Automatic extraction of tonal metadata from polyphonic audio recordings

Tonality is a relevant aspect of music perception, and then a main axis for music description. We need to represent this aspect of music using a set of features computed from an audio recording. These features can be used for content-based retrieval and navigation through digital music collections.

Emilia Gómez and P. Herrera
{emilia.gomez,perfecto.herrera}@iua.upf.es
http://www.iua.upf.es/mtg

Tonal Metadata

Description scheme:

(1) Temporal validity:
- Instantaneous descriptors: valid for a time point.
- Segment descriptors: defined within an audio segment.
- Global: representative of the whole excerpt or piece.

(2) Level of abstraction:
- Low-level: computed directly from the audio signal or from other low-level descriptors.
- High-level: it requires an inductive inference procedure.

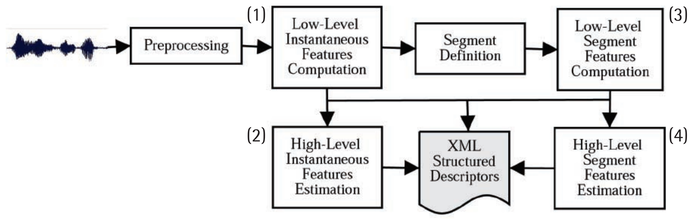

```
<AudioSegmentD>
  <MeanFramesD>
    <SpectrumD>
      <HPCP>0.280865 0.210845 0.0955102 0.048049 0.103835 0.211588 0.273482 0.279585 0.264783 0.243512
0.193451 0.122316 0.0948005 0.135353 0.212854 0.249327 0.194157 0.0857733 0.0301739 0.0720872 0.148929 0.182718
0.148403 0.0870518 0.0684743 0.13653 0.231203 0.270169 0.277118 0.293294 0.294509 0.217129 0.102392 0.0568156
0.124401 0.237221</HPCP>
    </SpectrumD>
  </MeanFramesD>
  <Melody>
    <Key>
      <Note>G</Note>
      <Mode>Major</Mode>
      <KeyStrength>0.863316</KeyStrength>
    </Key>
  </Melody>
  <ChildrenD>
    <SegmentDescriptors>
      <MeanFramesD>
        <SpectrumD>
          <HPCP>0.208107 0.145337 0.0478367 0.0185335 0.119326 0.275019 0.338163 0.284909 0.18745 0.130718
0.0813556 0.0259314 0.0132593 0.0980055 0.286679 0.408152 0.341707 0.153394 0.0265396 0.00675188 0.00382843
0.0029442 0.00282956 0.00301471 0.0176575 0.0898968 0.17107 0.207165 0.260252 0.350629 0.389508 0.28104 0.110348
0.0282654 0.0776006 0.173812</HPCP>
        </SpectrumD>
      </MeanFramesD>
      <Melody>
        <Key>
          <Note>G</Note>
          <Mode>Major</Mode>
          <KeyStrength>0.5616</KeyStrength>
        </Key>
      </Melody>
    </SegmentDescriptors>
    ....
  </ChildrenD>
</AudioSegmentD>
```


Name	Temporal Validity	Level of Abstraction	Data Type
HPCP	Instantaneous	Low	Float vector
Chord	Instantaneous	High	Textual label
Chord Strength	Segment/Global	High	Float value
Global HPCP	Segment/Global	Low	Float vector
Key	Segment/Global	High	Textual label
Key Strength	Segment/Global	High	Float value

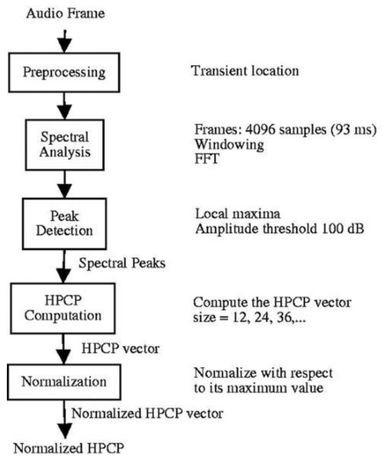
List of Descriptors

Feature Extraction

Block diagram for feature extraction:

