INFORMS Journal on Computing

Tonal Description of Polyphonic Audio for Music Content Processing

Emilia Gómez,

INFORMS Journal on Computing
Vol. 18, No. 3, Summer 2006, pp. 294–304
ISSN 1091-9856 | EISSN 1526-5528 | 06 | 1803 | 0294

DOI 10.1287/ijoc.1040.0126

Tonal Description of Polyphonic Audio for Music Content Processing

Emilia Gómez
Music Technology Group, Institut Universitari de l'Audiovisual, Universitat Pompeu Fabra,
Ocata 1, Barcelona, 08003, Spain, emilia.gomez@iua.upf.es

We present a method to extract a description of the tonal aspects of music from polyphonic audio signals. We define this tonal description using different levels of abstraction, differentiating between low-level signal descriptors and high-level textual labels. We also establish different temporal scales for description, defining some features as being attached to a certain time instant, and other global descriptors as related to a wider segment. The description is validated by estimating the key of a piece. We also propose the description as a tonal representation of the polyphonic audio signal to measure tonal similarity between audio excerpts and to establish the tonal structure of a musical piece.

Key words: tonal description; key estimation; content analysis and indexing; sound and music computing
History: Accepted by Elaine Chew, Guest Editor for Special Cluster on Computation in Music; received February 2004; revised June 2004; accepted November 2004.

1. Introduction

In the last few years, a great amount of audio material has been made accessible to the home user through networks and mass storage. This proliferation of digital music collections necessitates the development of technology that is able to interact with such collections in an easy and meaningful way. That is the goal of content retrieval and transformation systems; to retrieve and transform audio according to its content. There are many disciplines involved in this issue, such as signal processing, musicology, statistics, and information retrieval.

Tonality is one of the main aspects of western music, so it deserves a proper description. It is then necessary to develop methods that can extract information related to the tonal content of a piece of music.

In western tonal music, we define *key* as a system of relationships between a series of pitches having a *tonic*, or central pitch, as its most important element. Besides the tonic, one of the most important pitches of a key is the *dominant* degree, defined as the fifth degree of the scale. Another important degree is the *subdominant* degree, which is the fourth degree of the scale and lies below the tonic as much as the dominant lies above it, that is, a 5th.

There are two basic key *modes*: major and minor. Each of them has different musical characteristics regarding the position of tones and semitones within their respective scales. A *scale* is composed of a sequence of notes; each two notes form a certain interval (see Figure 1). An interval is defined by a ratio between two note frequencies f_1 and f_2. For an equal-tempered scale, a semitone (st) is always defined by a frequency ratio of $f_2/f_1 = 2^{1/12}$. An interval of n semitones is defined by a frequency ratio of $f_2/f_1 = 2^{n/12}$, that is, the interval in semitones n between two frequencies f_1 and f_2 is defined by $n = 12\log_2(f_2/f_1)$. Major and minor scales are represented in Figure 1.

Since each tonic manages both a major and a minor mode, there exist a total of 24 keys if we consider an equal-tempered scale and enharmonic equivalence (i.e., we do not distinguish between notes that sound the same but are spelled differently, such as D♯ and E♭). This corresponds to two different keys for each of the 12 semitones within the chromatic octave. These 24 keys can be arranged in a circle of fifths. Each pair of major and minor modes has the same collection of pitch classes and key signature, while the collections of neighboring, 5th-related pairs differ by a one sharp or flat. There are also some relationships between tonalities: the *parallel* major and minor (e.g., C major and C minor) share the same tonic but have different diatonic collections, while the *relative* major and minor (e.g., C major and A minor) share the same diatonic collection but have different tonics. A key is not limited to the pitch classes within its particular diatonic collection. In certain circumstances, the music can use pitch classes outside its tonic major or minor scale without weakening its sense of orientation towards the tonic. If the orientation towards the tonic is very strong, the music is considered to be very *tonal*, and in the opposite sense there is the concept of *atonality* (Sadie et al. 2004).

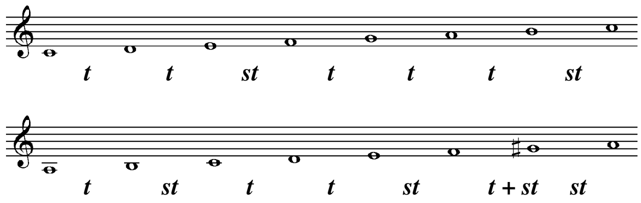


Figure 1 C Major and A Minor (Harmonic) Scales (*T*: Tonic; *SD*: Subdominant, *D*: Dominant Degrees)

Note. Intervals between consecutive notes are shown: *t*: Tone; *st*: Semitone.

Harmony is a term that denotes the simultaneous combination of notes, called *chords*, and over time, chord *progressions*. The term is used to describe notes and chords and also to denote a system of structural principles governing their combination. In the latter sense, harmony has its own body of theoretical literature (Sadie et al. 2004). In this paper we will consider only the aspects of the harmonic content related to the combination of notes into chords, and its relation to the tonality of the piece.

This paper presents a method to extract automatically a set of descriptors from polyphonic audio signals. These features represent the tonal content of the audio excerpt we analyze, including instantaneous descriptors related to harmony and global features related to tonality (as key note, mode, and tonal strength). The paper is organized as follows. Section 2 summarizes work related to tonal description of audio material, which situates the present study. In §3, we propose a scheme to structure tonal descriptors of audio material. Sections 4 and 5 propose a method to extract automatically these tonal descriptors from audio signals. Section 6 presents some results and an evaluation of the proposed methods in the context of key estimation. Finally, in §7, we present our conclusions from this study and plans for future work.

2. Related Work

Many efforts have been devoted to the analysis of chord sequences and key in MIDI representations of classical music, but little work has dealt directly with audio signals or other genres. The use of MIDI-oriented methods would need a previous step of automatic transcription of polyphonic audio, which is a very difficult task. See Klapuri (2003) for an extensive study on this issue.

Other approaches do not require a transcription, but work directly with audio recordings to extract information related to the pitch-class distribution of music. The pitch-class distribution of music is, somehow, directly related to the chords and the tonality of a piece. Chords can be recognized from the pitch-class distribution without precisely detecting which notes are played. Tonality can be also estimated from the pitch-class distribution without a previous chord-estimation procedure.

Fujishima (1999) proposed a system for chord recognition based on the pitch-class profile (PCP), a low-level feature. PCP, as formulated by Fujishima, is a twelve dimensional vector representing the intensities of the twelve semitone pitch classes. This chord-recognition system compares this vector with a set of chord-type templates to estimate the played chord. Based on this vector of features, Sheh and Ellis (2003) also introduced hidden Markov models to estimate the chords within an audio recording. They obtained a maximum of 26% frame accuracy using an evaluation corpus of 20 Beatles songs, which, according to the authors, was not yet sufficient to provide usable chord transcriptions of unkown audio.

In the context of beat estimation of drumless audio signals, Goto and Muraoka (1999) also introduced the computation of a histogram of frequency components, used to detect chord changes. This method did not require chord names to be identified, as the goal was to use these histograms to track beats at different rhythmic levels.

Constant Q profiles have also been used to characterize the tonal content of audio (Purwins et al. 2000). Constant Q profiles are twelve-dimensional vectors, each component referring to a pitch class, which are computed with the constant Q filter bank (Brown and Puckette 1992). Purwins et al. (2003) presented an example where constant Q profiles are used to track the tonal center of Chopin's C-minor Prélude op. 28, No. 20. They recently used these features to analyze the interdependence of pitch class and key as well as key and composer, building co-occurrence matrices from recorded music.

