ESTIMATING THE TONALITY OF POLYPHONIC AUDIO FILES: COGNITIVE VERSUS MACHINE LEARNING MODELLING STRATEGIES

Emilia Gómez *Perfecto Herrera*
Music Technology Group, Institut Universitari de l'Audiovisual
Universitat Pompeu Fabra
{emilia.gomez,perfecto.herrera}@iua.upf.es
http://www.iua.upf.es/mtg

ABSTRACT

In this paper we evaluate two methods for key estimation from polyphonic audio recordings. Our goal is to compare between a strategy using a cognition-inspired model and several machine learning techniques to find a model for tonality (mode and key note) determination of polyphonic music from audio files. Both approaches have as an input a vector of values related to the intensity of each of the pitch classes of a chromatic scale. In this study, both methods are explained and evaluated in a large database of audio recordings of classical pieces.

1. INTRODUCTION

Tonality and tonal aspects of musical pieces are very relevant for its appreciation. There have been attempts to relate those aspects with mood induction in listeners, and some kind of relatedness (or similarity) between different excerpts sharing tonality have been reported. Listeners are sensitive to key changes, which are also related to rhythm, structure, style and mood. Key changes can be used, for instance, as cues about the structure of a song, or as features to query for matching pieces in a database. Key and mode can also be used to navigate across digital music collections by computing similarities between the files or selected excerpts from them.

In western music, the term *key* (or *tonality*) is usually defined as the relationship between a set of pitches having a *tonic* as its main tone, after which the key is named. A key is then defined by both its tonic (also called *key note,* for example: *A*) and its mode (ex: *minor*). The tonic is one in an octave range, within the 12 semitones of the chromatic scale (ex: *A, A#/Bb, B, C, C#/Db, D, D#/Eb, E, F, F#/Gb, G*). The mode is usually minor or major, depending on the used scale. The major and minor keys then rise to a total set of 24 different tonalities.

Here we compare two approaches for computing the tonality from audio files containing polyphonic music. The first one is based on a tonality model that has been established after perceptual studies, and uses some musical knowledge to estimate the global key note and mode attached to a certain audio segment. The second one is based on machine learning algorithms trained on a database of labelled pieces. After a description of both approaches, we evaluate them, present the results and discuss some of our findings.

2. SYSTEM BLOCK DIAGRAM

The overall system block diagram is presented in Figure 1. In order to estimate the key from polyphonic recordings, we first extract a set of low-level features from the audio signal. These features are then compared to a model of tonality in order to estimate the key of the piece.

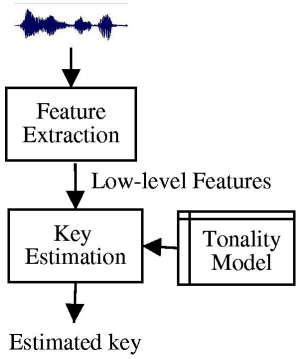

Figure 1. System block diagram.

In this study we have assumed that the key is constant over the considered audio segment. That means that the modulations we can find do not affect the overall tonality of the piece and we can estimate a tonality for the segment.

3. FEATURE EXTRACTION

The input of the key estimation block in Figure 1 is a vector of low-level features extracted from the audio signal. The features used in this study are the *Harmonic Pitch Class Profile* (HPCP), based on de *Pitch Class Profile* descriptor proposed by Fujishima in the context of a chord recognition system [1]. HPCP is a vector of low-level signal features measuring the intensity of each of the 12 pitch classes of the temperate scale within an analysis frame. The feature extraction procedure is summarized as follows. We refer to [2] for a detailed explanation.

1. Instantaneous HPCP vector is computed for each analysis frame using the magnitude of the spectral peaks that are located within a certain frequency band, considered as the most significant frequencies carrying harmonic information. We introduce a

weight into the HPCP computation to get into account differences in tuning, and the resolution is changed to less than one semitone. The HPCP vector is normalized for each analysis frame in order to discard energy information.

