Chroma-based estimation of musical key from audio-signal analysis

Geoffroy Peeters
Ircam - Sound Analysis/Synthesis Team, CNRS - STMS
1, pl. Igor Stravinsky
F-75004 Paris - France
peeters@ircam.fr

Abstract

This paper deals with the automatic estimation of key (keynote and mode) of a music track from the analysis of its audio signal. Such a system usually relies on a succession of processes, each one making hypotheses about either the signal content or the music content: spectral representation, mapping to chroma, decision about the global key of the music piece. We review here the underlying hypotheses, compare them and propose improvements over current state of the art. In particular, we propose the use of a Harmonic Peak Subtraction algorithm as a front-end of the system and evaluate the performance of an approach based on hidden Markov models. We then compare our approach with other approaches in an evaluation using a database of 302 baroque, classical and romantic music tracks.

Keywords: key estimation, pitch representation, Harmonic Peak Subtraction, hidden Markov model

1. Introduction

In the field of Music Information Retrieval, automatic estimation of musical key or of chord progression over time for a music track has received much attention in the recent years. This comes from the numerous applications it allows (search/ query music databases, automatic playlists generation and automatic accompaniment) and from the fact that recent studies have shown that reasonable results could be achieved without the need of a symbolic transcription (which is not always available) and without the necessity to extract such a transcription from the audio signal (audio to symbolic notes algorithms are still limited and costly).

In order to do that, most existing algorithms start by a front-end which converts the signal frames to the frequency domain (DFT or CQT [1] [13]) and then map it to the chroma domain [17] (or Pitch Class Profile [4]). Chroma/ PCP vectors represent the intensities of the twelve semitones of the pitch classes over time. Algorithms then try to find the key or chord progression that best explains the succession of chroma-vectors over time. For this, Chew [2] proposes the Spiral Array Model/ Center of Effect Generator, many others ([5] [6] [3] [10]) use key-chroma profiles derived from the probe-tone experiment of Krumhansl & Schmukler [8] or from the modified version proposed by Temperley [15]. These experiments were aimed at describing the perceptual importance of each semitone in a key resulting in a pitch-distribution profile.

Usually, key estimation systems rely on a succession of processes, each one making underlying hypotheses about either the signal content or the music content. The paper is organized according to these various processes: 1) extraction of information about periodicity or pitches from the audio signal, 2) mapping of this information to the chroma/ PCP domain, 3) decision of a global key for the music piece from the succession of chroma-vectors over time.

The first problem of current systems comes from the fact that we do not observe directly the various pitches in a spectral representation (DFT or CQT) but a mixture of their harmonics. This problem can be solved either by extracting the pitches (or removing the harmonics), or by considering the presence of these harmonics during the creation of the key-chroma profiles. The *first choice* is taken for example by Pauws [10] (using a model taking into account simultaneously the perceptual pitch and the musical background), Chuan [2] (using a fuzzy analysis system) or Cremer [3] (using an overtones removal process). The *second choice* is taken for example by Gomez [5] (extending the PCP to Harmonic Pitch Class Profiles by considering a theoretical amplitude contribution of the first 4 harmonics of each pitch of the three main triads in a given key) or Izmirli [6] (the contribution of the harmonics is there measured on a database of piano notes). While the solution in [5] provides a too rough approximation (a large part of musical instrument sounds does not behave as the proposed theoretical spectral envelope), the solution in [6] would require different spectral envelope measures for each specific instrument. In this paper we propose the use a Harmonic Peak Subtraction function, which allows reducing the influence of the higher harmonics of each pitch.

A second problem comes from the method used in order to decide on the global key. A hypothesis is often made about the existence of the key in the beginning of the track. Therefore, only the first part of the track is considered (first 20s of the first movement). Several approaches are taken to estimate the global key from the chroma-vectors over time. Gomez [5] chooses the key corresponding to the key-chroma profile which has the highest correlation with a global average chroma-vector. Izmirli [6] estimates at each time the key-chroma profile which is the most correlated with a cumulated over-time-chroma-vector, assigns to it a score and finally takes the key with the maximum average score. In this paper, we will test both methods. We will also test a proposed decision method based on modeling the succession of chroma-vectors over time by a set of hidden Markov models representing the various keys.

