Automatic Extraction of Tonal Metadata from Polyphonic Audio Recordings

Emilia Gómez[1] and Perfecto Herrera[1]

[1] *Universitat Pompeu Fabra, Barcelona, Spain*

Correspondence should be addressed to Emilia Gómez (emilia.gomez@iua.upf.es)

ABSTRACT

The goal of this paper is to present a system that automatically extracts metadata from polyphonic audio signals. This metadata is related to the tonal content of music. It can be used to characterize the *tonal space* of a piece, understood as the relationships between pitches, chords and keys, in accordance with the principles and procedures of tonality. These features are relevant in the context of music similarity and content search, retrieval and navigation across digital music collections.

1. INTRODUCTION

Harmony and tonality are relevant aspects of music perception, and for this reason, they are main axis of music description. It is, then, necessary to identify a set of features that could be representative of these aspects of music, and then to develop algorithms capable of automatically extracting these relevant features.

In this study, we consider harmony as a term to denote the combination of notes simultaneously, to produce chords, and along time, to produce chord progressions [4]. This combination of chords is related to the key or tonality of the piece. We present a system to automatically extract a representation of the tonal content of a piece of music, valid for content retrieval and navigation across digital music collections. First, we establish a way to structure the descriptors according to their temporal scale and their abstraction. Then, we explain how these descriptors are extracted from audio. We present some examples showing how the features are computed for a particular musical excerpt. Finally, we evaluate the system for a specific problem, key estimation, over a huge labeled audio database.

2. TONALITY RELATED METADATA

Chord and **key** labels are the metadata that have been mostly used to represent harmonic and tonal aspects of music when dealing with symbolic representations (ex: *C Major* key, *C#dim7* chord). Other relevant metadata that we could extract are related to the **scale** (ex: *diatonic, chromatic, pentatonic*) and **cadence** information of the piece. This metadata can also be defined for audio recordings instead of symbolic representations, which may imply a correspondence between audio and its symbolic notation obtained by means of an automatic transcription procedure.

In addition to these musical descriptors, we find other features which are not directly obtained by automatic transcription or musical analysis of the piece. Nevertheless, they can be very important for tonal perception and can be used to compare different pieces. First, let's consider some perceptual features, which are a measure of **dissonance** and **tonal strength**. We can also look at the **pitch class profile**, that is, the distribution of tones in the piece, which could also be very representative. This metadata might also be defined for symbolic representations, but we can compute them from audio without requiring a previous transcription. We believe there are some aspects that can be lost in the transcription stage, which are captured when working directly with audio.

Name	Temporal Validity	Level of abstraction (data type)
HPCP	Instantaneous	Low (float vector)
Chord	Instantaneous	High (textual label)
ChordStrength	Segment/global	Low (float value)
GlobalHPCP	Segment/global	Low (float vector)
Key	Segment/global	High (Textual label)
KeyStrength	Segment/global	Low (float value)

Table 1: List of descriptors.

Finally, there exist clear connections with other aspects of music description. This relationship must be considered for a proper characterization. First, rhythmic information is an important aspect to take into account. Tick and beat segments must be considered for harmonic characterization [3]. There is also a close connection to structure, that has to be considered in order to analyze tonal evolution (for instance, to study how the tonality is changing). On the other hand, some structure can be established by analyzing changes of harmonic features [9]. Harmonic features are also relevant for analyzing the complexity of a piece [12]. Finally, we find a connection with mood and style, so that these descriptors can act as a grounding for mood and style characterization [10].

In order to structure the descriptors, we define different **levels of abstraction** for the features, as well as different **temporal validities** for them:

- Each descriptor or feature is attached a certain temporal validity, which is defined by an audio segment (according to the MPEG-7 philosophy [7]). The descriptors can be classified into:
  - Instantaneous (valid for a time point),
  - Valid for a certain segment (as for example a phrase, a chorus, etc), or
  - Global (valid for the whole audio excerpt, representative of the whole piece)
- Each descriptor or feature is also defined to be:
  - Low-level: if it is computed directly from the audio signal or its frequency-domain representation. Low-level descriptors are float values, vectors or matrices of values.
  - High-level: if it requires an inductive inference procedure that can consist on the application of a tonal model or on the use of some machine learning techniques. They are textual labels or quantized values.

Table 1 shows a list of the features we are extracting from audio recordings, as well as their associated temporal validity and type.

