CHORD RECOGNITION USING MEASURES OF FIT, CHORD TEMPLATES AND FILTERING METHODS

Laurent Oudre[1], *Yves Grenier*[1], *Cédric Févotte*[2]

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

ABSTRACT

This paper presents an efficient method for chord transcription of music signals. A succession of chroma vectors is calculated from the signal in order to extract the musical content of the piece over time. We introduce a set of chord templates for several types of chords (major, minor, dominant seventh,...) : different chord models taking into account one or more harmonics of the notes of the chord are considered. In order to fit the chroma vectors to the chord templates, we analytically calculate a scale parameter. The detected chord over a frame is the one minimizing a measure of fit between a rescaled chroma vector and the chord templates. Several popular measures in the probability and signal processing field are considered for our task. In order to take into account the time-persistence, we perform a post-processing filtering over the recognition criteria which quickly smooths the results and corrects random errors. The system is evaluated on the 13 Beatles albums and compared to the state-of-the-art. Results show that our method outperforms state-of-the-art methods but more importantly is significantly faster.

Index Terms— chord recognition, music signal processing, music signal representation

1. INTRODUCTION

Chord representation is a compact musical writing which gives information on the harmonic content and structure of a song. Automatic chord transcription finds many applications in the field of Musical Information Retrieval such as song identification, query by similarity or structure analysis.

Automatic chord transcription usually consists in two steps : a feature extraction which captures the musical information over time and a recognition process which outputs chord labels from the calculated features.

The features used for chord recognition may differ from a method to another but are in most cases variants of the 12-dimensional *Pitch Class Profiles* introduced by Fujishima [1]. Every component represents the spectral energy of a semi-tone on the chromatic scale regardless of the octave. The calculation is based either on the *Constant Q Transform (CQT)* or on the *Short Time Fourier Transform (STFT)* and is performed either on fixed-length frames or variable-length frames. The succession of these chroma vectors over time is called *chromagram*.

The present paper focuses mainly on the second step of the chord transcription process, which consists in giving a chord label to every chromagram frame. Machine-learning methods such as Hidden Markov Models (HMMs) have been widely used for this task especially in the last years. The first HMM used in chord recognition [2] is composed of 147 hidden states each representing a chord (7 types of chords and 21 root notes). All the HMM parameters are learned by a semi-supervised training with an EM algorithm. This model is later improved by Bello & Pickens [3]. The number of hidden states is reduced from 147 to 24 by only considering major and minor chords for the 12 semi-tones root notes. Musical knowledge is introduced into the model by initializing the HMMs parameters with values inspired by musical and cognitive theory. All the parameters are learned with an unsupervised training through an EM algorithm except for the chord observation probability distributions : each chord is given predetermined structure driven by musical knowledge.

Yet, the first chord recognition system proposed by Fujishima [1] is not using HMM but chord dictionaries composed of 12-dimensional templates constituted by 1 (for the chromas present in the chord) and 0 (for the other chromas). 27 types of chords are tested and the transcription is done either by minimizing the Euclidean distance between *Pitch Class Profiles* and chord templates or by maximizing a weighted dot product. Fujishima's system is improved [4] by calculating a more elaborate chromagram including notably a tuning algorithm and by reducing the number of chords types from 27 to 4 (major, minor, augmented, diminished). Chord transcription is then realized by retaining the chord with larger dot product between the chord templates and the chromagram frames.

Our chord recognition system is based on the intuitive idea that for a given 12-dimensional chroma vector, the amplitudes of the chromas present in the chord played should be larger than the ones of the non-played chromas. By introducing chord templates for different chord types and roots, the chord present on a frame should therefore be the one whose template is the *closest* to the chroma vector according to a specific measure of fit. A scale parameter is introduced in order to fit the chroma vectors to the chord templates and finally the detected chord is the one minimizing the measure of fit between the rescaled chroma vector and the chord templates. In order to take into account the time persistence, some filtering methods which tend to smooth the results and correct the errors are used.

The paper is organized as follows. Section 2 gives a description of our recognition system : the rescaled chroma vectors, the chord templates, the measures of fit and the post-processing filtering methods. Section 3 presents an analysis of the results and a comparison with the state-of-the-art. Finally the main conclusions of this work are summarized in Section 4.

