Joint estimation of chords and downbeats from an audio signal

Hélène Papadopoulos, Geoffroy Peeters

Joint estimation of chords and downbeats from an audio signal

Hélène Papadopoulos* and Geoffroy Peeters

Abstract—**We present a new technique for joint estimation of the chord progression and the downbeats from an audio file. Musical signals are highly structured in terms of harmony and rhythm. In this paper, we intend to show that integrating knowledge of mutual dependencies between chords and metric structure allows us to enhance the estimation of these musical attributes. For this, we propose a specific topology of hidden Markov models that enables modelling chord dependence on metric structure. This model allows us to consider pieces with complex metric structures such as beat addition, beat deletion or changes in the meter. The model is evaluated on a large set of popular music songs from the Beatles that present various metric structures. We compare a semi-automatic model in which the beat positions are annotated, with a fully automatic model in which a beat tracker is used as a front-end of the system. The results show that the downbeat positions of a music piece can be estimated in terms of its harmonic structure and that conversely the chord progression estimation benefits from considering the interaction between the metric and the harmonic structures.**

Index Terms—**Chords, Downbeat, HMM**

I. Introduction

Within the last few years, the huge explosion of online music collections have become a great source of attention. Specific demands, such as asking an online store to find a song that fits his or her taste and musical expectation among millions of other tracks, became common requirements to music listeners. In this context, techniques for interacting with enormous digital music libraries at the song level are necessary. Content-based music retrieval is therefore a very active topic of research. Within the context of music information retrieval, many applications based on music content analysis have emerged, such as music classification or structural audio segmentation. These applications are mostly based on the use of musical descriptors that are extracted from the audio signal. For instance, two different versions of the same underlying musical piece generally share a similar harmonic structure. The detection of cover versions is thus frequently based on chord progression [1]. Manual annotation of the content of musical pieces is a very difficult and tedious task and requires an immense amount of effort. It is thus essential to develop techniques for automatically extracting musical elements from musical signals. We address this issue in this paper. Our research concentrates on musical descriptors related to the harmonic structure and the metric structure, which are some of the most important attributes of Western tonal music. More specifically, we focus on the problem of estimating two musical attributes: the chord progression, which is related to the harmony, and the downbeats, which are related to the metric structure. A piece of music can be characterized by its chord progression that determines the harmonic structure. The chord progression is closely related to the metric structure of the piece [2]. For example, chords will change more often on strong beats than on other beat positions in the measure. Most of the previous studies have dealt with various musical attributes independently. However, harmony and meter are deeply related to each other and their automatic estimation should be improved by exploiting their interrelationship. In this paper, we present a system that allows the simultaneous estimation of the chord progression and the downbeats from an audio file. Most of the previous work on downbeat detection have dealt with constant meter pieces. A contribution of this paper is that we consider the problem of complex meter (e.g. changes in the meter, addition or deletion of beats). We also consider the problem of imperfect beat tracking. The model is evaluated on a large set of popular music songs and gives very interesting results on pieces with complex metric structure.

II. Related Work

In this section, we review some previous works on chord and downbeat estimation that are related to the present work.

A. Chords

The first stage of a chord detection system consists in extracting some low-dimensional features from the audio signal that are appropriated for the task. Since their introduction in 1999, Pitch Class Profiles (PCP) [3] or Chroma-based representation [4] became common features for estimating chords or musical keys from audio recordings. PCP/chroma vectors are low-dimensional features that represent the intensity of the twelve semitones of the pitch classes. [3] proposes and uses this representation to derive a large set of chords using either a nearest neighbor or a weighted sum method. The system is successfully evaluated but only using synthetic sounds. Most of the works on chord detection are based on this representation. Recently, a new feature called the Tonal Centroid has been proposed [5]. [6] uses this feature in the

H. Papadopoulos is with the Sound Analysis/Synthesis Team, IRCAM / CNRS-STMS, Paris, FRANCE (corresponding author to provide; phone: +33 1 44 78 14 85; e-mail: Helene.Papadopoulos@ircam.fr).

G. Peeters is with the Sound Analysis/Synthesis Team, IRCAM / CNRS-STMS, Paris, FRANCE (e-mail: Geoffroy.Peeters@ircam.fr).

This work was supported by QUAERO project.

context of chord estimation and shows that his chord detection system performs better than when using chroma features.

Chord detection systems can be generally classified into two categories: template matching algorithms and machine learning algorithms. In template matching approaches, each feature vector computed from the audio signal is correlated with a set of chord templates that indicate the perceptual importance of the notes within a chord. The musical chord is obtained by selecting the template that gives the maximum correlation coefficient. [7] estimates chords using simple bit masks[1] that are compared to chroma features. Relying on the work of [8], [9] proposes chord templates that take into account the harmonics of the notes. Alternative to the template matching based approaches are the machine learning approaches. Hidden Markov models (HMMs) have been widely used in the context of chord estimation. [10] presents the first system for chord segmentation/recognition evaluated on natural sounds (whole pieces of music of commercial recordings). The system relies on HMM trained by the Expectation Maximization (EM) algorithm. [11] uses an approach similar to [10] but introduces musical knowledge in the HMM to improve the results. This work is extended in [9] and compared with other methods. [12] proposes a system that tracks chords and keys replacing the more traditional Gaussian emission distributions with Dirichlet distributions in the HMM. [13] proposes an acoustic chord transcription system that uses symbolic data to train HMMs, avoiding the tedious task of human annotation of chord names and boundaries.

B. Downbeats

A metric structure is a hierarchical structure. The most salient metrical level, called the *tactus* or *beat* level is a moderate level that corresponds to the foot-tapping rate. Here, we will also consider another common metrical level called *tatum*. The tatum level corresponds to the "shortest durational values in music that are still more than accidentally encountered " [14]. Musical signals are divided into units of equal time value called *measures* or *bars*. The relationship between measures and tactus/tatum is defined by the meter, which is usually indicated by a *time signature*. One important problem related to meter analysis is to find the position of the *downbeat* or the first beat of each measure.

Downbeat detection is an interesting problem that deserves to be carefully studied. Even if it has drawn less attention than beat tracking, there have been however a number of contributions dealing with various aspects of this problem. Most of the proposed approaches rely on prior knowledge such as tempo, time-signature of the piece or hand-annotated beat positions.

[15] presents a model that uses an autocorrelation technique to determine the downbeats in musical audio signals for which beat positions are known. The system relies upon the assumption that a piece of music will contain repeated patterns. It has been tested on 42 different pieces of music at various metrical levels, in several genres. It achieves a success rate of 81% for pieces in 4/4 time-signature and needs more testing on ternary time-signatures. [16] proposes an unbiased and predictive approach. The model is tempo independent and does not require beat tracking but requires some fair amount of prior knowledge acquired through listening or learning during a supervised training stage where downbeats are hand-labeled. The model has only been applied to music in 4/4 meter. [2] proposes two approaches to downbeat estimation. For percussive music, the downbeats are estimated using rhythmic pattern information. For non-percussive music, the downbeats are estimated using chord change information. [14] proposes a full analysis of musical meter into three different metrical levels: tatum, tactus and measure level. The downbeats are identified by matching rhythmic pattern templates to a mid-level representation. [17] uses a similar "template-based" approach in a drum-pattern classification task. In [18] an approach based on a spectral difference between band-limited beat-synchronous analysis frames is proposed. The sequence of beat positions of the input signal is required and the time-signature is to be known a priori. A recent method that segments the audio according to the position of the bar lines has been presented in [19]. The position of each bar line is predicted by using prior information about the position of previous bar lines as well as the estimated bar length. The model does not depend on the presence of percussive instruments and allows moderate tempo deviations.

C. Exploiting Relationships Between Musical Attributes

Although there is a strong relationship between the chord progression of a piece and other musical attributes such as the musical key or the metric structure, these attributes have typically been estimated separately in the past. However, some approaches that consider the interactions between chords and other musical attributes have already been proposed. [20] proposes a technique to estimate the predominant key in a musical excerpt from the chord progression. The key space is modeled on a 24-state hidden Markov model, where each state represents one of the 24 major and minor keys, and each observation represents a chord transition. It is argued that the tonal center can be better defined using chords instead of note distributions without regard for their position within musical phrases. [13] estimates simultaneously the chord progression and the musical key of an audio file. For each key, a key-dependent HMM is built, using the key information derived from symbolic data. The musical key is obtained by choosing a key model with the maximum likelihood, and the chord sequence is obtained from the optimal state path of the corresponding key model. [21] proposes a probabilistic framework devised to uniformly integrate bass lines extracted by using bass pitch estimation into hypothesis-search-based chord recognition. Bass lines are utilized as clues for improving chord recognition. [22] considers the interaction between harmonic and metric structure. It is related to our work in the sense that contextual information related to the meter is used for modeling the chord progressions. However, the approach is different. It is not based on a HMM but the strong relationship between chord progression and the meter of the piece is embedded in a tree structure that captures the chord

[1] A bit mask is a 12-dimensional vector corresponding to the 12 semitones of the pitch classes with 1 when the note belongs to the chord, 0 otherwise.

structure in a given musical style. The main assumption behind the model is that conditional dependencies between chords in a typical chord progression are strongly tied to the metric structure associated to it. In this model, a chord progression is seen as a two-dimensional architecture. Each chord in the chord progression depends both on its position in the chord structure (global dependencies) and on the surrounding chords (local dependencies).

III. Proposed Approach

Hidden Markov Models (HMM) have often been used to model the chord progression of an audio file (see for example [10], [11], [6]). One of the reasons why the chord progression is modeled by an HMM is that the observation of a given chord depends on the previous chord according to musical composition rules which can be modeled in a transition matrix. In this paper, we propose a specific topology of HMM that allows us to extract simultaneously the chord progression and the downbeats from an audio file. For this, we first extract a set of feature vectors that describe the signal. Here, we use the chroma features described above. The chroma vectors are averaged according to the tactus/tatum positions that have been extracted using the method proposed in [23] and checked by hand. The chord progression is represented using a hidden Markov model that takes into account global dependence on meter. We present a “double-states” HMM where a state is a combination of a chord type and a position of the chord in the measure. Harmonic and metric structure information are encoded in the transition matrix. In order to take into account several cases of metric structure, two transition matrices are proposed. Using a Viterbi decoding algorithm, the most appropriated matrix is selected. We then obtain simultaneously the most likely chord sequence and downbeat positions path over time. The flowchart of the system is represented in Fig. 1.