3. FEATURE EXTRACTION

Our approach for feature extraction includes the following steps, represented in figure 1:

1. First, we apply some preprocessing to the audio signal that consist in detecting the transients to avoid noisy features.
2. Then, we compute a vector of low-level instantaneous features, that we call HPCP (*Harmonic Pitch Class Profile*) vector. It represents the intensity of each pitch mapped to a single octave. It is based on the Pitch Class Profile proposed by Fujishima [5], and it is extracted by the following steps:
   (a) Spectral analysis of the input signal.
   (b) Spectral peak computation, defined as the local maximum of the spectrum magnitude.
   (c) HPCP vector computation.
3. Also related to a time point, we estimate the played chord and its strength by comparing the HPCP vector to a chord model built by a set of chord profiles representing the different chord types [5, 8].

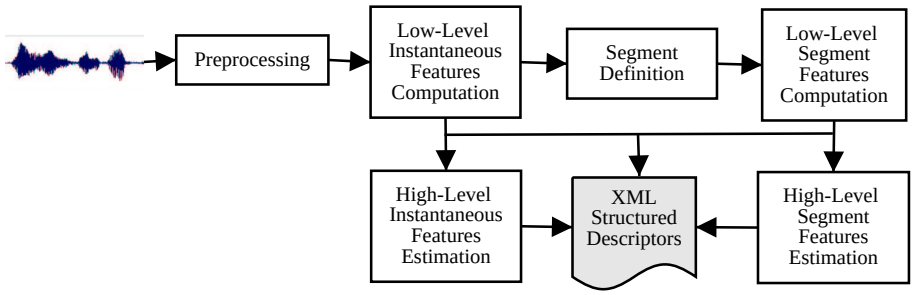

Fig. 1: Block diagram for feature extraction.

4. After having computed some instantaneous features, we can establish different segmentation strategies, as for example identifying beats. Segment descriptors (HPCP, chord or key) can be computed for each of the segments:
   (a) Statistics of the HPCP vector are computed over the segment boundaries.
   (b) We estimate the global high-level features chord or key and its strength by comparing to a chord or a key model, which is based on the one proposed by Krumhansl [6], which has been adapted to deal with polyphony.
5. Finally, global descriptors are computed as follows:
   (a) Statistics of the HPCP vector are computed for the whole audio excerpt.
   (b) We estimate the global high-level features key and its strength using the same key model that the one used for segments.

Further details about the methods are explained in [2].

4. CASE STUDY

In this section we analyze the features extracted from a certain audio excerpt. We see how the features are computed for a 20 seconds excerpt of the song *"Donde Estás Yolanda"*, by Pink Martini[1], corresponding to the chorus of the song.

[1] http://www.pinkmartini.com/

Figure 2 represents a 3D representation of the instantaneous HPCP vector evolution along time. Figure 3 represents a 2D representation of the same data. We can see that there is an increase of the HPCP value corresponding to a certain pitch class when the note is played.

This is shown with better resolution in figure 4, where values for several pitch classes increase at the same time due to polyphony and to the different harmonic frequencies.

Figure 5 represents the global HPCP vector for the chosen excerpt. We can see that there is a predominance of G, B and E pitch classes (and surroundings values). This vector represents the tonal content of the piece, the pitch distribution, equivalent to a note histogram. We can then use the vector to compare with other excerpt's values and find similar profiles corresponding to similar tonal content.

Figure 6 represents the global HPCP vector for the whole excerpt, and its correlation with the different key profiles defined using a key model based on [6]. There is a maximum value corresponding to E minor key (and equal to 0.66). This is the estimated key, which is also the correct one. We also find a local maximum for its major relative (A major).

5. IMPLEMENTATION

The system has been implemented within the CLAM framework [1]. Once computed, the descriptors are structured, so that they can be stored into an XML file. It makes possible to load them again for further computation or for comparison with other descriptor values.