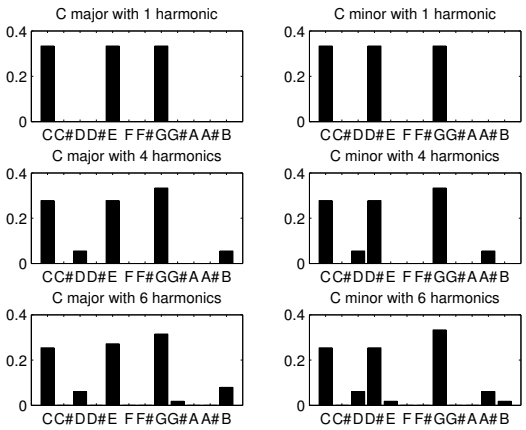

Figure 1: Chord templates for C major / C minor with 1, 4 or 6 harmonics.

2. SYSTEM

2.1. General idea

Let $\mathbf{C}$ denote the chromagram, with dimensions $12\times N$, composed of N successive chroma vectors $\{\mathbf{c}_n\}_n$. In practice, the chroma vectors are calculated from the music signal with the same method as Bello & Pickens [3]. 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 $\mathbf{p}_k$ be the 12-dimensional chord template defining chord k. We introduce a scale parameter $h_{k,n}$ whose role is to fit the chroma vector $\mathbf{c}_n$ with the chord template $\mathbf{p}_k$ according to the measure of fit used. The scale parameter $h_{k,n}$ is calculated analytically so as to minimize the measure between $h\,\mathbf{c}_n$ and $\mathbf{p}_k$:

$$h_{k,n} = \underset{h}{\operatorname{argmin}}\, D\left(h\,\mathbf{c}_n;\mathbf{p}_k\right), \tag{1}$$

such that :

$$\left[\frac{d\,D\left(h\,\mathbf{c}_n;\mathbf{p}_k\right)}{dh}\right]_{h=h_{k,n}} = 0. \tag{2}$$

We want to find the chord k whose template $\mathbf{p}_k$ is the *closest* to the rescaled chromagram frame $h_{k,n}\,\mathbf{c}_n$ for a specific measure of fit.

As such, we define $d_{k,n}$ as the measure of fit between the rescaled chromagram frame $h_{k,n}\,\mathbf{c}_n$ and the chord template $\mathbf{p}_k$:

$$d_{k,n} = D\left(h_{k,n}\,\mathbf{c}_n;\mathbf{p}_k\right). \tag{3}$$

The detected chord $\hat{k}_n$ for frame n is then the one minimizing the set $\{d_{k,n}\}_k$:

$$\hat{k}_n = \underset{k}{\operatorname{argmin}}\ \{d_{k,n}\}. \tag{4}$$

We have omit for sake of conciseness the expressions of $d_{k,n}$ and $h_{k,n}$ which are easily obtained.

2.2. Chord models

The chord templates are 12-dimensional vectors where each component represents the theoretical amplitude of each chroma in the chord. The intuitive model, which has already been used for chord recognition [1],[4], is a simple binary mask constituted of 1's for the chromas present in the chord and 0's for the other chromas.

Yet, the information contained in a chromagram captures not only the intensity of every note but a blend of intensities for the harmonics of every note. Like Gomez [5] and Papadopoulos [6], we assume an exponentially decreasing spectral profile for the amplitudes of the partials ; an amplitude of 0.6^{i-1} is added for the i^{th} harmonic of every note in the chord.

In our system three chord models are defined, corresponding to 1, 4 or 6 harmonics. Examples for C major and C minor chords are displayed on Figure 1.

From these three chord models we can build chord templates for all types of chords (major, minor, dominant seventh, diminished, augmented,...). By convention in our system, the chord templates are normalized so that the sum of the amplitudes is 1.

2.3. Measures of fit

We consider for our recognition task several measures of fit, popular in the field of signal processing.

- The **Euclidean distance** (*EUC*) defined by

$$D_{EUC}\left(\mathbf{x}|\mathbf{y}\right) = \sqrt{\sum_i \left(x_i - y_i\right)^2} \tag{5}$$

  has notably already been used by Fujishima [1] for the chord recognition task.

- The **Itakura-Saito divergence** defined by

$$D_{IS}\left(\mathbf{x}|\mathbf{y}\right) = \sum_i \frac{x_i}{y_i} - \log\left(\frac{x_i}{y_i}\right) - 1 \tag{6}$$

  is a measure of fit between two spectra, popular in the speech community [7]. Since it is not symmetrical, it is not a distance and it can therefore be calculated in two ways. $D_{IS}\left(h_{k,n}\,\mathbf{c}_n|\mathbf{p}_k\right)$ will later be referred as *IS1* and $D_{IS}\left(\mathbf{p}_k|h_{k,n}\,\mathbf{c}_n\right)$ as *IS2*.