The rest of the paper is organized as follows. First, in section IV-A, we present the front-end of our system, the extraction of a set of meter-related feature vectors that represent the audio signal. We then introduce in section IV-B a probabilistic model for simultaneous chord progression and downbeat position estimation. This model encodes contextual information in the state transition matrix (IV-E). In section IV-F, we present our approach to estimate the two considered musical attributes (chords and downbeats) using the Viterbi decoding algorithm. In sections V and VI, the proposed model is evaluated on a set of hand-annotated songs from the Beatles. A conclusion section closes the article.

IV. Model

A. Features Extraction

The front-end of our system is based on the extraction of a set of feature vectors (*chroma vectors*) that represent the audio signal.

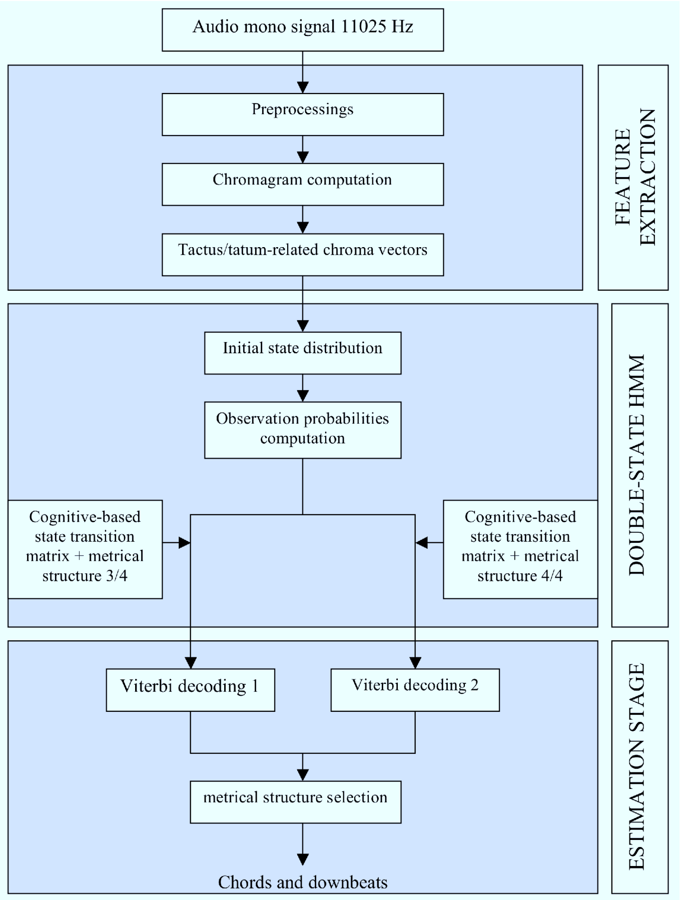
Audio mono signal 11025 Hz
Preprocessings
Chromagram computation
Tactus/tatum-related chroma vectors
FEATURE EXTRACTION
Initial state distribution
Observation probabilities computation
Cognitive-based state transition matrix + metrical structure 3/4
Cognitive-based state transition matrix + metrical structure 4/4
DOUBLE-STATE HMM
Viterbi decoding 1
Viterbi decoding 2
metrical structure selection
Chords and downbeats
ESTIMATION STAGE

Fig. 1. General flowchart of the proposed model for simultaneous chord progression and downbeat estimation.

1) Parameters: The audio signal is first down-sampled to 11025 Hz and converted to mono by mixing both channels. We consider frequencies between 60 Hz and 1000 Hz, which correspond to midi notes from B1 to B5. The upper limit is set to 1 kHz because the fundamentals and harmonics of the music notes in popular music are usually stronger than the non-harmonic components up to 1 kHz [24]. This choice is also supported by the fact that many of the higher harmonics, which are whole number multiples of the fundamental frequency, are far from any note of the Western chromatic scale. This is especially true for the 7^{th} and the 11^{th} harmonics.

2) Tuning: Because the energy peaks in the spectrogram are mapped to the chroma vectors, it is important that the peak frequencies of the spectrum correspond as close as possible to usual pitch values (262.6, 277.2, 293.7, ... Hz). Since the instruments may have been tuned according to a reference pitch different from the standard $A4 = 440$ Hz, it is necessary to estimate the tuning of the track. Here, the tuning is estimated using the method proposed in [25]. A set of candidate tunings between 427 Hz and 452 Hz (corresponding to the quarter-tones below and above $A4$) is tested. The amount of energy of the spectrum explained by the frequencies corresponding to the semitones based on each candidate tuning is measured. The candidate tuning that allows us to best explain the energy of the spectrum is selected as the tuning of the track. The estimated tuning A_{ref} is taken into account when computing the chromagram, as explained below. Concerning the database used in the evaluation (see part V), the estimated tunings of the tracks are comprised between 430 Hz and 444 Hz. Most of the songs are not based on a tuning of $A4 = 440$ Hz.

3) Chromagram Computation: The temporal sequence of chroma vectors over time is known as chromagram. Existing methods to compute a chromagram present some variances but follow, in general, two steps: first a semitone pitch spectrum is either computed from the Fourier transform or directly obtained by the constant Q transform (CQT) because the center frequencies of the CQT are spaced according to the frequencies of the equal-tempered scale; then the semitone pitch spectrum is mapped to the chroma vectors. As proposed in [26], smoothing the semitone pitch spectrum provides a reduction of transients and noise in the signal.

Here, we compute the chromagram using the constant Q transform, which was first introduced in 1991 by Brown [27]. It is a spectral analysis where frequency domain channels are not linearly spaced, as in Fourier transform-based analysis, but geometrically spaced (the frequency-resolution ratio remains constant), thus very similar to the frequency resolution of the human auditory system. The constant Q transform is closely related to the Fourier transform but gives a better representation of spectral data from a music signal. The constant Q transform of a temporal signal sample x(n) can be calculated as:

$$X^{cq}(k) = \sum_{n=0}^{N(k)-1} w(n,k)x(n)e^{-j2\pi f_k n} \tag{1}$$

where $X^{cq}(k)$ is the k^{th} component of the constant Q transform. For each value of k, the window function $w(n,k)$ is a function of the frequency. Let Q denote the constant ratio of frequency to resolution, $Q = \frac{f}{\delta f}$, and let S denote the sampling rate. The length of the window $w(n,k)$ in samples at frequency f_k is $N(k) = \frac{Q.S}{f_k}$. $N(k)$ is function of the frequency and thus of the bin position k.

Let β denote the number of bins per octave. Chroma features are usually represented in a 12-bin histogram, each bin corresponding to one of the 12 semitones of the equal-tempered scale. This corresponds to semitone spacing and in this case, $\beta = 12$. Very often, a higher resolution is used to get a finer pitch class representation. In our experiments, we found that using a 36-bin per octave resolution allows us to increase the accuracy of the results. In this case, each note in the octave is mapped to 3 bins in the chroma and $\beta = 36$. Let $f_{min,440}$ be the minimum frequency considered in the signal feature computation in the ideal case of a perfect tuning. The actual minimum frequency value f_{min} is chosen according to the estimated tuning of the track: $f_{min} = f_{min,440} * \frac{A_{ref}}{440}$. The center frequencies are geometrically spaced, according to the frequencies of the equal-tempered scale:

$$f_k = (2^{1/\beta})^k f_{min} \tag{2}$$

The constant Q transform increases time resolution towards higher frequencies. The length of the window $w(n,k)$ decreases with frequency. The hopsize is chosen to be equal to the smallest window length. In the case of $\beta = 12$, the center frequencies directly correspond to musical notes of the semitone pitch scale and the computation of the constant Q transform leads to a semitone pitch spectrum. When $\beta = 36$, it corresponds to a $\frac{1}{6}$-tone pitch spectrum. It is computed at each time instant t. The output signal of each filter $X^{cq}(k,t)$ is then smoothed over time using a 10-points median filtering[2]. This provides a reduction of transients and noise in the signal.

Finally, each bin b of a chroma vector computed at time instant n can be calculated as:

$$C_n(b) = \sum_{m=0}^{M} |(X^{cq}(b + m\beta))| \tag{3}$$

where $b \in [1;\beta]$ denotes the chroma bin index and M is the total number of octaves in the constant Q spectrum, defined by the upper and lower frequencies used for the analysis and set by the user. In our case $M = 4$.

We obtain a sequence of 12-dimensional vectors that are suitable feature vectors for our analysis.

4) Tactus/tatum-Synchronous Analysis: Since we want to study the relationship between chords and metric structure, we need to deal with observation features that are related to the meter. The frame by frame analysis does not fit our needs: we need to proceed to a beat synchronous analysis. To this end, the chromagram is averaged so that we obtain one feature per tactus/tatum[3]. In the present work, we use the beat tracker proposed in [23] as a front end of the system. Briefly, [23] proposes a method that aims at detecting tempo at the tactus level for percussive and non-percussive audio. The front-end of the system is based on a proposed reassigned spectral energy flux for the detection of musical events. The dominant periodicities of this flux are estimated by a combination of discrete Fourier transform and frequency-mapped autocorrelation function. The most likely meter, beat, and tatum over time are then estimated jointly using meter/beat subdivision templates and a Viterbi decoding algorithm. The beat tracking is then performed using a method adapted from a P-SOLA glottal closure instant detection using estimated tempo and local maxima of the onset-energy function. We refer the reader to [23] for more details.

For each tactus/tatum position p_k of the piece, we compute a chroma vector C_k. Each bin $C_k(b), b = [1;12]$ is obtained by computing the average of the K_k chroma vector bins $C_n(b)$

[2] Smoothing of the semitone pitch spectrogram strengthens spectral envelope continuity, a physical property, while smoothing on the chromagram does not rely on any physical property. That is why the filtering is performed on the notes rather than on the chroma vectors.

[3] The tactus/tatum positions are either considered as input to the system in the case of semi-automatic analysis or obtained using a beat tracker as a front-end of the system in the case of fully-automatic analysis.

over the considered tactus position and the following one:

$$C_k(b) = \frac{1}{K_k} \sum_{p_k \leq n < p_{k+1}} C_n(b) \tag{4}$$

The feature extraction stage is represented in Fig. 2.