2. Global HPCP is computed by averaging instantaneous HPCP within the considered segment.

4. TONALITY COMPUTATION USING A COGNITION-INSPIRED MODEL

This algorithm is based on a key estimation algorithm proposed by Krumhansl et al. and summarized in [3, pp. 77-111]: the *probe tone method*. It measures the expectation of each of the 12 tones of a chromatic scale after a certain tonal context. This measure is representative to quantify the hierarchy of notes in a given tonal context. The output of the model is a rating for each of the 12 semitones of a chromatic scale (starting from the tonic), shown in Figure 2. The data were produced by experienced musicians following tonal contexts that consisted of tonic triads and chord cadences. This profile is used to estimate the key of a MIDI melodic line, by correlating it with a vector containing the relative duration of each of the 12 pitch classes within the MIDI sequence [3].

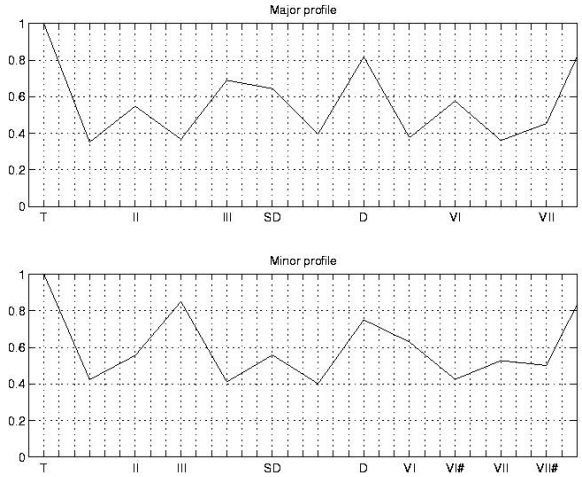

Figure 2. Probe tone ratings from the study by Krumhansl and Kessler (1982) shown with reference to a major key (top) and a minor key (bottom).

Our approach relies on extending this model to deal with audio recordings in a polyphonic situation. We consider the profile value for a given pitch class to represent also the hierarchy of a chord in a given key. Given this assumption, we consider all the chords containing a given pitch class when measuring the relevance of this pitch class within a certain key. For instance, the dominant pitch class (i=8) appears in both tonic and dominant chords, so that the profile value for i=8 adds the contribution of the tonic and the dominant chords of the key. We only consider the three main triads of the major/minor key as the most representative chords (tonic, subdominant and dominant).

We also adapt the method to work with audio features (HPCP related to energy) instead of MIDI. The spectrum of a note is composed of several harmonics, whose frequencies are multiples of the fundamental frequency *f* (*f, 2f, 3f, 4f,* etc.). When a note is played, HPCP increases at the pitch classes of the different harmonics. A note has then different associated pitch classes, one for each harmonic (not only the considered fundamental frequency). Each of the notes of the considered chords contributes to the profile values of its different harmonics. We make this contribution decrease along frequency using a linear function, in order to simulate that the spectrum amplitude decreases with frequency. Final profiles are represented in Figure 3.

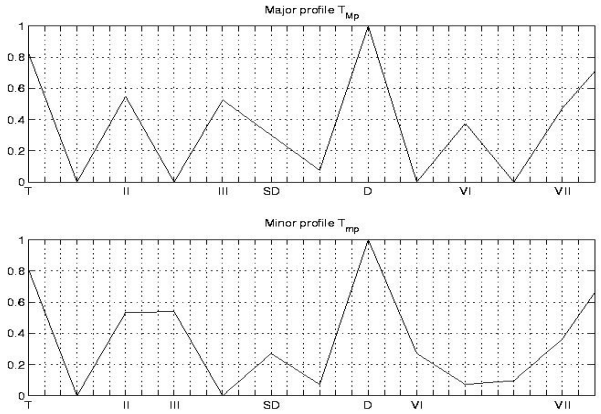

Figure 3. Profiles adapted to polyphony and HPCP shown with reference to a major key (top) and a minor key (bottom).

In order to build the profiles for the 24 different keys, we consider that the tonal hierarchy is invariant with respect to the chosen tonic. For instance, the B major profile is equal to the A major profile but shifted two bins (as A and B from a 2 semitones interval). The global HPCP vector is correlated with the different profiles, computed by circular shifting the adapted profiles. The maximum correlation gives the estimated key note and mode, as well as a correlation factor measuring the proximity of HPCP and the estimated key. More details on the method are found in [2].