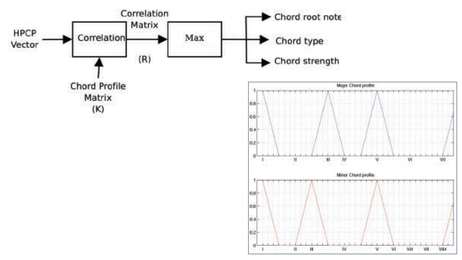


(1) Low-level Instantaneous feature computation: HPCP (*Harmonic Pitch Class Profile*). It represents the intensity of each pitch class mapped to an octave.

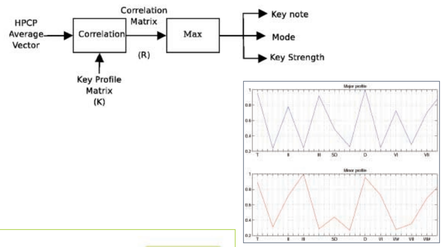


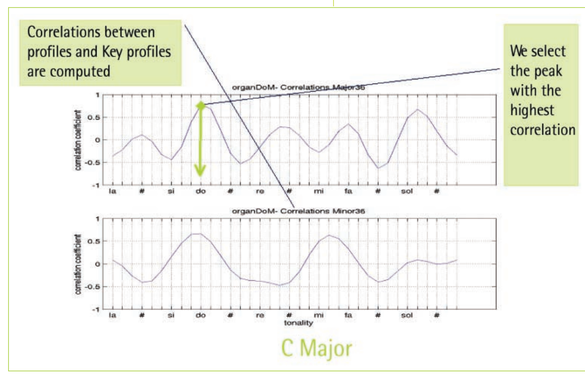
$$HPCP(n) = \sum_i^{nPeaks} w(n, f_i) \cdot ||a_i||^2 \quad n = 1...size \quad i = 1...nPeaks$$

(2) High-level Instantaneous feature computation: Chord, Chord Strength



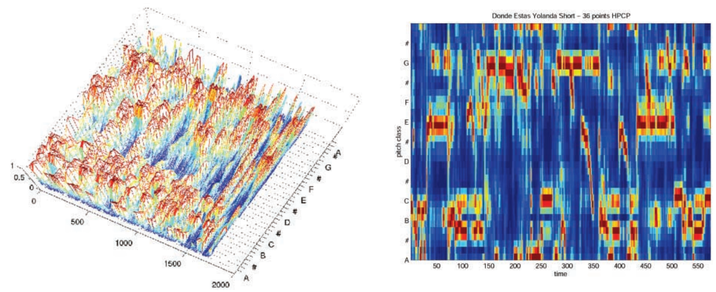
(3) Low-level Segment/Global descriptors computation: Average HPCP
(4) High-level Segment/Global descriptors computation: Key, Key Strength



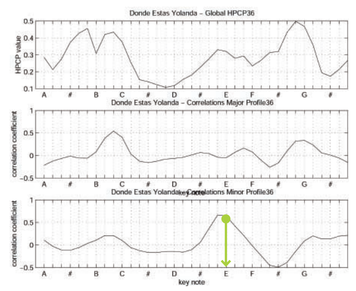


Results

Instantaneous HPCP for a 10 seconds excerpt of a song:



HPCP Global values and correlation for Major/Minor tonalities:



E Minor

Key Estimation Evaluation:

Small test database with different styles, key, mode. Labeled by hand (35 sounds).

Correct key note	65,5 %
Correct mode	83,24 %
Correct key	64,2 %

Database of 525 classical pieces labeled by their title.

Correct key	70 %
Mode error	3 %
Tuning error	6 %

References:
- *Fujishima, T. 1999. "Realtime chord recognition of musical sound: a system using Common Lisp Music". ICMC.*
- *Krumhansl, C.L. 1990. "Cognitive Foundations of Musical Pitch". Oxford University Press, New York.*
- *Sheh, A. and Ellis, D. 2003. "Chord Segmentation and Recognition using EM-Trained Hidden Markov Models ". ISMIR.*