Tzanetakis (2002) also proposed a set of features related to audio harmonic content in the context of musical genre classification. These features derive from a pitch histogram that is computed from MIDI or audio data obtained from MIDI. They represent the most common pitch class used in the piece, the frequency of occurrence of the main pitch class, and the range of the pitches of a song. These first two aspects are also captured by the features we propose in this article.

We see that tonal description of audio recordings is still not a solved problem. Chord recognition can be seen as a previous step to the characterizing of the tonal aspects of a piece of music. We would need to analyze chord sequences to represent the global tonality, and key estimation of audio recordings is still in the early stages of research. Current methods for tonal description have not been largely evaluated within audio musical collections, and there is the need to define proper evaluation frameworks and metrics.

Finally, semantic descriptors related to tonal aspects of music are now limited to key and mode, and therefore there is a need to explore new semantic descriptors measuring, for instance, how *tonal* a piece of music is. There is also motivation to study how tonality is related to other high-level concepts such as *mood* or *genre*.

The goal of this work is to contribute to the state of the art in tonal description. Our goal is to define a set of features to characterize different aspects of tonality, to present algorithms to extract these features automatically, and to evaluate the system for key estimation of a large audio database.

3. Description Scheme

We define different levels of abstraction and temporal scales to describe the tonal aspects of a music excerpt. This idea has already appeared in the literature (Dannenberg 1993, Leman 2002). Dannenberg proposed musical representations at different levels, from the most abstract (printed music) to the most concrete level (the audio signal). He mentioned the need for a multiple-hierarchy scheme to represent music, which should be extensible to support new concepts and structures. According to Leman et al. (2002), different representational levels for music exist, and they provide cues for content representation of musical audio.

In our system, we extract some features related to two main *temporal scales*:

- Instantaneous: we define features attached to an analysis frame, which is the minimal temporal unit that we consider. We define a set of instantaneous features.
- Global: here we define some features related to a wider audio segment, which, for example, could be a phrase, a chorus, or a whole song.

We also distinguish two *levels of abstraction*:

- Low-level signal descriptors: features related to the audio signal.
- High-level descriptors: textual labels computed from low-level descriptors after an inductive inference procedure, and related to the musical content of the signal (e.g., chord, key).

Table 1 shows examples of descriptors belonging to each of the defined categories. In this paper, we focus on instantaneous and global low-level descriptors and high-level descriptors representing the key or the tonal content of a piece of music. These representations are available for content retrieval and navigation across digital music collections. A user could, for instance, look in his or her collection for items in the *minor* mode, or search for similar songs to a chosen one regarding tonality. Another example would be to use tonality in combination with other musical aspects (as instrumentation or rhythm) to classify musical genres.

Table 1 **Example of Descriptors Classification**

Name	Temporal scale	Level of abstraction
HPCP	Instantaneous	Low
Chord	Instantaneous	High
Average HPCP	Global	Low
Key	Global	High

4. Computational Approach

Our computational approach for the extraction of tonal features includes the following steps:

1. *Preprocessing* of the audio signal
2. *Instantaneous descriptor computation* of the harmonic pitch class profile (HPCP) vector. This second step is illustrated in Figure 2. We compute a vector of features attached to an analysis frame following this procedure:
   (a) Spectral analysis of the input signal.
   (b) Calculation of the spectral peaks of the spectrum, which are the local maxima of the magnitude spectrum.
   (c) HPCP vector computation, being a vector of low-level descriptors (LLDs) or features.

In this frame level, we could perform chord recognition by comparing the HPCP vector to a set of chord templates, as presented in Fujishima (1999) and Sheh and Ellis (2003). On the other hand, in this paper we focus on the definition of some descriptors related to

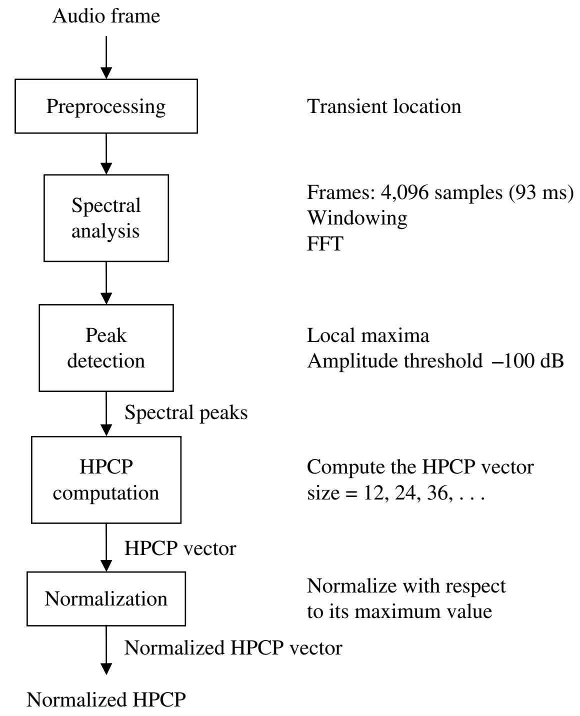


Figure 2 **Block Diagram for HPCP Computation**

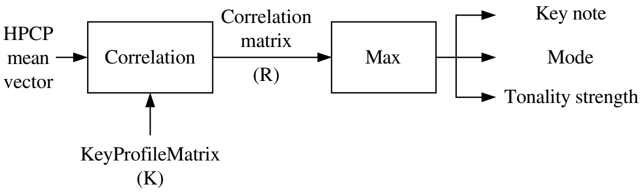


Figure 3 Block Diagram for Key Computation Using HPCP

tonality defined at a global level, i.e., features attached to the whole audio excerpt. Both approaches can be considered necessary and complementary in powerful music information retrieval (MIR) systems.

3. *Global descriptor computation*: the procedure to extract this set of descriptors is illustrated in Figure 3:

(a) Compute statistics of the HPCP vector over an audio excerpt. These statistics are the average and the variance of the instantaneous vector over the whole signal. We consider them global low-level features related to tonality. We assume the key is constant over the entire audio segment.

(b) Estimate the tonality using a key model based on Krumhansl (1990), which has been adapted here to deal with polyphony and audio features. We perform a correlation of the average HPCP vector with a set of key profiles to estimate the key note and mode, as well as the tonal strength or degree of *tonalness*.

These steps are explained in detail in the next sections.

4.1. Preprocessing

For preprocessing, we use a transient detection algorithm to eliminate regions where the harmonic structure is noisy, so that the areas located 50 ms before and after the transients are not analyzed. This preprocessing decreases the computational cost of the HPCP computation, and it degrades neither the resolution nor the effectiveness of the tonality descriptors. The transient detection algorithm used in this work was proposed by Bonada (2000) in the context of a time-scale audio modification algorithm.

4.2. Spectral Analysis and Peak Detection

The input signal is split into analysis frames, and each of the frames is multiplied by a window function. Then, we perform the fast Fourier transform (FFT) to get the frequency spectrum. The frames are long enough in order to have good frequency resolution: we use a frame size of $N_{frame} = 4{,}096$ samples, that is, $T_{frame} = 93$ ms for a sample rate of 44.1 KHz. Frames are overlapped with a hop size of $N_{hop} = 512$ samples (that is approximately $T_{hop} = 11$ ms). We use a *Blackman Harris 62 dB* window for analysis.

