# Probabilistic template-based chord recognition

Laurent Oudre, *Member, IEEE,* Cédric Févotte, *Member, IEEE,* Yves Grenier, *Member, IEEE*

***Abstract*—This paper describes a probabilistic approach to template-based chord recognition in music signals. The algorithm only takes chromagram data and a user-defined dictionary of chord templates as input data. No training or musical information such as key, rhythm or chord transition models is required. The chord occurrences are treated as probabilistic events, whose probabilities are learned from the song using an Expectation-Maximization (EM) algorithm. The adaptative estimation of these probabilities (together with an ad-hoc post-processing filtering) has the desirable effect of smoothing out spurious chords that would occur in our previous baseline work. Our algorithm is compared to various methods that entered the Music Information Retrieval Evaluation eXchange (MIREX) in 2008 and 2009, using a diverse set of evaluation metrics, some of which are new. The systems are tested on two evaluation corpuses; the first one is composed of the Beatles catalog (180 pop-rock songs) and the other one is constituted of 20 songs from various artists and music genres. Results show that our method outperforms state-of-the-art chord recognition systems.**

***Index Terms*—chord recognition, music signal processing, music signal representation, music information retrieval**

## I. Introduction

DESCRIPTION of music signals with relevant and compact representations has been one of the main fields of interest in Musical Information Retrieval (MIR) in the last few years. One of the most common representations of pop songs is chord transcription, which returns the musical content of a piece. This representation, whilst not precisely reproducing all the notes played by the instruments, allows musicians to easily playback songs. As such, a chord can be defined as a set of harmonically-related notes played simultaneously. In reality, there are many ways to define and classify chords, depending on the application context, time period or music type [1]. In this paper, we shall follow the conventions used for Western popular music, where a chord can be written by using two notions: *roots* and *types*. The *root* is the note upon which the chord is built, while the *type* gives the harmonic structure of the chord (i.e., the harmonic relationships between the notes within the chord). For example, a C major chord (composed of notes C, E and G) is described by its root note *C* and its type *major*, implying the presence of the major third *E* and the perfect fifth *G* in the chord construction. The chord transcription output by our automatic chord transcriber is a sequence of chord labels with their respective start and end times. This output can be used for song playback - which constitutes the main aim of our system - but also in other applications such as song identification, query by similarity or structure analysis.

Template-based chord recognition methods are based on the hypothesis that only the chord definition is necessary to extract chord labels from a musical piece. A chord template is a 12-dimensional vector representing the 12 semi-tones (or *chroma*) of the chromatic scale. Each component of the pattern is given a theoretical amplitude according to the chord definition. The most simple and intuitive chord template [2], [3], [4] has a binary structure, with amplitudes of 1 for the chromas within the chord definition and 0 for other chromas. More complex patterns have been considered, for example taking into account the harmonics of the chord notes [5], [6].

The first template-based system for audio chord recognition was developed by Fujishima [2]. This method is the first one that considers chords not only as sets of individual notes, but rather as entities whose structure is determined by one root and one type. The chord transcription process is based on the extraction from the signal of *Pitch Class Profiles (PCP)* or *chroma vectors*. The chroma vectors are 12-dimensional vectors where each component represents the energy or salience of one of the 12 semi-tones within the chromatic scale, regardless of the octave. The temporal evolution of these chroma vectors is called *chromagram*: it has been widely used in literature for chord or key estimation [5], [7]. In Fujishima's approach, 324 chords are detected, each of them modeled by a binary *Chord Type Template (CTT)*. The chord detection is performed by first calculating scores for every root and chord type, then selecting the best score. The scores are computed from chroma vectors and hand-tuned variations of the original CTT. Two matching methods between PCP and CTT are tested: the Nearest Neighbor Method (Euclidean distance between chroma vector and hand-tuned CTT) and the Weighted Sum Method (dot product between chroma vector and hand-tuned CTT). The hand-tuning is done by trial-and-error and accounts for the chord type probability and the number of notes within the chord type. Two post-processing methods are introduced in order to take into account the temporal structure of the chord sequence. The first attempt is to smooth over the past chroma vectors to both reduce the noise and use the fact that a chord usually lasts for several frames. The second attempt is to detect chord changes by monitoring the direction of the chroma vectors.

Harte & Sandler [3] use a very similar method to Fujishima's. The chromagram extraction is improved by applying a frequency tuning algorithm. They define binary chord templates for 4 chord types (major, minor, diminished and augmented) and then calculate a dot product between chroma vectors and chord templates. The temporal information is captured by applying low-pass filtering on the chromagram and median filtering to the detected chord sequence.

Lee [4] also uses binary chord templates, this time for the 24 major/minor triads. He introduces a new input feature called Enhanced Pitch Class Profile (EPCP) using the harmonic product spectrum. The chord recognition is then carried out by maximizing the correlation between chroma vectors and chord templates.

These template-based methods often have difficulties to capture long-term variations of chord sequences, as well as to generate compact chord transcriptions. In particular, these methods can give good frame-to-frame results but often produce fragmented transcriptions, hardly usable for immediate song playback. Complex probabilistic methods have been built in order to incorporate musical information such as key, chord transitions models, beats or structure. This high-level information can for instance be extracted from music theory and introduced in Hidden Markov Models (HMM) [8], [9], in Dynamic Bayesian Networks (BDN) [10], [11] or in rule-based systems [12], [13]. It can also be obtained from the training of HMM with audio data (annotated or not) [14], [15], [16], [17], [18]. Finally, some methods combine these two approaches, for instance using hypothesis search algorithms [19], [20].

The method presented in this paper builds on the deterministic template-based method described in [21], [22] while offering a novel statistical framework that explicitly models chord occurrences in songs as probabilistic events. The probability of each of the candidate chords in a song is learned directly from the song, i.e., in a data-driven way. Hence, the probabilistic approach allows one to extract a relevant and sparse chord *vocabulary* for every song. By vocabulary, we mean the subset of the *dictionary* that contains the chords played in the song. The term dictionary refers to the set of all user-defined chord candidates. The notion of vocabulary is not necessarily linked to key: for example, in the case of modulations, the chord vocabulary can contain chords from various keys. Our previous systems [21], [22] tended to produce fragmented chord transcriptions, and to detect chords that were not present in the ground-truth files. This phenomenon was mostly due to a large number of parallel errors (major-minor confusion). The main effect of the introduction of the chord probabilities in the model is the elimination of most spurious chords detected by our previous methods (i.e., the probability of chords absent from the song tends to zero). This leads to more compact and readable, in other words *sparser*, chord transcriptions while improving the detection scores. Contrary to other probabilistic chord recognition methods, our method can still be classified within the template-based methods, since the only information given to our system is the chord definition (i.e., the chord dictionary).

Besides the novel probabilistic chord transcription framework, another contribution of this paper is the large-scale comparison of our method with numerous state-of-the-art systems. Many metrics are considered, some of them new, and we propose a complete evaluation of several aspects of the chord recognition task.

Section II introduces notations and provides a short description of the main concepts of the deterministic baseline method [21], [22]. Section III describes our probabilistic approach and how it is built on the baseline method. Section IV presents the metrics and the two song corpus used for evaluation. Finally, Section V reports the results obtained by our probabilistic and deterministic methods, along with some state-of-the-art methods.

## II. Deterministic baseline method

This section describes the deterministic baseline method and introduces the main concepts of our chord recognition system. More details can be found in [21], [22].

### A. Principle and notations

Let $\mathbf{C}$ be a $12 \times N$ chromagram, composed of $N$ 12-dimensional successive chroma vectors $\mathbf{c}_n$. Let $\mathbf{W}$ be our $12 \times K$ chord dictionary, composed of $K$ 12-dimensional chord templates $\mathbf{w}_k$. Again, the dictionary is the set of all user-defined chord candidates. In this paper we will only consider major and minor chords built from the chromatic scale, hence $K = 24$.