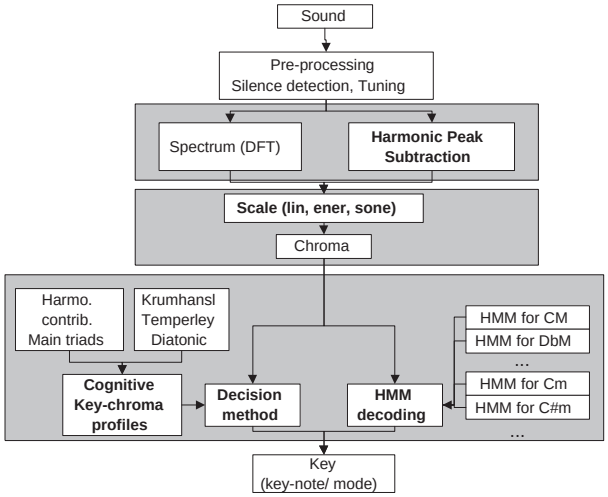

Figure 1. Global flowchart of the key estimation system.

The paper is organized as follows. In section 2.2, we propose our Harmonic Peak Subtraction function for spectral observation of periodicities. In section 2.3, we propose the mapping of it to the chroma domain. We show the importance of the scale used for the mapping and propose the use of a sone scale. In section 2.4, we present the various key decision methods and propose a method based on hidden Markov modeling of the keys. In section 3, we evaluate the performances of our system in comparison with various other systems. The evaluation is performed using a database of 302 baroque, classical and romantic music tracks.

2. Key estimation system

The global flowchart of the key estimation system is indicated in Figure 1. We detail it in the following.

2.1. Pre-processing stages

A set of pre-processing algorithms are first applied to the signal. The signal is first down-sampled to 11025Hz and converted to mono by mixing both channels. The exact *starting time* of the music piece in the sound file is estimated by a method based on loudness and spectral flatness measure. The *tuning of the track* is then found using the method we have proposed in [12]. In short, we test a set of candidate tunings between 427Hz and 452Hz (the quarter-tones below and above A4). For each candidate tuning, we estimate the amount of energy of the spectrum explained by the frequencies corresponding to the semitones based on this candidate tuning. For the database we will use in section 3, we have found tunings ranging from 438 to 447Hz with concentration at 440Hz and 443Hz. Using this estimation, the signal is re-sampled (using a polyphase filter implementation) in order to bring its tuning back to 440Hz. The rest of the system is based on a tuning of 440Hz.

2.2. Spectral observation: Harmonic Peak Subtraction

The front-end of most key estimation systems extracts a spectral representation from the signal. Since this representation will be mapped to the chroma domain, it is important that it represents only information about the pitches and not all their harmonics. Indeed, the presence of the harmonics of the pitches will distort the chroma representation (for example the harmonics $h = 3, 6$ will strengthen the presence of the fifth note and $h = 5$ the presence of the third) and induce error in the key estimation (especially the fifth up/down confusion). In this paper we propose the use of a Harmonic Peak Subtraction function, which allows reducing the influence of the higher harmonics of each pitch.

In the case of **mono-pitch signals**, we have proposed in [11] a function which combines a frequency representation $S(f_k)$ (the DFT or the Auto-Correlation of the DFT) with a temporal representation $r(\tau_l)$ (the Auto-Correlation of the signal or the Real-Cepstrum function) mapped to the frequency domain. The mapping consists in considering that the value of $r(\tau_l)$ is a measure of the periodicity at lag τ_l or at the frequency $1/\tau_l$. We interpolate the values of $r(\tau_l)$ in order to obtain the values of $r(\tau)$ at the same frequency as the DFT $\tau = 1/f_k$. Only the positive values of $r(1/f_k)$ are considered (Half Wave Rectification). We now have two measures of the periodicity at the same frequencies f_k and the final function is obtained by computing the product of both: $h(f_k) = S(f_k) \cdot r(1/f_k)$. This function has been tested in [11] for a task of pitch estimation. For this, we simply take the frequency corresponding to the maximum peak of $h(f_k)$ as the pitch estimation. This process has achieved 97% correct recognition over a large database.

