LARGE-SCALE STUDY OF CHORD ESTIMATION ALGORITHMS BASED ON CHROMA REPRESENTATION AND HMM

Hélène Papadopoulos and Geoffroy Peeters

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

ABSTRACT

This paper deals with the automatic estimation of chord progression over time of an audio file. From the audio signal, a set of chroma vectors representing the pitch content of the file over time is extracted. From these observations the chord progression is then estimated using hidden Markov models. Several methods are proposed that allow taking into account music theory, perception of key and presence of higher harmonics of pitch notes. The proposed methods are then compared to existing algorithms. A large-scale evaluation on 110 hand-labeled songs from the Beatles allows concluding on improvement over the state of the art.

1. INTRODUCTION

In Western tonal music, chord progression determines the harmonic structure of a piece of music. Analysis of chord progression therefore plays a crucial role in the understanding of this music. Music classification, music retrieval or in general all applications based on music content analysis benefit from the knowledge of chord progression.

However, manual labeling of each chord of a music piece is a difficult and tedious task, even for a well-trained person. Considering the recent availability of large music collections (online music catalogs, mobile devices) the automatic estimation of chord progression has become a major challenge.

In this paper, we present and compare several methods for estimating chord progression directly from acoustic signals of musical recordings. The methods are evaluated and compared to state of the art algorithms on a relatively large dataset (110 songs from the Beatles).

2. BACKGROUND

Since their introduction in 1999, Pitch Class Profiles (Fujishima [1]) or Chroma-based representation (Wakefield [2]) have become common features to automatically estimate chords or musical key from audio recordings ([3], [4], [5], [6], [7], [8]). PCP/chroma vectors represent the intensity of the twelve semitones of the pitch classes.

Fujishima [1] uses this representation to derive a large set of chords using either a nearest-neighbor or a weighted sum method. This system is successfully evaluated but only using synthetic sounds. The first system evaluated on natural sounds (whole pieces of music of commercial recordings) is the one by Sheh and Ellis [3]. Their system for chord segmentation/recognition relies on hidden Markov models (HMM) trained by EM algorithm. Another approach by Harte and Sandler [5] estimates chords using simple bit masks[1] compared to chroma features. Bello and Pickens [4] use an approach similar to [3] but introduce musical knowledge in the hidden Markov model. They show improvement over [3] using their system. It should be noticed that many other studies on chord estimation based on symbolic representation have been performed. In this paper we will refer to the recent studies made by [9].

The methods proposed in this paper start from the above-mentioned approaches. We systematically evaluate them and propose improvements to chord estimation systems. As most previous methods, the signal observations are the chroma features and the chord progression is represented using a hidden Markov model. Various ways of constructing the HMM are studied using either music theory, results from cognitive studies, smoothed training, multivariate Gaussian models or normalized-correlation.

Another major contribution of this paper is the use of harmonic extension of the PCP (Harmonic Pitch Class Profiles) in the case of chord estimation. Actually, most previous methods operate a direct mapping between the PCP/chroma values and the pitch of a note, i.e. a C note is represented by a single non-zero value in the chroma vector. In other words, the assumption is made that what we observe in the spectrum is directly the pitch of the notes. This is not true: each note produces a set of harmonics and thus a mixture of non-zero values in the chroma vector. Therefore, values at

[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.

pitch classes other than those of the notes will occur in the chroma vectors. For this reason, we propose to consider the presence of the harmonics in the model's parameters.

Harmonic Pitch Class Profiles (HPCP). To deal with this, in the case of key estimation, Gomez [8] proposes to take into account the harmonics of the notes using a theoretical spectral envelope. Izmirli measures the contribution of the harmonics on a piano database [10]. Peeters proposes in [6] the use of a Harmonic Peak Subtraction function which reduces the influence of the higher harmonics of each pitch.