Intuitively, the chord $\gamma_n \in [1, \ldots, K]$ detected at frame $n$ should be the one whose defining template $\mathbf{w}_{\gamma_n}$ is the *closest* to the chroma vector $\mathbf{c}_n$, given a certain measure of fit. Of course we assume that only one chord is played at each time frame. The fit between $\mathbf{c}_n$ and every possible template $\mathbf{w}_k$ has to be measured up to a scale parameter $h_{k,n}$ that accounts to energy variations, so that

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

Given a measure of fit $D\left( \,.\, ; \,.\, \right)$, the scale parameter is defined as

$$h_{k,n} = \underset{h}{\operatorname{argmin}} \; D\left(\mathbf{c}_n; h\, \mathbf{w}_k\right), \tag{2}$$

and must satisfy

$$\nabla_h D\left(\mathbf{c}_n; h\, \mathbf{w}_k\right)\big|_{h=h_{k,n}} = 0. \tag{3}$$

Given the set of computed scale parameters, the detected chord $\hat{\gamma}_n$ for frame $n$ is finally chosen as the one yielding best overall fit, i.e.,

$$\hat{\gamma}_n = \underset{k}{\operatorname{argmin}} \; \{D\left(\mathbf{c}_n; h_{k,n}\, \mathbf{w}_k\right)\}_k. \tag{4}$$

### B. Chord templates

Chord templates are 12-dimensional vectors that can be considered as theoretical chroma vectors, reflecting the contribution in amplitude of each chroma in the chord. Chord recognition methods often rely on chord templates, which are either fixed [2], [3], [4], [6], [8], [23] or learned from audio data [14], [16], [17]. We here consider fixed templates as they are easier to obtain (there is no need for annotated data) and do not depend on the training corpus. Hence, our chord templates are simple binary masks: an amplitude of 1 is given to the notes present in the chord and an amplitude of 0 is given to the other ones.[1] For example a *C major* chord is given an amplitude of 1 to chromas *C*, *E* and *G* while other chromas have an amplitude of 0. By convention, in our system the chord

[1] In practice a small value is used instead of 0, to avoid numerical instabilities that may arise with some measures of fit.

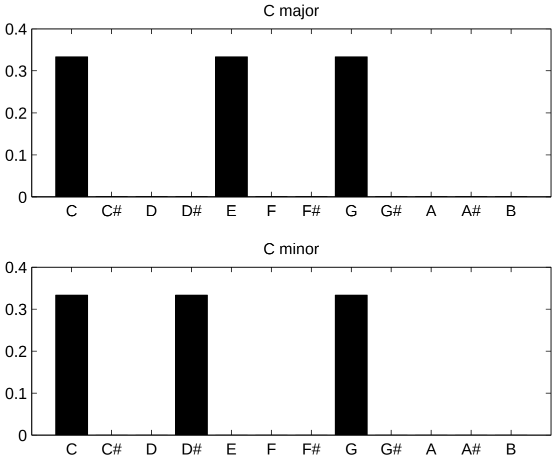


Fig. 1. Chord templates for C major and C minor (x-axis: note, y-axis: amplitude).

templates are normalized so that the sum of the amplitudes is 1 but any other normalization could be employed, as scale factors are refitted. Examples for *C major* and *C minor* chord are presented on Figure 1.

### C. Measures of fit

In [21], [22], we considered for our recognition task several measures of fit, which are popular in the field of signal processing. The well-known **Euclidean distance** defined by

$$D_{EUC}\left(\mathbf{x}|\mathbf{y}\right) = \sqrt{\sum_m \left(x_m - y_m\right)^2} \tag{5}$$

has already been used by Fujishima [2] for the chord recognition task. The **Itakura-Saito (IS) divergence** defined by

$$D_{IS}\left(\mathbf{x}|\mathbf{y}\right) = \sum_m \frac{x_m}{y_m} - \log\left(\frac{x_m}{y_m}\right) - 1 \tag{6}$$

was introduced in [24] and presented as a suitable measure of the goodness of fit between two spectra. It is popular in the speech community and has recently proven useful for source separation based on nonnegative matrix factorization of the spectrogram [25]. The **Kullback-Leibler divergence** [26] is a well-known measure of the dissimilarity between probability distributions. It has been widely used in information theory and has given rise to many variants: in the present paper, we use the generalized Kullback-Leibler (KL) divergence (also known as I-divergence) defined by

$$D_{KL}\left(\mathbf{x}|\mathbf{y}\right) = \sum_m x_m \log\left(\frac{x_m}{y_m}\right) - x_m + y_m. \tag{7}$$

These three measures of fit have probabilistic interpretations that will be discussed in Section III. Furthermore, they behave differently with respect to scale (taken here as the relative contribution of small energy observations with respect to higher energies), and in particular the IS divergence is scale-invariant (see [25] for further discussion).

### D. Post-processing filtering

Frame by frame chord recognition does not take into account the influence of adjacent frames, which may be considered suboptimal as it does not exploit the available information redundancy between the frames. In order to correct this, we introduced in our baseline work a post-processing filtering step that works upstream on the recognition criterion $\{D\left(\mathbf{c}_n; h_{k,n}\,\mathbf{w}_k\right)\}_n$. Two types of filtering have been tested for our baseline system: **low pass filtering**, that smoothes the output chord sequence and reflects the long-term trend in the chord changes, and **median filtering**, that has been widely used in image processing, is efficient to correct random errors while respecting transitions.

Note that this type of filtering is innovative, since it is applied to the recognition criterion itself, and not to the chromagram (as in previous work [2], [7], [8]) or to the detected chord sequence [8].

## III. Probabilistic framework

In this section we describe the main methodological contribution of this paper, a novel probabilistic template-based chord recognition system. Our approach is built on the deterministic baseline system described in Section II, but now the measures of fit are turned into likelihood functions and the chord occurrences are treated as probabilistic events. In particular, the probability of each chord is learned from the song, and this will be shown to *sparsify* the chord vocabulary (elimination of spurious chords), which in turn greatly improves transcription accuracy.

### A. Generative model for $\mathbf{c}_n$

When the Euclidean distance, the KL divergence or the IS divergence is used as the measure of fit, the criterion $D\left(\mathbf{c}_n; h_{k,n}\,\mathbf{w}_k\right)$ defined in Section II is actually a log-likelihood in disguise. Indeed, the latter measures of fit respectively underlie Gaussian additive, Poisson and Gamma multiplicative observation noise models and they may be linked to a log-likelihood such that

$$-\log p\left(\mathbf{c}_n|h_{k,n}, \mathbf{w}_k\right) = \varphi_1\, D\left(\mathbf{c}_n|h_{k,n}\,\mathbf{w}_k\right) + \varphi_2, \tag{8}$$

where $p\left(\mathbf{c}_n|h_{k,n}, \mathbf{w}_k\right)$ is the probability of chroma vector $\mathbf{c}_n$ (now treated as a random variable) given chord template $\mathbf{w}_k$ (a fixed deterministic parameter) and scale $h_{k,n}$ (treated as an unknown deterministic parameter), and where $\varphi_1$ and $\varphi_2$ are constants w.r.t. $h_{k,n}$ and $\mathbf{w}_k$. The exact correspondences between each measure of fit and its equivalent statistical observation model are given in Table I.

The distribution $p\left(\mathbf{c}_n|h_{k,n}, \mathbf{w}_k\right)$ represents the probability of observing $\mathbf{c}_n$ *given* that the chord played at frame $n$ is the $k^{th}$ one, i.e., the one modeled by template $\mathbf{w}_k$. Let us introduce the discrete state variable $\gamma_n \in [1, \ldots, K]$ which indicates which chord is played at frame $n$, i.e, $\gamma_n = k$ if chord $k$ is played at frame $n$. Hence, we write

$$p\left(\mathbf{c}_n|\gamma_n = k, h_{k,n}\right) = p\left(\mathbf{c}_n|h_{k,n}, \mathbf{w}_k\right). \tag{9}$$

