THE IRCAMKEYCHORD SUBMISSION FOR MIREX 2012

Johan Pauwels
STMS IRCAM-CNRS-UPMC
1 Place Igor Stravinsky
75004 Paris, France
pauwels@ircam.fr

Jean-Pierre Martens
ELIS/DSSP - Ghent University
Sint-Pietersnieuwstraat 41
9000 Gent, Belgium
martens@elis.ugent.be

Geoffroy Peeters
STMS IRCAM-CNRS-UPMC
1 Place Igor Stravinsky
75004 Paris, France
peeters@ircam.fr

ABSTRACT

This extended abstract presents the ircamkeychord system which was submitted to the MIREX 2012 tasks of Audio Chord Estimation and Audio Key Detection. It is a knowledge-based system that performs simultaneous estimation of chords and local keys, after which a global key is derived from the local keys. Multiple configurations were submitted that differ only in the musicological information that has been used.

1. OVERVIEW OF THE SYSTEM

The system can be divided into three parts: a feature extraction phase, a smoothing stage and a probabilistic model. First, input audio files are converted to mono, resampled to 8000 Hz and split into frames of 150 ms with a step size of 20 ms. Then a chroma representation is derived which aims to maximally couple higher harmonics to their fundamental frequency [5]. The resulting chroma profiles are sparse and in the ideal case, only contains chromas corresponding to the notes that are actually played. To achieve this, multiple pitch tracking techniques are used. Two chromagrams are computed this way, one where candidate fundamental frequencies are constrained between 100 Hz and 2000 Hz and another one between 55 Hz and 220 Hz. These chromagrams are subsequently smoothed by averaging them over inter-beat intervals as calculated by ircambeat [3]. The smoothed features are then fed into a probabilistic model.

This probabilistic model is an HMM where each state is composed of a key-chord combination. The probabilities of the HMM are not trained through EM or any other machine learning technique, but are derived from a number of submodels that are knowledge based. This decomposition into submodels allows us to explicitly set some dependencies between the different key-chord combinations and to reduce the parameters to a set that is no longer interdependent. The final goal is to derive a sequence of chords $\hat{C}$ and keys $\hat{K}$ that maximize the following expression:

$$\hat{K}, \hat{C} = \arg\max_{K,C} \prod_{n=1}^{N} P(k_n, c_n | k_{n-1}, c_{n-1}) \, P(\mathbf{x}_n | k_n, c_n)$$

where $\mathbf{x}_n$ stands for the feature vector at beat segment n, k is the key label and c the chord label.

The submodels that form the emission probabilities $P(\mathbf{x}_n|k_n, c_n)$ will be called acoustic models in the remainder of the text, while the submodels that make up the transition probabilities will be called musicological models. The probabilities generated by the former only take the current segment into account, while the latter consider the temporal dependency by means of prior musicological information.

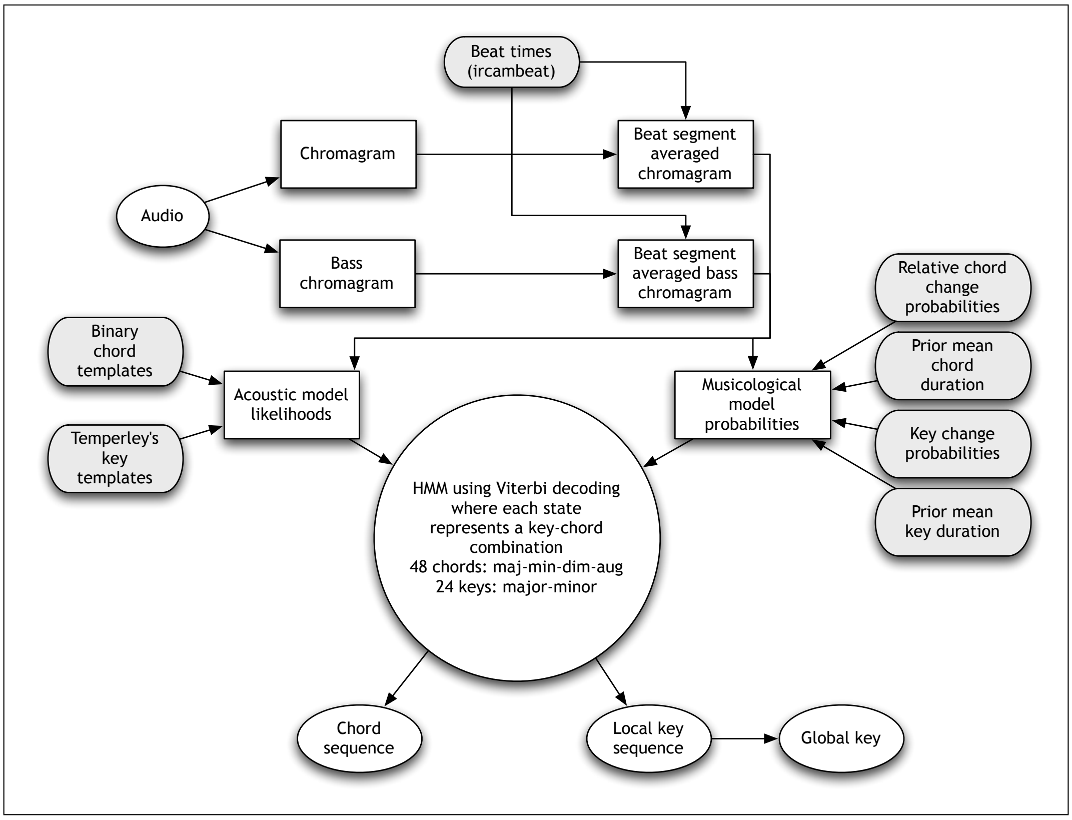
The acoustic model consists of 2 submodels, a key acoustic model and a chord acoustic model