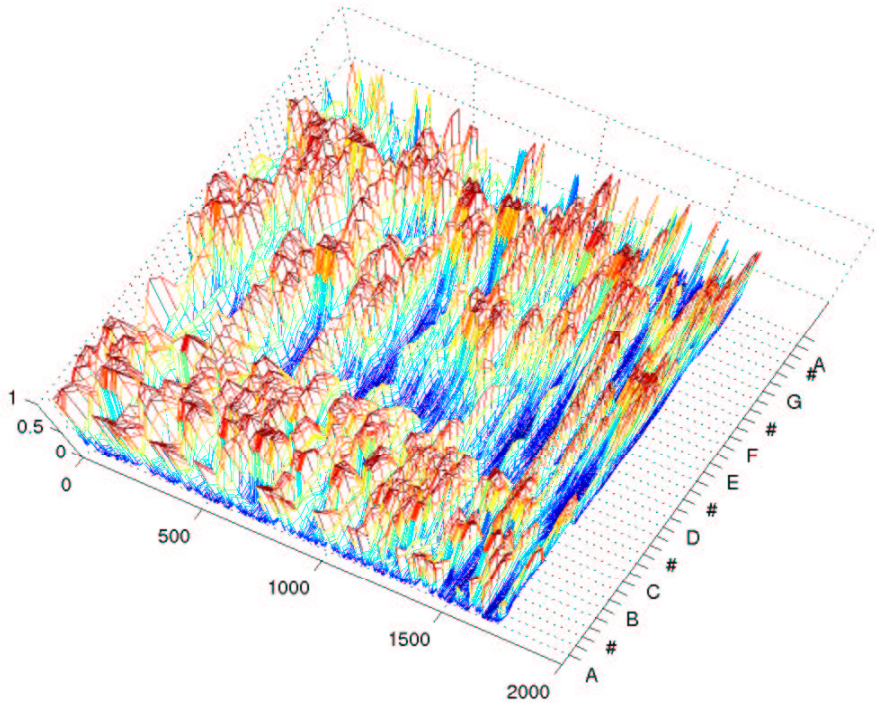

Fig. 2: HPCP Instantaneous Value.

Several processing objects have been implemented on top of the SMS model [11], mainly for HPCP computation and key and chord estimation. They can all be configured with different parameters, some of them related to spectral analysis (as for instance analysis frame size, window type and hop size) and others related to the descriptor computation (as for instance peak amplitude threshold, frequency limits for HPCP computation, HPCP size and resolution, etc).

6. EVALUATION

For initial evaluation, we have focused on key estimation from audio recordings, as it seems easy to measure the effectiveness of the method. Nevertheless, it has been difficult to build a representative test set of labeled pieces with varied styles and musical characteristics.

We have first set up a small test database of 35 sounds with different styles, key and modes, labeled by hand in terms of key. Each of them has a constant tonality. Using the system, we obtained 66.67 % of correct key estimation, 2.78 % of mode errors (and correct key notes), 13.89 % of key note errors (and correct mode) and 16.67 % of errors in both key note and mode (with 6 % tuning errors).

Then, we have evaluated the performance of the algorithm on a database of 525 complete classical music pieces segmented by track and labeled by their title (ex: *Mozart, Flute Concerto No 1 K313 G Major Andante non troppo*). There can be some modulations, as we have not verified by listening that there is a constant key for all the excerpts. This means that we assume that the modulations we can find do not modify the overall key of the piece. We obtained here 62.1 % of correct key estimation, 3.24 % of mode errors (and correct key note), 19.05 % of key note errors (and correct mode) and 15.62 % of errors in both key note and mode.

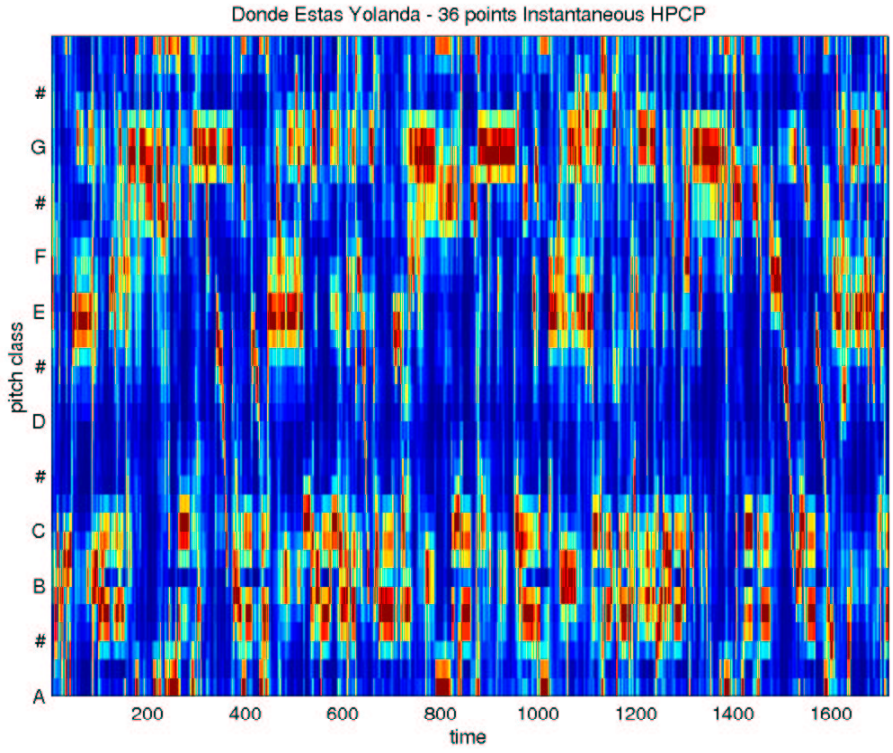

Fig. 3: HPCP Instantaneous Value.