In what follows, we rely on the model presented in [8]. This model extends the Pitch Class Profiles (PCPs) to the Harmonic Pitch Class Profiles (HPCPs). For this, a theoretical amplitude is attributed to each harmonic composing the spectrum of a note with an empirical decay factor set to 0.6 so that this contribution decreases with the frequency. The contribution for the first 6 harmonics of a note is given in Table 1. Therefore, higher harmonics contribute to the pitch class of their fundamental frequencies.

n	1	2	3	4	5	6
frequency	f	2.f	3.f	4.f	5.f	6.f
factor	1	s	s^2	s^3	s^4	s^5

Tab. 1. First 6 harmonics of a note and given amplitudes

Organization of the paper. The paper is organized as follows. In section 3.1 and 3.2, we detail the extraction of the chroma vectors from the audio signal. In section 3.3, we study several approaches to estimate the chords from the succession of chroma vectors over time using HMM and we describe in particular various configurations of the observation probabilities (section 3.3.2) and transition probabilities (section 3.3.3). In section 4, we evaluate our system and compare it to current existing systems.

3. SYSTEM

3.1. Pre-processings

3.1.1. Parameters

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 12.5% 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 music notes in popular music are usually stronger than the non-harmonic components up to $1kHz$ [11]. This choice is also supported by the fact that the mapping operated between the energy of the harmonics and the notes is only valid for the lowest harmonics, hence the lowest part of the spectrum.

3.1.2. Tuning

The energy peaks in the spectrogram will be mapped to the chroma vectors. It is therefore important that the peak frequencies 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 $A4 = 440Hz$, it is necessary to estimate the tuning of the track. Here, the tuning is estimated using the method proposed by Peeters in [7]. The signal is then re-sampled so that the rest of the system can be based on a tuning of $440Hz$.

3.2. Chromagram computation

The second stage of our analysis is the extraction of a sequence of observation vectors. The signal is converted from the frequency domain to the chroma domain. Chroma vectors are related to our perception of pitch [2]. They represent the intensity of the 12 semitone pitch classes over time. The temporal sequence of chroma vectors over time is known as chromagram. We compute the chromagram using the method proposed by Peeters in [6].

The chromagram is computed in three steps. First, the values of the DFT are mapped to a semitone pitch spectrum using the mapping function:

$$n(f_k) = 12\log_2(\frac{f_k}{440}) + 69, n \in \Re^+ \quad (1)$$

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 [2]. This provides a reduction of transients and noise in the signal. This smoothing allows us to significantly improve the overall results.

Finally, after this smoothing, the semitone pitches n are mapped to the the semitone pitch classes c within the mapping function:

$$c(n) = \mod(n, 12) \quad (2)$$

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

3.3. Chord estimation from the chroma vectors using hidden Markov models

We describe here several methods to estimate the chord progression over time of an audio signal. All these methods are

[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.

based on hidden Markov models (HMMs) [12]. The various methods differ in the way observation probabilities and transition probabilities are computed.

We consider an ergodic 24-states HMM, each state representing a single chord. Our chord lexicon is composed of 24 Major and minor triads (C Major, C# Major, ..., B Major, C minor, ..., B minor). Each state in the model generates an observation vector, the chroma feature, with some probability. This is defined by the **observation probabilities**. In part 3.3.2, we study three approaches to define these probabilities. The first one (Method 1) learns these probabilities by training a Gaussian model on chord-normalized chroma vectors. The second one (Method 2) does not use the training set but defines probabilities based only on music theory, considering the presence of higher harmonics (using the HPCPs). The third one (Method 3) is close to Method 2 but defines probabilities based on a normalized-correlation measure rather than a Gaussian model.

In music pieces, the transitions between chords result from musical rules that should be reflected in the **state transition matrix**. This is in fact the main reason why the problem is modeled using a Markov model. In part 3.3.3, we study four approaches to define the transition matrix. Method A is based on music theory: the closeness of chords in the doubly-nested circle of fifths. Method B uses the results of cognitive experiments: the closeness of chords using Krumhansl's key profiles. Method C learns the transitions probabilities from the HMM training. We finally propose a new method, D, which learns the transitions from score transcriptions.

In what follows, we denote by π and T, the initial state distribution and state transition probability distribution.

Given the observations, we estimate the most likely chord sequence over time in a maximum likelihood sense.