TABLE I

CORRESPONDENCES BETWEEN THE MEASURES OF FIT AND THEIR EQUIVALENT STATISTICAL OBSERVATION MODEL OF THE CHROMAGRAM

| Noise structure | Observation model $p\left(\mathbf{c}_n \mid h_{k,n}, \mathbf{w}_k\right)$ | Log-likelihood<br>$-\log\left(p\left(\mathbf{c}_n \mid h_{k,n}, \mathbf{w}_k\right)\right)$ | Scale parameter $h_{k,n}$ |
|---|---|---|---|
| Additive Gaussian noise<br>$\mathbf{c}_n = h_{k,n}\mathbf{w}_k + \epsilon$ | $\prod_{m=1}^{M} \mathcal{N}\left(c_{m,n}\,;\, h_{k,n}w_{m,k}, \sigma^2\right)$ | $\frac{1}{2\sigma^2} d_{EUC}^{2}\left(\mathbf{c}_n; h_{k,n}\mathbf{w}_k\right) + cst$ | $\frac{\sum_{m=1}^{M} c_{m,n}\, w_{m,k}}{\sum_{m=1}^{M} w_{m,n}^2}$ |
| Multiplicative Gamma noise $\mathbf{c}_n = (h_{k,n}\mathbf{w}_k) \cdot \epsilon$ | $\prod_{m=1}^{M} \frac{1}{h_{k,n}w_{m,k}} \mathcal{G}\left(\frac{c_{m,n}}{h_{k,n}w_{m,k}}\,;\, \beta, \beta\right)$ | $\beta\, d_{IS}\left(\mathbf{c}_n \mid h_{k,n}\mathbf{w}_k\right) + cst$ | $\frac{1}{M}\sum_{m=1}^{M} \frac{c_{m,n}}{w_{m,k}}$ |
| Poisson noise | $\prod_{m=1}^{M} \mathcal{P}\left(c_{m,n}\,;\, h_{k,n}w_{m,k}\right)$ | $d_{KL}\left(\mathbf{c}_n \mid h_{k,n}\mathbf{w}_k\right) + cst$ | $\sum_{m=1}^{M} c_{m,n}$ |

where $\mathcal{N}$, $\mathcal{G}$ and $\mathcal{P}$ are the probability distribution defined in Appendix
and $cst$ denotes terms constant w.r.t. $h_{k,n}\mathbf{w}_k$.

We are slightly abusing notations here as $\mathbf{w}_k$ should also appear on the left-hand side of Eq. (9), but as this is a fixed parameter as opposed to a parameter to be estimated, we will drop it from the notations. Now let us denote by $\alpha_k$ the probability of occurrence of chord $k$ in the song. Hence we have

$$P(\gamma_n = k) = \alpha_k, \tag{10}$$

where we so far assume that the frames are independent. Let us introduce the vector variables $\boldsymbol{\alpha} = [\alpha_1, \ldots, \alpha_K]^T$ (vector of all chord probabilities) and $\mathbf{h}_n = [h_{1,n}, \ldots, h_{K,n}]$ (vector of all scale parameters at frame $n$). Averaging over all possible states (chords), the statistical generative model of the chromagram defined by Eq. (9) and (10) is written more concisely as

$$p\left(\mathbf{c}_n|\boldsymbol{\alpha}, \mathbf{h}_n\right) = \sum_{k=1}^{K} \alpha_k\, p\left(\mathbf{c}_n|h_{k,n}, \mathbf{w}_k\right), \tag{11}$$

which defines a *mixture model.*

To recapitulate our model, given a dictionary of chords $\mathbf{W}$ with occurrence probabilities $\boldsymbol{\alpha}$, a chromagram frame $\mathbf{c}_n$ is generated by 1) randomly choosing chord $k$ with probability $\alpha_k$, 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 the vector $h_{k,n}\mathbf{w}_k$.

The only parameters to be estimated in our model are the chord probabilities $\boldsymbol{\alpha}$ and the set of amplitude coefficients $\mathbf{H} = \{h_{k,n}\}_{kn}$. Given estimates of these parameters, chord recognition at every frame $n$ may be performed by selection of the chord with largest posterior probability, i.e,

$$\hat{\gamma}_n = \underset{k}{\operatorname{argmax}}\ \{p(\gamma_n = k|\mathbf{c}_n, \hat{\boldsymbol{\alpha}}, \hat{\mathbf{h}}_n)\}_k. \tag{12}$$

Next we describe an EM algorithm for maximum likelihood estimation of parameters $\boldsymbol{\alpha}$ and $\mathbf{H}$.

## B. Expectation-Maximization (EM) algorithm

Let us denote by $\boldsymbol{\Theta} = (\boldsymbol{\alpha}, \mathbf{H})$ the set of parameters. Our task is to maximize the following objective function

$$\log p\left(\mathbf{C}|\boldsymbol{\Theta}\right) = \sum_n \log p\left(\mathbf{c}_n|\boldsymbol{\alpha}, \mathbf{h}_n\right), \tag{13}$$

which may routinely be done with an EM algorithm [27] using the set of chord state variables as missing data, which we denote $\boldsymbol{\gamma} = [\gamma_1, \ldots, \gamma_N]$. The EM algorithm involves computing (E-step) and maximizing (M-step) the following functional

$$Q\left(\boldsymbol{\Theta}|\boldsymbol{\Theta'}\right) = \sum_{\boldsymbol{\gamma}} \log p\left(\mathbf{C}, \boldsymbol{\gamma}|\boldsymbol{\Theta}\right) p\left(\boldsymbol{\gamma}|\mathbf{C}, \boldsymbol{\Theta'}\right) \tag{14}$$

where $\log p\left(\mathbf{C}, \boldsymbol{\gamma}|\boldsymbol{\Theta}\right)$ is referred to as the *complete data likelihood* and $p\left(\boldsymbol{\gamma}|\mathbf{C}, \boldsymbol{\Theta'}\right)$ is the *missing data posterior*. Each of the two EM steps is described next.

*a) E-Step:* Under the frame independence assumption the functional (14) can be written as

$$Q\left(\boldsymbol{\Theta}|\boldsymbol{\Theta'}\right) = \sum_{n=1}^{N}\sum_{k=1}^{K} \log p\left(\mathbf{c}_n, \gamma_n = k|\boldsymbol{\Theta}\right) p\left(\gamma_n = k|\mathbf{c}_n, \boldsymbol{\Theta'}\right). \tag{15}$$

Let us denote $\bar{\alpha}_{k,n}$ the posterior probability of state variable $\gamma_{k,n}$ (the notation is chosen in analogy with the notation chosen for its prior probability $\alpha_k$), i.e.,

$$\begin{aligned} \bar{\alpha}_{k,n} &= p\left(\gamma_n = k|\mathbf{c}_n, \boldsymbol{\Theta}\right) & (16) \\ &= \frac{\alpha_k\, p\left(\mathbf{c}_n|\gamma_n = k, \boldsymbol{\Theta}\right)}{\sum_{l=1}^{K} \alpha_l\, p\left(\mathbf{c}_n|\gamma_n = l, \boldsymbol{\Theta}\right)}, & (17) \end{aligned}$$

where the second equation comes naturally from the application of Bayes theorem and from the fact that the probabilities sum to 1. In the following, we denote by $\bar{\alpha}'_{k,n}$ the posterior state probabilities conditioned on parameter iterate $\boldsymbol{\Theta'}$. Hence, by expanding the complete data likelihood as

$$\log p\left(\mathbf{c}_n, \gamma_n = k|\boldsymbol{\Theta}\right) = \log p\left(\mathbf{c}_n|h_{k,n}, \mathbf{w}_k\right) + \log \alpha_k, \tag{18}$$

the E-step amounts to evaluating the EM functional as

$$Q\left(\boldsymbol{\Theta}|\boldsymbol{\Theta'}\right) = \sum_{n=1}^{N}\sum_{k=1}^{K} \left[\log p\left(\mathbf{c}_n|h_{k,n}, \mathbf{w}_k\right) + \log \alpha_k\right] \bar{\alpha}'_{k,n}, \tag{19}$$

which we recall is to be maximized w.r.t to $\boldsymbol{\Theta} = (\boldsymbol{\alpha}, \mathbf{H})$ and subject to $\sum_{k=1}^{K} \alpha_k = 1$.

*b) M-Step:* The derivative of $Q\left(\boldsymbol{\Theta}|\boldsymbol{\Theta'}\right)$ w.r.t to $h_{k,n}$ writes

$$\nabla_{h_{k,n}} Q\left(\boldsymbol{\Theta}|\boldsymbol{\Theta'}\right) = \bar{\alpha}'_{k,n} \nabla_{h_{k,n}} \log p(\mathbf{c}_n|h_{k,n}, \mathbf{w}_k), \tag{20}$$

so that updating $h_{k,n}$ amounts to solving

$$\nabla_{h_{k,n}} \log p(\mathbf{c}_n | h_{k,n}, \mathbf{w}_k) = 0, \tag{21}$$

which does not involve the current parameter estimate $\boldsymbol{\Theta}'$. Therefore, the parameter $\mathbf{H}$ can be precomputed and does not need to be updated during the EM iterations. Note that the estimation $\mathbf{H}$ is equivalent to that of Eq. (3) in the deterministic baseline method. The expressions of the scale parameters $h_{k,n}$ are presented on Table I.