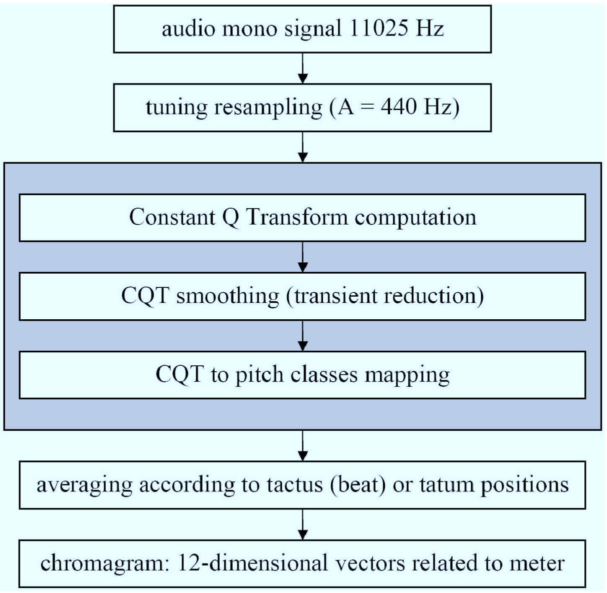
audio mono signal 11025 Hz
tuning resampling (A = 440 Hz)
Constant Q Transform computation
CQT smoothing (transient reduction)
CQT to pitch classes mapping
averaging according to tactus (beat) or tatum positions
chromagram: 12-dimensional vectors related to meter

Fig. 2. Chroma features extraction.

In our study, we have considered two cases. The chromagram has been averaged with respect to the beats or quarter notes (*tactus*) in the first case, and with respect to the eighth notes (*tatum*) in the second case.

B. Overview of the Model

We consider an ergodic $I * K$-states HMM where each state s_{ik} is defined as an occurrence of a chord c_i, $i \in [1; I]$ occurring at a "position in the measure" (position of a beat or tatum inside a measure) pim_k, $k \in [1; K]$:

$$s_{ik} = [c_i, pim_k].$$

In our experiments, our chord lexicon is composed of $I = 24$ Major and minor triads (C Major, ..., B Major, C minor, ..., B minor). The notation for chord types will be the following: CM, ..., BM, for major chords, Cm, ..., Bm for minor chords. In the proposed model, chord changes can only occur on beats or half beats, which corresponds respectively to the tactus and the tatum positions in the test set. In the rest of the text, the positions in the measure where chord changes occur will be referred to as "position in the measure" and denoted by *pim*. We consider here pieces predominantly in $3/4$ or predominantly in $4/4$ meters. In both cases, the transition matrix will allow 4 beat positions in the measure. $K = 4$ if we consider the tactus-level and $K = 8$ if we consider the tatum-level. Each state in the model generates with some probability an observation vector $\mathbf{O}(t_m)$ at time t_m. This is defined by the observation probabilities. Given the observations, we estimate the most likely chord sequence over time and the downbeat positions in a maximum likelihood sense.

We now describe in detail the characteristics of our HMM: initial state distribution, observation probability distribution and state transition probability distribution.

C. Initial State Distribution π

The prior probability π_{ik} for each state is the prior probability that a specific chord i occurring on pim_k has been emitted. Since we do not know *a priori* which chord the piece begins with and which *pim* the piece starts with, we initialize π_{ik} at $\frac{1}{I*K}$ for each of the $I * K$ states.

D. Observation Probabilities

The observation probabilities are computed in the following way. Let $P(\mathbf{O}(t_m)|s_{ik}(t_m))$ denote the probability that observation $\mathbf{O}$ has been emitted at time instant t_m given that the model is in state s_{ik}. Let $P(\mathbf{O}(t_m)|c_i(t_m))$ denote the one that has been emitted by chord c_i and $P(\mathbf{O}(t_m)|pim_k(t_m))$ the one that has been emitted given that the chord is occurring on pim_k. We now assume independence between *chord type* (CM, C#M, ..., Cm, ..., Bm) and the position of the chord in the measure. For instance, we consider that in any given song, even if we favor chord changes on *pim* = 1, we do not favor any *chord type*: a D major chord is as likely to occur at the beginning of a measure as a C major chord[4]. The observation probabilities are computed as:

$$P(\mathbf{O}(t_m)|s_{ik}(t_m)) = P(\mathbf{O}(t_m)|c_i(t_m))P(\mathbf{O}(t_m)|pim_k(t_m)) \tag{5}$$

1) Observation pim Probability Distribution: Equation 5 gives the observation probability for state s_{ik} depending on chord c_i and position in the measure pim_k. Here, the *pim* probability distribution $P(\mathbf{O}(t_m)|pim_k(t_m))$ is considered as uniform ($\frac{1}{K}$ for each *pim* in the measure). It is thus a constant multiplication that has no effect on the observation probability for state s_{ik}, which actually depends only on the chord type. We acknowledge that by doing so, we disregard signal information that could inform the downbeat tracking process. The system would benefit from downbeat information extracted from the signal, for instance by combining a rhythmic pattern approach with the proposed one. Future work will concentrate on the definition of a more elaborated *pim* probability distribution.

2) Observation Chord Symbol Probability Distribution: The probabilities $P(\mathbf{O}(t_m)|c_i(t_m))$ are obtained by computing the cosine distance between the observation vectors (the chroma vectors) and a set of chord templates which are the theoretical chroma vectors corresponding to the $I = 24$ major and minor triads.

[4] This is not strictly correct: for example some chords are more likely to occur than others on strong beats in the piece according to the musical key. We will not take into account these considerations here, they are left for future work.

TABLE I
NOTES AND HARMONICS COMPOSING A C MAJOR CHORD CONSIDERING 6 HARMONICS IN THE MODEL.

C major						
Note	Harmonics					
C	C	C	G	C	E	G
E	E	E	B	E	G#	B
G	G	G	D	G	B	D
Amplitude	1	0.6	0.36	0.216	0.1296	0.0778

Chord Templates:

We consider 24 chord templates corresponding to the 24 major and minor triads. The amplitude of a note in the template is non-zero if the note belongs to the considered chord (fundamental or harmonic). As proposed in [8], in the context of key estimation, the amplitude contribution of the h^{th} harmonic composing the spectrum of a note is set to be 0.6^{h-1}. In what follows, the chord template corresponding to chord c_i will be denoted by $\mathbf{CT}_i$. In Fig. 3, the chord templates for a CM and a Cm chord considering 6 harmonics in the model have been represented. The first six harmonics of the notes composing a C major chord and their corresponding amplitude are given in Table I. It can be seen that higher harmonics contribute to the pitch class of their fundamental frequencies. For instance, the amplitude of the G is very high in the C major chord (C-E-G) because, besides being a note of the chord, G is a strong harmonic of C. The chord templates for other chords (C#M, ..., BM, C#m, ..., Bm) are obtained from the CM and Cm chords by circular permutation.

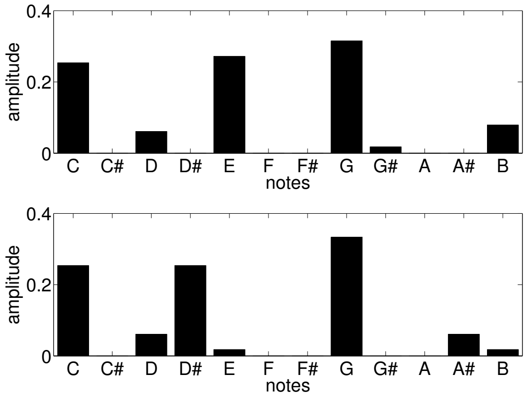

Fig. 3. Chord templates CT_1 (top) and CT_{13} (bottom) for C Major and C minor chords considering 6 harmonics in the model.

Chord Symbol Probabilities Computation:

At each time instant t_m, we compute the cosine distances between the observation vector $\mathbf{O}(t_m)$ and each of the 24 chord templates $\mathbf{CT}_i, i \in [1, 24]$.

$$For \quad i = 1 \dots 24, \quad P(\mathbf{O}(t_m)|c_i(t_m)) = \frac{\mathbf{O}(t_m).\mathbf{CT}_i}{\|\mathbf{O}(t_m)\|.\|\mathbf{CT}_i\|} \tag{6}$$

The 24 values $P(\mathbf{O}(t_m)|c_i(t_m))$ are normalized across components per template such that their components sum to unity.

E. State Transition Probability Distribution

In music pieces, the transitions between chords result from musical rules. Using a Markov model, we can model these rules in the state transition matrix. Let T denote the $I * K$-states transition matrix of our model. T takes into account both the chord transitions and their respective positions in the measure. The matrix T is derived from an I-state chord transition matrix denoted by T_c based on music-theoretical knowledge about key-relationships we used in [9]. In [28], Krumhansl studies the proximity between the various musical keys using correlations between key profiles obtained from perceptual tests. These key profile correlations have been used in [29] to derive a key transition matrix in the context of local key estimation as described below. Krumhansl gives numerical values corresponding to key profile correlations for C major and C minor keys. The values can be circularly shifted to give the transition probabilities for keys other than C major and C minor. In order to have probabilities, all the values are made positive by adding 1, and then normalized to sum to 1 for each key. The final key transition matrix size is 24 x 24. As we proposed in [9], the key transition matrix from [29] can be used as a chord transition matrix. This matrix is represented in Fig. 4.

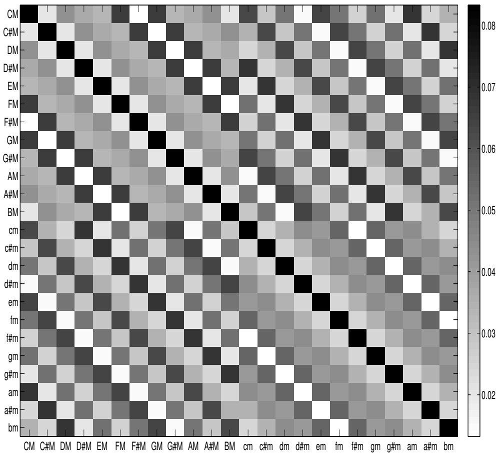

Fig. 4. State transition matrix between the 12 major and the 12 minor chords. Dark marks indicate high values in the transition matrix. Horizontal axis from left to right and vertical axis from top to bottom: chords (CM, C#M, BM, ..., cm, ..., Bm).

That main idea of the present model is that we favor chord changes on the beginning of the measures. In a piece of music, chord changes are in general related to the beats. As stated by [2]:

1) Chords are more likely to change on beat times than on other positions.

2) Chords are more likely to change on half-note times than on other positions of beat times.
3) Chords are more likely to change at the beginnings of measures than at other positions of half-note times.

Analysis of the evaluation dataset has shown that our data support these assumptions. Figure 5 shows the distribution of chord changes according to the position in the measure. It can be seen that the three statements reported above are corroborated by the chord annotation. In particular, about 90% of the chord transitions occur on a beat position (for most of them on the strong beats) and 76% of the chord transitions occur on a downbeat.

