Simultaneous estimation of chord progression and downbeats from an audio file

Hélène Papadopoulos, Geoffroy Peeters

SIMULTANEOUS ESTIMATION OF CHORD PROGRESSION AND DOWNBEATS FROM AN AUDIO FILE

Hélène Papadopoulos, Geoffroy Peeters

IRCAM
Sound Analysis/Synthesis Team, CNRS - STMS, Paris - France
papadopo@ircam.fr, peeters@ircam.fr

ABSTRACT

Harmony and metrical structure are some of the most important attributes of Western tonal music. In this paper, we present a new method for simultaneously estimating the chord progression and the downbeats from an audio file. For this, we propose a specific topology of hidden Markov models that allows us to model chords dependency on metrical structure. The model is evaluated on a dataset of 66 popular music songs from the Beatles and shows improvement over the state of the art.

***Index Terms*—** HMM, Chroma, Chord, Downbeat, Metrical structure

1. INTRODUCTION

Musical signals are highly structured in terms of harmony and rhythm. Thus, these two components are essential in the understanding of music. Harmonic analysis and rhythm analysis find many applications within the context of music information retrieval such as music classification, structural audio segmentation or in general all applications based on music content analysis. In Western tonal music, the chord progression determines the harmonic structure of a piece of music. It is strongly related to the metrical structure of the piece [1]. The meter is "the sense of strong and weak beats that arises from the interaction among hierarchical level of sequences having nested periodic components" [2]. In a piece of music, each chord is locally related to the surrounding chords according to the harmonic progression of the piece (local dependency). Furthermore, the position that a chord occupies in a measure or more generally in the global metrical structure has to be taken into account (global dependency). For example, chords will change more often on strong beats than on other positions. This musical characteristic has already been explored in previous works ([3], [4] [2], [5], [6]).

Metrical level is a hierarchical structure. The most salient metrical level, called the *tactus* or beat level, is a moderate metrical level which 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 " [7]. 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 metrical analysis is finding the position of the *downbeat* or the first beat of each measure.

In the last few years, there has been an increasing interest in modeling higher-level information with low-level signal features in the context of music analysis. Two paths have been explored. On the one hand, hierarchical frameworks based on rule-based approach have been proposed (see for instance [3], [4]). On the other hand, statistical framework including graphical models and Bayesian approach have been proposed (see [2], [8], [9], [10], [6]). Statistical approaches are more flexible than rule-based approaches and offer large opportunities to explore the interaction between low-level features with high-level music information. Our purpose is to show how the metrical information and the harmonic information of a piece of music interact and how this can be used into a mutually informing manner to improve both the estimation of the chord progression and the downbeat positions. For this, we propose a specific topology of HMM that allows us to extract simultaneously the chord progression and the downbeats from an audio file. Our approach is somehow related in spirit to Bayesian modeling. Indeed, we intend to model global dependencies within the chords. Although HMM usually concentrate on local dependencies, it is not the case here.

The paper is organized as follows. In section 2.1, we present the extraction of a set of meter-related feature vectors that represent the audio signal. We introduce a probabilistic model for simultaneous chord progression and downbeat positions estimation in section 2.2. In section 3, the proposed model is evaluated on a set of hand-labeled songs of the Beatles.

2. MODEL

In order to extract the chord progression and the downbeats from the audio signal, one first needs to extract a set of meter-related feature vectors that describe the signal. Pitch Class Profiles [11] or chroma-based representation [12] have become common features to automatically estimate chords or musical key from audio recordings ([13], [14], [15], [16], [17]). PCP/chroma vectors represent the intensity of the twelve semitones of the pitch classes. The chord progression is represented using a hidden Markov model that takes into account global dependencies on meter. The tactus/tatum positions have been extracted using the method proposed in [18]. In our evaluation, we have only considered songs with 100% tactus recognition rate.

Our model is general and could be applied to songs with any kind of time-signature (3/4, 4/4, ...). However, because of dataset availability, we have concentrated our evaluation on the case of popular music and limited our experiments to songs built on four-beat meters (most common case in popular music). We will assume that the time signature is known (4/4) and constant. We will also assume that chord changes can only occur on beats or after beats. Theses hypothesis correspond to the characteristics of a wide part of popular music. For instance, if we consider the first eight CDs of the Beatles (110 songs), only 3 songs do not fit the assumptions we made.