Regarding the optimization of parameter $\boldsymbol{\alpha}$, the sum constraint can be handled with the introduction of a Lagrangian term, leading to the following update:

$$\alpha_k = \frac{\sum_{n=1}^{N} \bar{\alpha}'_{k,n}}{\sum_{l=1}^{K} \sum_{n=1}^{N} \bar{\alpha}'_{l,n}}. \tag{22}$$

The resulting EM algorithm is summarized below. In the following we will refer to our probabilistic approach as PCR (standing for probabilistic chord recognition).

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

```
Input: Chromagram data $\mathbf{C} = [\mathbf{c}_1, \ldots, \mathbf{c}_N]$, chord
       templates $\mathbf{W} = [\mathbf{w}_1, \ldots, \mathbf{w}_K]$
Output: Chord probabilities $\boldsymbol{\alpha} = [\alpha_1, \ldots, \alpha_K]$
Initialize $\boldsymbol{\alpha}$
Compute scale parameters $\mathbf{H}$ as in Eq. (21)
for $i = 1 : n_{iter}$ do
    $\bar{\alpha}_{k,n}^{(i-1)} = \frac{p(\mathbf{c}_n|h_{k,n},\mathbf{w}_k)\,\alpha_k^{(i-1)}}{\sum_{l=1}^{K} p(\mathbf{c}_n|h_{l,n},\mathbf{w}_l)\,\alpha_l^{(i-1)}}$    // E-Step
    $\alpha_k^{(i)} = \frac{\sum_{n=1}^{N} \bar{\alpha}_{k,n}^{(i-1)}}{\sum_{l=1}^{K} \sum_{n=1}^{N} \bar{\alpha}_{l,n}^{(i-1)}}$    // M-Step
```

### C. Chord recognition under the probabilistic model

As already discussed in Section III-A, our chord recognition criterion is based on the frame-by-frame maximum state posterior probability, i.e.,

$$\hat{\gamma}_n = \underset{k}{\operatorname{argmax}} \, \{\bar{\alpha}_{k,n}\}_k. \tag{23}$$

Note that the state posterior probabilities are readily available from within the EM algorithm. Just like in the baseline method, this frame-by-frame chord recognition system can be improved by taking into account the long-term trend in the chord changes. We therefore propose to use an *ad hoc* filtering process that implicitly informs the system of the expected chord duration. The post-processing filtering is performed on the state posterior probabilities $\bar{\alpha}_{k,n}$ and not on the chromagram, as in [2], [7], [8], or on the detected chord sequence as in [8].

## IV. Evaluation

### A. Corpus

As for evaluation, we first consider the Beatles corpus, composed of all 13 albums of the Beatles (180 songs, PCM 44100 Hz, 16 bits, mono). This database has been extensively used for the evaluation of many chord recognition systems, in particular those presented at MIREX 2008 and 2009 for the Audio Chord Detection task [28], [29]. The annotation files are provided by Christopher Harte [30]. A total of 17 types of chords are used (maj, dim, aug, maj7, 7, dim7, hdim7, maj6, 9, maj9, sus4, sus2, min, min7, minmaj7, min6, min9) among with one 'no chord' label (N) corresponding to silences or untuned material. The alignment between annotations and audio files is performed with an algorithm also provided by Christopher Harte.

The second corpus was provided to us by the QUAERO project[2]. It consists of 20 musical pieces from commercial recordings annotated by IRCAM (PCM 22050 Hz, 16 bits, mono) from various artists (Pink Floyd, Queen, Buenavista Social Club, Dusty Springfield, Aerosmith, Shack, UB40, Fall Out Boy, Nelly Furtado, Justin Timberlake, Mariah Carey, Abba, Cher, Phil Collins, Santa Esmeralda, Sweet, FR David and Enya) and various genres (pop, rock, electro, salsa, disco,...). The corpus only contains major and minor labels. More details about this corpus can be found in [31].

### B. Chord dictionary

The evaluation protocol we use in this paper relies on the one used in MIREX 08 & 09 [28], [29]. Since major and minor chords are prominent in pop music, the evaluation is only based on a 25-chord dictionary: 12 major chord labels, 12 minor chords labels and one 'N' label corresponding to silences or untuned material.

The Beatles annotation files are therefore mapped to major and minor types following these rules (used in MIREX 08 & 09) [28], [29]:

- major: maj, dim, aug, maj7, 7, dim7, hdim7, maj6, 9, maj9, sus4, sus2
- minor: min, min7, minmaj7, min6, min9

Since the QUAERO corpus contains only major and minor chords already, no mapping is necessary.

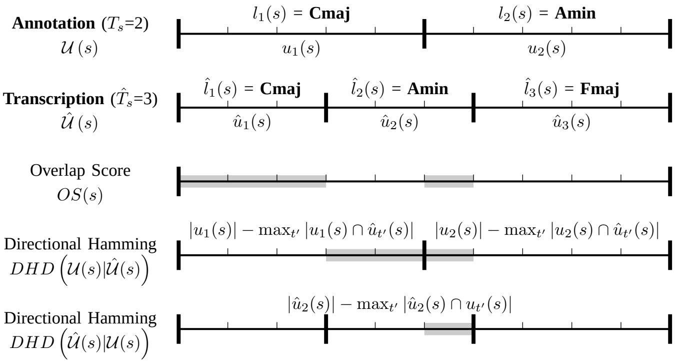
### C. Metrics

The chord transcription task is actually the fusion of two subtasks: a recognition task (find the correct label for each frame) and a segmentation task (find the correct chord boundaries). Also a good transcription is supposed to be compact and to use a sparse chord vocabulary. We here list some metrics in order to evaluate not only the quality of transcription but also the accuracy of segmentation and of chord vocabulary, as described below.

Let $\mathcal{S}$ be a corpus composed of $S$ songs. In the annotation files, each song $s$ is segmented with $T_s$ temporal segments $\mathcal{U}(s) = \{u_1(s), \ldots, u_{T_s}(s)\}$. For each segment $u_t(s)$, the annotation files provide a chord label $l_t(s)$.

Let us denote $|u|$ the duration of segment $u$ and $u \cap u'$ the intersection of segments $u$ and $u'$. Hence the total length of the song $s$ is $|s| = \sum_{t=1}^{T_s} |u_t(s)|$.

[2] QUAERO project: http://www.quaero.org


Annotation ($T_s$=2)
$\mathcal{U}(s)$
$l_1(s)$ = Cmaj
$l_2(s)$ = Amin
$u_1(s)$
$u_2(s)$
Transcription ($\hat{T}_s$=3)
$\hat{\mathcal{U}}(s)$
$\hat{l}_1(s)$ = Cmaj
$\hat{l}_2(s)$ = Amin
$\hat{l}_3(s)$ = Fmaj
$\hat{u}_1(s)$
$\hat{u}_2(s)$
$\hat{u}_3(s)$
Overlap Score
$OS(s)$
Directional Hamming
$DHD\left(\mathcal{U}(s)|\hat{\mathcal{U}}(s)\right)$
$|u_1(s)| - \max_{t'} |u_1(s) \cap \hat{u}_{t'}(s)|$
$|u_2(s)| - \max_{t'} |u_2(s) \cap \hat{u}_{t'}(s)|$
Directional Hamming
$DHD\left(\hat{\mathcal{U}}(s)|\mathcal{U}(s)\right)$
$|\hat{u}_2(s)| - \max_{t'} |\hat{u}_2(s) \cap u_{t'}(s)|$

Overlap Score $OS(s) = \frac{3+1}{10} = 0.4$

Hamming Distance $HD(s) = \frac{1}{2} \times \left(\frac{2+1}{10} + \frac{1}{10}\right) = 0.2$

Reduced Chord Length $RCL(s) = \frac{\frac{3+3+4}{3}}{\frac{5+5}{2}} = \frac{2}{3}$

Reduced Chord Label Number $RCLN(s) = \frac{3}{2}$

False Chord Label Number $FCLN(s) = 1$

Fig. 2. Example of calculation of Overlap Score, Hamming Distance, Reduced Chord Length, Reduced Chord Number and False Chord Label Number. The figure uses a discrete "clock" for purpose of illustration but in reality the time scale is continuous up to the sample period.

With our transcription method, each song $s$ is divided into $\hat{T}_s$ segments, and every segment $\hat{u}_t(s)$ is given a chord label $\hat{l}_t(s)$.

