Chord Estimation Using Compositional Hierarchical Model

Matevž Pesek and Matija Marolt

University of Ljubljana,
Faculty of Computer and Information Science
{matevz.pesek, matija.marolt}@fri.uni-lj.si

Abstract. Chroma feature vectors are widely used for automated chord estimation in the music information retrieval (MIR) field. We propose a compositional hierarchical model for unsupervised feature learning providing an alternative to chroma features. Both feature types are further modelled with two separate Hidden Markov models (HMMs) in order to estimate the chords. Further, a binary decision tree is proposed binding the HMM estimations into a new feature vector. The additional stacking of the classifiers provides a classification boost by 17.55 percent.

1 Introduction

There is a variety of tasks proposed in the MIR field, from computationally oriented automated chord estimation, beat tracking and tempo estimation to a higher-level based, e.g. mood estimation and an artist's influence. Automated chord estimation (ACE) is by no means a trivial task. Nor the solution for the ACE nor for any other task in the MIR field has proven to be a perfect one. Although reaching satisfying results on the databases available, the proposed approaches based on machine-learning algorithms gave only limited progress in the classification accuracy in the past years. The definition of a correct answer may not result in specific disjunct groups, as the category definition may vary for each individual. On the topic of transcription, the ground truth appears to be more uniform, as it is based upon the musicological rules which have stronger definitions, compared to defining the boundaries between genres.

We propose building two distinct HMMs based on chromagrams and CHM features. In addition, we propose stacked generalization method, combining the predictions of both HMMs and performing classification with a binary decision tree. The top layer of the classifier stack serves as a decision refiner, thus possibly overcoming the misses of each predictor by learning to differentiate between different interpretations.

2 Chroma vectors and HMM

The chroma feature or pitch class profile (PCP) vectors have long been used as an intermediate-level features for audio chord estimation tasks. A chroma feature vector consists of 12-dimensions, each representing a single pitch or a semi-tone within one octave. The pitch defined in a chroma vector is based on pitch classes of the frequencies in a harmonic spectrum. Wrapping a frequency spectrum into one octave provides a good representation of a music signal with little to no loss of information about the spectrum at an observed time. A chroma vector is calculated by using constant Q transform, which provides a spectrum based on a logarithmic scale. The frequencies are wrapped into an octave-invariant representation.

In MML 2013: International Workshop on Machine Learning and Music, ECML/PKDD, Prague, 2013.

A HMM is commonly used for harmony estimation in the MIR field [1,4]. The chroma vectors provide sufficient intermediate-level audio features and can be used for harmony classification with standard ML algorithms, e.g. a support vector machine. However, such classification eliminates information about the time series due to the observation of a single vector as an independent entity. The added value of HMMs as classifiers lies in the ability of time-dependent processing, hence the ability to model patterns of chord progressions in western music.

3 Compositional Hierarchical Model

The compositional hierarchical model (CHM) is based on the approach proposed in the field of computer vision and cognition [3]. We present the translated model adjusted for the sound signal manipulation. Sound filtering, emulating the outer and middle ear processing, is performed as the first stage of the process [2]. The process continues with the constant Q transform with 48 bins per octave, between 55 and 8000 Hz. The step size is 50 milliseconds with the maximum size of Hamming window of 100 milliseconds.

A spectrum of an audio recording is extracted from a source providing frequency bins for each frame. The A_0 layer represents the base for the CHM. A frequency bin represents the most basic piece of information. The energy of the audio signal at the given time is transformed into activations of the frequency bins. Co-occurrences of activations reveal patterns in the signal's structure. More frequent co-occurrences of activations are joined into a composition denoted part P_i, defined by an offset between the frequency bins. The activation of P_i depends on the activation of the frequency bins forming the part. An activation is defined as a sum of contained frequency activation magnitudes, whereas the part's location is not defined by the part itself, but rather by the location of the activation. Thus a single part may be activated across the spectrum at multiple locations for a given frame. Newly formed parts containing compositions of A_0 layer frequencies represent a new layer denoted as L_{r_0}. A CHM is built of several layers containing parts. The L_{r_0} layer can be used as the source for the additional part integration, thus creating an additional layer of composed parts L_{r_1} containing compositions based on the previous layer.

For the chord estimation task, we produce L_{r_0} and L_{r_1} layers and an additional L_{ior} layer representing pitch classes. This layer incorporates 48 intermediate-level features based on the activations' locations of lower-layer parts. We denote this layer invariant-or layer (L_{ior}) based on the method of parts' inference and the octave invariant activations of the layer. The activation's magnitude of each L_{ior} part depends on the strongest activation of a part on L_{r_1} layer.

We generated the hierarchical model on a small database of 88 piano-keys recorded on a grand piano. The candidate parts for each layer are learned unsupervised and form a larger set of candidates extracted from the audio. Nonetheless, a subset of parts successfully covers most of the signal, while others occur rarely and can be interpreted as noise or cover anomalies in the signal. The subset is refined by a greedy procedure of choosing candidates with greater coverage leading to a near-optimal solution. The magnitude of activation is evaluated along with the frequency, thus placing the signal's energy covered by a candidate as the second evaluation criterion. The candidate refinement procedure is done by a greedy approach of picking candidates with the largest added coverage with respect to previously selected parts.

For the chord estimation experiment performed in this paper, the L_{ior} part activations are exported as 12-dimensional feature vectors for signal representation. The down-scaling is achieved by joining the activation magnitudes of 48 parts into 12 components. We annotate the output as intermediate-level CHM features.