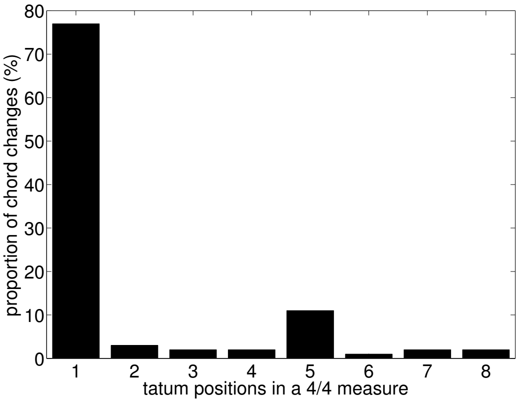

Fig. 5. Distribution of the chord changes according to the *pim* computed on the pieces in 4/4 of the Beatles evaluation dataset.

Because chords are more likely to change at the beginning of a measure than at other *pim*, we give lower self-transition probabilities for chords occurring on the last beat of a measure than on other *pim*. A self-transition is defined as a transition between two same chord symbols. For instance CM-CM is a self transition whereas CM-DM is not. The term self-transition here only refers to the spelling of the chord and is independent from its position in the measure. We extend the model we presented in [30] for constant $4/4$ meter to the case of variable meter. In the previous model, the meter was constrained to be constant and it was not allowed to jump over a pim (i.e. skip over or add one or several beats). Furthermore, the problem of imperfect beat tracking was not considered.

We will present in the next two parts of this paper the results of some experiments that we have carried out on a very large set of Beatles songs presenting various metric structures. A detailed analysis of the test set will be given in part V-A, but it is important to already note that many songs do not have a constant meter. We have picked out several cases of metric structure :

- constant 4/4 or 3/4 meter
- variable meter 4/4 with passages in 3/4 meter
- variable meter 3/4 with passages in 4/4 meter
- 1 or 2 added beats within a constant meter passage
- 1 or 2 deleted beats within a constant meter passage

However, because most of the songs have a predominant meter (3/4 or 4/4), we have chosen to simplify the problem considering two major cases. Two transition matrices, with same form but different values, are proposed. The first one corresponds to the case of songs in 4/4 meter with ternary passages and will be denoted as T_4. In this case, we favor measures of 4 beats but transitions to measures of 3 beats are allowed. The second transition matrix corresponds to the case of songs in 3/4 meter with passages in 4/4 and will be denoted as T_3. In this case, we favor measures of 3 beats but transitions to measures of 4 beats are allowed. We do not allow the algorithm to skip over or add one or several beats because this would reduce its robustness. Indeed addition or deletion of beats corresponds to exceptional situations that happen no more than a few times within a song.

T_3 and T_4 can be seen as block matrices where each block corresponds to a specific chord transition. They are derived from the I-state chord transition matrix T_c presented above. The transition probability between chord i and chord i' will be denoted as $T_c(i,i')$. This matrix is represented in the left part of Fig. 6.

T_3 and T_4 are related to both the metric and harmonic structures of a piece of music. The construction of T_3 and T_4 follows three steps. The first two concern the problem of the downbeats. The third step takes into account the chord type dimension.

Firstly two *pim* transition matrices T_{3pim} and T_{4pim} are defined, which represent the probability to transit from pim_k to $pim_{k'}$ in a song. According to our assumptions, only values $T_{3pim}(k,k')$/$T_{4pim}(k,k')$ such that $k' = (k+1)\ (mod\ 4)$[5] are non-zero, as well as $T_{3pim}(3,4)$/$T_{4pim}(3,4)$ so that transitions between measures in 4/4 and measures in 3/4 are allowed:

$$\left\{\begin{array}{rr} T_{3pim}(1,2) = & 1 \\ T_{3pim}(2,3) = & 1 \\ T_{3pim}(3,4) = & \alpha_3 \\ T_{3pim}(4,1) = & 1 \\ T_{3pim}(3,1) = & \beta_3 \end{array}\right. \left\{\begin{array}{rr} T_{4pim}(1,2) = & 1 \\ T_{4pim}(2,3) = & 1 \\ T_{4pim}(3,4) = & \alpha_4 \\ T_{4pim}(4,1) = & 1 \\ T_{4pim}(3,1) = & \beta_4 \end{array}\right. \quad (7)$$

with $\alpha_3 < \alpha_4$ and $\beta_3 > \beta_4$ so that measures in 3/4 are favored in the case of T_{3pim} and measures in 4/4 are favored in the case of T_{4pim}. In our experiments, we used $\alpha_3 = 0.6$, $\alpha_4 = 0.9$, $\beta_3 = 1.05$ and $\beta_4 = 0.85$. These values were manually selected in small scale simulations, starting from the value 1 and varying in a range of ± 0.5.

Secondly, we want to favor chord changes on downbeats, *i.e.* disfavor transition between identical chords at measure boundaries (between the last *pim* of a measure and the first *pim* of the next measure). For this self-transition case ($i' = i$), corresponding to the diagonal blocks of T_3 and T_4, we define a specific transition matrix, denoted by T'_{pim}. T'_{pim} is the same in the case of T_3 building and in the case of T_4 building. To favor chord changes on downbeats, we attribute a self-transition probability from beat 3 to beat 1 (3/4 time-signature) and from beat 4 to beat 1 (4/4 time-signature) lower than on other *pim* transitions:

[5] where $a = b\ (mod\ m)$ means that a and b have the same remainder for the Euclidian division by m.

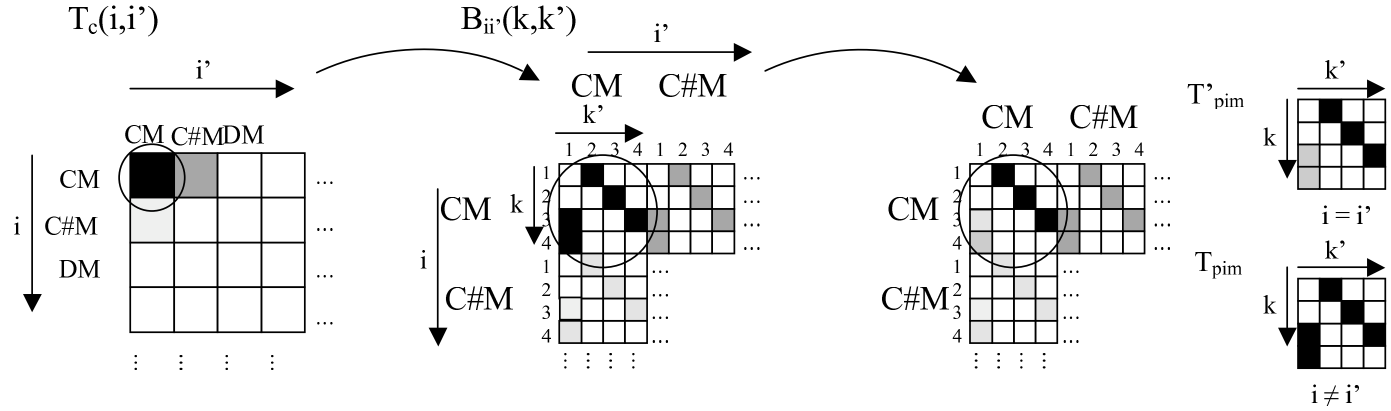

Fig. 6. In this figure, the darker the color, the higher the value in the transition matrix. The figures indicate: the chord transition matrix for a single-state HMM [left], the transition matrices for major to major chords in the case of double-states HMM, without taking into account the *pim* of the chord in the measure [middle left] and taking into account the *pim* of the chord [middle right] (lower value on transition pim_3-pim_1, pim_4-pim_1) and the *pim* transition matrices [right].

$$\begin{cases} T'_{pim}(1,2) = & \alpha \\ T'_{pim}(2,3) = & \beta \\ T'_{pim}(3,4) = & \gamma \qquad \text{s.t.} \quad \delta < \alpha, \beta, \gamma \\ T'_{pim}(4,1) = & \delta \\ T'_{pim}(3,1) = & \delta \end{cases} \tag{8}$$

The values α, β, γ and δ were again selected manually in small-scale simulations starting from the value 1 and varying in a range of ± 0.5 (testing values between 0.5 and 1.5).

It should be noted that, even if the model parameters were selected in part by hand and have an impact on the results, the exact values of these parameters is not critical. The same transition matrices with the same parameters have been used with success on a set of classical music pieces [31], which suggest that the values are not critical to the dataset. It should be possible to learn the parameters from the annotated files. However, until now, attempts to derive transition probabilities from training have given less accurate results than those obtained using values selected in part by hand.

Finally, we construct the global transition matrix T_3 from T_c, T_{3pim} and T'_{pim}, and normalize it so that the sum of each row is equal to 1 (Fig. 6 [middle]). Each block $B_{ii'}(k,k')$ of this matrix represents the transition from chord i at pim_k to chord i' at $pim_{k'}$:

$$B_{ii'}(k,k') = \begin{cases} T_c(i,i') \cdot T_{3pim}(k,k') & \text{if} \quad i \neq i' \\ T_c(i,i') \cdot T_{3pim}(k,k') \cdot T'_{pim}(k,k') & \text{if} \quad i = i' \end{cases} \tag{9}$$

The transition matrix T_4 is constructed in a similar way, using T_c, T_{4pim} and T'_{pim}.

