 8(b) and (e) shows the total performance results using the well-known Spectral Flux detection function [5] and Fig. 8(c) and (f) using the Resonator Time Frequency Image (RTFI) detection function [26]. These results are consistent with those shown in Fig. 8(a) and (d). It is also interesting to see that rhythmic decoding on Complex Domain and Spectral Flux perform similarly in terms of the maximum value

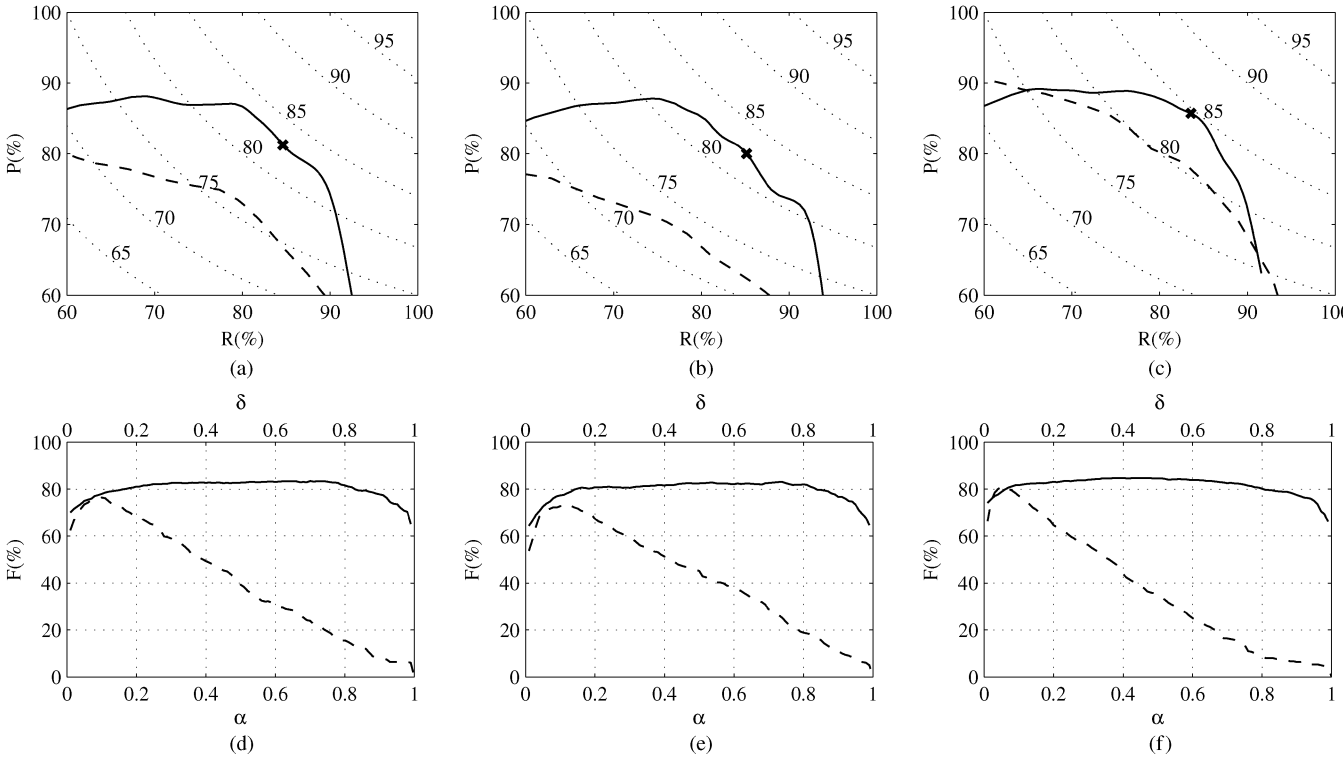

Fig. 8. Rhythmic decoding results using different onset detection functions and Davies et al. [30] tempo estimation algorithm. The influence of the rhythmic weighting parameter α (continuous lines) and the adaptive weighting parameter δ (dashed lines) is shown. Figures (a)–(c) plot precision (P) versus recall (R), where dotted lines are a contour plot of equal F-measure values. Figures (d)–(f) plot the F-measure (F) versus α and δ. The x-mark denotes condition $\alpha = 0.5$. The detection functions are: the Complex Domain [5], (a) and (d); the Spectral Flux [5], (b) and (e); and the Resonator Time Frequency Image [26], (c) and (f).