We define a "peak" as a local maximum in the magnitude spectrum, where the only constraints are that its frequency belongs to a certain range and that its magnitude is higher than a threshold. Due to the sampled nature of the spectra, each peak is accurate only to within half a spectral bin. A spectral sample represents a frequency interval of f_s/N Hz, where f_s is the sampling rate and N is the FFT size. Zero-padding and quadratic interpolation in the magnitude spectrum are used to refine the peak frequency estimate (Serra 1996).

4.3. HPCP Computation

The HPCP computation is based on PCP. As explained before, this vector measures the intensity of each of the twelve semitones of the diatonic scale, by mapping each frequency bin of the spectrum to a given pitch class. The HPCP introduces some modifications with respect to the computation proposed by Fujishima: first, we only use the spectral peaks in a certain frequency band to compute the HPCP intensity; second, we introduce a weight into the feature computation; third, we use higher resolution in the HPCP bins (decreasing the quantization level to less than a semitone). We only consider those spectral peaks whose frequency belongs to the interval $[100, 5{,}000]$ Hz, not considering very low and high frequencies in our analysis. One reason to ignore high and low frequencies is that the predominant audio objects are more noisy in these regions due to some percussion and instrumental noise (blowing, string frictions, etc.). If we discard these frequency regions, we optimize the speed of the algorithm (as the number of counted peaks decreases) while removing this noise from the HPCP values.

$$\mathrm{HPCP}(n) = \sum_{i}^{nPeaks} w(n, f_i) a_i^2 \qquad (1)$$

where $n = 1, \ldots, size$, a_i is the linear magnitude of the ith peak, and f_i is the frequency value of the ith peak. $i = 1, \ldots, nPeaks$, where $nPeaks$ is the number of spectral peaks that we consider, n is the HPCP bin, $size$ is the size of the HPCP vector (i.e., number of bins: 12, 24, 36, or other), and $w(n, f_i)$ is the weight of the frequency f_i when considering the HPCP bin n.

4.3.1. Weighting Function. Instead of contributing to a single bin (as, for example, the closest one), each frequency f_i contributes to the HPCP bin(s) that are contained in a certain window around its frequency value, as shown in Figure 4. For each of those bins, the contribution of the peak i with frequency f_i (which is the square of the peak linear amplitude a_i^2) is weighted using a $\cos^2$ function around the frequency of the bin. The value of the weight depends on the frequency distance between f_i and the center frequency of the bin n, f_n, measured in semitones, as follows:

Let the center frequency of the nth bin be:

$$f_n = f_{ref} 2^{n/size} \qquad (2)$$

where $n = 1, \ldots, size$. Let the distance in semitones between the peak frequency f_i and the bin center frequency f_n be

$$d = 12 \log_2\left(\frac{f_i}{f_n}\right) + 12\ m \tag{3}$$

where m is the integer that minimizes the magnitude of the distance $|d|$. Then, the weight is computed as follows:

$$w(n, f_i) = \begin{cases} \cos^2\left(\dfrac{\pi}{2}\dfrac{d}{0.5 \cdot l}\right) & \text{if } |d| \leq 0.5l \\ 0 & \text{if } |d| > 0.5l \end{cases} \tag{4}$$

where l is the length of the weighting window. This value is a parameter of the algorithm, and we have set it empirically to 4/3 semitones. Consider a HPCP of $size = 36$. Then, the bin resolution will be 1/3 semitone. If we consider $l = 4/3$ semitones, each spectral peak will contribute to four different HPCP bins with different weights, as illustrated in Figure 4.

This weighting procedure minimizes the estimation errors that we find when there are tuning differences and inharmonicity present in the spectrum. These factors can induce errors when mapping frequency values into HPCP bins.

4.4. Normalization

For each analysis frame, the HPCP values are normalized with respect to its maximum value, in order to store the relative relevance of each of the HPCP bins.

$$\text{HPCP}_{normalized}(n) = \frac{\text{HPCP}(n)}{\text{Max}_n(\text{HPCP}(n))} \tag{5}$$

where $n = 1, \ldots, size$.

4.5. Tonality Estimation

As shown in Figure 3, in order to estimate the key we compute a correlation of the HPCP vector with a matrix of HPCP profiles corresponding to the different keys K, with size $2 \times 12 \times size$. As it appears in (6), we obtain a correlation value $R(i, j)$ for each key note:

$$R(i, j) = r(\text{HPCP}, K(i, j)) \tag{6}$$

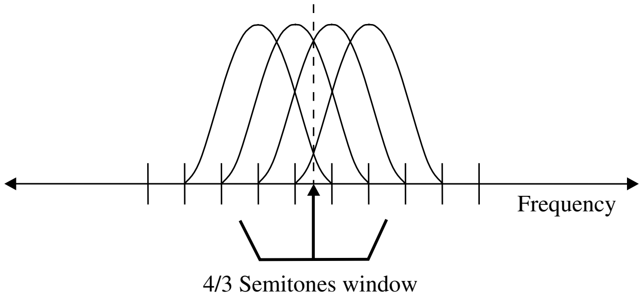


Figure 4 Weighting Function

$K(i, j)$ is the key profile. $i = 1, 2$, where 1 represents the major profile and 2 the minor profile. $j = 1, \ldots, 12$ for the 12 possible key notes. Both vectors (HPCP and $K(i, j)$) have *size* elements.

The maximum correlation value corresponds to the estimated key note and mode (represented by the indexes $i_{\max}$ and $j_{\max}$). We use the maximum correlation value $R(i_{\max}, j_{\max})$ as a measure of the *tonalness*, the degree of tonality or key strength:

$$R(i_{\max}, j_{\max}) = \max_{i,j}(R(i, j)) \tag{7}$$

4.5.1. Key Profile Definition.

To construct the key-profile matrix, we use a method based on the model proposed by Krumhansl and Schmuckler used for key estimation from score representations in musical instrument digital interface (MIDI) format (Krumhansl 1990). This technique considers that tonal hierarchy may be acquired through internalizing the relative frequencies and durations with which tones are played. The algorithm estimates the key from a set of note duration values, measuring how long each of the 12 pitch classes of an octave (C, C♯, etc.) have been played in a melodic line. In order to estimate the key of the melodic line, the vector of note durations is correlated to a set of key profiles or probe-tone profiles. These profiles represent the tonal hierarchies of the 24 major and minor keys, and each of them contains 12 values, which are the ratings of the degree to which each of the 12 chromatic scale tones fit a particular key. They were obtained by analyzing human judgements with regard to the relationship between pitch classes and keys (Krumhansl 1990, pp. 78–81).

In this study we have adapted this technique in two aspects: first to work with HPCP instead of note-duration values, and second to consider polyphony instead of melodic lines. In polyphonic music, more than one note can sound at the same time.

The HPCP value of a given pitch class $HPCP(i)$ represents the relative intensity of the pitch class for the audio excerpt in which it is computed. The intensity is also related to the relative importance of the pitch class within the given tonal context.