3.3.1. Initial state distribution

Since we do not know *a priori* which chord the piece begins with, we initialize π at $\frac{1}{24}$ for each of the 24 states. This choice has also been made in [4].

3.3.2. Observation symbol probability distribution

Method 1: Modeling by a multivariate Gaussian trained on a labeled dataset

With this method, the observation distribution is modeled by 24 (one for each state) single 12-dimensional multivariate Gaussian distributions defined by their mean vectors μ_i and covariance matrices Σ_i, with i denoting the i^{th} state.

In [3], the model is trained using the standard expectation maximization (EM) algorithm for HMM parameters estimation. The parameters μ and Σ are initialized with random values. According to [4], on the one hand, the template for a chord is almost universal and should not change from song to song. On the other hand, it is unlikely that every chord of the lexicon will be present in the training dataset. That is why it is proposed to selectively train the model, disallowing adjustments of μ and Σ while π and T are updated. We also believe that any reasonably sized training set will be insufficient to appropriately estimate the model's parameters. Indeed, since the number of observations in the dataset will likely differ among the 24 possible chords, training directly the model on the dataset may lead to overfit the model to a specific type of music (that means learning the characteristics of the dataset).

In order to learn the observation distribution for each of the 24 possible chords, we propose to first learn the model for the C Major chord and the C minor chord and then map the two trained models to all possible chords by circular permutation. A similar approach was proposed in [7] in the case of key estimation. We proceed as follows:

1. All the chroma vectors of the labeled training dataset are mapped to a root-note of C using circular permutation.
2. The mean vector and the covariance matrix for the C Major (C minor) chord are computed from all C Major (C minor) chroma vectors.
3. The mean vectors and covariance matrices for all chords are obtained from the two trained models by circular permutation.

The mean vectors for the C Major and C minor chords trained on the dataset presented in 4.1 are represented in the left part of Figure 1. Note that in this case we do not make any assumption on the signal (instrumentation, harmonics, etc.) and we do not introduce any musical knowledge. In what follows, we will call this method "Method 1".

Method 2: Modeling by a multivariate Gaussian based on music theory considering the presence of higher harmonics

In this case, the observation distribution does not rely on any training on a given dataset. As in [4], the observation distribution relies directly on music theory; however a major difference with [4] is that we consider the presence of the higher harmonics of the theoretical notes in the construction of the multivariate Gaussian models (by modifying the parameters μ and Σ). This consideration allows us to significantly improve the results over the method proposed in [4].

In [4], the mean vectors and covariance matrices reflect musical knowledge. The mean vectors are 12-dimensional vectors with 1 if the note belongs to the chord and 0 otherwise. For instance, for a C Major chord (C-E-G), the mean

vector will be 100010010000 (see middle-left part of Figure 1). In the covariance matrices, pitch which comprise the triad are more correlated than pitch which do not belong to the triad. The covariance between pitches which comprise the triad is thus given a non-zero value. The value is attributed with respect to music theory and empirical evidence from Krumhansl work [13], that is to say that the dominant (fifth degree) is more important than the mediant (third degree) in characterizing the root of a triad [3].

In this paper, we propose to take into account the contribution of the higher harmonics of the theoretical notes into the Gaussian parameters. We do this in the following way.
Mean vectors: For each note of a chord, we add the contribution of the harmonics in the mean vectors. The amplitude contribution of the h^{th} harmonic of a note is similar to the one proposed by [8]: 0.6^{h-1}. Table 2 indicates the considered harmonics and the corresponding amplitudes for the C Major and the C minor templates. We represent the corresponding mean vectors for C Major and minor (in the case of 4 harmonics) in the middle-right part of Figure 1.

