Probabilistic Framework for Template-Based Chord Recognition

Laurent Oudre [#], Cédric Févotte [*], Yves Grenier [#]

[#] *Institut TELECOM ; TELECOM ParisTech ; CNRS LTCI*
[*] *CNRS LTCI ; TELECOM ParisTech*
37-39 rue Dareau, 75014 Paris, FRANCE
{oudre,fevotte,grenier}@telecom-paristech.fr

***Abstract*—This paper describes a method for chord recognition from audio signals. Our method provides a coherent and relevant probabilistic framework for template-based transcription. The only information needed for the transcription is the definition of the chords : in particular neither annotated audio data nor music theory knowledge is required. We extract from the signal a succession of chroma vectors which are our model observations. We propose a generative model for these observations from chord distribution probabilities and fixed chord templates. The parameters are evaluated through an EM algorithm. In order to capture the temporal structure, we apply some post-processing filtering methods before detecting the chords.**

Our method is evaluated on two audio corpus. Results show that our method outperforms state-of-the-art chord recognition methods and also gives more relevant chord transcriptions.

I. Introduction

The description of musical signals with relevant and compact representations has been one of the main fields of interest in Musical Information Retrieval (MIR) for the last past years. One of the most common representations used for pop songs is the chord transcription, which consists in a sequence of chord labels with their respective lengths. This representation can be used in several applications such as song identification, query by similarity or structure analysis.

The features used in chord transcription may differ from a method to another but are in most cases variants of the *Pitch Class Profiles* introduced by Fujishima [1]. These features, also called *chroma vectors*, are 12-dimensional vectors. Every component represents the spectral energy of a semi-tone on the chromatic scale regardless of the octave on either a fixed-length or a beat-synchronous frame. The succession of these chroma vectors over time is called *chromagram*.

The chord recognition methods can be divided into 4 main categories : template-based, music-driven, data-driven and hybrid (combining the music- and data-driven approaches).

Template-based chord recognition methods are based on the hypothesis that only the definitions of the chords are needed in order to extract the chord labels from the musical piece. A chord template is a 12-dimensional vector representing the 12 semi-tone (or *chroma*) of the chromatic scale. Each component of the pattern is the theoretical amplitude of the chroma within the chord.

These templates can easily be defined for every chord : the detected chord on one given frame is the one whose template fits the best the chroma vector calculated for the frame. Many matching methods have been used : Euclidean distance [1], [2], [3], dot product [1], [4], correlation [5], Kullback-Leibler and Itakura-Saito divergences [2], [3]... The temporal structure is taken into account with post-processing filtering applied either on the chromagram [1], [4], [5] or on the calculated fit measures [2], [3].

These template-based methods often have difficulties to capture the long-term variations of the chord sequences, as well as giving harmonically-coherent sequences of chords. Complex probabilistic methods have been build in order to incorporate musical information such as key, chord transitions models, beats, structure, etc. This higher level information is either extracted from music theory [6], [7], [8], from training with audio data [9], [10], [11], [12] or combining these two approaches [13], [14], [15].

The method presented in this paper builds on the template-based method described in [2], [3] but gives a probabilistic framework by modeling the chord distribution probabilities of the song. Our model explicitly infers the probability of appearance of every chord in a given song. The introduction of this probabilistic approach allows to implicitly take into account the harmony and to extract for every song a relevant chord *vocabulary*[1]. Contrary to other probabilistic chord recognition methods, our method can still be classified within the template-based methods, since the only needed information is the definition of the chord templates.

Section II describes our system by introducing the notion of chord template, our probabilistic model and the algorithm used for the chord recognition. Section III presents the corpus used for evaluation and a qualitative and quantitative comparison of our method with the state-of-the-art.

[1]We mean by vocabulary a subset of the user-defined chord dictionary, which is expected to be representative of the harmonic content of the song.

MMSP'10, October 4-6, 2010, Saint-Malo, France.
978-1-4244-8112-5/10/$26.00 ©2010 IEEE.