F. Simultaneous Estimation of Chords and Downbeats

In order to find the optimal succession of states s_{ik} over time, the Viterbi decoding algorithm [32] is used successively with the two chord transition matrices T_3 and T_4. The algorithm provides the most likely path $\mathbf{Q}$ through the HMM states given the sequence of observations. The transitions matrix T which gives the greatest likelihood given the observation sequence $\mathbf{O}$ according to Equation 10 is selected. We obtain simultaneously the best sequence of chords over time and the downbeat positions.

$$T = \text{argmax}(P(\mathbf{O}, \mathbf{Q}|T_3), P(\mathbf{O}, \mathbf{Q}|T_4)) \tag{10}$$

V. Evaluation Method

A. Test Set

The proposed model has been tested on a set of hand-labeled Beatles songs. All the recordings are polyphonic, multi-instrumental songs containing drums and vocal parts. The chord annotations were kindly provided by C. Harte from Queen Mary University of London [33]. Since our chord lexicon only represents major and minor triads, we have performed a mapping from complex chords in the annotation (such as major and minor 6^{th}, 7^{th}, 9^{th}) to their root triads. The augmented chords, which include a major third, were mapped to major chords whereas the diminished chords, which include a minor third, were mapped to minor chords. The tactus positions and the ground-truth downbeats were manually annotated by the authors and checked by trained musicians. Meter information for each song was provided by the American musicologist Alan W. Pollack[6]. The original set comprises 180 songs of the Beatles, we reduced it to 162 songs removing songs having an overcomplicated metric structure and containing parts where downbeats were perceptually ambiguous and were extremely difficult to predict and annotate, even for a trained musician. For instance, the song *Good Morning, Good Morning* was not analyzed because, according to A.W. Pollack, the meter is "4/4 in intro, bridge and outro; anything but predictable in verse". For this reason, those files were not annotated. The songs of the testset can be classified according to their metric structure in the following way:

- 8 songs are in 3/4 meter
- 10 songs have a variable meter (presenting at least one change in time signature, more than two for most of them)

[6] http://www.icce.rug.nl/∼soundscapes/DATABASES/AWP/awp-notes_on.html

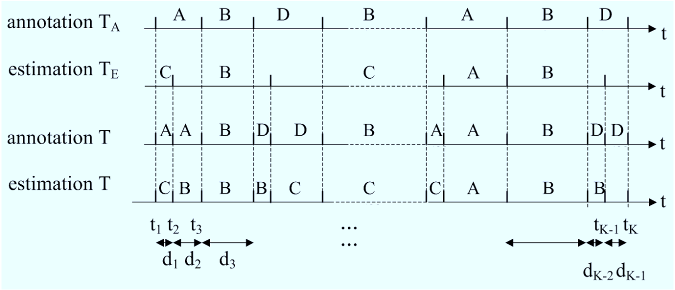

Fig. 7. Ilustration of chord label accuracy measure.

- 24 songs present some addition or deletion of beats
- The rest of the songs have a constant 4/4 meter

B. Evaluation Measures

To evaluate the performance of our system, we use the following evaluation measures:

1) Beat and Downbeat position Evaluation Measure: The evaluation of beat and downbeat estimation is performed using the standard Precision, Recall and F-measure. Precision P is defined as the ratio of detected beat/downbeat positions that are relevant. Recall R is defined as the ratio of relevant beat/downbeat positions detected. The F-measure F combines the two $F = 2RP/(R + P)$. For each song, we compare the estimated beat/downbeat positions against annotated beat/downbeat positions within a given tolerance window w depending on the local tempo. The tolerance window w is defined as 10% of the minimum distance between two successive beats in the track. It is centered on the estimated beats when computing the Precision and centered on the annotated beats when computing the Recall. The tolerance window depends on the local tempo (distance between two beat markers) in order to avoid drawing misleading conclusion from the results. Indeed, a fixed tolerance window of 0.166 s for instance would be very restrictive for slow tempi (half-beat duration of 0.5 s at 60 bpm) but would mean accepting counter-beats as correct for fast tempi (half-beat duration of 0.166 s at 180 bpm). The results indicated in Tables IV and III correspond to the mean and the standard deviation over all the tracks of the Precision, Recall and F-measures.

2) Chord Evaluation Measure: We consider two aspects of chord estimation: the label accuracy (how the estimated chord is consistent with the ground truth) and the segmentation accuracy (how the detected chord changes are consistent with the actual locations). The chord label accuracy measure is illustrated in Figure 7 and is defined as follows.

TABLE II
CHORD LABEL ACCURACY RESULTS (IN %) CONSIDERING SEVERAL CASES: NOT INTEGRATING/INTEGRATING METRIC STRUCTURE INFORMATION IN THE MODEL (NM/WM), TACTUS OR TATUM ANALYSIS (TAC/TAT), USING THEORETICAL BEAT POSITIONS (TB) OR AUTOMATICALLY ESTIMATED BEAT POSITIONS (EB).

no meter (NM)			
theoretical beats (TB)		estimated beats (EB)	
TAC	TAT	TAC	TAT
69.6 ± 13.9	71.2 ± 13.1	68.5 ± 14.0	71.2 ± 13.1

with meter (WM)			
theoretical beats (TB)		estimated beats (EB)	
TAC	TAT	TAC	TAT
71.5 ± 13.3	72.9 ± 13.3	70.4 ± 14.2	72.8 ± 13.3

	TAC	TAT	TAC	TAT
Improvement WM/NM (%)	2.7	2.4	2.8	2.2
Statistical significance	yes	yes	yes	yes

TABLE III
DOWNBEAT POSITION ESTIMATION RESULTS CONSIDERING SEVERAL CASES: THEORETICAL OR ESTIMATED BEATS (TB/EB), TACTUS/TATUM-SYNCHRONOUS ANALYSIS (TAC/TAT). PRECISION (PREC), RECALL (REC), F-MEASURE (F-M).

	theoretical beats (TB)		estimated beats (EB)	
	TAC	TAT	TAC	TAT
Prec	0.89 ± 0.20	0.84 ± 0.26	0.76 ± 0.30	0.80 ± 0.26
Rec	0.90 ± 0.20	0.86 ± 0.26	0.76 ± 0.31	0.79 ± 0.28
F-m	0.89 ± 0.20	0.85 ± 0.26	0.76 ± 0.31	0.79 ± 0.27

For each song s of the testset, let $T_A = (t_{A1}, t_{A2}, \ldots, t_{AM})$ denote time positions corresponding to the annotated (ground truth) chord changes and let $T_E = (t_{E1}, t_{E2}, \ldots, t_{EN})$ denote time positions corresponding to the estimated chord changes. We note $T = T_A \cup T_E$. For each segment $[t_k, t_{k+1}] \subset T$ of duration d_k, we compare the estimated chord $\hat{C}_k$ with the annotated chord C_k. The chord recognition rate is computed as:

$$\mu_s = 100 * \frac{\sum_{k \ so \ that \ C_k = \hat{C}_k} d_k}{\sum_{k=1}^{K-1} d_k} \tag{11}$$

Note that there is no stage in our HMM corresponding to "N" chords of the annotation (denoting noise, silent parts or non-harmonic sounds). They are counted as errors in the evaluation. The results we give in Table II correspond to the mean and standard deviation of correctly identified frames per song.

The chord segmentation accuracy is evaluated using the standard Precision P (ratio of detected chord changes that are relevant), Recall R (ratio of relevant chord changes detected) and f-measure F, using a tolerance window TW of 30% of the minimum distance between two beats in the track.

VI. ANALYSIS OF THE RESULTS

We provide in this section a detailed analysis of the results. This is illustrated through some examples that have been chosen for their relevance. Since the interrelationship between musical attributes is the main purpose of this work, special attention is devoted to this aspect. This section starts with a global presentation of the results. We then analyze in detail the downbeat estimation results. We continue with a comparison of the chord estimation results with other state-of-the-art chord detection systems through the Music

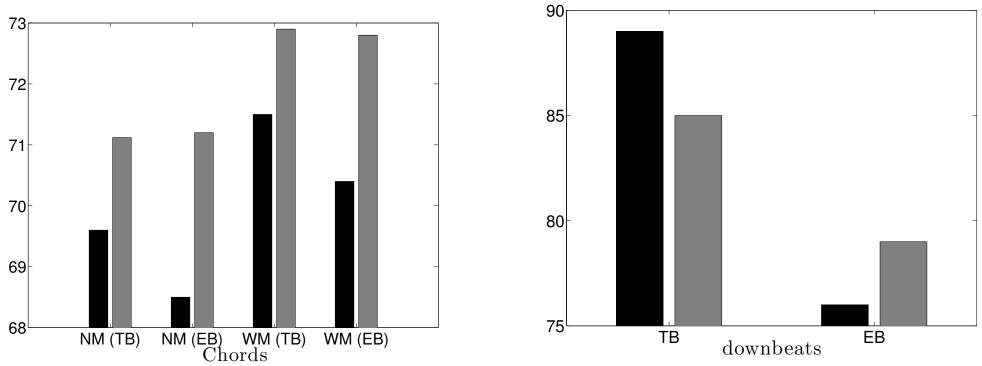

Fig. 8. Histogram of chord [left] and downbeat [right] estimation results (in %) considering several cases: not integrating/integrating metric structure information in the model (NM/WM); using theoretical beat positions (TB) or automatically estimated beat positions (EB). The results from the tactus-synchronous analysis are represented in black, the results from the tatum-synchronous analysis are represented in grey.

TABLE IV
BEAT POSITION ESTIMATION RESULTS.

Precision	Recall	F-measure
0.91 ± 0.22	0.88 ± 0.24	0.89 ± 0.23

Information Retrieval Evaluation eXchange (MIREX) 2008 results. This comparison is followed by a discussion about the influence of each musical attribute on the estimation of the other. We finish with some case study examples.

A. Chords and Downbeats Interaction

The results are presented in Tables II, and III and illustrated in Fig. 8. An accuracy result up to 79% (EB-TAT) suggests that relying on the chord structure of a piece is an appropriated approach for downbeat estimation. Conversely, taking into account the metric structure allows us to improve the chord recognition task by 2.8% relative improvement in the case of tactus-frame analysis and 2.2% relative improvement in the case of tatum-frame analysis. We performed a paired sample t-test to determine if there is a significant difference between the chord estimation result obtained without considering interaction with the metric structure (NM) and with consideration of interaction with the metric structure (WM). For the various situations (TB, EB, TAC, TAT), the null hypothesis could be rejected at the 5% significance level. We can conclude that there is a statistical difference on the chord estimation results when considering the metric structure in the model.

B. Downbeat Position Estimation

In this section, we evaluate the performance of downbeat estimation comparing the output of our algorithm to the ground truth downbeat times annotated by hand. Following the approach proposed in [34], we measure the performance of downbeat estimation considering two cases. On the one hand, we evaluate the upper limit of the model by estimating the downbeat positions using manual annotation of beat positions (referred to as theoretical beat positions (TB) in Table III). On the other hand, we measure the fully automatic performances of the system by using a beat tracker [23] as a front end of the system. The beat positions estimated with the beat tracker are referred to as estimated beat positions (EB) in Table III. With these two measures, we can distinguish between errors due to poor beat position estimation and errors due to the model.

1) Semi-automatic Downbeat Position Estimation: The results presented in Table III show that the system leads to a good estimation of downbeat positions. It achieves 89% of correct estimation in the case of tactus-synchronous analysis and 85% in the case of tatum-synchronous analysis. The encouraging results obtained in the case of tatum-synchronous analysis highlight the robustness of the presented approach. It can be remarked that the standard deviation is high. This can be explained by the fact that the downbeat estimation score is null for some pieces, in particular when there are many half-measure chord changes in the song. In this case, the downbeat positions may be located by the algorithm on the third instead of the first beats of the measures.