C major (Cminor)						
note	harmonics					
C	C	C	G	C	E	G
E (Eb)	E (Eb)	E (Eb)	B (Bb)	E (Eb)	G# (G)	B (Bb)
G	G	G	D	G	B	D
amplitude	1	0.6	0.62	0.63	0.64	0.65

Tab. 2. The first 6 harmonics and their amplitude for a C Major (C minor) triad

Covariance matrices: [4] only considers the correlation between the chroma vectors corresponding to the pitch of the notes belonging to a given chord. In our method, we also consider the correlation between the harmonics of each note. For example, for a C Major chord (C-E-G), D is the 3^{rd} harmonic of G. Hence, we attribute a non-zero value to the covariance between D and G. As in [4], the values we use are heuristic but we still respect the rule that the dominant is more important than the mediant in characterizing the root of a triad [4]. The covariance matrices we propose for a C Major and a C minor chord are represented in Figure 2 above the covariance matrices proposed in [4]. In what follows, we will call this method "Method 2".

[3] In [4], the covariance of the tonic with the dominant and of the dominant with the mediant is set to 0.8. The covariance of the tonic with the mediant is set to 0.6. Since we both use songs from the Beatles to evaluate our system, we will use the same values when testing method [4].

[4] The covariance of the tonic with the dominant is set to 0.6; the covariance of the dominant with the mediant is set to 0.5; the covariance of the tonic with the mediant is set to 0.3; the covariance of a note with its second harmonic is set to 0.1; the other non-zero values are set to 0.05. The matrix needs to be positive, semi-definite, so we set the non-triad diagonal members to 0.1.

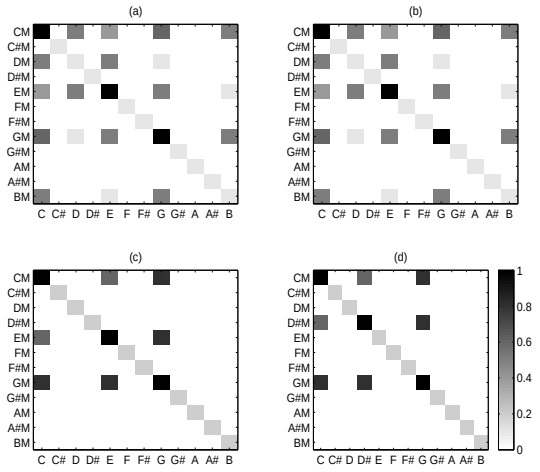

Fig. 2. Covariance matrices for a C Major and a C minor chord considering the presence of 4 harmonics (upper part, (a) and (b)) and proposed covariance matrices in [4] (bottom part, (c) and (d))

Method 3: Probability derived from correlation with chord templates

In this method, the observation probabilities are not modeled by a multivariate Gaussian distribution. They are obtained by computing the correlation between the observation vectors and a set of chord templates.

Chord templates: The chord templates are the theoretical chroma vectors corresponding to the 24 Major and minor triads. A chord template is a 12-dimensional vector which contains the theoretical amplitude values of the notes and their harmonics composing a chord. 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 in the case of the mean vectors in Method 2, we attribute an amplitude of 0.6^{h-1} to the harmonic h. In section 4, we will compare the system results without considering any harmonic ($nbh = 1$), with 4 harmonics ($nbh = 4$) and with 6 harmonics ($nbh = 6$). In the right part of Figure 1, we represent the chord templates of C Major and C minor when considering 6 harmonics.

Observation probabilities: For each chroma vector, we compute the correlation between the vector and the 24 chroma templates. We obtain 24 values $P(c_i)$, $i \in [1, 24]$, normalized so that $\sum_i P(c_i) = 1$. We now have 24 "pseudo-probabilities" which are used as observation probabilities in the HMM. In what follows, we will call this method "Method 3".