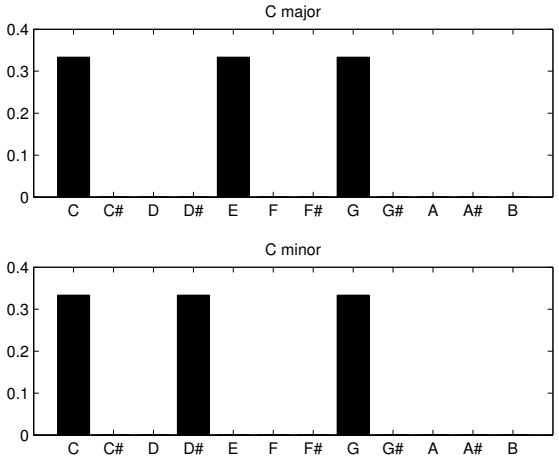

Fig. 1. Chord templates for C major and C minor.

II. System

A. Chord templates

Our chord templates are simple binary masks : an amplitude of 1 is given to the chromas present in the chord and an amplitude of 0 is given to the other chromas.[2] For example a *C major* chord will be given an amplitude of 1 to the chromas *C*, *E* and *G* while the other chromas will have an amplitude of 0. By convention in our system, the chord templates are normalized so that the sum of the amplitudes is 1 but any other normalization could be employed. Examples for C major and C minor chord are presented on Figure 1.

Extensive works have been done with other types of chord templates by taking into account higher harmonics for the notes of the chord [16], [14] but they did not seem to significantly improve the results for our system. We therefore decided to use only binary chord templates in this paper.

B. Generative model for the chroma vectors

Let $\mathbf{C}$ be a $12 \times N$ chromagram, composed of N M-dimensional (in practice M=12) successive chroma vectors $\mathbf{c}_n$. The chroma vectors are calculated from the music signal with the same method as Bello & Pickens [6], where the frame length is set to 753 ms and the hop size is set to 93 ms. We use the code kindly provided by the authors. Let $\mathbf{W}$ be our $12 \times K$ chord dictionary, composed of K M-dimensional chord templates $\mathbf{w}_k$.

Let us make the assumption that on frame n, the present chord γ_n is the one verifying :

$$\mathbf{c}_n \approx h_{\gamma_n,n}\mathbf{w}_{\gamma_n} \tag{1}$$

where $h_{\gamma_n,n}$ is a scale parameter.

The likelihood $p\left(\mathbf{c}_n|h_{k,n},\mathbf{w}_k\right)$ therefore describes the noise corrupting $h_{k,n}\mathbf{w}_k$ in the observation $\mathbf{c}_n$. Let us assume a Gamma multiplicative noise ϵ, that is to say :

$$\mathbf{c}_n = (h_{k,n}\mathbf{w}_k)\cdot\epsilon \tag{2}$$

Then, the observation model becomes :

$$p\left(\mathbf{c}_n|h_{k,n},\mathbf{w}_k\right) = \prod_{m=1}^{M}\frac{1}{h_{k,n}w_{m,k}}\mathcal{G}\left(\frac{c_{m,n}}{h_{k,n}w_{m,k}}\,;\,\beta,\beta\right) \tag{3}$$

where $\mathcal{G}$ is the Gamma distribution defined as :

$$\mathcal{G}\left(x;a,b\right) = \frac{b^a}{\Gamma\left(a\right)}e^{-bx} \tag{4}$$

and Γ is the Gamma function.

Let us denote $\gamma_n \in [1,\ldots,K]$ the discrete random state indicating the chord present on frame n and α_k the probability for the chord k to appear in the song. We consider the following state-model

$$\left\{\begin{array}{rcl} p\left(\mathbf{c}_n|\boldsymbol{\alpha},\mathbf{h}_n,\gamma_n=k\right) &=& p\left(\mathbf{c}_n|h_{k,n},\mathbf{w}_k\right) \\ p\left(\gamma_n=k\right) &=& \alpha_k \end{array}\right. , \tag{5}$$

which can equivalently be written as the following mixture model

$$p\left(\mathbf{c}_n|\boldsymbol{\alpha},\mathbf{h}_n\right) = \sum_{k=1}^{K}\alpha_k\, p\left(\mathbf{c}_n|h_{k,n},\mathbf{w}_k\right). \tag{6}$$

Under this model, a chromagram frame is in essence assumed to be generated by 1) randomly choosing chord k (with template $\mathbf{w}_k$) with probability α_k, 2) scaling $\mathbf{w}_k$ with parameter $h_{k,n}$ (to account for amplitude variations), and 3) generating $\mathbf{c}_n$ according to the assumed noise model and $h_{k,n}\mathbf{w}_k$.