2) Using Estimated Beat Positions Versus Theoretical Beat Positions: The downbeats estimation relies on the knowledge of the beat positions. For real applications of the system, we need to use automatically estimated beats. Errors in the beat tracking will be carried forward into the downbeat tracking stage. Beat tracking results evaluated on the testset are presented in Table IV and show that the beat tracking is not perfect. We thus expect a lower downbeat tracking performance using the estimated beats compared to downbeat tracking performance using the ground-truth beats. However, the decrease in the results from semi-automatic to fully-automatic analysis is lower in the case of tatum-frame analysis than in the case of tactus-frame analysis because some common beat tracking errors do not affect downbeat estimation at the tatum-level.

Some common errors in beat tracking algorithm are octave errors (e.g., halving or doubling the beat positions). In case

TABLE V
DOWNBEAT ESTIMATION RESULTS FOR PROPOSED APPROACH (PA), MEPD APPROACH (MEPD). RESULTS ACROSS THE WHOLE DATASET (WHOLE DATA), RESULTS ACROSS SONGS WITH PERFECT BEAT TRACKING (PERFECT BT), RESULTS ACROSS SONGS WITH IMPERFECT BEAT TRACKING (IMPERFECT BT). PRECISION (PREC), RECALL (REC), F-MEASURE (F-M).

	Whole Data	
	MEPD	PA
Precision	0.74 ± 0.36	0.81 ± 0.26
Recall	0.72 ± 0.37	0.79 ± 0.28
F-measure	0.72 ± 0.36	0.79 ± 0.27

	Perfect BT		Imperfect BT	
	MEPD	PA	MEPD	PA
Prec	0.90 ± 0.24	0.86 ± 0.26	0.36 ± 0.30	0.71 ± 0.24
Rec	0.90 ± 0.24	0.87 ± 0.26	0.30 ± 0.25	0.62 ± 0.25
F-m	0.90 ± 0.24	0.86 ± 0.26	0.32 ± 0.25	0.64 ± 0.23

of halved beat positions, a maximum downbeat tracking score recall of 0.5 using tactus-frame features can be expected, but a recall of 1 could be theoretically reached using tatum-frame features. Off-beat errors (tapping at the annotated metrical on the off-beat positions) in addition to a halved tempo estimation is another common beat tracking error. If such a beat estimation error is constant throughout the analyzed piece, we expect to have a null score for tactus-synchronous analysis but a score similar to the one obtained using the theoretical beat position for tatum-synchronous analysis. This was corroborated in our experiments. An interesting case of beat estimation errors concerns the addition or deletion of beats due to a tempo deviation (e.g., slowdown in the tempo). The presented system is supposed to tackle this situation as it does when there is beat addition or deletion within the music (see below).

3) Comparison With the State-of-the-art: We compare the performance of our algorithm (WM-EB-TAT) against those obtained using M.E.P. Davies's model [34], which we refer to as MEPD. In the MEPD approach, the downbeats are estimated based on spectral difference between band-limited beat synchronous analysis frames. The analysis is restricted to the cases where the time signature does not change. The algorithm requires a sequence of beat times and the time-signature of the input signal to be known a priori. For comparison with our system, we have used our beat tracker as input to the MEPD downbeat estimation system. We computed i) the results across the whole dataset, ii) the results across the songs for which the beat tracking was perfect, iii) the results across the songs for which the beat tracking was imperfect.

Results reported in Table V show that our system is globally more successful than the MEPD approach and thus compares favorably to the state-of-the-art. MEPD obtains better results across songs with perfect beat tracking. Most of those songs have a constant time-signature. For those files, The MEPD accuracy for each of those les will either be 0 or 100%, whereas our system may insert some additional downbeats. This highlights a shortcoming of our system: we need to make a compromise between favoring constant meter and allowing meter changes (see below). However, our system performs clearly better across the songs on which the beat tracking was imperfect. Any added or omitted beat in the beat tracking will irrecoverably degrade the MEPD downbeat tracking process whereas our system can handle those situations. Our system thus shows improvements over the state-of the-art.

4) Handling Variable Meter: Previous works on downbeat tracking have mostly focused on pieces with constant meter. The present work proposes an approach that considers some cases of variable meter. The results we obtain are encouraging. We obtain a score of 56% on tactus and tatum analysis on the 9 variable meter pieces of the testset. For each song, we can determine a predominant meter. The transition matrix corresponding to the predominant meter of the piece has been correctly chosen for all songs but one. Ideally, the system should remain in the new meter when a change in meter occurs. However, the model is built in order to favor constant meter within a music piece. For this reason, if the chord changes are not strongly enough marked in the chromagram (high spectral difference between frames), the system will not adapt to the meter change until there is a sufficiently clear chord change. In case of a meter change from a predominant meter $3/4$ to $4/4$, the proposed algorithm inserts measures in $4/4$ so that most of the downbeat positions are correctly detected when the predominant meter returns.

Let us illustrate this on an example. The song *I Me Mine* has a $3/4$ predominant meter with changes to $4/4$ meter. Due to percussive sounds, the chromagram is blurred and chord changes are not clear. Note that the beat positions are not perfectly estimated by the beat tracker (see the dashed rectangle). It can be seen in Fig. 9 that, during the $4/4$ meter passage (from $33s$ to $55s$), the system mostly remains in $3/4$. However, measures in $4/4$ are inserted (see the black circles) so that the downbeats are correctly estimated when the $3/4$ meter returns. Our model shows some adaptation to meter changes even if it is not perfect. On the provided example, the 11 bars of the 4/4 section cannot be divided into a whole number 3/4 bars. If the system had constantly remained in 3/4, the rest of the downbeats until the end of the song wouldn't have been correctly detected. Note that, even for many human listeners, it is very difficult to understand meter changes on this complex example. Experiments carried out by one of the authors on 6 trained musicians clapping their hands along with the music have shown that listeners needed between 2 and 3 measures before synchronizing with the correct downbeat positions of the $4/4$ meter passage.

The last line in Fig. 9 represents the downbeat tracking obtained by increasing the value of α in Eq.(7), so that constant 3/4 meter is less favored by the model. With this value, the algorithm shows more flexibility to the meter change. We plan to find methods to reduce the trade-off between favoring constant meter and allowing meter changes.

5) Handling Addition or Deletion of Beats: It is possible that there is an addition or an omission of beats within a constant-meter part of a song, either due to the music itself or due to a beat tracking error. This is illustrated in

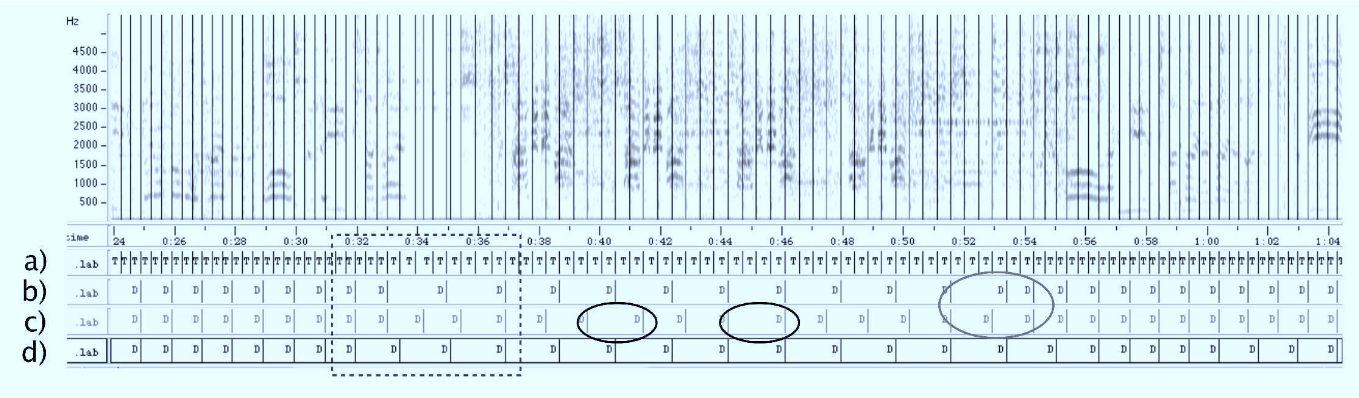

Fig. 9. Estimated downbeat positions of an excerpt of the song *I Me Mine*. Annotated beat positions [top, a)], annotated downbeat positions [middle top, b)], estimated downbeat positions [middle bottom, c)] with $\alpha_3 = 0.6$, estimated downbeat positions [middle bottom, d)] with $\alpha_3 = 0.85$. Measures in $4/4$ inserted by the model are indicated by the two black circles. Extra beats added by the Beatles at the end of the passage in $4/4$ meter are indicated by the grey circle. The dashed rectangle shows a region with errors in the beat tracking. The image has been obtained using the Open Source tool *Wavesurfer*.

Fig. 9: two extra beats have been added by the Beatles at the end of the passage in $4/4$ meter (see the grey circle). The estimated succession of beats is 1 2 3   1 2 3   1 2 3 instead of 1 2 3 4   1 2   1 2 3. This corroborates our expectations stated in part IV-E: the system synchronizes to the correct downbeat positions after a few beats following the added or deleted beat.

C. Chord Estimation

In this section, we analyze the performances of chord label estimation comparing the output of our algorithm to ground truth chord labels annotated by hand.

1) Comparison of the Results with MIREX 2008 "Audio Chord Detection": The authors of this paper participated to the first chord detection task in Music Information Retrieval Evaluation eXchange[7]. In the submitted system, the chords were estimated without considering interaction with downbeats. To set the algorithm presented in this article among other state-of-the-art chord detection algorithms, we first report and analyze the MIREX 2008 chord detection results.

The MIREX 2008 Audio Chord Detection task was divided into two subtasks. In the first subtask the systems were pretrained and tested against 176 Beatles songs. In the second subtask systems were trained on $2/3$ of the Beatles testset and tested on $1/3$. Our system does not need any training, we thus participated to the first subtask. An overlap score was calculated as the ratio between the overlap of the ground truth and detected chords and ground truth duration. Four songs were excluded from the original Beatles testset because of problems aligning the ground truth chords to the audio data.