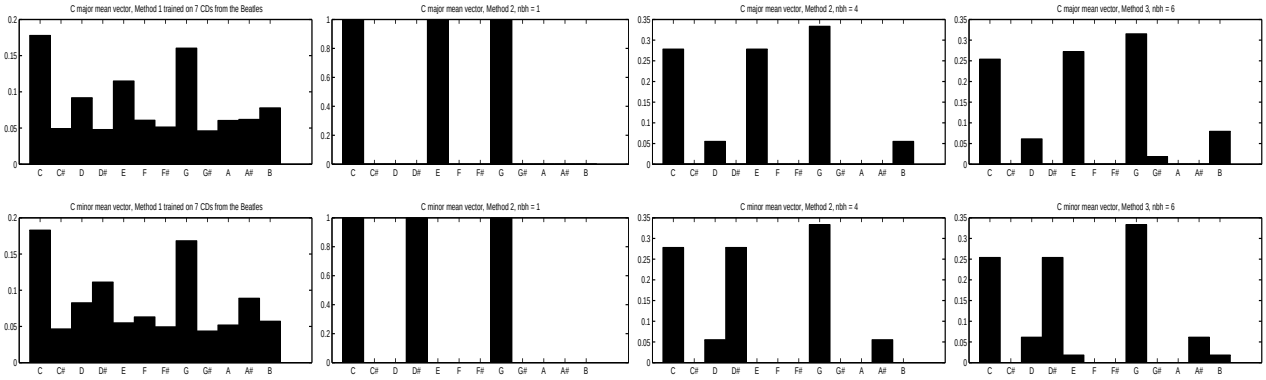

Fig. 1. Mean chroma vectors for the C Major (upper par of each figure) and C minor (lower part) chords using [from left to right]: Method 1 (trained using 7 CDs of the Beatles), Method 2 without harmonic contribution, Method 2 with 4 harmonics contribution, Method 3 with 6 harmonics contribution (in this case, the figures represent the chroma templates instead of the mean vectors).

3.3.3. State transition probability distribution

Method A: Theoretical approach using the doubly-nested circle of fifths

This method has been proposed by [4]. The transition probability between two chords is derived from musical knowledge: their distance in the doubly-nested circle of fifths (see Figure 3). The doubly-nested circle of fifths depicts relationships among the 12 equal-tempered pitch classes comprising the chromatic scale. Although we do not know which state is going to follow another one, musical rules allow us to make hypotheses that are more probable than others. For instance, especially in popular western music, an A major chord is more likely to be followed by a F# minor or D major chord than by a G# Major chord. The corresponding state transition matrix is represented in the left part of Figure 4.

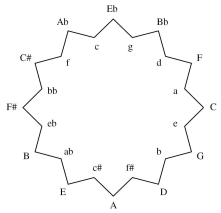

Fig. 3. Doubly-nested circle of fifths. From [4]

Method B: Cognitive approach using correlation between key profiles

Krumhansl studies the proximity between the various musical keys using correlations between key profiles obtained from perceptual tests. These correlations have been used by [9] to derive a key transition matrix in the context of local key estimation. In our experiments, we have achieved good results for chord estimation using the key transition matrix from [9] as a chord transition matrix (see middle part of Figure 4).

Method C: Trained approach using the EM algorithm

This approach uses the transition matrix provided by the training of the HMM using the Baum-Welsh (EM) algorithm, i.e. the system is trained using on the one side the succession of chroma vectors extracted from the audio signal and on the other side the corresponding chord labels. Although this approach is the usual one and the one used for example in [3] [4], it did not provide satisfactory results in our evaluation.

Method D: Trained approach using the chord transcription

As opposed to the previous method, this approach is only based on symbolic information, i.e. the chord labels transcription of the training set. From the succession of transcribed chord labels over time, we derive an "annotation" transition matrix which is, as in the previous case, specific to the training set (in our case the Beatles corpus).

We want to learn from the training set the probabilities of transiting from one chord to another. We achieve this by counting the number of occurrences of each chord transition in the training set. Our goal is to construct a 24-dimensional matrix T that indexes all the chord transitions. However, be-