$$P(\mathbf{x}|k_n, c_n) = P(\mathbf{x_n}|k_n) P(\mathbf{x_n}|c_n)$$

Both use template matching where the cosine similarity is used as similarity measure between the smoothed features and a set of templates. For the key acoustic model these are Temperley's key profiles. The chord acoustic model templates are binary and derived from music theory: chromas that should in theory belong to the chord have a value of 1, other chromas 0. Both chromagrams are concatenated to form a 24-dimensions feature vector.

The musicological model consists of 4 components, a key and a chord duration model and a key and a chord change model. The key and chord duration models are simply expressed as self-transition probabilities (respectively P_k and P_c).

For the key change model $P(k_n|k_{n-1})$, we assume that the probability of a key change does not depend on the absolute keys itself, but only on the interval between the two tonics and on their modes. In order to make this more explicit, a variable transformation is introduced. Given that a key k consists of a tonic t and a mode m, we require that the distance between keys $d(k_x, k_y)$ is only a function of m_x, m_y and $i_{x,y}$, where the latter represents the interval between roots. The key change model $P(k_n|k_{n-1})$ can then be reduced to $P(i_{n,n-1}, m_n|m_{n-1})$. As the number of key changes in an annotated corpus is relatively small compared to the current corpus sizes, these values are theoretically derived, based on Lerdahl's regional distance (a more detailed explanation can be found in [1]).



Beat times (ircambeat)
Chromagram
Beat segment averaged chromagram
Audio
Bass chromagram
Beat segment averaged bass chromagram
Relative chord change probabilities
Binary chord templates
Prior mean chord duration
Acoustic model likelihoods
Musicological model probabilities
Temperley's key templates
Key change probabilities
HMM using Viterbi decoding where each state represents a key-chord combination 48 chords: maj-min-dim-aug 24 keys: major-minor
Prior mean key duration
Chord sequence
Local key sequence
Global key

Figure 1. Flowchart of the ircamkeychord system.

For the chord change model $P(c_n|c_{n-1}, k_{n-1})$, another assumption is made, namely that the probability of chord change depends only on the relative chords as expressed in a key (mirroring scholarly analysis, where one speaks about movements between scale degrees). An extra notation is introduced for representing this concept of a relative chord c'_y expressed in the context of a key k_x. The chord change model thus gets reduced to $P(c'_n|c'_{n-1}, m_{n-1})$. The values of this chord change model can either be derived from music theory or can stem from co-occurrence count on a symbolic data set, depending on the configuration used (see below).

