AUDIO CHORD EXTRACTION USING A PROBABILISTIC MODEL

Johan Pauwels, Matthias Varewyck, Jean-Pierre Martens
Department of Electronics and Information Systems
Ghent University, Belgium
{johan.pauwels, matthias.varewyck, jean-pierre.martens}@elis.ugent.be

ABSTRACT

This paper presents our submission to the MIREX 2008 Audio Chord Detection task. The front-end of our system incorporates a novel feature extractor which uses multiple pitch tracking techniques to extract for each frame a chroma profile that is more robust against chroma contributions not originating from fundamental frequencies but from harmonics thereof. The back-end of our system implements a probabilistic framework for the simultaneous recognition of chords and keys. The system works with probabilities and density functions derived from Lerdahl's tonal distance metric and consequently, it needs no explicit training.

1 IMPLEMENTATION OVERVIEW

Input wavefiles are converted to mono, resampled to 8 kHz and split into frames. The frame length is 150 ms and the hopsize is 20 ms. For each frame, the front-end calculates a chroma profile. Consecutive frames are grouped per 10 in so-called segments to improve the stability of the output and to speed up the calculation. The average chroma profiles of these segments are then supplied to the back-end.

The back-end generates a chord label for each segment. This label represents one of four triads (major, minor, diminished and augmented) that can be defined for each of the 12 chromas. However, for the MIREX task only the major and minor triads were withheld and the diminished and augmented triads were mapped to a no-chord. The key output of the back-end has been discarded as well.

The present implementation works offline, but it could be changed into a streambased system with little or no performance loss. It runs 96% real-time on an Intel Pentium M 1.86 GHz with 1GB of RAM. On average 13% of the time is spent on the resampling step, 22% on the front-end and 65% on the back-end. Lots of opportunities for speed up are available and have not yet been exploited.

2 THE FRONT-END OF THE SYSTEM

As in many other systems, the acoustic observations are chroma profiles, but the calculation of these profiles differs from what is commonly used. In its simplest form, such a profile is just a log-frequency representation of the spectral content folded into a single octave. However, the problem with such a representation is that e.g. the third harmonic of a pitch folds into a chroma that is located at +7 or -5 semitones with repect to the fundamental, thus adding evidence to a second pitch class that is not necessarily present in the signal.

Our front-end uses the novel implementation proposed by Varewyck et al [3]. It aims at maximally coupling the higher harmonics to their fundamental frequency by the application of multiple pitch tracking techniques. Ideally, if that coupling were perfect, the chroma profile would only represent notes that are actually played, and the chord detection would mainly be a matter of pattern matching.

The values of the chroma profile are scaled such that they add up to 1, making them insensitive to the intensity of the sound. Fundamental frequencies lower than 100 Hz are considered to be bass-notes and are not allowed to contribute to the profile. Although such bass-notes could make a significant addition to the chord, mostly they just repeat a note from the higher registers or they do not contribute to the chord (e.g. a walking bass), and therefore we argue that it does more harm than good to include them.

A consequence of using a pitch tracker for chroma profile generation is that if no frequency is supported as a fundamental frequency by the presence of higher harmonics, the chroma profile will be a vector of zeros. At the moment, such a profile does not yet cause the back-end to generate a no-chord, but this is one of the planned improvements of our system.

3 THE BACK-END OF THE SYSTEM

3.1 Overview

The back-end follows a unified probabilistic framework for the simultaneous recognition of chords and keys. It was introduced by Catteau et al. [1], and slightly modified since then. The input is a sequence of chroma profiles each representing one segment. The profiles form a sequence of length N of acoustic observations, denoted as $\mathbf{X} = \{\mathbf{x}_1, \ldots, \mathbf{x}_N\}$.

The back-end is expected to retrieve the key label sequence $\hat{\mathbf{K}} = \left\{\hat{k}_1, \ldots, \hat{k}_N\right\}$ and the chord label sequence

$\hat{\mathbf{C}} = \{\hat{c}_1, \ldots, \hat{c}_N\}$ which meets the following condition

$$\hat{\mathbf{K}}, \hat{\mathbf{C}} = \arg\max_{\mathbf{K},\mathbf{C}} P(\mathbf{K}, \mathbf{C}) P(\mathbf{X}|\mathbf{K}, \mathbf{C})$$

The term $P(\mathbf{X}|\mathbf{K}, \mathbf{C})$ is computed by an acoustic model and $P(\mathbf{K}, \mathbf{C})$ by an a priori tonality model. By assuming $\mathbf{x}_i$ to be independent of $k_j, c_j \ \forall i \neq j$ and by using a bigram tonality model, this formula can be factorized into

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

The solution can then be found by a Dynamic Programming search which retains at every segment index the optimal path to each of the 1152 eligible key-chord pairs: 48 chords (4 types of triads for 12 pitch classes) times 24 keys (major and minor key for 12 pitch classes). The final result is then identified as the path ending in the key-chord pair with the highest probability at the final segment index.