of F, 83.5% and 83.1% respectively, and their dependence on the value of α but adaptive thresholding on Complex Domain outperforms Spectral Flux.

As can be seen by comparing Fig. 8(a), (b), and (c), adaptive thresholding on RTFI outperforms Complex Domain and Spectral Flux detection functions. Interestingly, rhythmic decoding on RTFI also improves total performance compared to Complex Domain and Spectral Flux. In fact, the maximum F-measure for adaptive thresholding on RTFI is 80.9% and rhythmic decoding on RTFI achieves a maximum F-measure of 84.8%, which is 1.3% larger than Complex Domain. It can be seen in the top-left and bottom-right part of Fig. 8(c) that the rhythmic decoding curve is below adaptive thresholding. Still, the F-measure on the region where adaptive thresholding outperforms rhythmic decoding is below 76.0% and the performance curve of rhythmic decoding is mostly above adaptive thresholding.

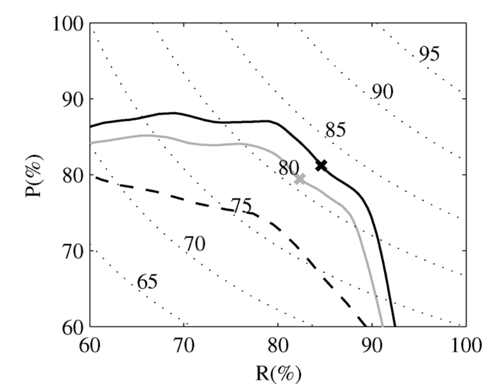

Fig. 9. Comparison of tempo estimation algorithms using the Complex Domain detection function [5]: adaptive thresholding (dashed line), rhythmic decoding with Davies *et al.* algorithm [30] (continuous line) and rhythmic decoding with Ellis's algorithm [36] (gray line). The x-mark denotes condition $\alpha = 0.5$.

C. Tempo Estimation Dependence Analysis

The effect of tempo estimation in the accuracy of the decoded onsets is now analyzed by comparing the proposed probabilistic model when using Davies *et al.* method [30] and Ellis's algorithm [36] to estimate the tempo. Fig. 9 shows the curve of total performance, precision versus recall, of our rhythmic decoding system using the Complex Domain detection function to estimate the peak candidates and Davies *et al.* method [30] and Ellis's algorithm [36] for tempo estimation. Results of the adaptive thresholding peak-picking algorithm are also included as the baseline performance. The figure shows that rhythmic decoding works better than adaptive thresholding using any of the tempo estimation algorithms. The maximum F-measure value is 81.4% using Ellis's tempo estimator, 83.5% using Davies *et al.* method and 76.5% for adaptive thresholding. The proposed probabilistic framework works better with Davies *et al.* tempo estimator since its curve of performance is above rhythmic decoding using Ellis's approach. This is something we could expect since, as shown in [39], Davies *et al.* tempo estimator performs statistically better than Ellis's algorithm. These results suggest that the proposed probabilistic framework obtains good results provided the tempo estimator provides a fair estimate of the beat period.