A total of 8 algorithms were submitted to the first subtask, and our algorithm obtained the fourth place. Note that silent or no-chord segments were not estimated with our algorithm. The differences in the results between the participants are very small, probably because the approaches are similar (using HMM). The four highest results were the following: Bello and Pickens [11] obtained 66% of correct detected chords, Mehnert [35] 65% correct, Ryynanen and Klapuri [36] 64% correct, Papadopoulos and Peeters [37] 63% correct. Our system compares favorably to the trained-systems. Indeed, 7 algorithms were submitted to the second subtask. The approach proposed by Uchiyama, Miyamoto, and Sagayama [38] gave results that were significantly better than the other submitted algorithms (72% correct). Ellis obtained [39] 66% correct results. All the remaining algorithms gave results above 62%.

Using MIREX's exact methodology (chord evaluation measure and dataset), we have re-computed the score obtained with our MIREX 2008 algorithm and computed the score obtained with the newly proposed algorithm (EB-WM-TAT). We obtained a statistically significant relative improvement of 2.4%[8].

2) Chord segmentation: Table VI presents the chord segmentation accuracy results. It can be seen that jointly estimating the downbeats with the chords allows us to improve significantly the chord segmentation. Chord estimation results presented in Table II may seem contradictory since, in the TB case, tatum-based features result in better chord estimation whereas tactus-based features result in better downbeat tracking. It is worth noting that chord estimation is better on tatum-based results even without joint estimation of chords/downbeats. This may be explained by the fact that tatum-based analysis allows us to take into account chord changes on off-beats whereas tactus-frame analysis only allows chord changes on beats. However, improvement of chord segmentation accuracy corresponds with downbeat estimation accuracy. For instance, downbeat estimation based on the beat tracking is better on tatum-frame analysis than on

[7] http://www.music-ir.org/mirex/2008/

[8] Note that despite using MIREX methodology, we did not recover the results reported by MIREX. We obtained 68.8% for our MIREX algorithm and 70.5% for the newly proposed algorithm. A deeper analysis of MIREX's results have shown that it is very likely that there are some errors in the evaluation. For instance, all participants obtained a score close to zero for songs number 23, 35, 76, 97 and 104.

TABLE VI
Chord segmentation accuracy results (in %) considering several cases: not integrating/integrating metric structure information in the model (NM/WM), tactus or tatum analysis (TAC/TAT), using theoretical beat positions (TB) or automatically estimated beat positions (EB). Rel. Imp. (%) indicates the relative improvement between the two approaches. Precision (Prec), Recall (Rec), F-measure (F-m).

	no meter (NM)			
	theoretical beats (TB)		estimated beats (EB)	
	TAC	TAT	TAC	TAT
Prec	61.6 ± 16.8	43.7 ± 15.0	55.6 ± 21.4	43.6 ± 15.1
Rec	59.1 ± 17.1	56.5 ± 16.7	52.7 ± 21.3	56.4 ± 16.6
F-m	59.0 ± 15.2	48.4 ± 15.0	52.8 ± 19.8	48.2 ± 14.8

	with meter (WM)			
	theoretical beats (TB)		estimated beats (EB)	
	TAC	TAT	TAC	TAT
Prec	68.3 ± 17.7	57.4 ± 18.0	61.3 ± 23.2	56.8 ± 18.3
Rec	72.5 ± 17.1	73.5 ± 18.4	64.4 ± 23.8	71.1 ± 18.7
F-m	69.1 ± 15.8	63.3 ± 17.2	61.6 ± 22.1	62.0 ± 17.2
Rel. Imp.	17.12	30.8	16.7	28.6

TABLE VII
Proportion (in%) of chord errors corresponding to harmonically related chords.

Relative	Dominant	Sub-dominant
10	13	34

tactus-frame analysis and consequently, chord segmentation is slightly better on tatum-frame analysis than on tactus-frame analysis.

3) Analysis of Chord Detection Errors: In this part, we focus only on chord estimation results and analyze chord detection errors. The results indicated in Table II show that we obtain up to 72.8% of correctly identified chords on our testset. As can be seen, the standard deviation of the results is relatively high (around 13%). A deeper analysis of the results shows that the errors come from a subset of songs which possess specific characteristics described below.

- **Chord confusion due to chord lexicon mapping:** As mentioned earlier, because of our limited chord dictionary, a mapping was performed between complex chords and their root triad. The chord type distribution in the testset is unbalanced and whereas the majority of songs in the evaluation testset are composed of triad chords, some of them contain many partial or complex (non-triads) chords. The system sometimes recognises other triads than the root triad of the complex chord analyzed, which decreases the recognition rate. For instance, the song *Ask Me Why* contains many G#min7 chords (G#-B-D#-F#). This complex chord comprises a G# minor chord (G#-B-D#) and a B major chord (B-D#-F#). The theoretically correct answer depends on the tonal function of the chord in the harmonic progression. Modeling chord sequences using longer dependencies between chords, using for instance probabilistic N-grams, would help characterize the complexities of harmonic progressions in western tonal music.

- **Neighboring triad confusion:** Table VII shows that a large portion of chord errors (about 57%) correspond to harmonically close triad confusion: relative chords (Am being confused with CM); dominant chords (CM being confused with GM) or subdominant chords (CM being confused with FM). Parallel major/minor chords (EM being confused with Em) account for 13%. The distribution of the type of errors is similar for all the configurations of the system (TAC, TAT, TB, EB). Note that their is a notable predominance of sub-dominant errors in the results. This may be due to the high value given to transitions between subdominant chords in the cognitive-based transition matrix. We have found that diminishing this value decreases the sub-dominant errors rate. However, this also decreases the global results. This shows some limitations of our approach that is not based on training but on theoretical and cognitive-based music knowledge. If the system does not recognize exactly a chord but makes such confusions, the result can still be useful for higher-level structural analysis such as key estimation, harmony progression or segmentation. The results obtained by the system when taking into account these harmonically close chords are quite high (80%). The harmonically close chord errors do not have all the same qualitative weight. Parallel errors, for instance, may badly affect key recognition. However, the most common harmonically close errors are dominant and sub-dominant chords (having a perfect fifth relationship between the estimated and ground-truth chord), which should not affect key estimation. Neighboring triad confusion may not be critical to downbeat estimation. A relevant example of this assessment here concerns the detection of metric structure. We obtain a score of 57% correctly detected chords on the song *Don't Bother Me*, which is rather low compared to the other songs. However, most of the errors correspond to neighboring chords and the harmonic structure has been well-preserved (chord changes occur according to the measures), as illustrated in Fig. 10. For that reason, the downbeat positions of the song have been correctly detected.

- **Passing tones, missing notes:** In the song *Till There Was You*, there is a repeating pattern beginning by an F major chord lasting two beats. The system estimates the following chords: FM-Dm. If we listen to the music, we can hear that on the first two beats, the guitar is playing a broken F major chord (F-A-C). On the second beat, the C note is not present any more. A musician would naturally label the two chords as an F major chord, ignoring the fact that there are missing notes (because it is the same harmony). However, the signal features only take into account notes which are present in the signal. As a result, the estimated chords do not match exactly those of the ground truth. This example leads to the relevant question of how to evaluate the performances of a chord estimation system. The ground truth is provided

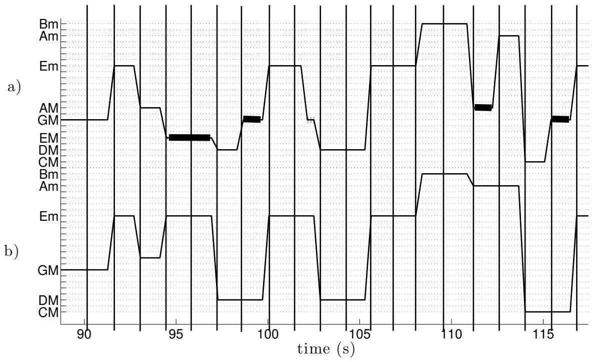

Fig. 10. Estimated chord progression of an excerpt of the song *Don't Bother Me* [a)] and ground truth [b)]. The downbeat positions are represented by vertical lines.

- by trained musicians who not only take into account the notes present in the signal but also the harmonic context to label the chords, ignoring the addition or the deletion of some notes in their annotation. This complicates the evaluation of the algorithm.
- **Limitation of the chroma-based approach for inharmonic sounds:** It is interesting to notice that we obtain much better results for the five first Beatles albums than for the others (from the "Norwegian Wood (This Bird Has Flown)" on 1965's Rubber Soul). The reason for this may come from the extended use of the Indian sitar instrument[9] and various percussive instruments such as bells, wood blocks or congas that cause transients. Since the chroma-based approach strongly relies on the presence of harmonic sounds, the use of chroma-based signal features would ideally require a pre-processing step that effectively reduces transients and noise. We plan to concentrate on this point in future work.

4) Tactus-synchronous Versus Tactum-synchronous Analysis: Table II indicates that the tatum-frame analysis performs slightly better in general than the tactus-frame analysis. This may be due to the fact that tatum-based analysis allows us to take into account chord changes on off-beats whereas tactus-frame analysis only allows chord changes on beats.

D. Case Study Examples

In this part, we present some examples that illustrate some important advantages from estimating simultaneously the chords and the downbeat positions.

1) Boundary Errors: Taking into account the position of the downbeats when estimating the chord progression allows us to enhance the accuracy of the estimation. Indeed, when this information is not taken into account, the chord change may be detected a beat before or after its theoretical position, because of the smoothness of chord transition. This is illustrated in Fig. 11. The ground-truth is indicated by the truth-line c). When the chords are estimated independently from downbeat positions, errors often occur around *pim*. When they are taken into account, chord changes on the correct position are favored (see line b)).

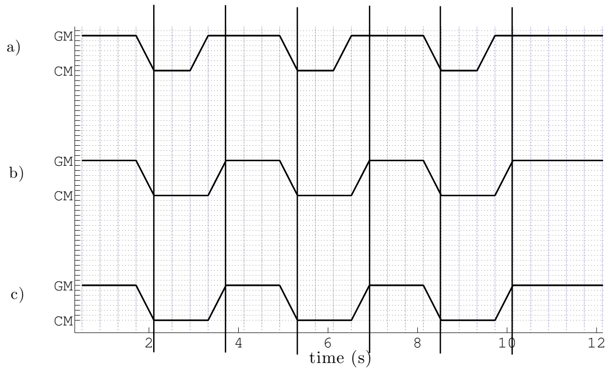

Fig. 11. Chord progression of the first few seconds of the song *Love Me Do* without taking into account the downbeat positions [a)] and taking into account the downbeat positions [b)]. Ground truth [c)]. The downbeat positions are represented by vertical lines.

2) Chord Changes: The example in Fig. 12 clearly shows how the chord progression estimation task can benefit from modelling chord dependencies to the metric structure. This piece is in C Major key and it changes between C Major and G Major chords about every two measures (ground-truth line c)). Without taking into account global dependencies (line a)), chord transitions are badly detected and the estimated chord progression remains almost all the time on the G Major chord instead of transiting between GM and CM. The knowledge of downbeat positions (line b)) allows us to better detect transitions.