3.2 Acoustic model

The acoustic model expresses the likelihood of an observation given a proposed key-chord combination. The components of the observation vector $\mathbf{x}_n$ are assumed to be independent of each other and of the key k_n. This way the resulting acoustic probability reduces to the product of the probabilities for all pitch classes. Since a pitch class does either belong to the proposed chord or not, there are two probability distributions to distinguish. These distributions are modeled by single-sided Gaussians centered around $X = 1/3$ or 0 for a pitch class that does or does not belong to the chord respectively. The reason for the factor 3 is that we expect three pitch classes to contribute to the chroma profile of a chord.

3.3 Tonality model

The tonality model describes the probability of different transitions between chord-key pairs in the output sequence. We can further convert the model into a product of a key transition and a chord transition model:

$$P(k_n, c_n|k_{n-1}, c_{n-1}) = P(k_n|k_{n-1}, c_{n-1}) P(c_n|k_n, k_{n-1}, c_{n-1})$$

Both transition models are derived from Lerdahl's distance metric [2] for measuring the dissimilarity between two key-chord pairs. The underlying assumption of our system is thus that transitions between similar key-chord combinations tend to occur more frequently than transitions between dissimilar combinations. This may be not the best possible premise but it has the advantage of not requiring any training of the tonality model, and consequently, of not risking to create a model whose quality depends too much on the selection of the training set.

We assume on intuitive grounds that the influence of c_{n-1} on the key transition probability will be less than that of k_{n-1}, and therefore we simply ignore it.

The probability of staying in the same key is fixed (system parameter), and the probabilities for going to one of the different other keys are derived from the Lerdahl distance between the chords on the first degree of the these keys. An exponential is used to convert distances into probability estimates.

For the chord transition probability we again assume intuitively that k_{n-1} accounts for less than c_{n-1} and k_n, and therefore we ignore it. We further make a distinction between chord pairs (c_{n-1}, c_n) that are both diatonic in k_n and others. The diatonic transition probabilities are derived from the Lerdahl distance between chords in the same key, but weighted by a function that favours chords comprising the key tonic or dominant. Again an exponential is used to convert distances to probability estimates. The probability of all non-diatonic transitions is fixed and set to a value that is lower than the smallest of all diatonic transitions.

4 RESULTS

In the MIREX evaluation, our system obtained a score of 59% in the pre-trained subtask, while the best performing system by Bello & Pickens achieved 66%. All other systems were trained on (part of) the test set. Therefore, we expect the trained models to have learned Beatles-specific chord progressions which will not necessarily generalize to other music genres.

Ignoring the major-minor variations of a chord improved our result by 3%, which is consistent with the other submissions. While our algorithm was the slowest in the test, it still ran at 26% real-time without optimizations, which shows that a streaming implementation should be feasible.

Comparing the results of the pre-trained subtask to those of the train-test subtask leads to interesting conclusions. While we see an expected performance loss for systems that are submitted to both categories (Khadkevich & Omologo and, to a lesser extent, Lee), the best two performing systems are as good or better than the one of Bello & Pickens. Looking at the description of the winning system by Uchiyama, Miyamoto & Sagayama, it is safe to attribute at least a part of the lead to their preprocessing step which filters the audio input of its percussive components. This shows the importance of the feature calculation step in chord extraction systems.

5 FUTURE WORK

One planned improvement is to extend the back-end such that it can also generate no-chord labels. Another option

is to investigate the benefits of working with trained probabilistic models that are tuned to a specific collection of music, e.g. the Beatles albums. Finally, it is our intention to replace the ad-hoc grouping of frames by a more intelligent segmentation, or to move to a purely frame-based system once we have exploited some opportunities for speeding up the back-end.

6 ACKNOWLEDGEMENTS

This work was conducted in the context of the Semantic description of musical audio (GOASEMA) project, which is funded by the Bijzonder Onderzoeksfonds (BOF), Ghent University.

We would like to thank the people at IMIRSEL for organizing MIREX and Chris Harte for providing the ground truth chord labels that made the Audio Chord Detection task possible.

7 REFERENCES

[1] Catteau, B.; Martens, J.-P. and Leman, M. "A probabilistic framework for audio-based tonal key and chord recognition", *Advances in data analysis* Springer, Berlin, Germany, 2007.

[2] Lerdahl, F. *Tonal pitch space*, Oxford University Press, New York, 2001.

[3] Varewyck, M.; Pauwels, J. and Martens, J.-P. "A novel chroma representation of polyphonic music based on multiple pitch tracking techniques", *Proceedings of the ACM International Conference on Multimedia (to appear)*, Vancouver, BC, Canada, 2008.