We consider the original probe tones (Krumhansl 1990) $T_M(i)$ and $T_m(i)$ for major (represented by the subscript M) and minor scales (represented by the subscript m), $i = 1, \ldots, 12$, as the strength of the pitch class i in a given key. The index i corresponds to the pitch class that is $i - 1$ semitones higher than the key note. Then, in a given major key, $i = 1$ will correspond to the tonic (that is the maximum value), $i = 8$ to the dominant, etc. Major and minor profiles are represented in Figure 5.

The tonal hierarchy within a melodic line should be maintained in a polyphonic situation, where the melodic line is converted into a chord sequence. In this way, we can consider that T_M and T_m represent the

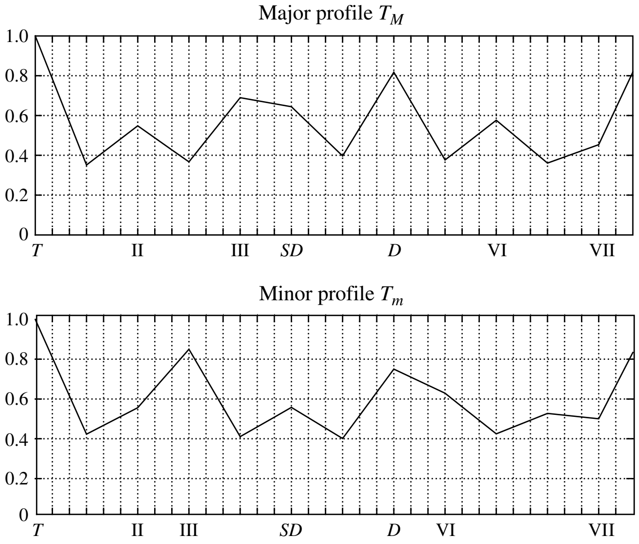


Figure 5 Major and Minor Profiles, T_M and T_m, as Proposed by Krumhansl (1990) Normalized by their Maximum Values

strength of the chord i (tonic, subdominant, dominant chord, etc.) in a given key. Given this assumption, we should 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 ($i = 1$) and dominant ($i = 8$) chords, so that the profile value of the dominant pitch class will be the sum of the contribution of the tonic and the dominant chords of the key. This can be expressed by the equation $T_{Mp}(8) = T_M(1) + T_M(8)$, where p stands for *polyphonic*. We compute the probability of finding a given pitch class in a given major key $T_{Mp}(i)$, as the weighted sum of the probability of finding each of the chords (of a major key) to which this pitch class belongs:

$$T_{Mp}(i) = \sum_{j=1}^{12} \alpha_M(i, j) T_M(j) \tag{8}$$

where $i = 1, \ldots, 12$. In the same way, we compute the probability of finding a given pitch class in a given minor key $T_{mp}(i)$, as the weighted sum of the probability of finding each of the chords (of a minor key) to which this pitch class belongs:

$$T_{mp}(i) = \sum_{j=1}^{12} \alpha_m(i, j) T_m(j) \tag{9}$$

where $i = 1, \ldots, 12$. These equations can be also expressed as follows:

$$T_{Mp} = \alpha_M T_M^t \tag{10}$$

$$T_{mp} = \alpha_m T_m^t \tag{11}$$

where T_{Mp}, T_M, T_{mp}, and T_m are 1×12 vectors, and α_M, α_m are 12×12 square matrices. Within a major key, $\alpha_M(i, j)$ represents the weight of the pitch class i when considering the chord whose root is pitch class j. $\alpha_m(i, j)$ represents the weight of the pitch class i when considering the chord whose root is pitch class j within a minor key. After performing some comparative tests (see §6), we consider only the three main triads of the key as the most representative ones: tonic ($i = 1$), dominant ($i = 8$), and subdominant ($i = 6$), defining the following matrices for major (see (12)) and minor (see (13)) keys:

$$\alpha_M = \begin{pmatrix} 1&0&0&0&0&1&0&0&0&0&0&0\\ 0&0&0&0&0&0&0&0&0&0&0&0\\ 0&0&0&0&0&0&0&1&0&0&0&0\\ 0&0&0&0&0&0&0&0&0&0&0&0\\ 1&0&0&0&0&0&0&0&0&0&0&0\\ 0&0&0&0&0&1&0&0&0&0&0&0\\ 0&0&0&0&0&0&0&0&0&0&0&0\\ 1&0&0&0&0&0&0&1&0&0&0&0\\ 0&0&0&0&0&0&0&0&0&0&0&0\\ 0&0&0&0&0&1&0&0&0&0&0&0\\ 0&0&0&0&0&0&0&0&0&0&0&0\\ 0&0&0&0&0&0&0&1&0&0&0&0 \end{pmatrix} \tag{12}$$

$$\alpha_m = \begin{pmatrix} 1&0&0&0&0&1&0&0&0&0&0&0\\ 0&0&0&0&0&0&0&0&0&0&0&0\\ 0&0&0&0&0&0&0&1&0&0&0&0\\ 1&0&0&0&0&0&0&0&0&0&0&0\\ 0&0&0&0&0&0&0&0&0&0&0&0\\ 0&0&0&0&0&1&0&0&0&0&0&0\\ 0&0&0&0&0&0&0&0&0&0&0&0\\ 1&0&0&0&0&0&0&1&0&0&0&0\\ 0&0&0&0&0&1&0&0&0&0&0&0\\ 0&0&0&0&0&0&0&0&0&0&0&0\\ 0&0&0&0&0&0&0&0&0&0&0&0\\ 0&0&0&0&0&0&0&1&0&0&0&0 \end{pmatrix} \tag{13}$$

Finally, some modifications are made to the profiles to take into account the fact that we work with audio features (HPCP values) instead of MIDI. The spectrum of a note is composed of several harmonics, whose frequencies are multiples of the fundamental frequency (f, $2f$, $3f$, $4f$, etc.). Then, when a note is played, the intensity increases at the frequency of the different harmonics. This fact appears in the HPCP values, where the value increases for different indices.

If we consider an equal-tempered scale, the index i_n associated to the nth harmonic of a note (we can call it the nth *harmonic pitch class*) can be computed as follows:

$$i_n = \text{mod}\,[(i_1 + 12\log_2(n)), 12]; \tag{14}$$

Table 2 Contribution for the First Six Harmonics of a Note

i_n	Frequency	Factor
1	f	1
2	$2f$	s
3	$3f$	s^2
4	$4f$	s^3
5	$5f$	s^4
6	$6f$	s^5

where i_1 is the pitch class that corresponds to the note's fundamental frequency.

The profile values for a given pitch class i ($T_{Mp}(i), T_{mp}(i)$) are equal to the sum of contributions of all the pitch classes containing i as a harmonic pitch class. That means that each note of the chords under consideration (associated to each of the "1" values within α_M and α_m matrices) contributes to the profile values of its i_n harmonics ($n = 1, 2, \ldots, nHarmonics$). We make the contribution decrease with the frequency by multiplying this contribution with a function with a certain decay s, in order to simulate that the spectrum amplitude decreases with frequency. This is illustrated in Table 2.

The spectral decay factor s has been empirically set to 0.6, and we have considered the first four harmonics for computation ($nHarmonics = 4$). If a harmonic is located between two different pitch classes, we use the same weighting scheme as for HPCP computation. The final profiles T_{Mp} and T_{mp} are represented in Figure 6.