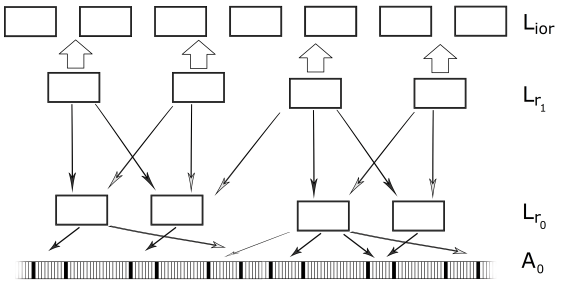

Fig. 1. The lower part of the CHM including A_0 - spectral representation, L_{r_0} and L_{r_1} layers with compositions based on co-occurrences; and L_{ior} layer providing the octave-invariant representation for further classification.

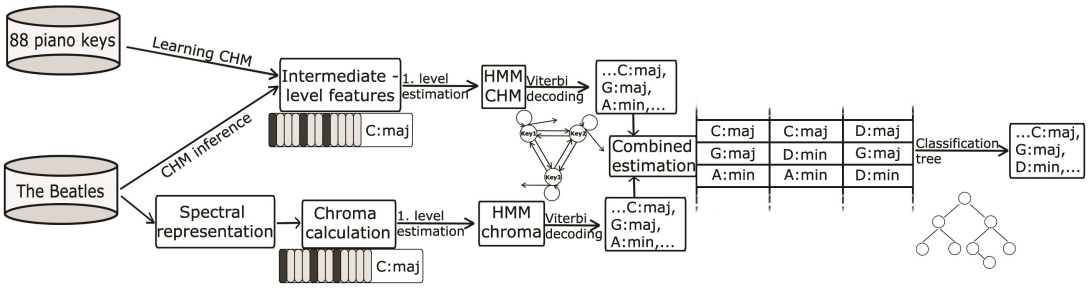

Fig. 2. The processing begins with chroma and the CHM features extraction, HMM learning and stacking performed by combining the HMM estimations and re-estimating results with a binary decision tree.

4 Experiment

We performed the calculation of chroma feature vectors on the data set consisting of twelve *The Beatles* albums kindly provided by C. Harte. The CHM was built on a set of 88 piano-key recordings, whereas the intermediate-level features were calculated for the Beatles data set.

For each type of features, an ergodic 24-state HMM is built with each hidden state representing a chord from a set of minor and major chords $\{C, ..., B, Cm, ..., Bm\}$. The data set annotations consist of a variety of chords, thus the more complex chords are translated to the root major or minor chord, e.g. C:maj7/E is translated to C:maj. Without information about the starting chords for each song, the a-priori values for matrix π are set to $\frac{1}{24}$ for each hidden state. The initial values for state-transition matrix are based on a doubly-nested circle of fifths, thus providing some musicological know-how based on western music.

A cross-validation for this experiment was performed using one album for the training set while the rest of the data set is being estimated. The Viterbi algorithm is used for the path estimation, thus producing the most probable path through the HMM. Estimated chord sequences are combined for additional classification as shown in Figure 2.

The outputs of both HMMs are combined into a new estimation vector with both estimations as parameters and ground truth. The stacking was performed with a five times two-fold cross validation.

Model	$\overline{CA}$	σ
HMM_{chm}	50.43 %	0.1371
HMM_{chroma}	49.03 %	0.2140
Binary Decision Tree	67.28%	

Table 1. The average classification accuracy of HMM_{chm} and HMM_{chroma} models with standard deviances. The classification accuracy of both approaches differs minimally. The classification accuracy of binary decision tree is significantly greater.

The initial per-frame classification accuracy results are displayed in Table 1. The classification accuracy results of both feature types are comparable. The comparability of the results does not automatically imply the ability to boost results by combining them due to the minimal difference between the results, which can reflect the similarity of both feature types. Nevertheless, the stacking process including a binary decision tree was performed. Our hypothesis is formed on an assumption that two different approaches include different interpretations, each with an ability to distinguish some chords better than others. However, combining the interpretations will provide best of each estimation.

With stacking we have improved the classification by 17.55 percent on average to 67.28%. Despite the comparable results of both HMMs and the similarity of both feature types the accuracy boost can be explained by the unsupervised CHM feature learning. The L_{r_1} parts model the pitch features learned in an unsupervised manner. Therefore, the chroma vector representation directly representing pitch classes can significantly differ from the CHM feature vector.

5 Conclusions

We have implemented the effectiveness of combining the chord estimations based on different feature types as one of the possibilities of improving the classification accuracy for the ACE task. The initial HMM results were significantly boosted by stacking the classifiers. Two HMMs were built taking chroma and CHM features as an input. An additional classification with a binary decision tree was provided as a second layer of stacking providing estimation refinement, hence observing different feature-type classifiers as expert models with different background.

References

1. Juan P. Bello and Jeremy Pickens. A robust mid-level representation for harmonic content in music signals. In *Proceedings of ISMIR*, London, 2005.
2. Hugo Fastl and Eberhard Zwicker. *Psychoacoustics: Facts and models.* 2007.
3. Aleš Leonardis and Sanja Fidler. Towards scalable representations of object categories: Learning a hierarchy of parts. *Computer Vision and Pattern Recognition, IEEE*, pages 1–8, 2007.
4. Helene Papadopoulos and Geoffroy Peeters. Large-case Study of Chord Estimation Algorithms Based on Chroma Representation and HMM. *Content-Based Multimedia Indexing*, 53-60, 2007.