The underlying process of this method is that the ACF (or Real-Cepstrum) $r(\tau)$ can be considered as the decomposition of the power spectrum (log-amplitude spectrum), $A(f_k)$, on a cosine function $g_\tau(f_k) = \cos(2\pi f_k \tau)$ and therefore measures the periodicity of the peaks of $A(f_k)$. This is illustrated in Figure 2 where we superimposed $g_\tau(f_k)$ on $A(f_k)$ for various lags: $\tau = T_0/5$, $\tau = T_0$ and $\tau = 2T_0$. We decompose $g_\tau(f_k)$ into its positive and negative part: $g_\tau(f_k) = g_\tau^+(f_k) - g_\tau^-(f_k)$. *Positive values* of $r(\tau)$ occur only when the contribution of the projection of $A(f_k)$ on $g_\tau^+(f_k)$ is greater than the one on $g_\tau^-(f_k)$ (this is the case for the sub-harmonics of f_0, $\tau = k/f_0,\ k \in \mathbb{N}^+$). *Non-positive values* occur when the contribution of $g_\tau^-(f_k)$ is larger than or equal to the one of $g_\tau^+(f_k)$ (this is the case for the higher harmonics of f_0, $\tau = 1/(kf_0),\ k > 1, k \in \mathbb{N}^+$). On the other side, energy in the spectrum $S(f_k)$ only exist for $f = f_0, 2f_0$, ... so that when multiplying $S(f_k)$ and $r(1/f_k)$ only the peak at $f = f_0$ remains.

This function is not a pitch detection algorithm but a representation that strengthens the energy at the pitch frequency and reduces the energy at the other harmonics. Because of that we would like to use this method as a front-end for key estimation which would therefore avoid the effect we have mentioned above about the presence of higher harmonics.

However in the case of **multi-pitch signals**, the above-mentioned function cannot be applied directly. For multi-pitch signal, the relationship between $r(\tau)$ and the periodicity of the various pitches becomes intricate. We therefore use the same underlying process but without the use of the projection on cosine functions. This process can be summarized as testing the hypothesis that f_k is a pitch (value given by the projection of $A(f_k)$ on $g_\tau^+(f_k)$) against the hypothesis that f_k is a higher harmonic (projection on $g_\tau^-(f_k)$) or a lower harmonic of another pitch (multiplication by $S(f_k)$) [1] .

[1] It should be noted that this method does not allow to solve the missing

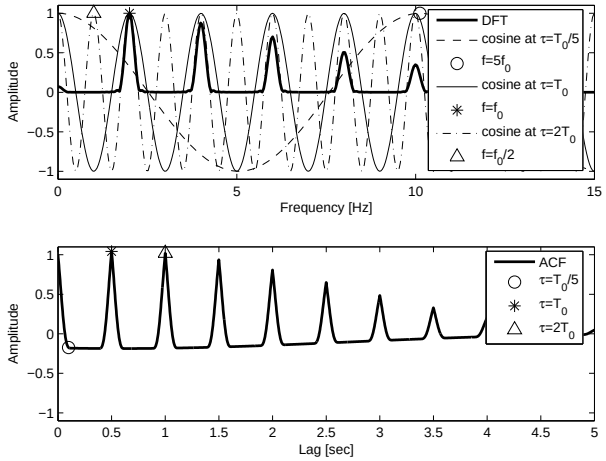

Figure 2. [Top] Magnitude of the DFT of the signal; superimposed: cosine at $\tau = T_0/5, T_0, 2T_0$ and $f = 5f_0, f_0, f_0/2$; [bottom] ACF function; superimposed: $\tau = T_0/5$, $\tau = T_0$, $\tau = 2T_0$ positions; on a periodic signal at f_0=1/T_0=2Hz.

We first compute at each frame the energy spectrum (or the log-amplitude spectrum) $A(k)$. For each frame and each frequency f_k we then compute a score [2] defined as

$$\hat{r}(f_k) = \sum_{h=1}^{H} A(hf_k) - \max\{\alpha(f_k), \beta(f_k), \gamma(f_k)\} \quad (1)$$

$$\alpha(f_k) = \sum_{h=0}^{H-1} A\left(\left(h + \frac{1}{2}\right) f_k\right)$$

$$\beta(f_k) = \min_{h \in \{\frac{1}{3}, \frac{2}{3}, \frac{4}{3}, \frac{5}{3}\}} A(hf_k) \quad (2)$$

$$\gamma(f_k) = \min_{h \in \{\frac{1}{5}, \frac{2}{5}, \frac{3}{5}, \frac{4}{5}\}} A(hf_k)$$

$\hat{r}(f_k)$ is the sum of the energy (log-amplitude) of the spectrum explained by the hypothesis that f_k is a pitch, penalized by the hypothesis that f_k is an even (α), third (β) or fifth (γ) harmonic of a lower pitch.