In order to build the profile matrix for the 24 different keys, K (in (6)), we consider that the tonal hierarchy (related to major and minor keys) is invariant with respect to the chosen key note. This means, for instance, that the profile for B major will be the same as A major but shifted two bins (corresponding to the two semitones between A and B). We perform a linear interpolation between bins to compute *size* bins from the original 12 values in $T_{T_{Mp}}$ and $T_{T_{mp}}$. As explained before (and shown in Figure 3), a correlation value is computed between the global HPCP value and each of the 24 key profiles. The one with the highest correlation value is chosen as the estimated tonality.

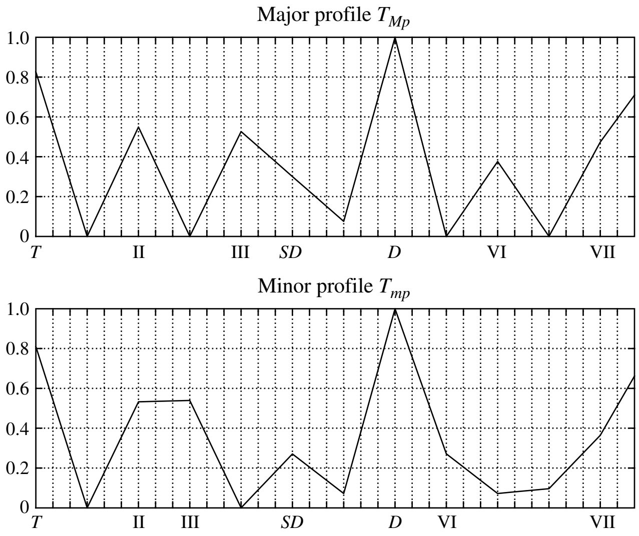


Figure 6 T_{Mp} and T_{mp} Profiles

5. Case Study

In this section, we present some examples of how the algorithm works. Figure 7 shows the results of the analysis of a polyphonic excerpt in C major. Average HPCP values, as well as correlation with major and minor models, are shown. In the HPCP profile, we can identify peaks in the tonic C and dominant G notes, as well as in the major third E. The correlation shows a maximum value of 0.84 in C major. Other peaks appear a 5th above within the circle of fifths (which is G major and its relative minor E minor), as well as in C minor, which shares the dominant and subdominant chords (considering harmonic minor scale). In both examples we can check that the features represent some aspects of tonal content. These aspects are considered in other models that represent the relationships within keys (Temperley 1999, Chew 2000).

Figure 8 shows the results of the analysis of a polyphonic excerpt in F♯ minor. As in the last example, mean HPCP values, as well as correlation with major and minor models are shown. In the HPCP profile, we can identify peaks in the tonic F♯ and dominant C♯ notes, as well as on the minor third A. The correlation shows a maximum value 0.79 in F♯ minor. Other peaks appear a 5th below within the circle of fifths (which is D major and its relative minor B minor), as well as in F♯ major, which shares the dominant chord (considering harmonic and melodic minor scales) and the subdominant chord (considering melodic minor scale).

Figure 9 shows the results of the analysis for a polyphonic percussion excerpt, where we cannot identify any key. In the HPCP profile there is no clear peak, while the correlation peaks barely exceed 0.5. In this case, the low correlation values indicate that the key is not well defined.

These examples show how the system works. Some additional material can be found in the Online Supplement to this paper on the journal's website.

6. Evaluation

For evaluation we have set up a small test database with 35 samples (passages of polyphonic musical pieces) of different styles, and in various key and modes, labeled by hand in terms of key. Each excerpt has a constant tonality. The results are presented in Table 3 and in Figure 10. For this test database, we