- The **Kullback-Leibler divergence** measures the dissimilarity between two probability distributions. In the present paper we use the generalized Kullback-Leibler divergence defined by

$$D_{KL}\left(\mathbf{x}|\mathbf{y}\right) = \sum_i x_i \log\left(\frac{x_i}{y_i}\right) - x_i + y_i. \tag{7}$$

  Just like the Itakura-Saito divergence, the Kullback-Leibler divergence is not a distance so that we can build two measures of fit : $D_{KL}\left(h_{k,n}\,\mathbf{c}_n|\mathbf{p}_k\right)$ *(KL1)* and $D_{KL}\left(\mathbf{p}_k|h_{k,n}\,\mathbf{c}_n\right)$ *(KL2)*.

2.4. Filtering methods

So far our chord detection is done frame by frame without taking into account the results on the adjacent frames. Yet in practice, it is rather unlikely for a chord to last only one frame. Furthermore the information contained in the adjacent frames can help decision : it is one of the main advantages of the methods using

HMM, where the introduction of transition probabilities naturally leads to a smoothing effect. In order to take into account the time-persistence we introduce some post processing filtering methods which works upstream on the calculated measures.

We introduce new criteria $\tilde{d}_{k,n}$ based on L successive values of the $\{d_{k,n'}\}_{n'}$ previously calculated, centered on frame n (L is then odd). In our system two types of filtering are used.

- The **low-pass filtering** defined by

$$\tilde{d}_{k,n} = \frac{1}{L} \sum_{n'=n-\frac{L-1}{2}}^{n+\frac{L-1}{2}} d_{k,n'} \quad (8)$$

  tends to smooth the output chord sequence and to reflect the long-term trend in the chord change.

- The **median filtering** defined by

$$\tilde{d}_{k,n} = \text{med}\left\{d_{k,n'}\right\}_{n-\frac{L-1}{2} \leq n' \leq n+\frac{L-1}{2}} \quad (9)$$

  has been widely used in image processing and is particularly efficient to correct random errors.

In every case, the detected chord $\hat{k}_n$ on frame n is the one that minimizes the set of values $\left\{\tilde{d}_{k,n}\right\}_k$:

$$\hat{k}_n = \underset{k}{\text{argmin}} \left\{\tilde{d}_{k,n}\right\} \quad (10)$$

Since these two filtering methods act in different ways, they can also be used one after the other in order to combine their advantages. We shall later refer to the successions of the two methods respectively as *LP+M* and *M+LP*.

3. EVALUATION AND RESULTS

3.1. Protocol of evaluation

The evaluation method used in this paper corresponds to the one used in MIREX 08 for the Audio Chord Detection task[1].

Our evaluation database is made of 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 [8]. In these annotation files, 17 types of chords (maj, dim, aug, maj7, 7, dim7, hdim7, maj6, 9, maj9, sus4, sus2, min, min7, minmaj7, min6, min9) and one 'no chord' label (N) corresponding to silences or untuned material are present.

Since most of the state-of-the-art methods can only detect major and minor chords, the 17 types of chords present in the annotation files are mapped to the major and the minor.

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 the 180 songs is then called *Average Overlap Score (AOS)*.

3.2. Influence of the parameters

The previously described five measures of fit (*EUC*, *IS1*, *IS2*, *KL1* and *KL2*), three chord models (1, 4 or 6 harmonics) and two filtering methods (low-pass and median) with neighborhood sizes from $L = 3$ to $L = 25$ are tested, which constitutes 360 parameter sets. Only major and minor chords models are used for these simulations.

In order to understand the influence, the role and the robustness of every parameter, below is presented an analysis of the previously described AOSs calculated for the 360 parameter sets.

3.2.1. Measures of fit

On Table 1 are presented statistics on the 360 calculated Average Overlap Scores (72 per measure of fit).

	Max	Min	Mean	Std dev.
EUC	0.693	0.569	0.641	0.033
IS1	0.689	0.380	0.500	0.125
IS2	0.691	0.162	0.503	0.240
KL1	0.685	0.130	0.427	0.219
KL2	0.696	0.588	0.656	0.032

Table 1: Influence of the measure of fit on the Average Overlap Scores.

The different measures give close results but differ in robustness. The most efficient and robust measures are the generalized Kullback-Leibler divergence KL2 and the Euclidian distance. For these two measures, the standard deviation is low, which means that the other parameters (number of harmonics, filtering methods, neighborhood size) do not have a strong influence on the results and that the choice of the measure is here decisive.