Finally, the four submodels are combined as follows, where the extra requirement has been added that a key change can only take place together with a chord change. A balance parameter ($\alpha, \beta, \gamma, \delta$) for each of the four submodels has been introduced in order to regulate the relative importance of each model.

$$\begin{aligned}
P(k_n, & c_n|k_{n-1}, c_{n-1}) \\
&= P_c^{\delta} && (k_n = k_{n-1} \,\&\, c_n = c_{n-1}) \\
&= P_k^{\beta} P\left(c'_n|c'_{n-1}, m_{n-1}\right)^{\gamma} (1-P_c)^{\delta} \\
& && (k_n = k_{n-1} \,\&\, c_n \neq c_{n-1}) \\
&= 0 && (k_n \neq k_{n-1} \,\&\, c_n = c_{n-1}) \\
&= P\left(i_{n,n-1}, m_n|m_{n-1}\right)^{\alpha} (1-P_k)^{\beta} \\
&\quad P\left(c'_n|c'_{n-1}, m_{n-1}\right)^{\gamma} (1-P_c)^{\delta} \\
& && (k_n \neq k_{n-1} \,\&\, c_n \neq c_{n-1})
\end{aligned}$$

The scope of the system consists of 48 chords, namely 4 triads (maj–min–dim–aug) for each chroma, and 24 keys (major and minor modes). The local keys for every beat-synchronized segment are finally combined into one global key by means of majority voting. A flowchart of the system can be found in Figure 1.

2. SYSTEM CONFIGURATIONS

Six different versions have been submitted to MIREX'12, three each for the Audio Key Detection (AKD) and the Audio Chord Estimation (ACE) task. Feature extraction and acoustic models are the same for all 6. Two of these, PMP3 and PMP4 for ACE and AKD respectively, are degenerate configurations that do not perform joint key and chord estimation. Rather, the outputs of the acoustic models are simply smoothed using an HMM where each state represents either a chord or a key and whose transition matrix only contains two values, one for diagonal and one for off-diagonal elements. They are meant as baseline systems that provide a reference to evaluate the other configurations to. The other four configurations are only differing in their chord change models. The models of PMP1, PMP2 and PMP6 are obtained using CO-occurrence counting on symbolic data, respectively the isophonics annotations and the “Popular” and “Classical” subsets of the 9GDB data set [4]. PMP5 uses the theoretically derived model based on Lerdahl's simple chord distance as described in [1].

3. DISCUSSION OF THE RESULTS

Looking at the relative performance of our own systems, PMP1 respectively PMP6 are the best for chord and for key estimation. The fact that PMP1 has the best chord estimation performance on the MIREX'09 data set is expected, as the chord change model was also derived from this set, but it also performs best on the McGill set, even though the 9GDB-Popular set has a broader variety of artists. For key estimation, the co-occurrence counting chord change model of PMP6 outperforms the theoretically based one of PMP5. The reason is probably that the theory on which the latter is based very much restricts changes to chords that are non-diatonic, which seems even for classical music a too strong assumption. The inclusion of musicological information leads to a significant improvement for the task of key estimation, whereas it is not that important for chord estimation, confirming our findings in [1, 2].

With respect to other systems, our system attains scores that are among the top results for key estimation, but that fall somewhat behind others for chord estimation. However, a positive result is that the difference is smaller for the unseen McGill data, indicating better generalization capabilities than some other systems.

As some final critical notes on the tasks themselves, we would like to point out that the key estimation task uses very synthetic data generated from MIDI files by a subpar virtual instrument and thus might not be entirely representative for scores on real world data. Also, the current evaluation measure for the chord estimation results does not penalize the generation of chords with superfluous chromas.

4. ACKNOWLEDGMENTS

This work was partly supported by the Quaero Program funded by Oseo French agency and by the MIReS project funded by EU-FP7-ICT-2011.1.5-287711.

5. REFERENCES

[1] Johan Pauwels and Jean-Pierre Martens. Integrating musicological knowledge into a probabilistic system for chord and key extraction. In *Proceedings of the 128th Convention of the AES*, London, UK, 2010.

[2] Johan Pauwels, Jean-Pierre Martens, and Marc Leman. Modeling musicological information as trigrams in a system for simultaneous chord and local key extraction. In *Proceedings of the IEEE International Workshop on Machine Learning for Signal Processing (MLSP)*, Beijing, China, 2011.

[3] Geoffroy Peeters. Template-based estimation of time-varying tempo. *EURASIP Journal of Advances in Signal Processing*, 2007(1):067215, 2007.

[4] Carlos Pérez-Sancho, David Rizo, and José M. Iñesta. Genre classification using chords and stochastic language models. *Connection science*, 21(2–3):145–159, 2009.

[5] Matthias Varewyck, Johan Pauwels, and Jean-Pierre Martens. A novel chroma representation of polyphonic music based on multiple pitch tracking techniques. In *Proceedings of the 16th ACM International Conference on Multimedia (MM'08)*, pages 667–670, 2008.