7. CONCLUSIONS

We have proposed a structured set of metadata related to harmony and tonality, and proposed a way to extract them in an automatic way. The validity of the system has been evaluated in the context of key estimation of polyphonic audio.

We have identified the need of having an annotated test database to evaluate our descriptors, possibly complemented with some user ratings of perceptual descriptors, so that we could evaluate perceptual descriptors as chord or tonal strength.

8. ACKNOWLEDGMENTS

This research has been partially funded by the EU-FP6-IST-507142 project SIMAC (Semantic Interaction with Music Audio Contents) and by the Spanish national project TIC2003-07776-C02-02 Promusic.

Emilia Gómez would like to thank Takuya Fujishima and Jordi Bonada for useful discussions and advices on the methods.

9. REFERENCES

[1] CLAM 2004. "C++ Library for Audio and Music", MTG-IUA, Universitat Pompeu Fabra. http://www.iua.upf.es/mtg/clam/

[2] Gómez, E. 2004. "Tonal description of polyphonic audio for music content processing". Submitted for publication. February 2004.

[3] Gouyon, F. Meudic, B. 2003. "Towards Rhythmic Content Processing of Musical Signals: Fostering Complementary Approaches". Journal of New Music Research Vol.32.1.

[4] "Grove Music Online: the New Grove Dictionary of Music and Musicians", second edition, "The New Grove Dictionary of Opera and The New Grove Dictionary of Jazz", second edition. S. Sadie, J. Tyrrell, B. Kernfeld, eds. Online publication. http://www.grovemusic.com

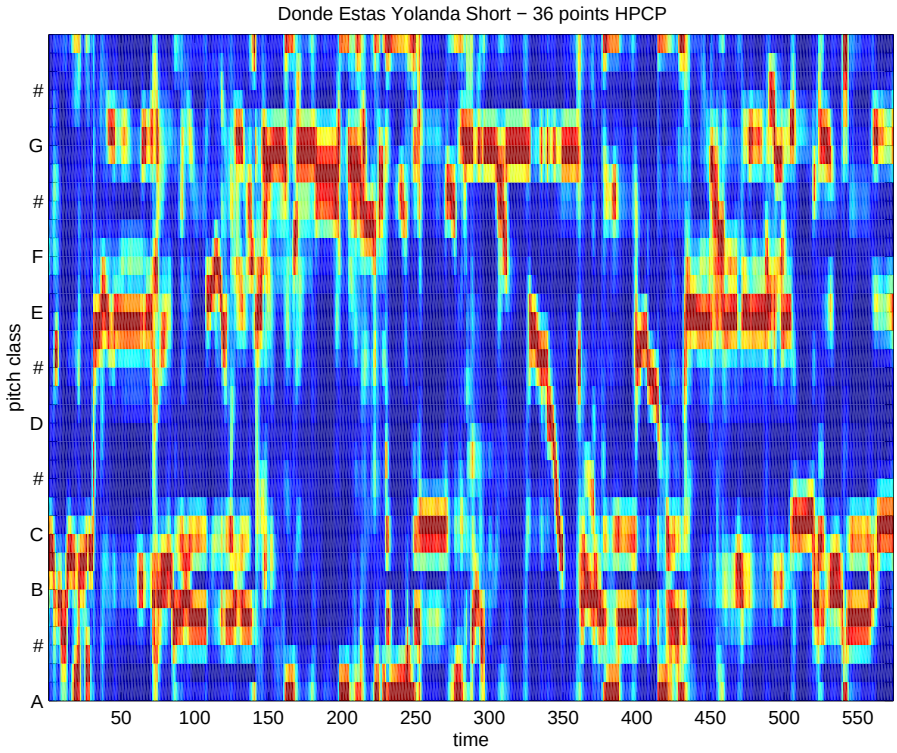

Fig. 4: HPCP Instantaneous Value for 10 seconds excerpts.

[5] Fujishima, T. 1999. "Realtime chord recognition of musical sound: a system using Common Lisp Music". Proceedings of International Computer Music Conference, Beijin, China, 1999, ICMA, San Francisco, pp. 464–467.

[6] Krumhansl, C. L. 1990. "Cognitive foundations of musical pitch". Oxford University Press, New York.