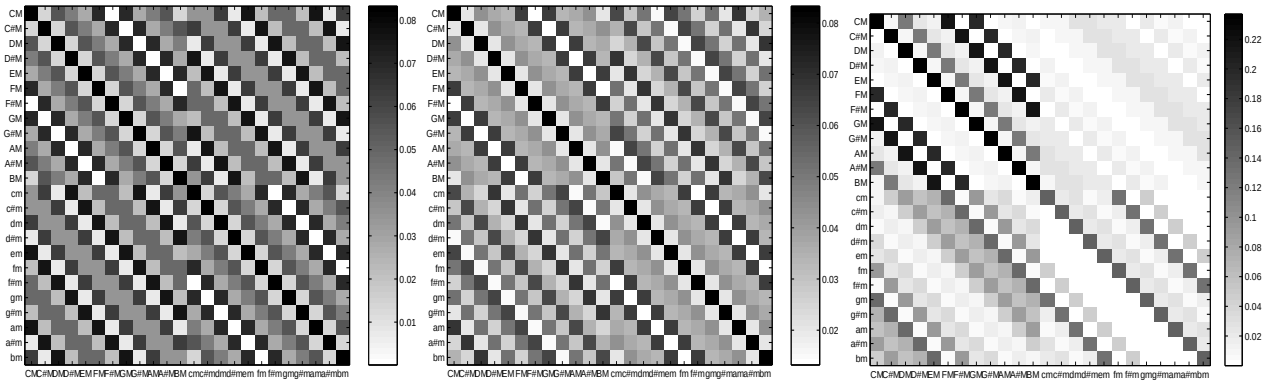

Fig. 4. State transition probability distribution between the 12 Major and minor chords using [from left to right]: method A, method B, and method D

cause the distribution of musical keys is not homogeneous in the training set, we are likely to favor specific chord transitions[5], and therefore the transition matrix will be imbalanced. In order to face this problem, we only consider **relative chord transitions** (GM $\rightarrow$ CM transition is considered equivalent to CM $\rightarrow$ FM). We denote by $T(i,j)$ the value of the transition matrix that represents the probability of transiting between chord i at time $t-1$ to chord j at time t. The indexes $i,j \in [1,12]$ represent the Major (M) chords, $i,j \in [13,24]$ the minor (m) chords. The matrix is therefore composed by four sub-matrices which represent transitions between M to M, m to m, M to m and m to M chords. These four cases are processed separately.

1. We first select from the training set all chord transitions belonging to a specific case (MM, mm, Mm, mM).
2. For each chord belonging to a given subset, we then compute the relative chord transitions. Each chord transition $i \rightarrow j$ is characterized by the equivalent transition from/to a root-note of C. We denote it by $f(i,j)$.
3. We then form a 12-dimensional vector $\tau(k)$ by counting the number of relative chord transitions $f(i,j) = k$.
4. Using these vectors, we form the $T(1, k \in [1,12])$ (MM), $T(13, k \in [13,24])$ (mm), $T(1, k \in [13,24])$ (Mm), $T(13, k \in [1,12])$ (mM).
5. The diagonal of the sub-matrices (self-transition) is processed in a separate way. We set the diagonal values to $1.1 \max(\tau(k))$
6. The rest of the sub-matrices, are constructed by circular permutation.
7. We finally normalize the matrix T so that the sum of each row is equal to 1.

The resulting matrix trained on the dataset presented in section 4.1 is represented in the right part of Figure 4. It is interesting to observe the predominance (high transition values in the matrix) of typical transitions in the matrix, such as the II/V/I (transition between dm, GM and CM) which seems usual in this set of Beatles albums. However, the amount of transitions between Major and minor chords is much lower than the amount of transitions between two Major chords in this training set. The consequence of that, is a lower recognition rate for tracks with Major to minor chords.

3.3.4. Chord progression over time detection

In all cases (Method 1, 2, 3, A, B, C or D), the optimal succession of chords over time is found using the Viterbi decoding algorithm [14] which gives us the most likely path trough the HMM states given our sequence of chroma observations.

4. EVALUATION AND RESULTS

4.1. Test set and protocol

The system has been tested on a set of 110 hand-labeled files from the first eight albums of the Beatles. We have used this dataset since it allows a direct comparison to other publications. The chord annotations where kindly provided by C. Harte [6]. All the recordings are polyphonic, multi-

[5]For instance, if 90% of the training set is in C Major we are more likely to observe a II/V/I transition in C Major, i.e. dM/GM/CM, than a II/V/I transition in F#M, i.e. g#m/D#M/F#M.

[6]www.elec.qmul.ac.uk/digitalmusic/

instrumental songs containing drums and vocal parts. To our knowledge, it is the first attempt to evaluate a chord detection system on such a large dataset.