*1) Performance metrics:* Our primary goal is to evaluate the accuracy of the chord labels attributed by our method. *Overlap Score* $OS(s)$ is defined as the ratio between the length of the correctly analyzed chords and the total length of the song, i.e.,

$$OS(s) = \frac{\sum_{t=1}^{T_s} \sum_{t'=1}^{\hat{T}_s} |u_t(s) \cap \hat{u}_{t'}(s)|_{l_t(s)=\hat{l}_{t'}(s)}}{|s|}. \tag{24}$$

This Overlap Score ranges from 0 to 1. The higher the score, the better the performance.

The *Average Overlap Score (AOS)*, which has been used for MIREX 2008 [28], is the mean of the Overlap Scores $OS(s)$ of the corpus:

$$AOS = \frac{1}{S} \sum_{s=1}^{S} OS(s). \tag{25}$$

The chord recognition task can be seen as the joint recognition of chord root and chord type. Another metric can also be defined: the *Average Root Overlap Score (AROS)*, which is defined just like the AOS, but only assesses root detection.

*2) Segmentation metrics:* In order to evaluate the segmentation quality, recent publications [11] have used the *Hamming Distance (HD)* calculated from the *Directional Hamming Divergence (DHD)* [32]. The DHD reflects the unfitness of one segmentation to another. The directional Hamming divergence between the annotation segmentation $\mathcal{U}(s)$ and the transcription segmentation $\hat{\mathcal{U}}(s)$ is defined as:

$$DHD\left(\mathcal{U}(s)|\hat{\mathcal{U}}(s)\right) = \frac{\sum_{t=1}^{T_s} |u_t(s)| - \max_{t'} |u_t(s) \cap \hat{u}_{t'}(s)|}{|s|}. \tag{26}$$

The inverse directional Hamming divergence is defined as:

$$DHD\left(\hat{\mathcal{U}}(s)|\mathcal{U}(s)\right) = \frac{\sum_{t=1}^{\hat{T}_s} |\hat{u}_t(s)| - \max_{t'} |\hat{u}_t(s) \cap u_{t'}(s)|}{|s|}. \tag{27}$$

Finally, the Hamming distance between the two segmentations is defined as the mean of the two directional Hamming divergences:

$$HD(s) = \frac{DHD\left(\mathcal{U}(s)|\hat{\mathcal{U}}(s)\right) + DHD\left(\hat{\mathcal{U}}(s)|\mathcal{U}(s)\right)}{2}. \tag{28}$$

The Hamming Distance tends to reflect the dissimilarity of two segmentations: this metric takes values between 0 and 1 and the lower the value, the better the segmentation quality. In particular, a value of 0 is obtained when both segmentations are exactly the same. The mean of all the Hamming Distances of the corpus is called Average Hamming Distance (AHD).

*3) Fragmentation metrics:* A chord transcription is expected to display "compactness". Indeed, the presence of numerous fragmented chords can lead to noisy and hardly understandable transcriptions. In order to evaluate whether a chord recognition method produces fragmented transcriptions or not, we propose a new metric that we coin *Average Chord Length (ACL)*. Let us first define for a song $s$, the *Reduced Chord Length* $RCL(s)$ as the ratio between the experimental average chord duration and the duration of the ground truth, i.e.,

$$RCL(s) = \frac{\frac{1}{\hat{T}_s} \sum_{t=1}^{\hat{T}_s} |\hat{u}_t(s)|}{\frac{1}{T_s} \sum_{t=1}^{T_s} |u_t(s)|}. \tag{29}$$

Note that this score can also be defined as the ratio between $T_s$ and $\hat{T}_s$. This metric should be as close to 1 as possible; when lower than 1, the transcriber tends to overfragment the piece. We define the Average Chord Length as the mean of all the Reduced Chord Lengths of the corpus.

*4) Chord vocabulary metrics:* Another indicator of the quality of a chord transcription is the compactness of the chord vocabulary used for the transcription. Remember that for each song the chord vocabulary is the subset of the chord dictionary needed to transcribe that song. When the song relates to a specific key, it reflects by extension the tonal context of the piece, but by chord vocabulary we mean a wider notion than key. We define two metrics for assessing the correctness of the estimated chord vocabulary: the *Average Chord Label Number (ACLN)* and the *Average False Chord Label Number (AFCLN)*.

Given a song $s$, we define the *Reduced Chord Label Number* $RCLN(s)$ as the ratio between the number of different chord labels used for the transcription and in the ground truth annotation. Better results are obtained when this metric approaches 1. When greater than 1, the transcription uses a too wide chord vocabulary. The Average Chord Label Number is the mean of all Reduced Chord Label Numbers of the corpus.

The Average False Chord Label Number is the average number of chord labels that do not belong to the annotation files. It should be as low as possible: an AFCLN of 0 would indicate that the method always detects the correct chord vocabulary.

These metrics are illustrated on a small example in Figure 2.

## V. Results

### A. Experimental setup

We use the chromagram proposed by Bello & Pickens [8] as the input to our system. Other chromagrams were considered in preliminary studies (in particular the one described in [8], [7], [33]) but we found Bello & Pickens' chromagram to give the best results with our system. We used the code kindly provided by the authors. The window length is 743 ms and the hop size is set to 93 ms. Their implementation also performs a silence ('no chord') detection using an empirically set threshold on the energy of the chroma vectors. More details about the calculation of the chromagram can be found in [8].

For our new probabilistic methods, two sets of parameters are to be chosen: the probability distribution parameters ($\sigma^2$ for the Gaussian model and $\beta$ for the Gamma model), and the post-processing filtering parameters. For each of these observation distributions, extensive simulations have been done in order to find the optimal 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 the chromagrams.

The post-processing methods and neighborhood sizes used here are chosen in order to optimize the value of the Average Overlap Score on the Beatles corpus. Nevertheless, there is not much difference between two filtering methods or close neighborhood sizes.

The experimental parameters used for our PCR methods are as follows.

- **Gaussian additive noise model**: $\sigma^2 = 0.02$ and median filtering on 17 frames (2.23s) ;
- **Gamma multiplicative noise model**: $\beta = 3$ and low-pass filtering on 15 frames (2.04s) ;
- **Poisson noise**: median filtering on 13 frames (1.86s).

We will respectively refer to the method based on the above models as PCR/Gaussian, PCR/Gamma, PCR/Poisson.

### B. Example on one Beatles song

Before giving the overall results, we propose here to investigate the differences between the deterministic and probabilistic approaches on one example. We chose the Beatles' song *Run for your life* from the album *Rubber Soul* as its transcription demonstrates the improvements brought by the probabilistic approach over the deterministic one. As such, Figure 3 displays the transcription obtained with PCR/Gamma and with OGF1, the latter being one of the two deterministic algorithms that we submitted at MIREX'09.[3] A first observation is that PCR/Gamma gives better results. Indeed, the Overlap Scores are respectively 0.733 and 0.919 for OGF1 and PCR/Gamma, showing a clear improvement. A careful examination of both transcriptions suggests three explanations for this improvement:

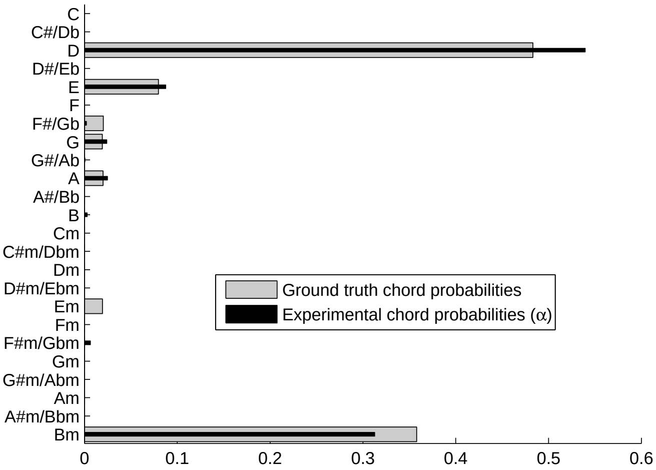


Fig. 4. Ground-truth and estimated chord probability distributions $\hat{\alpha}$ for the Beatles song *Run for your life*.

- PCR/Gamma detects chord boundaries more accurately than OGF1.
- PCR/Gamma seems to detect longer chords while OGF1 gives very fragmented results.
- The chord vocabulary used in the transcription output by PCR/Gamma is sparser than the one returned by OGF1, preventing in particular from some major-minor confusions.