Thanks to ANR Projet Ecoute and AII Project Quaero for funding.

2.1. Features extraction

We work directly on the audio signal. In our analysis, the signal is down-sampled to $11025Hz$, converted to mono and converted to the frequency domain by a DFT using a Blackman window of length $0.48s$ with 25% overlap. Because of frequency resolution limits (the frequency distance between adjacent semitone pitches becomes small in low frequencies), we only consider frequencies above $60Hz$. The upper limit is set to $1kHz$ because the fundamentals and harmonics of the notes in popular music are usually stronger than the non-harmonic components up to $1kHz$ [3]. This choice is also supported by the fact that the mapping operated between the energy of the harmonics and the chroma vectors is only valid for the lowest harmonics, hence the lowest part of the spectrum. The tuning of the track is estimated using the method proposed in [16]. The signal is then re-sampled so that the rest of the system can be based on a tuning of the standard $A4 = 440Hz$. The temporal sequence of chroma vectors over time is known as chromagram. It is computed using the method proposed in [16]. First, the values of the DFT are mapped to a semitone pitch spectrum using the mapping function: $n(f_k) = 12\log_2(\frac{f_k}{440}) + 69, n \in \Re^+$, where f_k are the frequencies of the Fourier transform and n correspond to the semitone pitch scale values. Then, the semitone pitch spectrum is smoothed over time using a median filtering. This provides a reduction of transients and noise in the signal. Finally, after this smoothing, the semitone pitches n are mapped to the the semitone pitch classes c using the mapping function: $c(n) = \mod(n, 12)$. We obtain a sequence of 12-dimensional vectors that are suitable feature vectors for our analysis.

Tactus/tatum-related chroma vectors: Since we want to study the relationship between chords and metrical 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 related analysis. To this end, the chromagram is averaged so that we obtain one feature per tactus/tatum[1]. 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. We will further discuss the relevance of both approaches.

2.2. Chord progression and downbeat estimation from the chroma vectors using a "double state" HMM

2.2.1. 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, $i \in [1 : I]$ 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). We assume that chord changes can only occur on beats or on after beats. The positions in the measure where chord changes occur will be referred to "position in the measure" and denoted by *pim*. For a song built on four-beats meter, $K = 4$ if we consider the tactus-level ($k \in [1; 4]$ for a 4/4 measure) and $K = 8$ if we consider the tatum-level ($k \in [1; 8]$ for a 4/4 measure). If there are K possible *pim* in a measure, the total number of states is thus I chords by K *pim* i.e. $I * K$ states. 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. In what follows, we denote by π and T the initial state distribution and state transition probability distribution.

[1]The tactus/tatum positions are considered as inputs of our system.

2.2.2. Initial state distribution π

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

2.2.3. Observation probabilities $P(\mathbf{O}(t_m)|s_{ik})$

The observation probabilities are computed in the following way. Let $P(\mathbf{O}(t_m)|s_{ik})$ denote the probability that observation $\mathbf{O}(t_m)$ has been emitted at time instant t_m given that the model is in state s_{ik}. Let $P(\mathbf{O}(t_m)|c_i)$ denote the one that it has been emitted by chord c_i and $P(\mathbf{O}(t_m)|pim_k)$ the one that it has been emitted given that the chord is occurring on *pim* k. As said before, we rely upon the assumption that *chord changes* are more likely to occur at the beginning of measures than at other *pim*. We now assume independence between *chord type* (CM, C#M, ..., cm, ..., bm) and *pim*. 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. The observation probabilities are computed as:

$$P(\mathbf{O}(t_m)|s_{ik}) = P(\mathbf{O}(t_m)|c_i)P(\mathbf{O}(t_m)|pim_k) \quad (1)$$

Observation chord symbol probability distribution: The observation chord symbol probabilities $P(c_i|\mathbf{O}(t_m))$ are obtained by computing the correlation 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. For more details, see [19].