[7] Martínez, J. M. Editor, MPEG-7 Overview, ISO/IEC JTC1/SC29/WG11N5525 document. http://www.chiariglione.org/mpeg/standards/mpeg-7/mpeg-7.htm

[8] Sheh, A. and D. Ellis. 2003. "Chord segmentation and recognition using EM-Trained Hidden Markov Models". Proceedings of International Symposium on Music Information Retrieval, Baltimore, 2003.

[9] Ong, B. and Herrera, P. 2004. "Computing Structural Descriptions of Music through the Identification of Representative Excerpts from Audio Files". Proceedings of 25th International AES Conference, London, UK.

[10] Gabrielsson, A. and Lindström, E. "The Influence of Musical Structure on Emotional Expression". In Juslin P. N. and Sloboda J. A. editors, "Music and Emotion", Oxford University Press, 2001.

[11] Serra, X. and Smith, J. 1990."Spectral Modeling Synthesis:A Sound Analysis/Synthesis Based on a Deterministic plus Stochastic Decomposition". Computer Music Journal Vol.14 .4 pp. 12-24.

[12] Streich, S. and Herrera, P. 2004. "Toward Describing Perceived Complexity of Songs: Computational Methods and Implementation". Proceedings of 25th International AES Conference, London, UK.

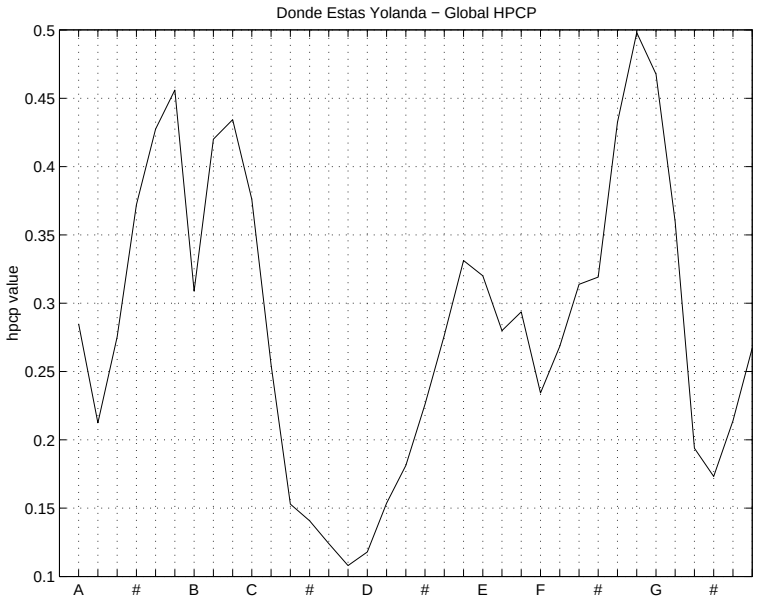

Fig. 5: HPCP Global Value.

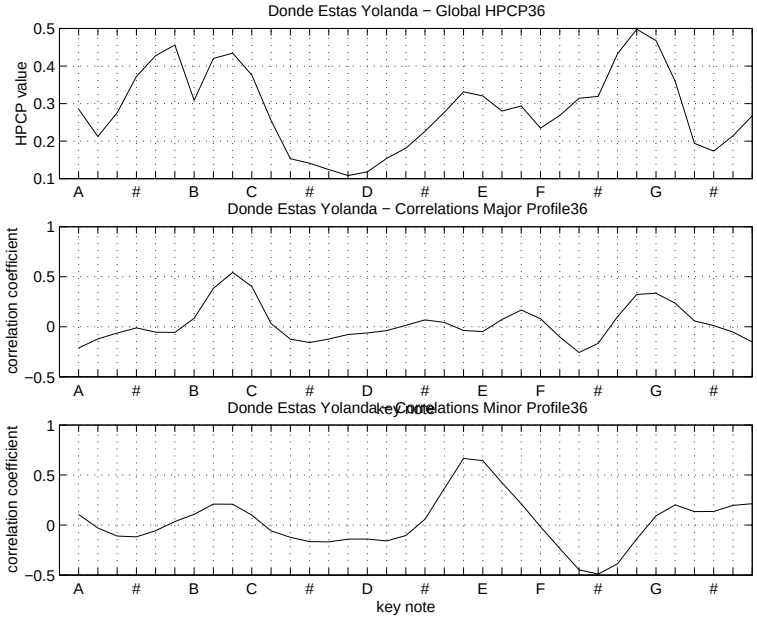

Fig. 6: HPCP Global Value and Key Correlations for Major and Minor Tonalities.