The evaluation metrics computed for this song confirm these observations:

- The Hamming Distances are respectively 0.206 and 0.060, which reflects the fact that the segmentation provided by PCR/Gamma is very similar to the one described in the annotation files.
- The Reduced Chord Lengths are respectively 0.345 and 1.085, which shows that PCR/Gamma better evaluates the chord length.
- The chord vocabulary used by PCR/Gamma is smaller than OGF1's one: the Reduced Chord Label Numbers for the two methods are 1.75 and 0.75 respectively. Since the second value is closer to 1 than the first one, the number of chords used by PCR/Gamma is the most accurate. The calculation of the False Chord Label Numbers confirms this: they are respectively equal to 6 and 0, which means that the transcription provided by PCR/Gamma does not use any chord labels that were not present in the annotation files.

PCR/Gamma efficiently estimate the chord vocabulary thanks to the parameter $\boldsymbol{\alpha}$ that models, for each song, the chord probability distribution. This is illustrated on Figure 4 which displays the estimated chord probabilities (vector $\boldsymbol{\alpha}$) on top of the empirical normalized chord length histogram computed from the annotation file. We may see that they closely fit, which confirms our assumption regarding the

[3] In essence, OGF1 corresponds to the baseline system described in Section II based on a KL divergence and median filtering with smoothing window of size 15 and a dictionary composed only of major and minor chords. For details see [21], [22], [29].

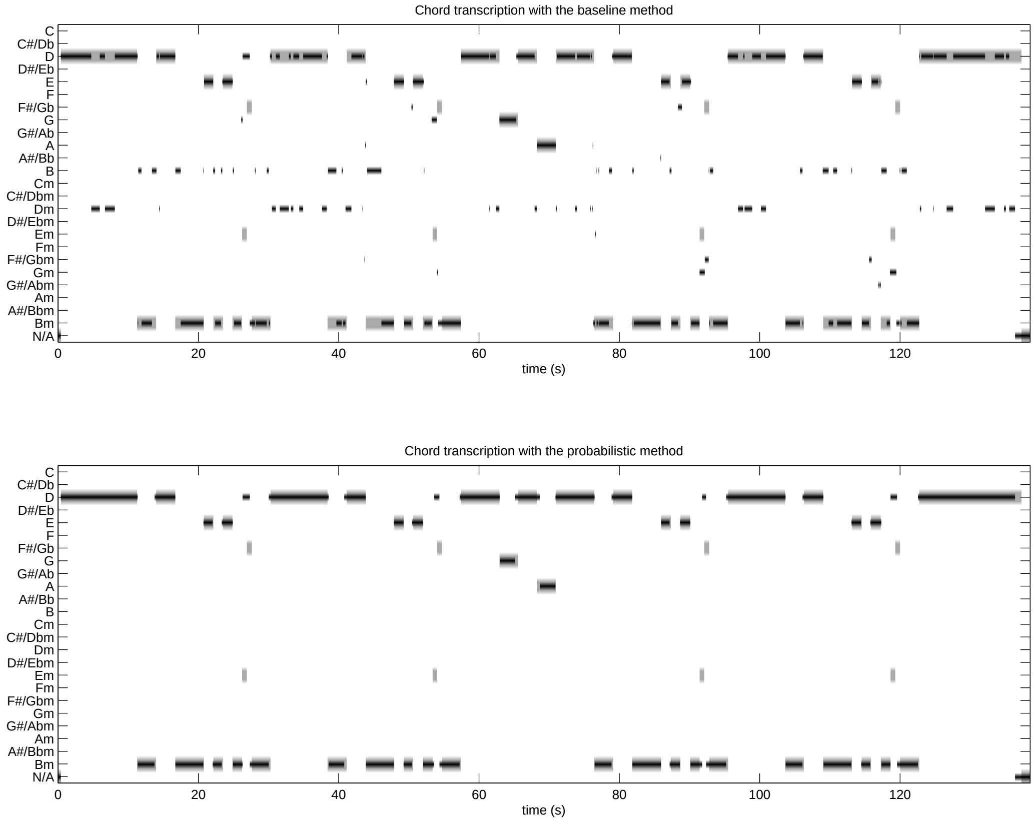


Fig. 3. Examples of chord transcription on the Beatles song *Run for your life*. The estimated chord labels are displayed in black while the ground-truth chord annotation is in gray. Top figure: transcription with the baseline deterministic approach (OGF1), bottom figure: transcription with the probabilistic approach (PCR/Gamma).

TABLE II
COMPARISON WITH THE STATE-OF-THE-ART ON THE BEATLES CORPUS

| | MIREX 2008 | | | MIREX 2009 | | | | | PCR methods | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | BP | RK | PVM | KO1 | KO2 | DE | OGF1 | OGF2 | Gaussian | Gamma | Poisson |
| Average Overlap Score (AOS) | 0.707 | 0.705 | 0.648 | 0.722 | 0.734 | 0.738 | 0.714 | 0.724 | 0.749 | **0.758** | 0.744 |
| Average Root Overlap Score (AROS) | 0.740 | 0.763 | 0.680 | 0.754 | 0.761 | 0.772 | 0.775 | 0.783 | 0.785 | **0.787** | 0.775 |
| Average Hamming Distance (AHD) | 0.153 | 0.146 | 0.209 | 0.152 | 0.150 | 0.156 | 0.163 | 0.152 | **0.146** | 0.149 | 0.156 |
| Average Chord Length (ACL) | 0.941 | 1.074 | 0.422 | 1.169 | 1.168 | 0.890 | 0.552 | 0.717 | 0.872 | 0.920 | **1.057** |
| Average Chord Label Number (ACLN) | 1.441 | 1.414 | 2.285 | 1.507 | 1.319 | 1.667 | 2.070 | 1.693 | 1.314 | 1.185 | **1.012** |
| Average False Chord Label Number (AFCLN) | 3.560 | 3.330 | 8.490 | 3.760 | 2.590 | 4.590 | 7.390 | 4.990 | 2.640 | 1.860 | **1.060** |
| Run time (in seconds) | 1619 | 2241 | 12402 | 6382[1] | 6382[1] | 1403 | 790 | 796 | 480 | 482 | 486 |

accurate estimation of the chord vocabulary.

### C. Comparison with state-of-the-art

Our methods are now compared to some state-of-the-art systems according to the metrics defined in Section IV. These methods have all been tested with their original implementations and have all participated in MIREX 2008 [28] or 2009 [29].

**MIREX 2008:**

- BP: Bello & Pickens [8]
- RK: Ryynänen & Klapuri [17]
- PVM: Pauwels, Verewyck & Martens [34]

**MIREX 2009:**

- KO1 & KO2: Khadkevich & Omologo [18]
- DE: Ellis [14]
- OGF1 & OGF2: our *baseline method* [21], [22]

More details about these methods can be found in the given references or from the corresponding MIREX websites [28], [29].

*1) Beatles corpus:* Table II presents the results obtained by these 11 chord recognition systems on the Beatles corpus.

Quantitative scores such as AOS or AROS show that our probabilistic approach slightly outperforms state-of-the art: the AOS we obtain with PCR/Gamma is indeed 2% larger than the best score (DE). Since all the scores are close, it is interesting

[1] As the KO1 and KO2 methods are run together because they share common resources, we here report the total running time.

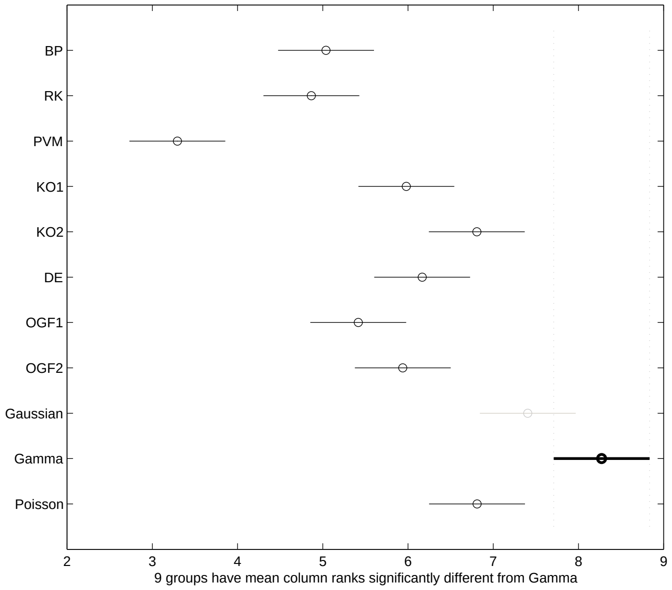


Fig. 5. Tukey-Kramer's test performed on Overlap Scores calculated on the Beatles corpus. The x-axis shows the average rank for each chord recognition method along with the comparison interval.

to figure out whether the methods are significantly different from each other. We therefore propose to perform a Tukey-Kramer's test [35], [36] on the Overlap Scores. This test was notably run in MIREX 2008 & 2009 and compares the average rank of every chord recognition method to the average rank of every other system. Results of the Tukey-Kramer's test are displayed on Figure 5. It shows that the improvement brought by our PCR approach is significant. Indeed PCR/Gamma is significantly better than all the other tested methods, except for PCR/Gaussian. The latter method performs significantly better than all the methods except for KO2, PCR/Gamma and PCR/Poisson. Finally, PCR/Poisson is significantly better than BP, RK, PVM and OGF1. In particular, PCR/Gamma outperforms OGF1 on 148 songs over 180.