TABLE II
MAXIMUM F-MEASURE AND SELECTED TEMPLATES FOR EACH FILE OF THE DATASET

	F-measure			Template	
No.	Adapt.	Rhythmic	Oracle	Rhythmic	Oracle
1	96.8	100.0	100.0	{1,1/2}	{1,1/2}
2	94.1	87.5	87.5	{1}	{1}
3	89.1	93.0	93.0	{1,1/2}	{1,1/2}
4	85.3	95.4	98.5	{1,2/3,1/3}	{1}
5	88.9	89.4	89.4	{1,2/3,1/3}	{1,2/3,1/3}
6	90.9	93.3	93.3	{1,1/2}	{1,1/2}
7	87.0	87.5	89.2	{1,2/3,1/3}	{1}
8	93.6	94.3	94.3	{1,2/3,1/3}	{1,2/3,1/3}
9	86.0	90.4	90.4	{1,1/2}	{1,1/2}
10	63.2	66.7	66.7	{1,2}	{1,2}
11	90.9	94.9	94.9	{1,1/2}	{1,1/2}
12	60.0	64.4	79.0	{1,2/3,1/3}	{1,2}
13	80.7	82.8	82.8	{1,1/2}	{1,1/2}

D. Detailed Analysis

For a more detailed analysis, results of the individual test signals are presented in Table II. This table shows the maximum F-measure value obtained using the best overall δ and α parameters for adaptive thresholding (Adapt.) and rhythmic decoding (Rhythmic). To explore the limitations of the proposed probabilistic framework in more detail, an *oracle* approach is also introduced. The oracle makes a rhythmic decoding of the onsets as described in Section II-B4 but, instead of selecting the rhythmic template according to the criterion proposed in (13), it selects the template that achieves the maximum F-measure. Obviously, we cannot use the actual performance to automatically select templates in practice, but this gives us insight on the accuracy of the selection of templates. To this end, Table II also shows the maximum F value achieved by the oracle approach (Oracle) and the rhythmic templates selected by the proposed rhythmic decoding system and the oracle approach.

It is interesting to verify that rhythmic decoding performs consistently better than adaptive thresholding. This agrees with the total performance results shown in Fig. 8 and suggests that rhythmic information is correctly exploited. The maximum F value obtained by rhythmic decoding is larger than the one achieved by adaptive thresholding for all the files of the dataset except for the second excerpt. In this excerpt, four of the annotated onsets are less than 50 ms away from each other. This inter-onset-time difference is of the order of the estimated inaccuracy of the hand labeling process [5]. In this case, the annotated times are expected to be very noisy ground-truth estimates. This is confirmed by the adaptive thresholding algorithm which achieves its maximum F-measure for a very small value of $\delta = 0.01$. This value of δ is far from the optimal threshold value $\delta^* = 0.1$ shown in Fig. 8(a) and it will not help to improve the total performance of adaptive thresholding on the whole dataset. As we could expect, these onsets are not rhythmically related and the rhythmic decoding algorithm is not able to correctly detect them.

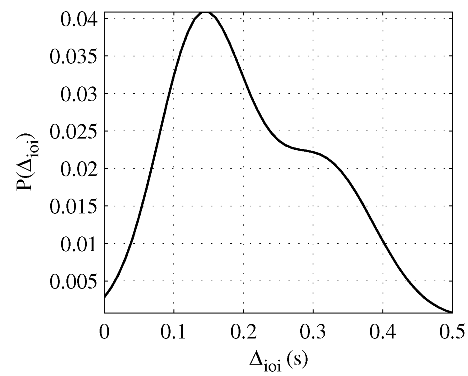

Fig. 10. Probability distribution of the inter-onset time interval of the jazz song "Unquity Road" by Pat Metheny.

As can be also seen in Table II, the rhythmic template selected according to the likelihood of the model defined in (13) generally agrees with the oracle selection. This suggests an appropriate behavior of the selection criterion. However, we find differences on the selection of the templates for excerpts 4, 7, and 12. Test signal number 4 is a pop excerpt where the oracle approach selects the single template $M_{\rm single} = \{1\}$ and achieves a 3% more of F value than the proposed rhythmic decoding system which chooses an odd template $M_{\rm odd} = \{1, 2/3, 1/3\}$. In any case, the periodicity defined by the single template, {1}, is included in the odd template, {1,2/3,1/3}, and the performance achieved by rhythmic decoding is still relevant, 10.0% higher than the reference system defined by the adaptive thresholding approach.