Given parameters $\boldsymbol{\alpha} = [\alpha_1,\ldots,\alpha_K]$ and $\mathbf{H} = \{h_{k,n}\}_{kn}$, we choose for frame n the chord with highest state posterior probability :

$$\hat{\gamma}_n = \underset{k}{\operatorname{argmax}}\ \lambda_{k,n}^{post} \tag{7}$$

where $\lambda_{k,n}^{post} = p\left(\gamma_n = k|\mathbf{c}_n,\boldsymbol{\alpha},\mathbf{h}_n\right)$.

C. EM algorithm

Let us summarize the notations :

- $\mathbf{C} = [\mathbf{c}_1,\ldots,\mathbf{c}_N]$ is the $M \times N$ matrix containing the chromagram observations,
- $\boldsymbol{\Theta} = (\boldsymbol{\alpha},\mathbf{H})$ is the set of parameters,
- $\boldsymbol{\gamma} = [\gamma_1,\ldots,\gamma_N]$ is the vector of dimension N containing the chord state variables.

The log-likelihood $\log p\left(\mathbf{C}|\boldsymbol{\Theta}\right)$ can typically be maximized using an EM algorithm based on missing data $\boldsymbol{\gamma}$, where the following functional needs to be iteratively computed (E-step) and maximized (M-step) :

$$Q\left(\boldsymbol{\Theta}|\boldsymbol{\Theta}'\right) = \sum_{\boldsymbol{\gamma}}\log p\left(\mathbf{C},\boldsymbol{\gamma}|\boldsymbol{\Theta}\right)p\left(\boldsymbol{\gamma}|\mathbf{C},\boldsymbol{\Theta}'\right) \tag{8}$$

For sake of conciseness, calculations are not displayed on this paper. One of the main results of the algorithm derivation is that the parameter $\mathbf{H}$ does not need to be updated during the

[2] In practice a small value is used instead of 0, to avoid numerical instabilities that may arise.

EM iterations and can therefore be precomputed by calculating $h_{k,n}$ such as :

$$\left[\frac{d\ \log p\left(\mathbf{c}_n|h,\mathbf{w}_k\right)}{dh}\right]_{h=h_{k,n}} = 0, \tag{9}$$

The resulting EM algorithm is summarized below.

Algorithm 1: EM algorithm for probabilistic template-based chord recognition

Input: chromagram $\mathbf{C} = [\mathbf{c}_1, \ldots, \mathbf{c}_N]$, chord templates $\mathbf{W} = [\mathbf{w}_1, \ldots, \mathbf{w}_K]$ and $h_{k,n}$ such that $\frac{d \log p(\mathbf{c}_n|h,\mathbf{w}_k)}{dh} = 0$

Output: a posteriori probability $\lambda_{k,n}^{post}$ and $\boldsymbol{\alpha} = [\alpha_1, \ldots, \alpha_K]$

Initialise $\boldsymbol{\alpha}$

for $i = 1 : n_{iter}$ **do**

$$\left[\lambda_{k,n}^{post}\right]^{(i-1)} = \frac{p(\mathbf{c}_n|h_{k,n},\mathbf{w}_k)\ \alpha_k^{(i-1)}}{\sum_{k'=1}^{K} p\left(\mathbf{c}_n|h_{k',n},\mathbf{w}_{k'}\right)\ \alpha_{k'}^{(i-1)}} \quad \texttt{// E-Step}$$

$$\alpha_k^{(i)} = \frac{\sum_{n=1}^{N}\left[\lambda_{k,n}^{post}\right]^{(i-1)}}{\sum_{k'=1}^{K}\sum_{n=1}^{N}\left[\lambda_{k',n}^{post}\right]^{(i-1)}} \quad \texttt{// M-Step}$$

end

D. Chord recognition with the probabilistic model

The matrix $\lambda_{k,n}^{post}$ represents the state posterior probabilities of every chord k of the dictionary for every frame n. Let us assume that the matrix has been calculated with the algorithm previously presented. As seen in II-B, the detected chord $\hat{\gamma}_n$ for frame n is finally :

$$\hat{\gamma}_n = \underset{k}{\operatorname{argmax}}\ \lambda_{k,n}^{post}. \tag{10}$$