The introduction of other evaluation metrics allows us to compare chord recognition methods according to several criteria. Indeed, the 4 other metrics tend to evaluate the segmentation, the fragmentation and the good detection of the chord vocabulary.

The segmentation, i.e., the detection of the chord boundaries, is measured by the AHD (that better be as low as possible). We notice that, except for the PVM method, all the AHD values are close (around 0.15). Indeed, statistical tests are rather inconclusive: a Tukey-Kramer's test shows that except for the PVM method, they are no strong differences between the chord recognition methods. For example, the method obtaining the best AHD (PCR/Gaussian), is only significantly different from PVM, OGF1 & DE.

The fragmentation is evaluated thanks to the ACL (that better be as close to 1 as possible). Some methods seem to slightly overestimate the chord length (RK, KO1, KO2 & PCR/Poisson), but most of them tend to over-fragment the chords. Some methods (PVM, OGF1) even detect chords with half their real duration. On the contrary, our PCR methods seem to avoid this fragmentation effect and the best results are obtained with PCR/Gamma.

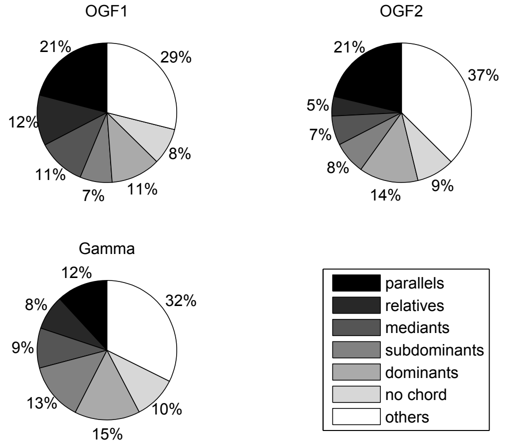


Fig. 6. Error distribution of OGF1, OGF2 and PCR/Gamma on the Beatles corpus. As a reminder, the scores of each is 0.714, 0.724 and 0.758 respectively. Refer to [22] for a detailed description of the error types.

One of the main contributions of the probabilistic framework is the explicit evaluation of the song chord vocabulary. The good detection of this chord vocabulary is described by two metrics: the ACLN (that better be as close to 1 as possible) and the AFCLN (that better be as low as possible). We notice that all methods seem to over-evaluate the number of chords. The two methods PVM and OGF1 seem to be particularly penalized by this phenomenon. Our 3 PCR methods, on the contrary, reliably evaluate the chord vocabulary: they obtain the 3 best scores. AFCLN scores confirm these results: the introduction of chord probabilities allows to correctly capture the chord vocabulary of every song.

As the PCR approach does not require any training nor side-information (besides the chromagram data and the statistical model specification) its computation time is quite low. Thanks to some code optimization, our PCR methods perform even faster than the baseline methods OGF1 & OGF2, and are therefore twice as fast as other state-of-the-art methods.

In previous work [22], we have presented an analysis of the errors commonly made by chord recognition methods: we propose here to conduct the same analysis for our probabilistic approach. Figure 6 presents the distribution of error sources for PCR/Gamma and for the two deterministic methods OGF1 & OGF2. Five types of common errors are emphasized, corresponding either to structural similarity or harmonic proximity situations (see [22] for details). We notice that parallel errors, which were very common with OGF1 & OGF2, do not seem to be as important with PCR/Gamma. This is an interesting observation as these errors were related to the template-based character of our methods, in which chords were likely to be mistaken one for another when they had notes in common. The improvement on this aspect is probably explained by the capacity of our system to efficiently evaluate the chord vocabulary, leading to a lower number of major-minor confusions.

TABLE III
COMPARISON WITH THE STATE-OF-THE-ART ON THE QUAERO CORPUS

| | MIREX 2008 | | | MIREX 2009 | | | | | PCR methods | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | BP | RK | PVM | KO1 | KO2 | DE | OGF1 | OGF2 | Gaussian | Gamma | Poisson |
| Average Overlap Score (AOS) | 0.699 | 0.730 | 0.664 | 0.670 | 0.665 | 0.719 | 0.707 | 0.682 | 0.739 | **0.773** | 0.760 |
| Average Hamming Distance (AHD) | 0.142 | **0.117** | 0.175 | 0.153 | 0.156 | 0.127 | 0.142 | 0.137 | 0.131 | 0.124 | 0.130 |
| Average Chord Length (ACL) | 0.903 | **1.021** | 0.494 | 1.084 | 1.109 | 0.823 | 0.565 | 0.683 | 0.835 | 0.896 | 0.806 |
| Average Chord Label Number (ACLN) | 1.559 | 1.516 | 2.323 | 1.549 | 1.351 | 1.906 | 2.297 | 1.970 | 1.529 | 1.336 | **1.138** |
| Average False Chord Label Number (AFCLN) | 3.650 | 3.250 | 7.850 | 3.600 | 2.550 | 5.300 | 7.700 | 5.850 | 3.150 | 2.150 | **1.150** |

*2) QUAERO corpus:* Recent publications such as [11] have discussed the necessity of testing chord recognition methods on other corpus than the popular Beatles corpus. We have therefore also run all the tested systems on the QUAERO corpus and the results are displayed on Table III.

A first observation is that except for RK, PVM, PCR/Gamma and PCR/Poisson, all the methods get lower AOS on this corpus than on the Beatles data. Once again, our probabilistic methods give the best results: in particular, PCR/Gamma performs even better than on the Beatles corpus. Although the small number of songs in the corpus does not allow one to perform a real significant difference test, the AOS obtained by PCR/Gamma is 4% higher than the best state-of-the-art method (RK), which is a large difference when looking at the scores. As far as segmentation is concerned, the RK method gives the best results but nevertheless PCR/Gamma gives the second best AHD result. Once again, most of the chord recognition methods tend to underestimate the chord length: however PCR/Gamma gives the fourth best ACL score. Finally, we observe with the ACLN and AFCLN metrics that our PCR methods still outperform other state-of-the-art methods on the chord vocabulary estimation.

More importantly, these results tend to answer the overfitting concern related to the Beatles corpus, since our methods achieve better performances on the QUAERO corpus than on the Beatles one. In addition, music genre and style do not seem to influence our systems, as all the calculated scores for the Beatles corpus and the QUAERO corpus are close.

## VI. CONCLUSION

In this paper we presented a novel probabilistic framework for template-based chord recognition. In comparison with our previous work the key ingredient of our new approach is the introduction of chord probabilities, learned from the song. This tends to produce more accurate chord vocabulary (and in particular more compact), i.e., to eliminate many of the spurious chords that appeared in the transcriptions produced with our baseline deterministic approach. Indeed, the probability of occurrence of the spurious chords is automatically driven to zero and the chords are hence smoothed out of the transcription. This translates into both better recognition and segmentation scores. Interestingly, the vector of estimated chord probabilities reflects the “harmonic profile” of the song and may be of interest for applications such as key estimation or can serve as a descriptor for MIR tasks.

As for perspective we envisage the following lines of work. Firstly, of the three considered observation noise models, the Gamma multiplicative noise model appeared to lead to best results. This model requires the tuning of an extra shape parameter that we handled by trial-and-error. The automatic estimation of this parameter is not trivial but could be envisaged with numerical methods. This may improve even more recognition scores and also yield another relevant descriptor of the song. Secondly, in this work we have taken into account the long-term variations in the song using an ad-hoc post-processing filtering of the states posterior distributions. Future work will consider more sophisticated models that improve the probabilistic model so as to more adequately model the rhythmic structure of music.

