MIREX CHORD RECOGNITION SYSTEM
A PROBABILISTIC TEMPLATE-BASED CHORD RECOGNITION METHOD

Laurent Oudre[1], **Cédric Févotte**[2], **Yves Grenier**[1]
[1]Institut TELECOM ; TELECOM ParisTech ; CNRS LTCI
[2]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.

1. 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. [1] 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.

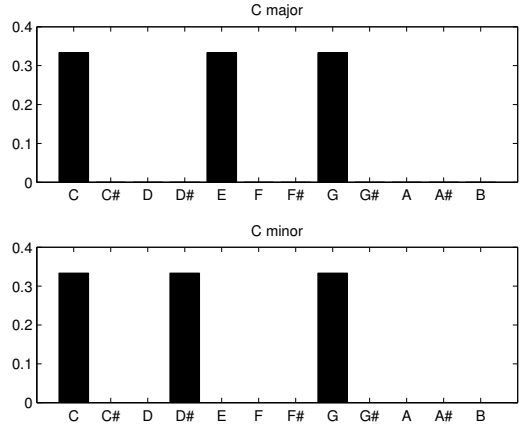

Figure 1. Chord templates for C major and C minor.

2. GENERATIVE MODEL FOR THE CHROMA VECTORS

Let $\mathbf{C}$ be a $12 \times N$ chromagram, composed of N 12-dimensional successive chroma vectors $\mathbf{c}_n$. In practice, the chroma vectors are calculated from the music signal with the same method as Bello & Pickens [1]. 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 12-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 :

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

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 :

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

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

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

the following state-model

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

which can equivalently be written as the following mixture model

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

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 :

$$\hat{\gamma}_n = \underset{k}{\operatorname{argmax}} \, \alpha_{k,n}^{post} \quad (7)$$

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

3. 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) :

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

4. CHORD RECOGNITION WITH THE PROBABILISTIC MODEL

Let the matrix $\alpha_{k,n}^{post}$ 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 2, the detected chord $\hat{\gamma}_n$ for frame n is finally :

$$\hat{\gamma}_n = \underset{k}{\operatorname{argmax}} \, \alpha_{k,n}^{post}. \quad (9)$$

This frame-to-frame chord recognition system can be improved by taking into account the temporal context. We 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 $\alpha_{k,n}^{post}$ in order to take into account the time persistence.

5. MIREX SUBMISSION

The experimental parameters used for our probabilistic method are $\beta = 3$ and low-pass filtering on 15 frames.

6. ACKNOWLEDGMENT

The authors would like to thank C. Harte [2] for his very useful annotation files used during the development phase.

7. REFERENCES

[1] J.P. 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)*, pages 304–311, London, UK, 2005.

[2] 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)*, pages 66–71, London, UK, 2005.