5. MACHINE LEARNING FOR TONALITY MODELLING

Different experiments have been performed, all of them involving comparisons between different inductive strategies, including the most usual ones like binary trees, bayesian estimation, neural networks, or support vector machines, but also some interesting meta-learning schemes such as boosting, or bagging. Meta-learning can be defined as the enhancement or extension of basic learning algorithms by means of incorporating other learners [5], which, in general, improve the performance and generalization capabilities of the base learners. Most of the experiments were carried out using Weka[1]:

[1] http://www.cs.waikato.ac.nz/ml/weka/

1. Learning the tonic or key note using the low-level descriptors HPCP as input.
2. Learning the mode using HPCP information.
3. Learning the key note and the mode altogether.
4. Learning tonality (key note and mode) using the HPCP vector and the estimation derived from the perceptual/cognitive model (Section 4), which is considered as a mixed approach.

6. RESULTS

6.1. Audio material

We have built an audio database of 878 excerpts of classical music for evaluation, including many composers as, for instance, Mozart, Chopin, Scarlatti, Bach, Brahms, Beethoven, Handel, Pachelbel, Tchaikovsky, Sibelius, Dvorak, Debussy, Telemann, Albinoni, Vivaldi, Pasquini, Glenn Gould, Rachmaninoff, Schubert, Shostakovich, Haydn, Benedetto, Elgar, Bizet, Liszt, Boccherini, Ravel, Debussy, etc. We also include some jazz versions of classical pieces (e.g. Jacques Lousier, The Swingle Singers, etc). Most of the included titles were first movement (in case that the piece is a multi-movement form as sonata or symphony). All the key note and mode annotations were taken from the FreeDB database[1]. Some additional manual corrections were made to include other movements or because of FreeDB wrong metadata, although systematic checking has not been performed.

We divided the database in two sets: the training set, consisting of 661 audio files and the holdout set including the remaining 217 titles. We kept this holdout in order to test the generalization capabilities of the models using none of the instances used in the training phase. The tonality models were then derived using the 661 instances not assigned to the holdout. Most of the tests involved between 10 and 20 instances for each tonality.

6.2. Model for tonality perception

The results of the evaluation over the holdout database are presented in Figure 4, with 59,5% of correct tonality, 82% of correct mode and 65% of correct key note estimation. The confusion matrix is found in the author's web page[2].

We find that the 19% of the estimation errors correspond to confusions between minor/major relatives (e.g., C major confused with A minor), and other 24% correspond to tuning errors (e.g., E minor confused with Eb minor). It can also be seen that the 5,7 % of the errors have been made by estimating the upper 5th within the circle of fifths (e.g., C major confused with G major) or the key whose tonic form a 5th ascending interval with the correct one (e.g., D minor confused with A minor). 19% of the keys were confused with the near key down on the circle of fifths (A major confused with D major) or the key whose tonic is located at a 5th descending interval (e.g., A minor confused with D minor). Only 44% of the errors correspond to non-related tonality confusions.

6.3. Machine learning models

We present the results according to the addressed subproblems: mode induction, key note induction, and combined key note and mode induction. We observe, among other things, that there is no single "best learner" capable of optimally approximating the solutions for all of them.

6.3.1. Mode induction

The best results for mode induction were obtained using an instance-based learner which bases its decision on the class assigned to the five nearest neighbour cases (84% of correct decisions). Surprisingly, the rest of studied methods scored far below this family of models. The second best method was a multilayer *perceptron* with one hidden layer containing 20 units, which achieved 71% of correct decisions. In all cases, there were much more errors because of wrong assignment of minor mode than the other way round.

6.3.2. Key note induction

The application of "agnostic" machine learning strategies to the problem of assigning an overall key to a music piece yielded slightly better results than the perceptual/cognitive strategy. In this case, a Bayesian classifier with Density Estimation was the best of the set (72% of correct decisions). The Sequential Maximum Optimization algorithm (a kind of Support Vector Machine) scored close to that (70%), and again a 5 Nearest-Neighbour provided good results (69%).

6.3.3. Simultaneous key note and mode induction

Achieving a combined answer for key note and mode is the most complex problem addressed in this series, as there were 24 different classes to classify the input patterns. Here, the best approach was that of a multilayer back-propagated *perceptron* with a single hidden layer of 20 units (63% of correct decisions). Again, instance-based strategies scored among the best (60%), although the best results were not quite far from those from the perceptual/cognitive model (59%). The confusion matrix for this approach is shown in the author's web page[3].