Since our chord lexicon only represents Major an 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 triad. This point is important when analyzing the results.

4.2. Results

The results are indicated in Table 3.

• *Res* row corresponds to the exact recognition rate on all the frames (approximately 135.000 with the chosen parameters).

• *Rct* row represents the "close triads" recognition rate as discussed below.

In this table, we compare the various methods for observation distribution and the choice of parameters:

• (Method 1) Gaussian observation distribution with training. For this method, the evaluation has been performed using a 8-folds cross-validation (each album was evaluated using the seven remaining albums as training data).

• (Method 2, $nbh = 1$) Gaussian observation distribution with music theory as proposed in [4].

• (Method 2, $nbh = 4$) Our proposal: Gaussian observation distribution with music theory considering the presence of four higher harmonics .

• (Method 3, $nbh = 1, 4, 6$) Our proposal: Observation distribution from correlation with templates combined with music theory considering the presence of one, four or six higher harmonics.

Note that we only present here the results obtained using method B for the transition matrix (see explanation below).

	Method1	Method2		Method3		
	training	nbh=1	nbh = 4	nbh = 1	nbh = 4	nbh = 6
Res	69.95±14.90	61.57±14.72	69.28±114.2	67.54±13.54	70.24±17.01	70.96±19.23
Rct	84.08±9.87	74.67±10.47	81.82±9.91	81.22±9.64	82.57±10.49	86.18±8.67

Tab. 3. Chord recognition rate using method 1, 2 and 3 for the observation distribution and method B (theoretical transition matrix based on correlation between key profiles) for the transition matrix. Res: exact chord recognition rate. Rcl: chord recognition rate including close triads. nbh: number of harmonics considered in the model

4.3. Analysis of results

Chord estimation method: The results obtained with the various methods are pretty close to each other. In our experiments, the best results were obtained with Method 3 (70.96%). Note that there was no training of the observation distribution in this case. Despite the fact that Method 1 uses training (and is therefore likely to fit very well to the characteristics of the Beatles), Method 2 with $nbh = 4$ (which does not use training at all) gives as high results[7].

Transition matrix: The best results were obtained using the theoretical transition matrix based on correlation between key profiles (Method B). The transition matrix based on the doubly-nested circle of fifths (Method A) gives slightly lower results. We do not present the results obtained with the two trained matrices (Methods C and D) because they are much lower for reasons explained in section 3.3.3 (overfitting to the dataset).

Number of harmonics: Considering the presence of higher harmonics in the creation of the parameters clearly improves the results. For instance, for Method 3, considering 6 harmonics in the templates brings about 5% relative improvement. This is even clearer in the case of Method 2 where considering harmonics in the model's parameters brings about 12.5% relative improvement.

4.4. Discussion

Chords confusion due to ambiguous mapping: As it can be seen in Table 3, the standard deviation of the results is relatively high (up to 19%) independently from the chosen method. A deeper analysis of the results would show that the number of bad recognition comes from a reduced set of songs with partial or complex (non-triads) chords. For instance, for *Love You To* from *Revolver*, we obtain less than 3% of chords correctly identified. Provided annotation indicates that almost all the chords of this song but a few are Cmin(*b3) chords, i.e. a triad without the third note (C-G). In such case, it is difficult to make a decision between Major and minor chords in the absence of musical key information. Our system in fact has recognized for all cases Major chords which makes the recognition rate decrease. In future works, in order to avoid that, we plan to include information about the tonal context as it has been proposed in [15] or [11].

As mentioned before, because of our limited chord dictionary, a mapping was performed between complex chords and their root triad. Chords of four notes often contain other triads than their root triad. For instance, a Cmin7 (C-Eb-G-Bb) contains C minor (C-Eb-G) chord and Eb Major (Eb-G-Bb) chord. Whereas the majority of songs in the evaluation dataset are composed of triad chords, some of them contain a lot of more complex chords, and it sometimes happens that the system recognizes other triad than the root triad of the complex chord analyzed, which makes also the recognition rate decrease.