Observation *pim* probability distribution: The *pim* probability distribution $P(pim_k|\mathbf{O}(t_m))$ is considered here as uniform ($\frac{1}{K}$ for each *pim* in the measure). Note that this probability distribution could be derived from information given by the signal. Future works will concentrate on that.

2.2.4. State transition probability distribution T

The main reason why the problem is modeled using a Markov model is that 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 T. According to [1], chords are more likely to change at the beginning of a measure than at other *pim*. Starting from this statement, we detect the downbeats by giving lower self-transition probabilities in the state transition matrix for chords occurring on the K^{th} beat.

The $I * K$-states transition matrix T used in our HMM takes into account both the chord transitions and their respective positions in the measure. It is derived from a I-states chord transition matrix T_c based on music-theoretical knowledge about key-relationships. We refer the reader to [19] for more details. We note $T_c(i, i')$ the transition probability between chord i and chord i'. This matrix is represented in Figure 1 [left].

We also define a *pim* transition matrix T_{pim} which represents the probability to transit from *pim* k to *pim* k'. Since we do not allow our present system to jump over a pim (i.e. skip over or add one or several beats), only the values $T_{pim}(k, k')$ for $k' = k + 1[K]$ are

non-zero. All non-zero values are set to the same value. This matrix is represented in Figure 1 [right, top].

We need here to distinguish between two cases: the first case concerns transitions between two different chords ($i' \neq i$), the second case concerns self-transitions ($i' = i$) and corresponds to the diagonal blocks of T. Since we want to favor chord changes on downbeats, *i.e.* disfavor self-transition between the last *pim* of a measure and the first *pim* of the next measure, we need to define a specific transition matrix for the self-transition case ($i' = i$). This specific matrix is denoted by T'_{pim} . This matrix is represented in Figure 1 [right, bottom]. As one can see T'_{pim} differs from T_{pim} only in the value $T'_{pim}(K, 1)$ which is lower than $T_{pim}(K, 1)$. The consequence of this lower value is that T'_{pim} disfavors transition between identical chords (self-transition) at measure boundaries. In our experiments (case $4/4$ time-signature), we have attributed empirical values to $T'_{pim}(k, k'), k, k' \in [1; 4]$ with respect to the fact that we want to favor chord changes on downbeats[2]. Note that these values could be learned from the dataset by counting the proportion of chord changes on each measure position in the dataset.

From T_c, T_{pim} and T'_{pim}, we construct the global transition matrix T normalized so that the sum of each row is equal to 1 (Figure 1 [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':

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

2.2.5. Chord progression and downbeats detection

The optimal succession of states $[c_i, pim_k]$ over time is found using the Viterbi decoding algorithm [20] which gives us the most likely path trough the HMM states given our sequence of observations. It gives us simultaneously the best sequence of chords over time and the downbeat positions.

3. EVALUATION AND RESULTS

3.1. Test set

The proposed model has been tested on a set of 66 hand-labeled songs of the Beatles[3]. All the songs are built on four-beat meter with constant time signature. The chord annotations were kindly provided by C. Harte from QMUL. Note that since our chord lexicon only represents major and minor triads, we have performed a mapping of complex chords in the annotation (such as major and minor 6^{ths}, 7^{ths}, 9^{ths}, augmented and diminished chords) to their root triads. The tactus were obtained using the method proposed in [18]. The ground-truth downbeats have been annotated by hand by the authors. All the recordings are polyphonic, multi-instrumental songs containing drums and vocal parts.

3.2. Overall results

The results are indicated in Table 1. Let S denote the total number of songs in the dataset and let consider a song s divided into N frames. Each signal frame (tactus-frame or tatum-frame) of the ground truth has been mapped to a chord of our lexicon. Let $\hat{C}_n, n \in [1; N]$ denote the theoretical chord corresponding to frame n and let C_n denote the estimated chord at that frame. We compute the correct chord recognition rate for song s as:

$$\mu_s = \frac{1}{N} \sum_{n \in [1;N]} (C_n = \hat{C}_n) \quad (2)$$

The results we give in Table 1 correspond to the mean and standard deviation of correctly identified frames per song:

$$\mu = \frac{1}{S} \sum_{s \in [1;S]} \mu_s \text{ and } \sigma = \frac{1}{S} \sqrt{\sum_{s \in [1;S]} (\mu - \mu_s)^2} \quad (3)$$

	NM		WM	
	TAC	TAT	TAC	TAT
DU	70,5±12,7	73,3±12,3	73,9±12,4	75,3±12,1
DK	SAME		73,9±12,3	75,9±12,4

D	
TAC	TAT
92,42	78,79

Table 1. Chord estimation rate and downbeat estimation rate.