6.4. Combination of approaches

As it is the case in some meta-learning approaches, the combination of two different algorithms can improve the

[1] http://freedb.freedb.org/

[2] http://www.iua.upf.es/~egomez/TonalDescription/GomezHerrera-ISMIR2004.html

performance provided both generate different error patterns. Our experiments using the output of the perceptual/cognitive model as an additional input for the best machine learning algorithm has yielded no improvement to the presented results except in the case of key estimation, where the Bayesian learner yielded 77% when we included the predicted key, mode and strength from the perceptual/cognitive model. This addition amounts to an improvement of 5% (11% compared to the performance of the perceptual/cognitive model alone).

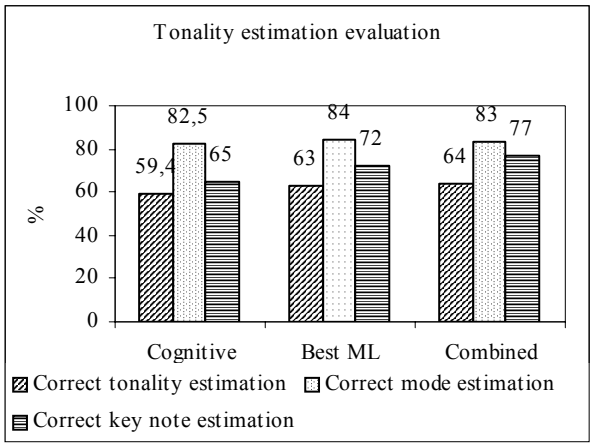

Figure 4. Evaluation results. % of correct estimation.

7. DISCUSSION

Comparing the tonal cognition-inspired approach to the machine learning techniques that we can consider as "tools of the trade", modest improvements in performance can be achieved by the latter (7% when computing the key note) or by embedding the former into the latter (12% for key note computation).

As it is pointed out by Krumhansl, the tonal descriptors we have considered are severely restricted, in the sense that they do not capture any musical structure. These features take into account neither order information nor the chords' position in the event hierarchy, as for instance, its place in the rhythmic or harmonic structure [3, pp. 66]. In fact, some of the estimation errors may be caused by tonality changes that affect the overall key measures and labelling. We will work on these structural and rhythmic aspects along future research.

8. CONCLUSIONS

We have presented a comparison between two different approaches for tonality estimation from polyphonic audio. The first one is inspired in the *probe tone method* and considers some aspects of tonality cognition. The second one uses "blind" machine learning techniques to model key by analyzing a training annotated collection. We have evaluated both methodologies over a large audio database, achieving a 64% of correct overall tonality estimation. Very small improvements were found by only using machine learning algorithms, which is somehow a puzzling observation that requires further experiments with different data representations and more intensive parameter tweaking of the algorithms. We have still room for improvement in order to come up with a robust technique that allow us to exploit tonality information for retrieval in a general-purpose popular music database and also for aiding the discovery of music information in a similar way to what Purwins et al. [4] have recently presented.

9. ACKNOWLEDGMENTS

The authors would like to thank Takuya Fujishima and Jordi Bonada for their advices on the feature extraction procedure. This research has been partially funded by the EU-FP6-IST-507142 project SIMAC (Semantic Interaction with Music Audio Contents)[1] and by the Spanish Government project TIC2003-07776-C02-02 Promusic.

10. REFERENCES

[1] Fujishima, T. "Realtime chord recognition of musical sound: a system using Common Lisp Music", *ICMC*, Beijing, China, 1999, pp. 464–467.

[2] Gómez, E. "Tonal description of polyphonic audio for music content processing". *INFORMS Journal on Computing. Special Cluster on Music Computing.* Chew, E., Guest Editor, 2004.

[3] Krumhansl, C. L. *Cognitive foundations of musical pitch*. Oxford University Press, New York, 1999, pp. 16-49.

[4] Purwins, H., Blankertz, B., Dornhege, G., and Obermayer, K. "Scale degree profiles from audio investigated with machine learning", *116th AES Convention*, Berlin, Germany, 2004.

[5] Witten, I. H. and Frank, E. *Data mining: practical machine learning tools and techniques with Java implementations*. Morgan Kaufmann, San Francisco, 2000.

[1] http://www.semanticaudio.org