## ACKNOWLEDGMENT

The authors would like to thank J. Bello, D. Ellis, M. Khadkevich, J. Pauwels and M. Ryynänen for making their code available. We also wish to thank C. Harte for his very useful annotation files.

This work was supported in part by ANR-09-JCJC-0073-01 TANGERINE (Theory and applications of nonnegative matrix factorization) and by the Quaero Programme, funded by OSEO, French State agency for innovation.

## APPENDIX
## EXPRESSIONS OF STANDARD PROBABILITY DISTRIBUTIONS

| | |
|---|---|
| Gaussian | $\mathcal{N}\left(x;\mu,\sigma^2\right) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$ |
| Gamma | $\mathcal{G}\left(x;\alpha,\beta\right) = \frac{\beta^\alpha}{\Gamma(\alpha)} e^{-\beta x}$ |
| Poisson | $\mathcal{P}\left(x;\lambda\right) = \frac{\lambda^x}{\Gamma(x+1)} e^{-\lambda}$ |

where $\Gamma$ is the Gamma function.

## REFERENCES

[1] E. Taylor, *The AB Guide to Music Theory Part 1.* Associated Board of the Royal Schools of Music, 1989.

[2] 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.

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

[4] K. Lee, “Automatic chord recognition from audio using enhanced pitch class profile,” in *Proceedings of the International Computer Music Conference (ICMC)*, New Orleans, USA, 2006.

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

[6] 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.

[7] G. Peeters, "Musical key estimation of audio signal based on hidden Markov modeling of chroma vectors," in *Proceedings of the International Conference on Digital Audio Effects (DAFx)*, Montreal, Canada, 2006, pp. 127–131.

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

[9] H. Papadopoulos and G. Peeters, "Chord estimation using chord templates and HMM," Abstract of the Music Information Retrieval Evaluation Exchange, 2008.

[10] M. Mauch, K. Noland, and S. Dixon, "Using musical structure to enhance automatic chord transcription," in *Proceedings of the International Society for Music Information Retrieval Conference (ISMIR)*, Kobe, Japan, 2009, pp. 231–236.

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

[12] 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.

[13] C. Sailer and K. Rosenbauer, "A bottom-up approach to chord detection," in *Proceedings of the International Computer Music Conference (ICMC)*, New Orleans, USA, 2006, pp. 612–615.

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

[15] J. Burgoyne, L. Pugin, C. Kereliuk, and I. Fujinaga, "A cross-validated study of modelling strategies for automatic chord recognition in audio," in *Proceedings of the International Conference on Music Information Retrieval (ISMIR)*, Vienna, Austria, 2007, pp. 251–254.

[16] 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.

[17] 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.

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

[19] 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 Conference on Music Information Retrieval (ISMIR)*, Barcelona, Spain, 2004, pp. 100–105.

[20] 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 Conference on Music Information Retrieval (ISMIR)*, Philadelphia, USA, 2008, pp. 39–44.

[21] 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.

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

[23] B. Pardo and W. Birmingham, "Algorithms for chordal analysis," *Computer Music Journal*, vol. 26, no. 2, pp. 27–49, 2002.

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

[25] C. Févotte, N. Bertin, and J. Durrieu, "Nonnegative matrix factorization with the Itakura-Saito divergence. With application to music analysis," *Neural Computation*, vol. 21, no. 3, pp. 793–830, 2009.

[26] S. Kullback and R. Leibler, "On information and sufficiency," *Annals of Mathematical Statistics*, vol. 22, no. 1, pp. 79–86, 1951.

[27] A. Dempster, N. Laird, and D. B. Rubin, "Maximum likelihood from incomplete data via the EM algorithm (with discussion)," *Journal of the Royal Statistical Society. Series B (Methodological).*, vol. 19, no. 1, pp. 1–38, 1977.

[28] http://www.music-ir.org/mirex/wiki/2008:Audio_Chord_Detection.

[29] http://www.music-ir.org/mirex/wiki/2009:Audio_Chord_Detection.

[30] 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 Conference on Music Information Retrieval (ISMIR)*, London, UK, 2005, pp. 66–71.

[31] L. Oudre, "Template-based chord recognition from audio signals," Ph.D. dissertation, TELECOM ParisTech, 2010.

[32] S. Abdallah, K. Noland, M. Sandler, M. Casey, and C. Rhodes, "Theory and evaluation of a Bayesian music structure extractor," in *Proceedings of the International Conference on Music Information Retrieval (ISMIR)*, London, UK, 2005, pp. 420–425.

[33] Y. Zhu, M. Kankanhalli, and S. Gao, "Music key detection for musical audio," in *Proceedings of the International Conference on Multimedia Modeling (MMM)*, Melbourne, Australia, 2005, pp. 30–37.

[34] J. Pauwels, M. Varewyck, and J.-P. Martens, "Audio chord extraction using a probabilistic model," Abstract of the Music Information Retrieval Evaluation Exchange, 2008, http://www.music-ir.org/mirex/abstracts/2008/mirex2008-audio_chord_detection-ghent_university-johan_pauwels.pdf.

[35] J. Tukey, "The problem of multiple comparisons." Princeton University, 1953, unpublished manuscript.

[36] Kramer, "Extension of multiple range tests to group means with unequal number of replications," *Biometrics*, vol. 12, no. 3, pp. 307–310, 1956.

**Laurent Oudre (M'10)** was born in France in 1985. He graduated from Supélec, Gif-sur-Yvette, France in 2007 and received the M.Sc. degree in Communications and Signal Processing at Imperial College London, UK in 2007. He received his Ph.D. degree in Signal Processing at TELECOM ParisTech, Paris, France in 2010. Since November 2010, he is a post-doctoral researcher in statistics at TELECOM ParisTech, Paris, France.

His research interests forcus on signal processing and its applications (sound and image).

**Cédric Févotte (M'09)** obtained the State Engineering degree and the MSc degree in Control and Computer Science from École Centrale de Nantes (France) in 2000, and then the PhD degree in 2003. As a PhD student he was with the Signal Processing Group at Institut de Recherche en Communication et Cybernétique de Nantes where he worked on time-frequency approaches to blind source separation. From 2003 to 2006 he was a research associate with the Signal Processing Laboratory at University of Cambridge (Engineering Dept) where he got acquainted with Bayesian approaches to audio signal processing tasks such as audio source separation and denoising. He was then a research engineer with the start-up company Mist-Technologies (now Audionamix) in Paris, working on mono/stereo to 5.1 surround sound upmix solutions. In Mar. 2007, he joined TELECOM ParisTech, first as a research associate and then as a CNRS tenured research scientist in Nov. 2007.

His research interests generally concern statistical signal processing and unsupervised machine learning with audio applications.

**Yves Grenier (M'81)** Yves Grenier was born in Ham, Somme, France, in 1950. He received the Ingénieur degree from Ecole Centrale de Paris, in 1972, the Docteur-Ingénieur degree from Ecole Nationale Supérieure des Télécommunications (now called Télécom ParisTech), Paris, in 1977, and the Doctorat d'Etat es Sciences Physiques, from University of Paris-Sud in 1984.

He has been with Télécom ParisTech since 1977, as Assistant Professor, and since 1984 as Professor. He has been Head of the Signal and Image Processing Department since january 2005.

From 1974 to 1979, his interests have been in speech recognition, speaker identification and speaker adaptation of recognition systems. Between 1979 and 1988, he has been working on signal modeling, spectral analysis of noisy signals, with applications to speech recognition and synthesis, estimation of nonstationary models, time frequency representations. He created ARMALIB, a signal processing software library that has been incorporated in SIMPA, the signal processing software proposed by GDR-PRC CNRS ISIS.

Since 1988, his research has been devoted to multichannel signal processing: beamforming, source localisation, source separation. He concentrated in particular on applications to audio and acoustics, and to microphone arrays. During this period, he has been involved in European ESPRIT projects (2101 ARS from 1989 to 1992, 6166 FREETEL from 1992 to 1994).

Since 1996, he has been interested in audio signal processing (acoustic echo cancellation, noise reduction, signal separation, microphone arrays, loudspeaker arrays) and in music information retrieval (multi-pitch estimation, chord recognition). He participated to the European project K-Space. He is now participating to: the European NoE 3D-Life, the French-German Quaero project, and the French project Romeo, among others.

He has been co-chairman of the "10th International workshop on acoustic echo and noise control" IWAENC 2006. He has been technical co-chair of the "2010 IEEE International Workshop on Multimedia Signal Processing" MMSP 2010.