On the contrary, rhythmic decoding does not obtain a large increase in performance on excerpts 7 and 12 when compared to adaptive thresholding. Test signal number 7 is an excerpt of the jazz tune "Unquity Road" by Pat Metheny and Fig. 10 shows an estimation of the probability distribution of the inter-onset time interval of this excerpt. As can be seen, this musical excerpt has a complex rhythmic structure and the potential increase given by the oracle approach with respect to the reference adaptive thresholding is not very large. In this case, the assumption of a constant rhythm structure and the simplicity of the rhythmic information integrated in the probabilistic model makes the system unable to deal with very complex temporal relationships. For this reason, the maximum performance of the rhythmic decoding system does not significantly improve the performance of the traditional thresholding approach. Test signal 12 is a classical excerpt where both adaptive thresholding and rhythmic decoding do a poor job detecting onset. The maximum performance obtained by rhythmic decoding is 4% larger than adaptive thresholding but the oracle approach achieves a maximum F value of 79.0%. Although the potential increase in performance that we can obtain by using rhythmic information is large, the criterion defined in (13) is not able to correctly select the most appropriate rhythmic template in terms of maximum F-measure. A HMM is a generative model and the criterion defined in (13) selects the model that best fits to the observations. Thus, this template selection approach does not imply the maximization of a performance measure such as the F-measure.

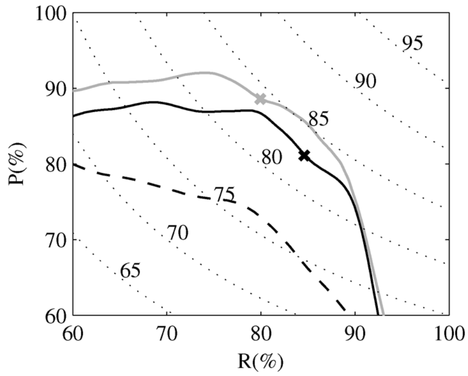

Fig. 11. Results for the Complex Domain detection function [5] showing precision (P) versus recall (R) for adaptive thresholding δ (dashed line), rhythmic decoding (continuous line), and the oracle approach (gray line). The x-mark denotes condition $\alpha = 0.5$.

In summary, the proposed approach is limited by the simplicity of the rhythmic information integrated in the probabilistic model and the accuracy of the selection of the rhythmic templates. As expected, results suggest that the proposed system seems to work best for music with simple and regular onset patterns.

In this section, we have studied the performance on each of the files that comprises the dataset, but it is also interesting to compare the curve of total performance of the proposed system with the oracle approach. Fig. 11 presents a comparison of the total performance in terms of precision (P) and recall (R) of both approaches. The oracle curve represents the performance that could be obtained if we knew the annotations and selected the rhythmic template according to the maximum F-measure. As can be seen, the performance of the rhythmic decoding approach is very close to that of the oracle approach. In fact, the maximum F value achieved by the oracle is 85.4% versus 83.5% of the proposed rhythmic decoding system. This suggests that, in general, the likelihood-based selection criterion presented in (13) does a reasonable work selecting the rhythmic template.

V. Conclusion and Future Work

We have proposed a method for onset event decoding in complex mixtures. The method explicitly integrates rhythmic contextual knowledge and information extracted from the signal using a probabilistic formulation. The proposed algorithm exploits the temporal expectations associated with a rhythmic hypothesis and makes musically meaningful event decisions. A further benefit of this probabilistic approach is that the method defines a specific optimality criterion and estimates the onsets events that best explain the extracted information from the audio signal.