This frame-to-frame chord recognition system can be improved by taking into account the temporal context. Most of the methods using HMM assume an exponentially temporal distribution, which does not suit well the rhythmic structure of pop songs. We therefore propose to use a low-pass filtering process as an *ad hoc* processing which implicitly inform the system of the appropriate durations of the expected chords. The post-processing filtering method is applied to $\lambda_{k,n}^{post}$ in order to take into account the time persistence.

III. Evaluation and results

A. Corpus and evaluation

The evaluation method used in this paper corresponds to the one used in MIREX 09 for the Audio Chord Detection task[3].

Our evaluation database is constituted by 2 corpus :

- The Beatles corpus constituted by the 13 Beatles albums (180 songs, PCM 44100 Hz, 16 bits, mono). The evaluation is realized thanks to the chord annotations of the 13 Beatles albums kindly provided by Harte and Sandler [17]. In these annotation files, 17 types of chords and one 'no chord' label corresponding to silences or untuned material are present. The 17 types of chords present in the annotation files are mapped to the major and the minor according rules already used by MIREX.
- The second corpus has been provided by the QUAERO project. It contains 20 songs (PCM 22050 Hz, 16 bits, mono) from various artists (Pink Floyd, Queen, Buenavista Social Club, Justin Timberlake, Mariah Carey, Abba, Cher, etc.) and various genres (pop, rock, electro, salsa, disco,...). This corpus only contains major and minor labels.

For each song an *Overlap Score* is calculated as the ratio between the sum of the lengths of the well detected chords and the total length of the song. The mean of the Overlap Scores over all the songs of the corpus is then called *Average Overlap Score (AOS)*.

B. Results

For our new probabilistic method, two parameters are to be set : the probability distribution parameters (β) and the post-processing filtering parameters. Extensive simulations have been done in order to find good probability distribution parameters. These parameters are chosen in order to fit the model to the chord recognition task, i.e. to model the type of noise present in audio signals. The post-processing filtering neighborhood sizes used here are chosen in order to optimize the value of the Average Overlap Score on the Beatles corpus. Nevertheless, there are not much differences between close neighborhood sizes. The experimental parameters used for our probabilistic methods are $\beta = 3$ and low-pass filtering on 15 frames.

Figure 2 presents an example of the results obtained with two chord transcription methods on one Beatles song. The new method is compared to the baseline method described in [2], [3] detecting only major and minor chords, which was submitted to MIREX 2009 as *OGF1*. The estimated chord labels are in black while the ground-truth chord annotation is in gray. The first observation is that the probabilistic transcription seems here to give better quantitative results. The transcription also seems to be more musically and temporally relevant. These good results can be explained by at least two hypothesis :

- The probabilistic transcription seems to detect longer chords while the baseline method give very segmented results.
- The chord vocabulary used by the probabilistic transcription is sparser than the one used in the baseline method, preventing the detection of off-key chords.

These good results are confirmed by the AOS calculated on the two corpus. Table I presents the scores for our new method, and several state-of-the art methods that entered MIREX in 2008 or 2009. All algorithms are tested with the author's implementations. The following methods are tested :

MIREX 2008 :

- BP : Bello & Pickens [6]