[7] It should be noted however that we did not recover the high results given in [4] with Method 1 and $nbh = 1$ which is very close to the one presented in [4] and tested on the same dataset.

Neighboring triad confusion: Most of the errors that occur correspond to harmonically close triad confusions: - parallel Major/ minor chords (EM being confused with em), - relative chords (am being confused with CM), - dominant chords (CM being confused with GM) or - subdominant chords (CM being confused with FM). 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. Table 3 shows that if we consider close triads recognition as correct, the recognition rate of method 3 reaches up to 86%. It also becomes now the method with the smallest standard deviation, 9%.

Limitation of 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 last three (from the "Norwegian Wood (This Bird Has Flown)" on 1965's Rubber Soul). The reason for this, comes from the extended use of the Indian sitar instrument[8] which produces highly inharmonic sounds. Since the chroma-based approach strongly relies on the presence of harmonic sounds it is not appropriate to use it for such music.

5. CONCLUSION AND FUTURE WORKS

In this paper we have proposed and compared several methods for the automatic estimation of chord progression of a music piece. All the methods are based on chroma representation of the audio signal and on modeling of the sequence of observation using a hidden Markov model. The methods have been compared on a large-scale evaluation. The best results were obtained with the modeling of the observation probabilities using a normalized correlation with a set of extended chord templates and a cognitive-based transition matrix. The templates are extended by considering the presence of higher harmonics of each pitch note of a chord. The transition matrix is derived from cognitive experiments on the perception of musical key. Current limitations of our system mainly come from the confusion between the various interpretation one can make about chords. A solution to that is to integrate extra (context) information such as musical key information. The integration of metrical information could also increase the robustness of the system.

6. ACKNOWLEDGEMENTS

Part of this work was conducted in the context of the French RIAM project "Ecoute".

[8] Sitar instrument is a stringed instrument that uses sympathetic strings along with regular strings. This produces very lush sound with inharmonic components.

7. REFERENCES

[1] Takuya Fujishima. Real-time chord recognition of musical sound: A system using common lisp music. In *ICMC*, pages 464–467, Beijing, China, 1999.

[2] Gregory H. Wakefield. Mathematical representation of joint time-chroma distribution. In *SPIE*, volume 3807, July Denver, Colorado, 1999.

[3] Alexander Sheh and Daniel P.W. Ellis. Chord segmentation and recognition using em-trained hmm. In *ISMIR*, pages 183–189, Baltimore, MD, 2003.

[4] Juan P. Bello and Jeremy Pickens. A robust mid-level representation for harmonic content in music signal. In *ISMIR*, pages 304–311, London, UK, 2005.

[5] Christopher A. Harte and Mark B. Sandler. Automatic chord identification using a quantised chromagram. In *AES 118th Convention*, Barcelona, Spain, 2005.

[6] Geoffroy Peeters. Chroma-based estimation of tonality from audio-signal analysis. In *ISMIR*, pages 115–120, Victoria, Canada 2006.

[7] Geoffroy Peeters. Musical key estimation of audio signal based on hmm modeling of chroma vectors. In *In DAFX, McGill*, pages 127–131, Montreal, Canada, 2006.

[8] Emilia Gomez. Tonal description of polyphonic audio for music content processings. *INFORMS Journal on Computing*, 18(3), 2006.

[9] Katy Noland and Mark Sandler. Key estimation using a hidden markov model. In *ISMIR*, pages 121–126, Victoria, Canada, 2006.

[10] Özgür Izmirli. Template based key finding from audio. In *ICMC*, pages 211–214, Barcelona, Spain, 2005.

[11] Namunu C.Maddage. Automatic Structure Detection for Popular Music. *IEEE MultiMedia*, 13(1):65–77, 2006.

[12] L. Rabiner. A tutorial on hidden markov model and selected applications in speech. *IEEE*, 77(2):257–285, 1989.

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

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

[15] Arun Shenoy and Ye Wang. Key, Chord and Rhythm Tracking of Popular Music Recordings. *Computer Music Journal*, 3(29):75–86, 2005.