The detection accuracy has been evaluated using a hand-labeled dataset of real music signals and a comparison with the commonly used adaptive thresholding algorithm has been provided. Results showed that rhythmic information can be successfully exploited for onset detection. It has been also shown that, in terms of total performance, the optimality criterion proposed for onset decoding is robust to the selection of the parameter introduced to weight the influence of the rhythmic information.

In addition, the system defines a general framework to be used in combination with any onset detection function and rhythmic structure estimator. This differs from standard onset detection research which focuses on the definition of a feature that works for multiple signals of different nature. It has been shown that rhythmic information can be successfully exploited over a range of onset detection functions and tempo estimation algorithms. Finally, a detailed analysis of the results showed that the method works best for music with simple and regular onset patterns.

As part of our future work we plan to adapt the system for beat tracking by defining an appropriate transition distribution. We will also explore the effect of variable tempo in our model and the possibility of inferring the rhythmic template along the duration of the input signal. The effects of a causal decoding of the onsets will be also studied. We are also interested in studying the dependence on the rhythmic information weighting parameter and how accurate the tempo estimation is. The definition of a specific set of rhythmic templates according to the genre of the music could be also explored. Another interesting extension includes the integration of multiple features as inputs. In this case, a large dataset for training will be required in order to learn the dependencies between the different features.

Acknowledgment

The authors would like to thank Enrique Argones Rúa for his useful and interesting comments and Juan Pablo Bello for kindly providing the dataset.

References

[1] D. Rudoy, P. Basu, and P. Wolfe, "Superposition frames for adaptive time-frequency analysis and fast reconstruction," *IEEE Trans. Signal Process.*, vol. 58, no. 5, pp. 2581–2596, May 2010.

[2] R. Zhou, J. D. Reiss, M. Mattavelli, and G. Zoia, "A computationally efficient method for polyphonic pitch estimation," *EURASIP J. Adv. Signal Process.*, vol. 2009, 2009.

[3] S. Dixon, "Evaluation of audio beat tracking system BeatRoot," *J. New Music Res.*, vol. 36, no. 1, pp. 39–51, 2007.

[4] A. Robertson and M. D. Plumbley, "B-keeper: A beat-tracker for live performance," in *Proc. Int. Conf. New Interfaces for Musical Expression (NIME)*, New York, Jun. 6–9, 2007, pp. 234–237.

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

[6] S. Dixon, "Onset detection revisited," in *Proc. 6th Int. Conf. Digital Audio Effects (DAFx-06)*, Montreal, QC, Canada, Sep. 18–20, 2006, pp. 133–137.

[7] N. Collins, "A comparison of sound onset detection algorithms with emphasis on psychoacoustically motivated detection functions," in *Proc. AES Conv. 118*, Barcelona, Spain, May 2005, no. 6363.

[8] A. Lacoste and D. Eck, "A supervised classification algorithm for note onset detection," *EURASIP J. Appl. Signal Process.*, vol. 2007, no. 1, pp. 153–153, 2007.

[9] B. S. F. Eyben, S. Böck, and A. Graves, "Universal onset detection with bidirectional long short-term memory neural networks," in *Proc. 11th Int. Conf. Music Inf. Retrieval (ISMIR'10)*, Utrecht, The Netherlands, Aug. 2010.

[10] A. Klapuri, "Introduction to music transcription," in *Signal Processing Methods for Music Transcription*, A. Klapuri and M. Davy, Eds. New York: Springer, 2006, pp. 3–20.

[11] F. Gouyon and S. Dixon, "A review of automatic rhythm description systems," *Comput. Music J.*, vol. 29, no. 1, pp. 34–54, 2005.

[12] A. Hazan, R. Marxer, P. Brossier, H. Purwins, P. Herrera, and X. Serra, "What when causal expectation modeling applied to audio signals," *Connection Sci.*, vol. 21, pp. 119–143, 2009, 06/2009.