3.2.2. Number of harmonics

On Table 2 is presented an analysis of the 360 Average Overlap Score according to the number of harmonics considered.

	Max	Min	Mean	Std dev.
1 harmonic	0.696	0.614	0.670	0.016
4 harmonics	0.696	0.420	0.583	0.104
6 harmonics	0.635	0.130	0.383	0.206

Table 2: Influence of the number of harmonics on the Average Overlap Scores.

The single harmonic and 4 harmonics chord models both gives the best results but only the single harmonic model gives very robust results. The introduction of other harmonics leads to less robust results.

Considering more than one harmonic leads to give non-null values to chromas which are not present in the chord (see Figure 1). Itakura-Saito and Kullback-Leibler divergence both contain a logarithm component which is very sensitive to the zeros in the chord templates. The null components are therefore very important for the chord discrimination. We believe that the less null chromas there are in the model, the harder the chord discrimination is and so the worse the results are.

3.2.3. Filtering methods

In order to analyze the influence of the filtering methods, on top of the low-pass and median filtering, we also test here the median+low-pass filtering, the low-pass+median filtering, and the

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

no filtering case. Table 3 presents the Best Average Overlap Scores obtained for each filtering method with the corresponding neighborhood size used.

	Best AOS	Best neighborhood size
no filtering	0.651	N/A
low-pass filtering	0.693	$L = 13$
median filtering	0.696	$L = 17$
LP+M filtering	0.698	$L_{LP} = 5, L_M = 17$
M+LP filtering	0.698	$L_M = 15, L_{LP} = 5$

Table 3: Influence of the filtering methods and the neighborhood sizes on the Best Average Overlap Score.

Results show that filtering clearly enhances the results. All the filtering methods give good and very close results.

The combination median+low-pass gives the best results probably because it enables to exploit the properties of both filtering methods : the low-pass filtering tends to smooth the chord sequence while the median filtering reduces the random errors.

The optimal neighborhood sizes correspond to a length of approximately 2 seconds with the parameters used for the computation of the chroma vectors : interestingly, the average chord length for the 180 songs of the Beatles corpus is also around 2 seconds.

3.3. Variants of the method

The simplicity of our chord models allows to easily introduce other types of chords than major and minor, such as dominant seventh, diminished, augmented, etc. As the evaluation method only takes into account major and minor chords, once the chords have been detected with their appropriate models, they will then be mapped to the major and minor following the same rules already used for the annotation files.

Among all the chord types considered, the best results are obtained by considering major, minor and dominant seventh chords (AOS : 0.711). This can be explained by the high number of dominant seventh chords in the Beatles corpus (third most present chord type after major and minor).

3.4. State-of-the-art

Our method is now compared with the method of Bello & Pickens [3] which obtained the best results for the Audio Chord Detection task for pretrained systems at MIREX 08. This method has been tested with its original implementation on the same Beatles corpus and evaluated with the same protocol.

Table 4 presents the results of our recognition method. Among with the major-minor method used so far (Maj-Min), we display the results obtained by considering major, minor and dominant seventh chord types (Maj-Min-7).

	AOS	Time	Parameters
Maj-Min-7	0.711	809 s	*1 harm., KL2, M+LP* $L_M = 15, L_{LP} = 5$
Bello & Pickens	0.700	1619 s	N/A
Maj-Min	0.698	801 s	*1 harm., KL2, LP+M* $L_{LP} = 5, L_M = 15$

Table 4: Comparison with the state-of-the-art.

Our Maj-Min-7 method gives the best results and more importantly with a very low computational time. It is indeed twice as fast as the Bello & Pickens method.

4. CONCLUSION

In this paper we have presented an efficient, simple and fast chord transcription method. Instead of using HMMs, the joint use of popular measures of fit and filtering methods enables to pertinently and quickly recognize both the structure and the time-persistence of the chords. The combination of low-pass and median filtering performs both a smoothing of the results over time and a correction of random errors. Since our method is based on the musical content extracted in the chromagram, it does not require neither training nor information about the style, rhythm and instruments. The simplicity of the system enables to keep the computation time low.

5. ACKNOWLEDGMENT

The authors would like to thank C. Harte for his very useful annotation files.

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

6. REFERENCES

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

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

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

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

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

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

[7] F. Itakura and S. Saito, "Analysis synthesis telephony based on the maximum likelihood method," in *Proc. of the International Congress on Acoustics*, Tokyo, Japan, 1968, pp. 17–20.

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