[3]http://www.music-ir.org/mirex/2009/

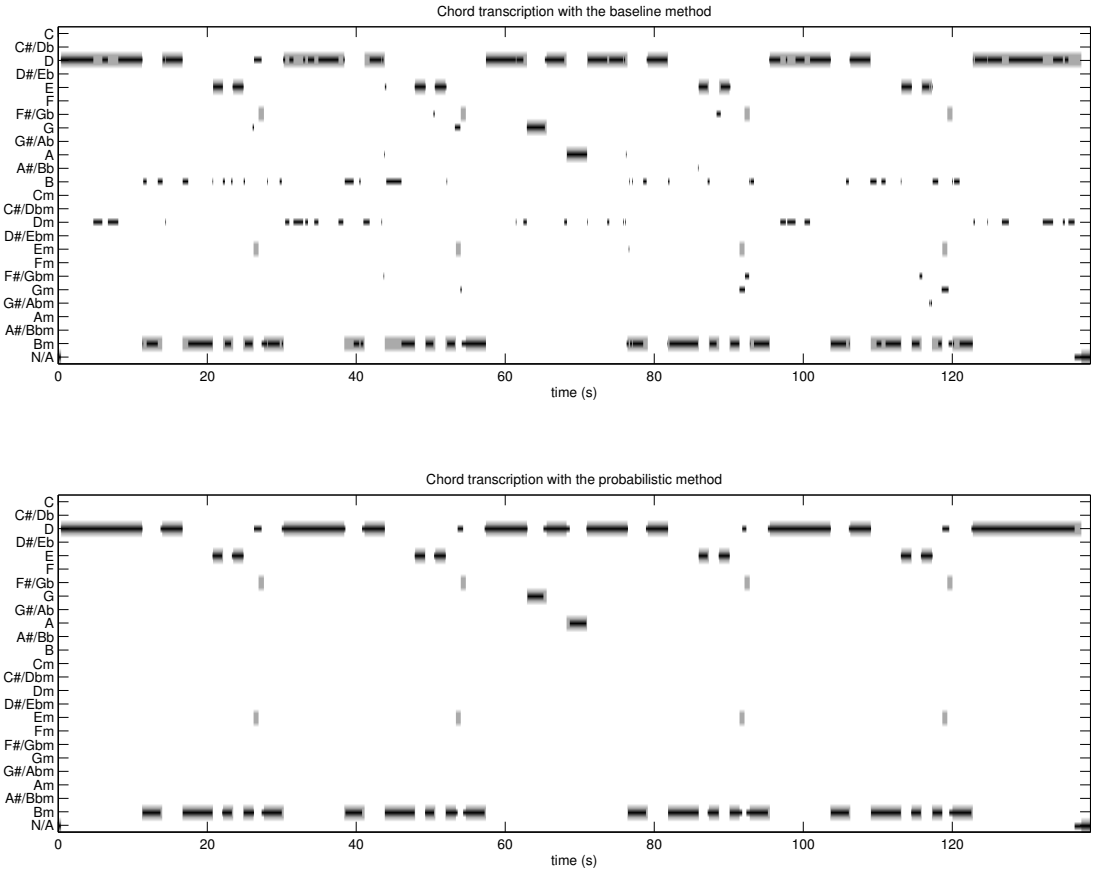

Fig. 2. Examples of chord transcription on the Beatles song *Run for your life*. The estimated chord labels are in black while the ground-truth chord annotation is in gray. At the top is represented the baseline method and at the bottom the new probabilistic method.

TABLE I
COMPARISON WITH THE STATE-OF-THE-ART : AVERAGE OVERLAP SCORES ON THE BEATLES AND QUAERO CORPUS

	Beatles corpus	Quaero corpus
Our method	**0.758**	**0.773**
OGF1	0.718	0.706
OGF2	0.724	0.682
DE	0.738	0.719
BP	0.707	0.699
RK	0.705	0.730

- RK : Ryynänen & Klapuri [11]

MIREX 2009 :

- DE : Ellis [9]
- OGF1 & OGF2 : our *baseline method* [2], [3].

These results shows that our new method outperforms the state-of-the-art, not only on the Beatles corpus, but also on another corpus containing other types of music styles and genres. This shows that while not using training data, our model manages to capture the harmonic content of the songs. The Friedman's test for significant differences[4] show that on the Beatles corpus, our new method is significantly better that OGF1, BP and RK.