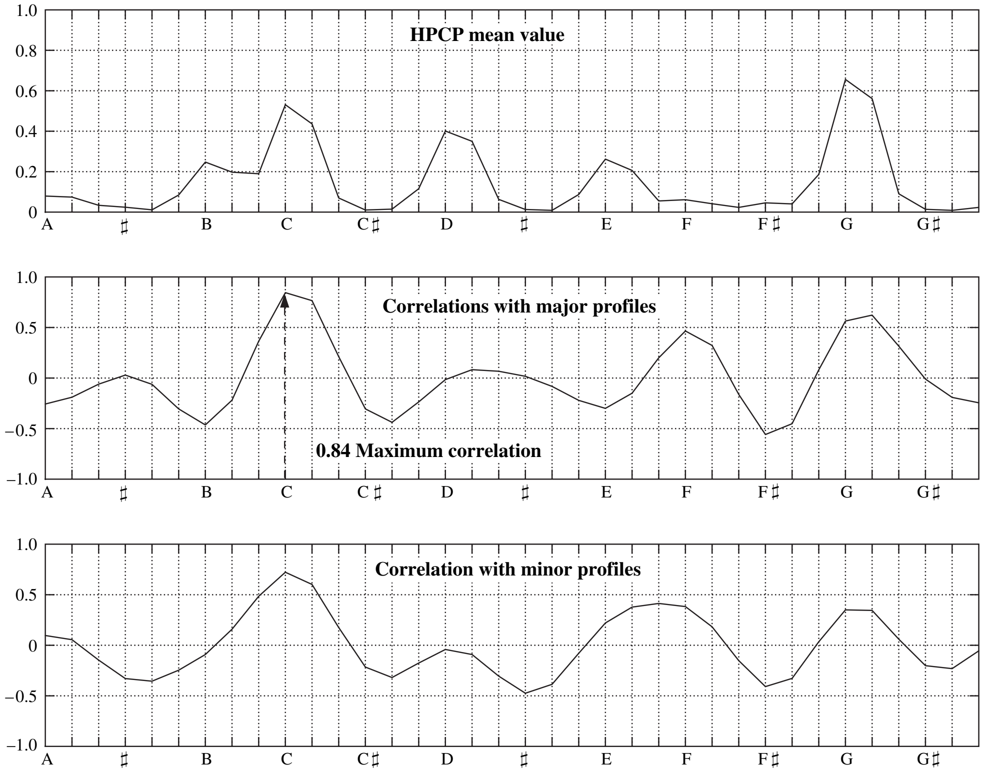


Figure 7 Example of HPCP Profile for a C Major Key Audio Excerpt

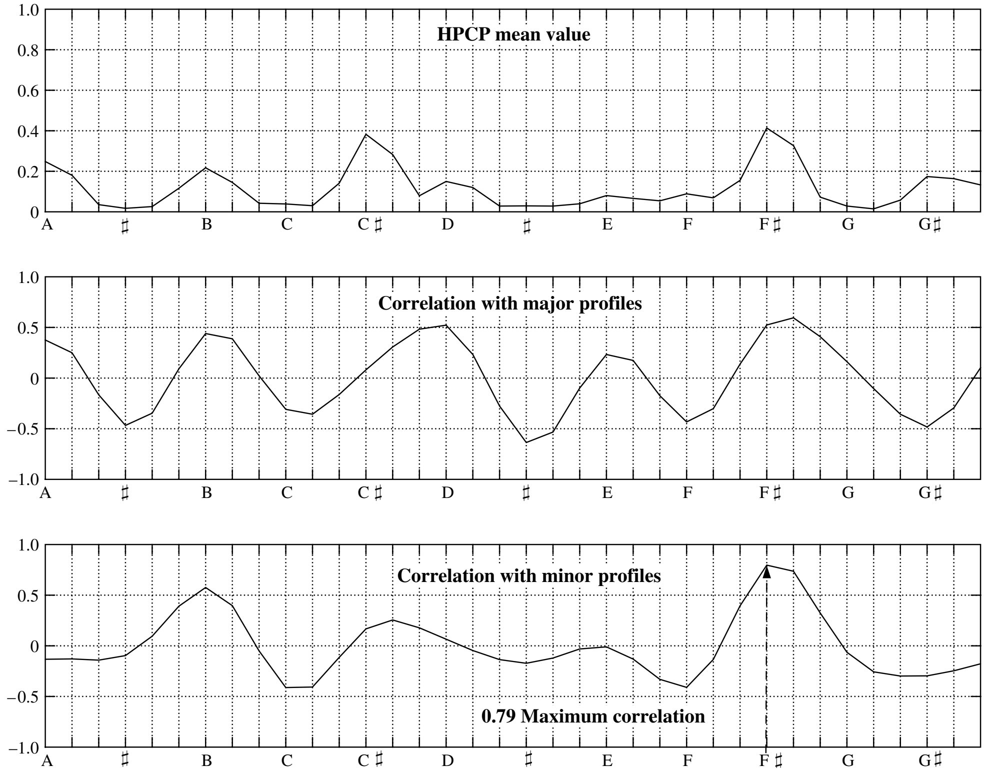


Figure 8 Example of HPCP Profile for an F♯ Minor Key Audio Excerpt

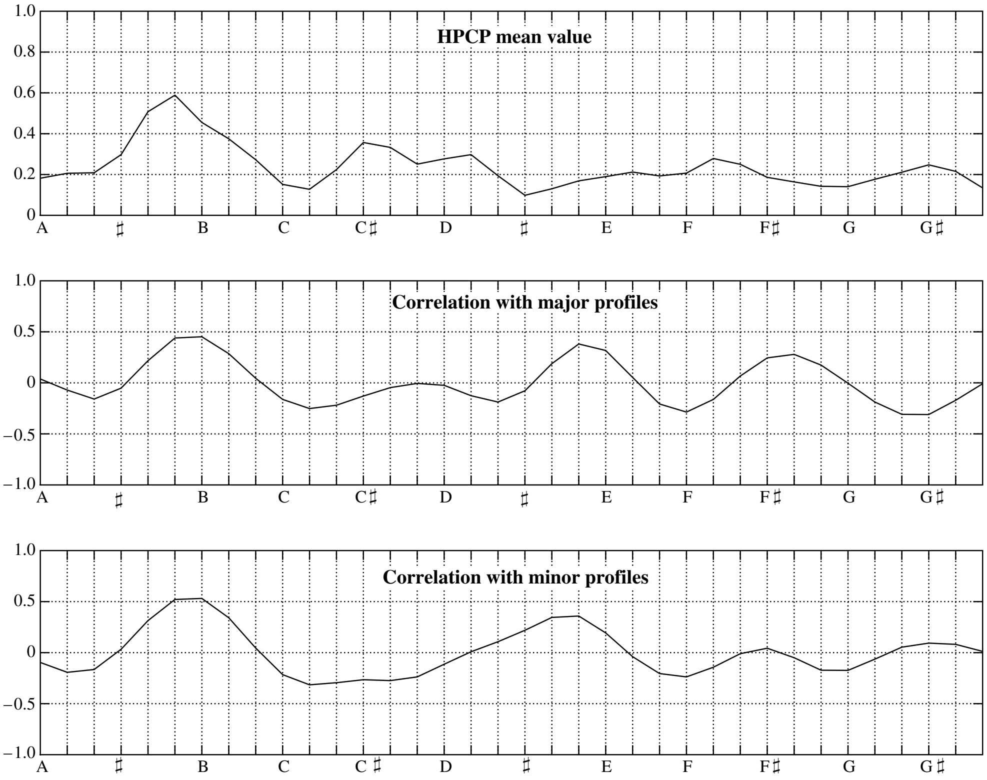


Figure 9 HPCP Profile for a Percussive Sound (No Clear Tonality)

have tested different configurations in order to measure the improvements of the following aspects:

1. Comparison of PCP (Fujishima 1999) or HPCP for key estimation. The original PCP algorithm was modified to compute 36 values instead of 12, in order to get a better frequency resolution.

2. Comparison of the original profiles T_M and T_m (Krumhansl 1990) against the proposed $T_{Mchords}$ and $T_{mchords}$, built either considering all the chords of the key or only the three most relevant chords (i.e., tonic, dominant, and subdominant chords), as explained in §4.

Results are presented in Table 3, where several evaluation metrics are shown. The first ones are the percentage of correct key estimation (both tonic and mode), the percentage of correct mode estimation and the percentage of correct key note estimation. These three metrics evaluate how the algorithm performs for key, mode, and tonic (key note) estimation. Percentages are also shown for the cases where the error is only in mode estimation (i.e., confusions with major/minor relatives) and when the error is only in key-note estimation. Finally, we show the percentage of the items where both key note and mode were not correctly estimated.

Table 3 Results of Evaluation of a Small Test Database (%)

			HPCP	HPCP	PCP
	PCP	HPCP	All	Main	Main
Percentage	Krumhansl	Krumhansl	chords	chords	chords
Correct	36.4	51.5	**75.8**	**75.8**	42.4
Correct mode	69.7	75.8	87.9	**90.9**	66.7
Correct key note	48.5	54.5	75.8	**78.8**	51.5
Only mode errors	33.3	24.2	12.1	15.2	24.2
Only key note errors	12.1	3	0	3	9.1
Key note and mode errors	18.2	21.2	12.1	6.1	24.2

We observe that there are significant improvements when using HPCP and modified Krumhansl profiles, obtaining a 75.8% of correct estimation, 90.9% of correct mode estimation, and 78.8% of correct key note estimation when using the three main chords of the major and minor keys. When the window size is increased to 371 ms (16,384 samples), the performance also increased to 81.8%. This performance was achieved when considering the main chords to build the profiles, and the same results were obtained using HPCP and PCP ($size = 36$ bins). That suggests that HPCP is more robust than PCP with smaller window sizes, i.e., lower frequency resolution for FFT.

We also evaluated the performance of the algorithm on a database of 883 classical music pieces segmented by track and labeled by title (e.g., *Mozart, Flute Concerto No. 1 K313 G Major Andante non troppo*). They include composers such a Mozart, Chopin, Scarlatti, Bach, Brahms, Beethoven, Handel, Pachelbel, Tchaikovsky, Sibelius, Dvorak, Debussy, Telemann, Albinoni, Vivaldi, Pasquini, Glenn Gould, Rachmaninoff, Schubert, Shostakovich, Haydn, Benedetto,

Table 4 **Results of Evaluation of a Classical Test Database**

	Percentage
Correct	66.1
Correct mode	87.5
Correct key note	68.9
Only mode errors	2.72
Only key note errors	21.4
Key note and mode errors	9.74

