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 2009 Audio Chord Detection task. It is an optimized version of last year's. The front-end of our system 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 20 in non-overlapping 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. The key output of the back-end has been discarded. The present implementation works offline, but it could be changed into a streambased system with little or no performance loss.

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 implementation proposed by Varewyck et al [1]. 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.

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.

This chroma extractor is now publicly available in Marsyas [2] or on request as a C++-library or Matlab-function.

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. [3], 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

$$\hat{\mathbf{K}}, \hat{\mathbf{C}} = \arg\max_{\mathbf{K},\mathbf{C}} P(\mathbf{K}, \mathbf{C}) P(\mathbf{X}|\mathbf{K}, \mathbf{C}) \qquad (1)$$

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

$$\hat{\mathbf{K}}, \hat{\mathbf{C}} = \arg\max_{\mathbf{K},\mathbf{C}} \prod_{n=1}^{N} P(\mathbf{x}_n|k_n, c_n) P(k_n, c_n|k_{n-1}, c_{n-1}) \qquad (2)$$

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.

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:

$$P(k_n, c_n|k_{n-1}, c_{n-1}) = \quad (3)$$

$$P(k_n|k_{n-1}, c_{n-1})P(c_n|k_n, k_{n-1}, c_{n-1}) \quad (4)$$

Both transition models are derived from Lerdahl's distance metric [4] 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 transitions ending in a chord c_n that is diatonic in k_n or not. The balance between both is adjusted by a system parameter (set to a 0.8-0.2 in favour of diatonic endings). The probability of transitions between chords c_{n-1} and c_n both diatonic in k_n is further divided based on the Lerdahl distance between chords in the same key, but weighted by a function that gives preference to chords with the key tonic or dominant as root. Again an exponential is used to convert distances to probability estimates. The probability of all non-diatonic transitions is uniformly distributed such that

$$\forall X, Z : \sum_Y P(c_n = X|c_{n-1} = Y, k_n = Z) = 1 \quad (5)$$

4. RESULTS

The system described above entered the MIREX competition as "PVM1". A baseline system was submitted as well under the name "PVM2". This baseline is a template matcher which uses the described features averaged over a sliding window and calculates the cosine similarity between the extracted chroma profiles and 48 binary templates. The profile with the highest similarity then gives its name to the frame. The inclusion of the baseline extractor in the contest allows for a more precise comparison between the performance of our main submission on the MIREX test set and on our own data.

When looking at the results, we see that at 68%, the performance of our main submission is about average. The best submissions all lie closely together, with a maximum of 71% for Mauch & Dixon and Oudre, Grenier & Fevotte. The result of the baseline (65%) is somewhat better than expected, because the difference between the two systems is more than 5 % on our own data. Unfortunately, no comparison with last year's results can be made at this moment, since the results for just the Beatles dataset are not (yet?) available.

5. FUTURE WORK

We plan on evaluating our system on the Beatles dataset, in particular paying attention to the issue of tunings deviating from the reference 440 Hz, which recently came up on the MUSIC-IR mailinglist. In its current implementation, no tuning algorithm is included because this did not prove to be an issue on our test data. Also, using the Beatles dataset will allow us to adapt our transition probabilities to this specific case.

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 and Matthias Mauch for providing the ground truth chord labels that made the Audio Chord Detection task possible.

7. REFERENCES

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

[2] Tzanetakis, G., and P. R. Cook: "MARSYAS: A framework for audio analysis", *Organized Sound*, Vol. 4, No. 3, 2000.

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

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