• *NM* (No Meter)/*WM* (With Meter) columns correspond to the exact recognition rate on all the frames without/when taking into account chords dependency on the metrical structure in the model. TAC corresponds to a tactus-frame analysis, TAT to a tatum-frame analysis.

• *DU* (Downbeat Unknown) row corresponds to the case where downbeats are estimated simultaneously with the chords. *DK* (Downbeat Known) row corresponds to the case where downbeats are given by manual annotation. In this case, we only intend to evaluate the influence of the knowledge of downbeats on chord recognition.

• *D* corresponds to the percentage of songs where the downbeats have been correctly estimated. Note that the dataset contains only pieces without skipped or added beats. For a given song, the estimated downbeats are then either all correct or all incorrect.

3.3. Analysis of the results

Downbeat estimation: The percentage of downbeats correctly estimated is encouraging. It achieves 92% (79%) of correct estimation in the case of tactus-frame (tatum-frame) analysis. Note that without chord information, the downbeats estimation would be 25% for tactus-frame analysis and 12.5% for tatum-frame analysis.

Importance of the knowledge of the downbeat positions: In Table 1 [left], we can see that the chord estimation task benefits from the knowledge of the downbeat positions either given manually (DK) or estimated through the model (DU). Taking into account the *pim* of the chords in the measures allows to improve the chord recognition task by 4.9% relative improvement in the case of tactus-frame analysis and 3.5% relative improvement in the case of tatum-frame analysis. It is important to note that when we perform simultaneous estimation of chord progression and downbeats, the global rate for chord recognition is better than when we do not take into account the influence of the metrical structure, even if the downbeats are not correctly estimated in all the pieces.

Tactus-frame/tatum-frame analysis: Table 1 indicates that the tatum-frame analysis performs better in general than the tactus-frame analysis. Some chords in our dataset do not change exactly on the beats (voice effects, after beats). The tatum-frame analysis allows to take into account chord changes on more positions than the tactus-frame analysis and thus gives better results.

Chord changes and boundaries errors: The example in Figure 2 clearly shows how the chord progression estimation task can benefit from modeling chords dependency on the metrical structure. This piece is in C major key and it transits between C major and G major

[2] $T'_{pim}(1,2) = T'_{pim}(2,3) = T'_{pim}(3,4) = 1.1$, $T'_{pim}(4,1) = 0.85$, $T'_{pim}(k,k') = 0$ otherwise.

[3] The list of the tracks can be found at the following URL: http://recherche.ircam.fr/equipes/analyse-synthese/papadopoulos/.

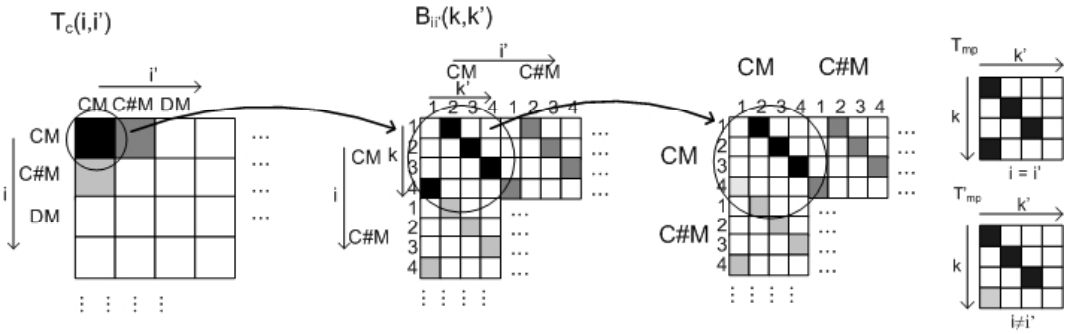

Fig. 1. Chord transition matrix for a singles-state HMM [left], transition matrices for Major to Major chords in the case of double-states HMM, without taking into account the *pim* of the chord in the measure [middle left] and taking into account the *pim* of the chord [middle right], *pim* transition matrices [right].