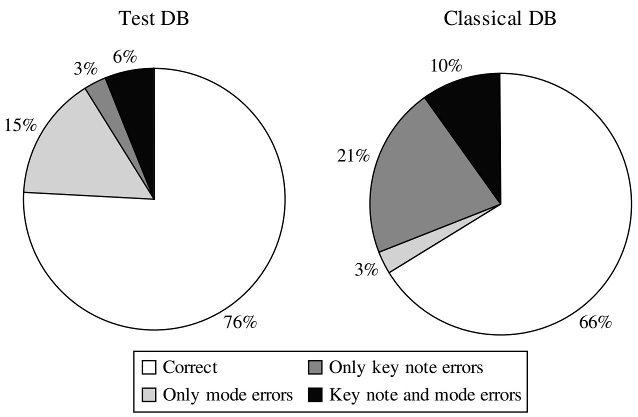
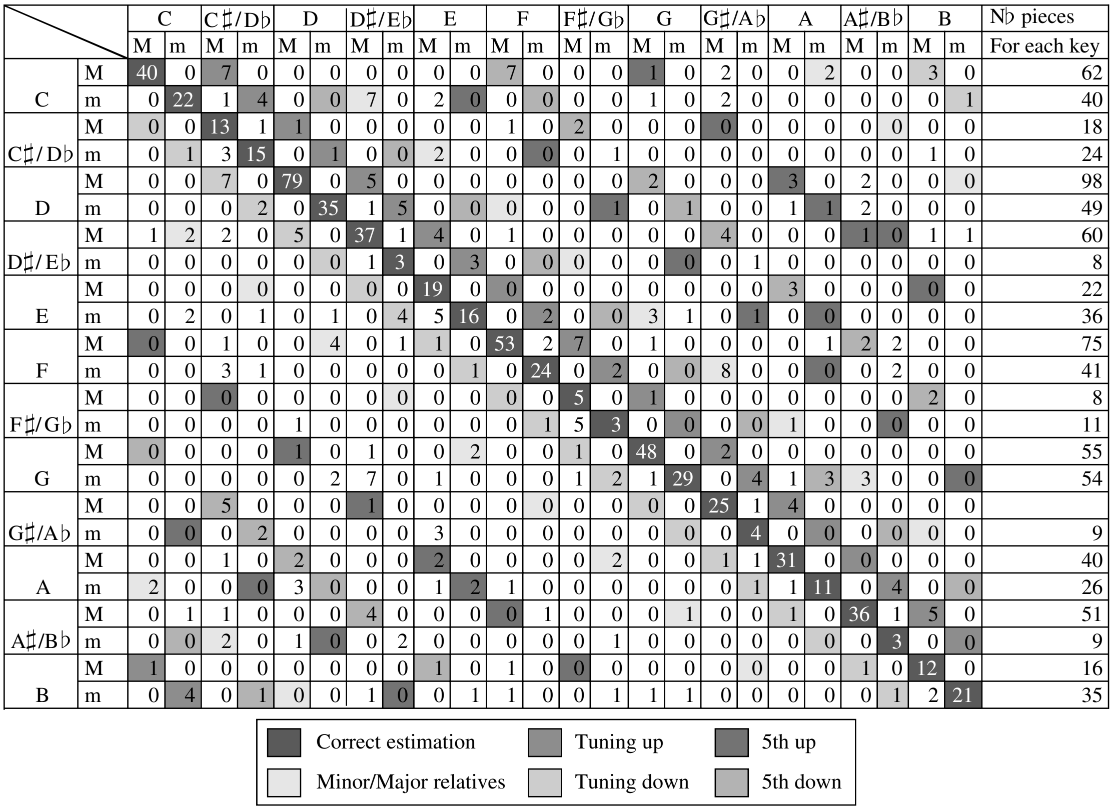


Figure 10 **Evaluation Results**

Elgar, Bizet, Listz, Boccherini, Ravel, Debussy, etc. We also included some jazz versions of classical pieces (e.g., Jacques Lousier, The Swingle Singers). Most of the included pieces were first movements (in the case that the piece is a multi-movement form as a sonata or symphony). All the key note and mode annotations were taken from the FreeDB database (FreeDB 2004). Some additional manual corrections were done because of incorrect FreeDB metadata, although systematic checking was not performed. We assumed that the key is constant within the whole piece. That means that the modulations we find do not modify the overall key of the piece. We considered the parameters that yield better results with the small test database: HPCP and three main chords, using a window size of 4,096 samples (93 ms). The results are presented in Table 4 and in Figure 10. Correct estimation is obtained for 66.1% of the items, and mode is estimated correctly for 87.5% of the pieces. The majority of errors occur when the mode is correctly estimated but not the key note.

In Figure 11, we present the confusion matrix for the estimation. Rows represent the correct tonality, and columns the estimated tonality. We find that 5% of the errors were due to minor/major relatives, 11% to tuning errors (one semitone higher or lower), 14% of the errors estimated the upper 5th within the circle of fifths, and 40% a 5th lower within the



		C		C♯/D♭		D		D♯/E♭		E		F		F♯/G♭		G		G♯/A♭		A		A♯/B♭		B		N♭ pieces
		M	m	M	m	M	m	M	m	M	m	M	m	M	m	M	m	M	m	M	m	M	m	M	m	For each key
C	M	40	0	7	0	0	0	0	0	0	0	7	0	0	0	1	0	2	0	0	2	0	0	3	0	62
	m	0	22	1	4	0	0	7	0	2	0	0	0	0	0	1	0	2	0	0	0	0	0	0	1	40
C♯/D♭	M	0	0	13	1	1	0	0	0	0	0	1	0	2	0	0	0	0	0	0	0	0	0	0	0	18
	m	0	1	3	15	0	1	0	0	2	0	0	0	0	1	0	0	0	0	0	0	0	0	1	0	24
D	M	0	0	7	0	79	0	5	0	0	0	0	0	0	0	2	0	0	0	3	0	2	0	0	0	98
	m	0	0	0	2	0	35	1	5	0	0	0	0	0	1	0	1	0	0	1	1	2	0	0	0	49
D♯/E♭	M	1	2	2	0	5	0	37	1	4	0	1	0	0	0	0	0	4	0	0	0	1	0	1	1	60
	m	0	0	0	0	0	0	1	3	0	3	0	0	0	0	0	0	0	1	0	0	0	0	0	0	8
E	M	0	0	0	0	0	0	0	0	19	0	0	0	0	0	0	0	0	0	3	0	0	0	0	0	22
	m	0	2	0	1	0	1	0	4	5	16	0	2	0	0	3	1	0	1	0	0	0	0	0	0	36
F	M	0	0	1	0	0	4	0	1	1	0	53	2	7	0	1	0	0	0	0	1	2	2	0	0	75
	m	0	0	3	1	0	0	0	0	0	1	0	24	0	2	0	0	8	0	0	0	0	2	0	0	41
F♯/G♭	M	0	0	0	0	0	0	0	0	0	0	0	0	5	0	1	0	0	0	0	0	0	0	2	0	8
	m	0	0	0	0	1	0	0	0	0	0	0	1	5	3	0	0	0	0	1	0	0	0	0	0	11
G	M	0	0	0	0	1	0	1	0	0	2	0	0	1	0	48	0	2	0	0	0	0	0	0	0	55
	m	0	0	0	0	0	2	7	0	1	0	0	0	1	2	1	29	0	4	1	3	3	0	0	0	54
G♯/A♭	M	0	0	5	0	0	0	1	0	0	0	0	0	0	0	0	0	25	1	4	0	0	0	0	0	
	m	0	0	0	2	0	0	0	0	3	0	0	0	0	0	0	0	0	4	0	0	0	0	0	0	9
A	M	0	0	1	0	2	0	0	0	2	0	0	0	0	2	0	0	1	1	31	0	0	0	0	0	40
	m	2	0	0	0	3	0	0	0	1	2	1	0	0	0	0	0	0	1	1	11	0	4	0	0	26
A♯/B♭	M	0	1	1	0	0	0	4	0	0	0	0	1	0	0	0	1	0	0	1	0	36	1	5	0	51
	m	0	0	2	0	1	0	0	2	0	0	0	0	0	1	0	0	0	0	0	0	0	3	0	0	9
B	M	1	0	0	0	0	0	0	0	1	0	1	0	0	0	0	0	0	0	0	0	1	0	12	0	16
	m	0	4	0	1	0	0	1	0	0	1	1	0	0	1	1	1	0	0	0	0	0	1	2	21	35

■ Correct estimation ▨ Tuning up ■ 5th up
□ Minor/Major relatives ▨ Tuning down ▨ 5th down

Figure 11 **Confusion Matrix for the Classical Database**
Note. Rows represent the correct key whereas columns represent the estimated one. The number of confusions is written in the corresponding cell.

circle of fifths. The confusion matrix reveals the need to consider, both in the key model and in the analysis of the correlation functions, some aspects of perceptual similarities, already pointed out in the literature (Temperley 1999, Chew 2000, Janata et al. 2002).