[13] P. Grosche and M. Müller, "Computing predominant local periodicity information in music recordings," in *Proc. IEEE Workshop Applicat. Signal Process. Audio Acoust.*, New Paltz, NY, 2009.

[14] P. Grosche and M. Mueller, "Extracting predominant local pulse information from music recordings," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 19, no. 6, pp. XXX–XXX, Aug. 2011.
[15] N. Degara, A. Pena, M. E. P. Davies, and M. D. Plumbley, "Note onset detection using rhythmic structure," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, Dallas, TX, Mar. 2010, pp. 5526–5529.
[16] A. Klapuri, A. J. Eronen, and J. T. Astola, "Analysis of the meter of acoustic musical signals," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 14, no. 1, pp. 342–355, Jan. 2006.
[17] G. Peeters, "Beat-tracking using a probabilistic framework and linear discriminant analysis," in *Proc. 12th Int. Conf. Digital Audio Effects (DAFx-09)*, 2009.
[18] G. Peeters and H. Papadopoulos, "Simultaneous beat and downbeat-tracking using a probabilistic framework: Theory and large-scale evaluation," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 19, no. 6, pp. XXX–XXX, Aug. 2011.
[19] C. Raphael, "Automatic segmentation of acoustic musical signals using hidden Markov models," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 21, pp. 360–370, 1999.
[20] M. Mauch and S. Dixon, "Simultaneous estimation of chords and musical context from audio," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 18, no. 6, pp. 1280–1289, Aug. 2010.
[21] M. P. Ryynänen and A. P. Klapuri, "Automatic transcription of melody, bass line, and chords in polyphonic music," *Comput. Music J.*, vol. 32, no. 3, pp. 72–86, 2008.
[22] C. Raphael, "A hybrid graphical model for rhythmic parsing," *Artif. Intell.*, vol. 137, pp. 217–238, May 2002.
[23] S. A. Abdallah and M. D. Plumbley, "Unsupervised onset detection: A probabilistic approach using ICA and a hidden Markov classifier," in *Cambridge Music Process. Colloquium*, Cambridge, U.K., 2003.
[24] H. Thornburg, R. Leistikow, and J. Berger, "Melody extraction and musical onset detection via probabilistic models of framewise STFT peak data," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 15, no. 4, pp. 1257–1272, May 2007.
[25] H. Thornburg, "Detection and modeling of transient audio signals with prior information," Ph.D. dissertation, Center for Comput. Res. in Music and Acoust., Stanford, CA, 2005.
[26] R. Zhou, M. Mattavelli, and G. Zoia, "Music onset detection based on resonator time frequency image," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 16, no. 8, pp. 1685–1695, Nov. 2008.
[27] A. Holzapfel, Y. Stylianou, A. Gedik, and B. Bozkurt, "Three dimensions of pitched instrument onset detection," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 18, no. 6, pp. 1517–1527, Aug. 2010.
[28] E. Benetos and Y. Stylianou, "Auditory spectrum-based pitched instrument onset detection," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 18, no. 8, pp. 1968–1977, Nov. 2010.
[29] L. Rabiner, "A tutorial on hidden Markov models and selected applications in speech recognition," *Proc. IEEE*, vol. 77, no. 2, pp. 257–286, Feb. 1989.
[30] M. E. P. Davies and M. D. Plumbley, "Context-dependent beat tracking of musical audio," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 15, no. 3, pp. 1009–1020, Mar. 2007.
[31] J. Forney, "The Viterbi algorithm," *Proc. IEEE*, vol. 61, no. 3, pp. 268–278, Mar. 1973.
[32] R. O. Duda, P. E. Hart, and D. G. Stork, "Nonparametric techniques," in *Pattern Classification*. New York: Wiley-Interscience, 2000.
[33] L. Daudet, G. Richard, and P. Leveau, "Methodology and tools for the evaluation of automatic onset detection algorithms in music," in *5th Int. Conf. Music Inf. Retrieval (ISMIR)*, Barcelona, Spain, Oct. 10–14, 2004.
[34] G. Peeters, "Template-based estimation of time-varying tempo," *EURASIP J. Appl. Signal Process.*, vol. 2007, no. 1, pp. 158–158, 2007.
[35] M. E. P. Davies, N. Degara, and M. D. Plumbley, "Evaluation methods for musical audio beat tracking algorithms," Centre for Digital Music, Queen Mary Univ., Tech. Rep. C4DM-TR-09-06, 2009.
[36] D. P. W. Ellis, "Beat tracking by dynamic programming," *J. New Music Res.*, vol. 36, pp. 51–60, 2007.
[37] N. Degara, A. Pena, and S. Torres-Guijarro, "A comparison of score-level fusion rules for onset detection in music signals," in *Proc. 10th Int. Conf. Music Inf. Retrieval (ISMIR'09)*, Kobe, Japan, Oct. 2009.
[38] The Music Information Retrieval Evaluation Exchange (MIREX). [Online]. Available: http://www.music-ir.org/mirex/
[39] M. F. McKinney, D. Moelants, M. E. P. Davies, and A. Klapuri, "Evaluation of audio beat tracking and music tempo extraction algorithms," *J. New Music Res.*, vol. 36, no. 1, pp. 1–16, 2007.