VII. Conclusion and Future Work

In this paper, we have presented a system that allows the simultaneous estimation of the chord progression and the downbeat positions of an audio file. The key idea behind our approach is that the harmonic structure is closely related to the metric structure of a piece of music. Relying on this idea, we have built a specific topology of HMM where each state is a combination of an occurrence of a chord and a position of the chord in the measure. Each state is thus related on the one hand to the harmonic structure and on the other hand to the metric structure of the piece. Harmonic structure information and metric structure information are encoded in the state transition matrix. The chord progression and the downbeats are estimated jointly based on the assumption that chords are more likely to change on the beginning of a measure than on other positions. An important contribution of this article is that we consider the case of pieces with varying time-signatures.

The system has been evaluated and compared to the state-of-the-art on a large set of hand-labeled files. We have demonstrated that considering the interaction between the two musical attributes allows their simultaneous estimation and

[9] The sitar is a stringed instrument that uses sympathetic strings in addition to regular strings. This produces a very lush sound with complex, competing harmonic components.

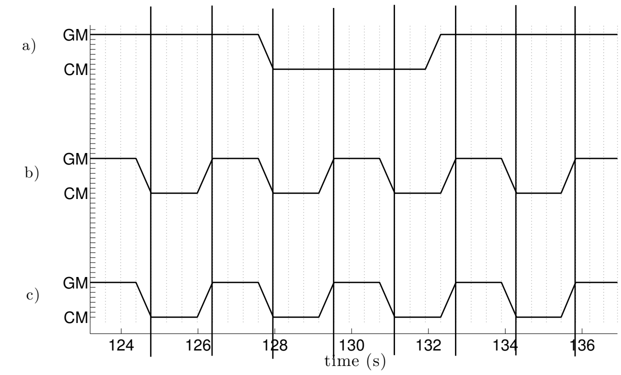

Fig. 12. Chord progression of the last few seconds of the song *Love Me Do* without taking into account the downbeat positions [a)] and taking into account the downbeat positions [b)]. Ground truth [c)]. The downbeat positions are represented by vertical lines.

that the robustness and the chord estimation accuracy is higher when estimated jointly with downbeats.

We have provided a detailed analysis of the results illustrated by case studies that suggest that some points need further improvement that include a pre-processing step that removes transients and noise and the use of longer dependencies between chords (using, for instance, probabilistic N-grams).

We have considered the problem of using imperfect beat positions obtained from beat tracking. Results show that using a tatum-synchronous analysis instead of a tactus-synchronous analysis might temper the effects of imperfect beat tracking on downbeat tracking. The model allows us to take into account pieces with complex metric structure. The downbeat tracking results for pieces in variable meter are encouraging even if they need further improvement. For the moment, the system is built so that it remains in a single predominant meter along the analyzed track. It would be highly desirable that the system shows more flexibility to the meter changes. Future work will concentrate on this point.

An analysis of the results shows that the harmonic structure of a piece is an important clue for determining the downbeat positions. However, it has been noticed that in some cases (such as when chords change every two beats), the relationship between chord changes and downbeats is ambiguous. This model would benefit from a more complete functional chord analysis. Combining the present system, which is based on harmony, with a rhythmic pattern approach would probably also allow improvement of the downbeat tracking process.

VIII. Acknowledgment

Many thanks to Matthew E. P. Davies for making his source code available, to Leigh Smith for his precious suggestions and to the anonymous reviewers for their very helpful comments.

References

[1] J. Serrà, E. Gómez, P. Herrera, and X. Serra, “Chroma binary similarity and local alignment applied to cover song identification,” *IEEE Transactions on Audio, Speech and Language Processing*, vol. 16, pp. 1138–1151, 2008.

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

[3] T. Fujishima, “Real-time chord recognition of musical sound: A system using common lisp music,” in *ICMC*, Beijing, China, 1999, pp. 464–467.

[4] G. Wakefield, “Mathematical representation of joint time-chroma distribution,” in *SPIE Conf. Advanced Sig. Proc. Algorithms , Architecture and Implementation*, vol. 3807, July Denver, Colorado, 1999, p. 637645.

[5] C. Harte, M. Sandler and M. Gasser, “Detecting harmonic change in musical audio,” in *AMCMM*, Santa Barbara, 2006.

[6] K. Lee, “A system for chord transcription, key extraction, and cadence recognition from audio using hidden Markov models,” Ph.D. dissertation, Stanford University, CA, USA, May 2007.

[7] C. Harte and M. Sandler, “Automatic chord identification using a quantised chromagram,” in *AES 118th Convention*, Barcelona, Spain, 2005.

[8] E. Gómez, “Tonal description of polyphonic audio for music content processing,” *INFORMS J. on Computing*, vol. 18, no. 3, pp. 294–304, 2006.

[9] H. Papadopoulos and G. Peeters, “Large-scale study of chord estimation algorithms based on chroma representation and HMM,” in *CBMI*, Bordeaux, France, 2007.

[10] A. Sheh and D. Ellis, “Chord segmentation and recognition using em-trained HMM,” in *ISMIR*, Baltimore, MD, 2003, pp. 183–189.

[11] J. Bello and J. Pickens, “A robust mid-level representation for harmonic content in music signal,” in *ISMIR*, London, UK, 2005, pp. 304–311.

[12] J. Burgoyne and L. Saul, “Learning harmonic relationships in digital audio with Dirichlet-based hidden Markov models,” in *ISMIR*, London, 2005.

[13] K. Lee and M. Slaney, “Acoustic chord transcription and key extraction from audio using key-dependent HMMs trained on synthesized audio,” *Audio, Speech, and Language Processing, IEEE Transactions on [see also Speech and Audio Processing, IEEE Transactions on]*, vol. 16, no. 2, pp. 291–301, 2008.

[14] A.P. Klapuri, A. Eronen and J. Astola, “Analysis of the meter of acoustic musical signals,” *IEEE Transactions on Audio, Speech and Language Processing*, vol. 14, no. 1, pp. 342–355, 2006.

[15] H. Allan, “Bar lines and beyond - meter tracking in digital audio,” Master’s thesis, School of Informatics, University of Edinburgh, 2004.

[16] T. Jehan, “Downbeat prediction by listening and learning,” in *WASPAA*, New Paltz, NY, October 2005.

[17] D. Ellis and J. Arroyo, “Eigenrhythms: Drum pattern basis sets for classification and generation,” in *ISMIR*, Barcelona, 2004.

[18] M. Davies and M. Plumbley, “A spectral difference approach to downbeat extraction in musical audio,” in *EUSIPCO*, Florence, Italy, 2006.

[19] M. Gainza, D. Barry and E. Coyle, “Automatic bar line segmentation,” in *AES 123rd Convention*, New York, NY, USA, October 2007.

[20] K. Noland and M. Sandler, “Influences of signal processing, tone profiles, and chord progressions on a model for estimating the musical key from audio,” *Comput. Music J.*, vol. 33, no. 1, pp. 42–56, 2009.

[21] K. Sumi and K. Itoyama and K. Yoshii and K. Komatani and T. Ogata and H.G. Okuno, “Automatic chord recognition based on probabilistic integration of chord transition and bass pitch estimation,” in *ISMIR*, Philadelphia, Pennsylvania US, 2008.

[22] J.F. Paiement, D. Eck, S. Bengio and D. Barber, “A graphical model for chord progressions embedded in a psychoacoustic space,” in *ICMC*, Bonn, Germany, 2005.

[23] G. Peeters, “Template-based estimation of time-varying tempo,” *EURASIP Journal on Advances in Signal Processing*, vol. 2007, pp. Article ID 67 215, 14 pages, 2007, doi:10.1155/2007/67215.

[24] N.C. Maddage, “Automatic structure detection for popular music,” *IEEE MultiMedia*, vol. 13, no. 1, pp. 65–77, 2006.

[25] G. Peeters, “Musical key estimation of audio signal based HMM modeling of chroma vectors,” in *In DAFX, McGill*, Montreal, Canada, 2006, pp. 127–131.

[26] ——, “Chroma-based estimation of tonality from audio-signal analysis,” in *ISMIR*, Victoria, Canada 2006, pp. 115–120.

[27] J. Brown, “Calculation of a constant Q spectral transform,” *J. Acoustical Society of America*, vol. 89, no. 1, pp. 425–434, 1991.

[28] C. Krumhansl, *Cognitive Foundations of Musical Pitch.* Oxford University Press, New York, 1990.

[29] K. Noland and M. Sandler, “Key estimation using a hidden Markov model,” in *ISMIR*, Victoria, Canada, 2006, pp. 121–126.

[30] H. Papadopoulos and G. Peeters, “Simultaneous estimation of chord progression and downbeats from an audio file,” in *ICASSP*, Las Vegas, 2008.

[31] H. Papadopoulos and G. Peeters, “Local key estimation based on harmonic and metric structures,” in *Dafx*, Como, Italie, 2009.

[32] B. Gold and N. Morgan, *Speech and audio Signal Processing: Processing and Perception of Speech and Music*. John Wiley & Sons, Inc., 1999.

[33] C. Harte, M. Sandler, S. Abdallah, and E. Gómez, “Symbolic representation of musical chords a proposed syntax for text annotations,” in *ISMIR*, London, UK, 2005.

[34] M. Davies, “Towards automatic rhythmic accompaniment,” Ph.D. dissertation, Queen Mary University of London, London, UK, August 2007.

[35] M. Mehnert, G. Gatzsche, D. Arndt, and T. Zhao, “Circular Pitch Space Based Chord Analysis,” in *MIREX*, Philadelphia, Pennsylvania USA, 2008.

[36] M. Ryynänen and A. Klapuri, “Chord Detection Method for Mirex 2008,” in *MIREX*, Philadelphia, Pennsylvania USA, 2008.

[37] H. Papadopoulos and G. Peeters, “Chord Estimation Using Chord Templates And HMM,” in *MIREX*, Philadelphia, Pennsylvania USA, 2008.

[38] Y. Uchiyama, K. Miyamoto, N. Ono, and S. Sagayama, “Automatic Chord Detection Using Harmonic Sound Emphasized Chroma From Musical Acoustic Signal,” in *MIREX*, Philadelphia, Pennsylvania USA, 2008.

[39] D. Ellis, “Simple Trained Audio Chord Recognition,” in *MIREX*, Philadelphia, Pennsylvania USA, 2008.