7. Conclusions and Further Work

In this paper, we have presented a system that can automatically extract a set of features related to tonality from polyphonic audio. These features can be used to establish similarity between audio excerpts and to navigate through digital collections.

It has been quite difficult to evaluate the system, as we could find neither labeled databases nor other algorithms evaluated in the same way, in order to compare the results. To solve the problem, we built two polyphonic audio databases for evaluation, with different instrumentations and styles, using editorial annotated metadata and some manual corrections. Correct estimation is achieved for 66% of the items, and mode is correctly found for 87.5% of the cases. Many of the estimation errors are associated with neighboring tonalities. These results reveal that the system might be useful within a music information retrieval framework, in combination with other musical aspects such as instrumentation, rhythm, and structural description.

There are many things to improve and ideas for future work. We would like to compare the proposed set of features with others found in the literature (Tzanetakis 2002, Purwins et al. 2000). We also want to compare the key model used with other existing ones (Temperley 1999, Chew 2000) and with some machine-learning strategies. In addition, in order really to analyze the key structure of a piece, we need to trace key changes or *modulations*, performing tonality tracking on the data.

There is also a need to introduce some other aspects of tonal perception into our current model. Another idea is to combine the system with a chord-estimation algorithm as a way to enrich the description and to support the key-estimation process. Finally, interesting work remains to be done regarding tonal similarity and the use of the tonal strength or *"tonalness"* as a proper semantic descriptor.

Acknowledgments

The author thanks Takuya Fujishima for his explanations on the PCP, Perfecto Herrera for his help with the document and ideas for the key model, and Jordi Bonada for his transient detection algorithm for supporting the implementation and for his ideas regarding the HPCP computation. Finally, the author acknowledges the help of the journal reviewers and the following people from MTG who have also served as reviewers of the document: Pedro Cano, Oscar Celma, Alexandra Gardner, Fabien Gouyon, Enric Guaus, and Sebastian Streich. This work has been partially funded by European IST Project 507142 SIMAC (Semantic Interaction with Music Audio Contents, http://www.semanticaudio.org) and Spanish National Project TIC2003-07776-C02-02 Promusic.

References

Bonada, J. 2000. Automatic technique in frequency domain for near-lossless time-scale modification of audio. *Internat. Comput. Music Conf., Berlin.* ICMA, San Francisco, CA, 396–399.

Brown J. C., M. S. Puckette. 1992. An efficient algorithm for the calculation of a constant *Q* transform. *J. Acoustic Soc. America* **92** 2698–2701.

Chew, E. 2000. Towards a mathematical model of tonality. Ph.D. thesis, Operations Research Center, MIT, Cambridge, MA.

Dannenberg, R. B. 1993. A brief survey of music representation issues, techniques and systems. *Comput. Music J.* **17** 20–30.

FreeDB. 2004. Free DB database, online publication, accessed February 2004. http://www.freedb.org.

Fujishima, T. 1999. Realtime chord recognition of musical sound: A system using common lisp music. *Internat. Comput. Music Conf., Beijing, China.* ICMA, San Francisco, CA, 464–467.

Goto, M., Y. Muraoka. 1999. Real-time beat tracking for drumless audio signals: Chord change detection for musical decisions. *Speech Comm.* **27** 311–335.

Janata, P., J. L. Birk, J. D. Van Horn, M. Leman, B. Tillmann, J. J. Bharucha. 2002. The cortical topography of tonal structures underlying western music science. *Science* **298** 2167–2170.

Klapuri, A. 2003. Signal processing methods for the automatic transcription of music. Ph.D. thesis, Signal Processing Laboratory, Tampere Institute of Technology, Tampere, Finland.

Krumhansl, C. L. 1990. *Cognitive Foundations of Musical Pitch.* Oxford University Press, New York.

Leman, M. 2002. Musical audio mining. J. Meij, ed. *Dealing with the Data Flood Symposium, Rotterdam, The Netherlands.* STT Netherlands Study Centre for Technology Trends, Rotterdam, The Netherlands.

Leman, M., L. P. Clarisse, B. De Baets, H. De Meyer, M. Lesaffre, G. Martens, J. P. Martens, D. Van Steelant. 2002. Tendencies, perspectives, and opportunities of musical audio-mining. A. Calvo Manzano, A. Pérez López, J. Salvador Santiago, eds. *Forum Acusticum, Session SS-MUS-01*, Seville, Spain. http://www.ia.csic.es/sea/4375cd.html.

Purwins, H., B. Blankertz, K. Obermayer. 2000. A new method for tracking modulations in tonal music in audio data format. S. I. Amari, C. L. Giles, M. Gori, V. Piuri, eds. *Neural Networks-IJCNN*, Vol. 6. IEEE Computer Society, New York, 270–275.

Purwins, H., T. Graepel, B. Blankertz, K. Obermayer. 2003. Correspondence analysis for visualizing interplay of pitch class, key, and composer. E. L. Puebla, G. Mazzola, T. Noll, eds. *Perspectives in Mathematical Music Theory*. Verlag epOs-Music, Osnabruck, Germany, 432–454.

Sadie, S., J. Tyrrell, B. Kernfeld, eds. 2004. *Grove Music Online: The New Grove Dictionary of Music and Musicians*, 2nd ed. *The New Grove Dictionary of Opera and The New Grove Dictionary of Jazz*, 2nd ed. http://www.grovemusic.com.

Serra, X. 1996. Musical sound modeling with sinusoids plus noise. G. D. Poli, A. Picialli, S. T. Pope, C. Roads, eds. *Musical Signal Processing*. Swets & Zeitlinger, Lisse, The Netherlands, 91–122.

Sheh, A., D. Ellis. 2003. Chord segmentation and recognition using EM-trained hidden Markov models. H. H. Hoos, D. Bainbridge, eds. *Proc. 4th Internat. Sympos. Music Inform. Retrieval*. Baltimore, MD, 183–189. http://www.ismir.net.

Temperley, D. 1999. What's key for key? The Krumhansl-Schmuckler key finding algorithm reconsidered. *Music Perception* **17** 65–100.

Tzanetakis, G. 2002. Pitch histograms in audio and symbolic music information retrieval. M. Fingerhut, ed. *Internat. Sympos. Music Inform. Retrieval*, Paris, France, 31–38. http://www.ismir.net.