Norberto Degara (S'10) received the telecommunications engineer degree from the University of Vigo, Vigo, Spain, in 2001 and the M.S. degree in electrical engineering from the University of Texas (UT), Austin, in 2007. He is currently pursuing the Ph.D. degree at the University of Vigo.

From 2004 to 2005, he was a Research Engineer at LMS International, Belgium, as a Marie Curie Fellow. He was recipient of a Pedro Barrié de la Maza Foundation fellowship for continuation of studies at UT Austin. In 2009, he visited the Centre for Digital Music, Queen Mary University of London, London, U.K. His research focuses on audio and music signal processing, including onset detection, beat tracking, and rhythm analysis.

Matthew E. P. Davies received the B.Eng. degree in computer systems with electronics from King's College London, London U.K., in 2001 and the Ph.D. degree in electronic engineering from Queen Mary University of London, London, in 2007.

From 2007 until 2011, he was a Postdoctoral Researcher in the Centre for Digital Music. He has recently joined the Sound and Music Computing Group at INESC, Porto, Portugal. His research interests include beat tracking and rhythm analysis, evaluation methods, music therapy, and sparse representations.

Antonio Pena (M'93) received the M. S. and Ph.D. degrees in telecommunication engineering from the Universidad Politécnica de Madrid, Madrid, Spain, in 1990 and 1994, respectively.

He has been an Associate Professor with the Universidad de Vigo, Vigo, Spain, since 1995. His research activity was closely related to real-time implementations of MPEG audio coders for broadcast equipment from 1992 to 2001. Nowadays, research on sound signal analysis for applications on acoustics, including sound source separation and subjective evaluation, and the coordination of the trademark "Sonitum", providing both acoustic consulting and teaching, are his main activities.

Mark D. Plumbley (S'88–M'90) received the B.A. (Hons.) degree in electrical sciences and the Ph.D. degree in neural networks from the University of Cambridge, Cambridge, U.K., in 1984 and 1991, respectively.

From 1991 to 2001, he was a Lecturer at King's College London. He moved to Queen Mary University of London in 2002, where he is now an EPSRC Leadership Fellow and Director of the Centre for Digital Music. His research focuses on the automatic analysis of music and other audio sounds, including automatic music transcription, beat tracking, and audio source separation, and with interest in the use of techniques such as independent component analysis (ICA) and sparse representations.

Prof. Plumbley chairs the ICA Steering Committee, and is a member of the IEEE SPS TC on Audio and Acoustic Signal Processing.