[4]This test has been for MIREX 2008 & 2009 : see http://www.music-ir.org/mirex/2008/index.php/Audio_Chord_Detection_Results for details

IV. CONCLUSION

We have presented in this paper a new probabilistic framework for the template-based chord recognition, which allows to correct some of the issues caused by these types of methods. In particular, by evaluating the chord vocabulary for every song, our method can efficiently extract from the song the harmonic context, and therefore gives more relevant chord transcriptions. Furthermore, since our method does not need any information on the song nor training, its performances do not depend on the music genre.

ACKNOWLEDGMENT

The authors wish to thank C. Harte for his very useful annotation files and J. Bello, M. Ryynnen and D. Ellis for their codes.

This work was realized as part of the Quaero Programme, funded by OSEO, French State agency for innovation.

REFERENCES

[1] T. Fujishima, "Realtime chord recognition of musical sound: a system using Common Lisp Music," in *Proceedings of the International Computer Music Conference (ICMC)*, Beijing, China, 1999, pp. 464–467.

[2] L. Oudre, Y. Grenier, and C. Févotte, "Chord recognition using measures of fit, chord templates and filtering methods," in *Proceedings of the IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, New York, USA, 2009, pp. 9–12.

[3] ——, "Template-based chord recognition : influence of the chord types," in *Proceedings of the International Symposium on Music Information Retrieval (ISMIR)*, Kobe, Japan, 2009, pp. 153–158.

[4] C. Harte and M. Sandler, "Automatic chord identification using a quantised chromagram," in *Proceedings of the Audio Engineering Society*, Barcelona, Spain, 2005.

[5] K. Lee, "Automatic chord recognition from audio using enhanced pitch class profile," in *Proceedings of the International Symposium on Music Information Retrieval (ISMIR)*, Victoria, Canada, 2006.

[6] J. Bello and J. Pickens, "A robust mid-level representation for harmonic content in music signals," in *Proceedings of the International Symposium on Music Information Retrieval (ISMIR)*, London, UK, 2005, pp. 304–311.

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

[8] M. Mauch and S. Dixon, "Simultaneous estimation of chords and musical context from audio," *IEEE Transactions on Audio, Speech and Language Processing*, 2010.

[9] A. Sheh and D. Ellis, "Chord segmentation and recognition using EM-trained hidden Markov models," in *Proceedings of the International Symposium on Music Information Retrieval (ISMIR)*, Baltimore, MD, 2003, pp. 185–191.

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

[11] M. Ryynänen and A. Klapuri, "Automatic transcription of melody, bass line, and chords in polyphonic music," *Computer Music Journal*, vol. 32, no. 3, pp. 72–86, 2008.

[12] M. Khadkevich and M. Omologo, "Use of hidden markov models and factored language models for automatic chord recognition," in *Proceedings of the International Symposium on Music Information Retrieval (ISMIR)*, Kobe, Japan, 2009, pp. 561–566.

[13] T. Yoshioka, T. Kitahara, K. Komatani, T. Ogata, and H. Okuno, "Automatic chord transcription with concurrent recognition of chord symbols and boundaries," in *Proceedings of the International Symposium on Music Information Retrieval (ISMIR)*, Barcelona, Spain, 2004.

[14] H. Papadopoulos and G. Peeters, "Large-scale study of chord estimation algorithms based on chroma representation and HMM," in *Proceedings of the International Workshop on Content-Based Multimedia Indexing*, Bordeaux, France, 2007, pp. 53–60.

[15] K. Sumi, K. Itoyama, K. Yoshii, K. Komatani, T. Ogata, and H. Okuno, "Automatic chord recognition based on probabilistic integration of chord transition and bass pitch estimation," in *Proceedings of the International Symposium on Music Information Retrieval (ISMIR)*, Philadelphia, USA, 2008, pp. 39–44.

[16] E. Gómez, "Tonal description of polyphonic audio for music content processing," in *Proceedings of the INFORMS Computing Society Conference*, vol. 18, no. 3, Annapolis, MD, 2006, pp. 294–304.

[17] C. Harte, M. Sandler, S. Abdallah, and E. Gomez, "Symbolic representation of musical chords: A proposed syntax for text annotations," in *Proceedings of the International Symposium on Music Information Retrieval (ISMIR)*, London, UK, 2005, pp. 66–71.