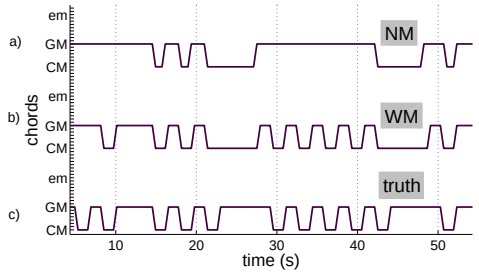

Fig. 2. Chord progression for the song *Love Me Do* not considering a)/considering b) chords dependency on meter, ground truth c).

chords about every two measures (truth line). Without taking into account global dependencies (NM line), chord transitions are badly detected and the estimated chord progression remains almost all the time on G major chord instead of transiting between G major and C major. The knowledge of downbeat positions (WM line) allows to better detect transitions. Furthermore, using the chords dependency on the metrical structure also allows to improve the exact location of chord changes (boudaries). Without taking into account this dependency, chord changes are often detected a beat before or after their theoretical positions.

4. CONCLUSION AND FUTURE WORK

In this paper, we have proposed a method that allows to estimate simultaneously the chord progression and the downbeats from an audio file. This method has been evaluated on a large set of hand-labeled files and gives very encouraging results. From this evaluation, we can state that the chord progression of a piece of music benefits from the knowledge of downbeat positions and conversely that the downbeats of popular music songs can be estimated using harmonic information. An interesting result from our evaluation is that tatum-related analysis is better than tactus-related analysis for the estimation of chord progression. Future works will consist in evaluating the performances of our model on songs with other time signatures. Including a time-signature estimation algorithm in our system will allow us to deal with pieces with meter changes.

5. REFERENCES

[1] 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.

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

[3] N.C. Maddage, "Automatic Structure Detection for Popular Music," *IEEE Multi-Media*, vol. 13, no. 1, pp. 65–77, 2006.

[4] A. Shenoy and Y. Wang, "Key, chord and rhythm tracking of popular music recordings," *Computer Music Journal*, vol. 3, no. 29, pp. 75–86, 2005.

[5] C. Raphael and J. Stoddard, "Harmonic analysis with probabilistic graphical models," in *ISMIR*, Baltimore, Maryland, 2003, pp. 177–181.

[6] H. Thornburg, *Detection and Modeling of Transient Audio Signals with Prior Information*, Ph.D. thesis, Stanford University, september 2005.

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

[8] A.T. Cemgil, *Bayesian Music Transcription*, Ph.D. thesis, Radboud University of Nijmegen, 2004.

[9] D. Temperley, "Bayesian models of musical structure and cognition," *Musicae Scientiae*, vol. 8, pp. 175–205, 2004.

[10] J.O. Smith R.J. Leistikow, H.D. Thornburg and J. Berger, "Bayesian identification of closely-spaced chords from single-frame stft peaks," in *DAFx*, Naples, Italy, October 2004.

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

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

[13] A. Sheh and D.P.W. Ellis, "Chord segmentation and recognition using em-trained hmm," in *ISMIR*, Baltimore, MD, 2003, pp. 183–189.

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

[15] C.A. Harte and M.B. Sandler, "Automatic chord identification using a quantised chromagram," in *AES 118th Convention*, Barcelona, Spain, 2005.

[16] G. Peeters, "Chroma-based estimation of tonality from audio-signal analysis," in *ISMIR*, Victoria, Canada 2006, pp. 115–120.

[17] E. Gomez, "Tonal description of polyphonic audio for music content processings," *INFORMS Journal on Computing*, vol. 18, no. 3, pp. 294–304, 2006.

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

[19] H. Papadopoulos and G. Peeters, "Large-scale study of chord estimation algorithms based on chroma representation and hmm," in *CBMI*, Bordeaux, France, 2007.

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