- The first term of (1) is equivalent to the projection of $A(f_k)$ on $g_\tau^+(f_k)$ but using a narrower and constant-over-τ frequency bandwidth basis [3]. It is the sum of the harmonic of the current frequency f_k.
- α penalizes frequencies which are even harmonics of a lower pitch (the current frequency is potentially the second, fourth, sixth, ... harmonic of a lower pitch).
- β penalizes frequencies which have third harmonic relationship with a lower frequency. We make the underlying assumption of spectral envelope continuity by taking the minimum over the considered harmonics: if f_k was the 3rd harmonic of a pitch $f_k/3$ then energy should be present at $\frac{1}{3}, \frac{2}{3}, \frac{4}{3}, \frac{5}{3}$.
- γ is the same as β but for the fifth harmonic.

fundamental problem.

[2] This score plays the same role as $r(f_k)$ in the previous method; however $\hat{r}(f_k)$ is directly computed at the frequencies f_k of the DFT bins. Therefore no interpolation is required.

[3] The frequency bandwidth corresponding to the positive-valued part of $g_\tau(f)$ varies with τ.

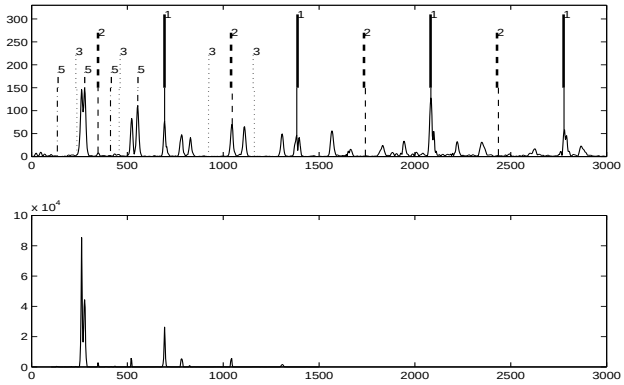

Figure 3. [Top] Magnitude spectrum $A(f_k)$; [Bottom] Harmonic Peak Subtraction function $\hat{h}(f_k)$ (see text).

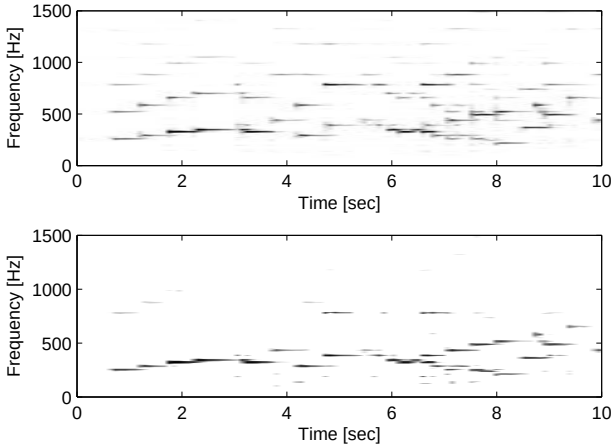

Figure 4. [Top] Spectrogram $S(f_k, t)$; [Bottom] Harmonic Peak Subtraction function over time $\hat{h}(f_k, t)$ (see text).

We finally take the maximum over the three penalties. As for the mono-pitch case, we then compute the product of both measures: $\hat{h}(f_k) = S(f_k) \cdot \hat{r}(f_k)$.

We illustrate the computation of the Harmonic Peak Subtraction function in Figure 3 for a multi-pitch signal which is the superposition of C4 (261.6Hz), C#4 (277.2Hz), and F5 (698.5Hz) viola sounds. The upper part represents $A(f_k)$. The frequency being analyzed with $\hat{r}(f_k)$ is 698Hz. We superimposed the various possible interpretation of this frequency: - 1^{st} harmonic of a F5 note (1), - 2^{nd} harmonic of a F4 (2) - 3^{rd} harmonic of a A#3 (3) or - 5^{th} harmonic of a C#3 (5). The resulting function $\hat{h}(f_k)$ is represented in the lower part of the figure showing emphasis on the C4, C#4, and F5 frequencies. In Figure 4, we illustrate the results of the Harmonic Peak Subtraction function over time in comparison with the spectrogram on the first 10s of J.S. Bach, Well-Tempered Clavier, 02 Fugue in CM.

In the rest of the paper, the results are presented for $A(f_k)$ corresponding to the log-amplitude spectrum and $S(k)$ to the magnitude spectrum. The use of a log-amplitude scale allows reducing the influence of the spectral envelope of each instrument in the computation of $\hat{h}(f_k)$.

2.3. Mapping to chroma scale

Shepard [14] proposes to represent the pitch as a two dimensional structure: the tone height (octave number) and the chroma (pitch class). Based on that, the chroma spectrum or Pitch Class Profile (PCP) has been proposed in order to map the values of the Fourier transform (or Constant-Q transform) frequencies to the 12 semi-tones pitch classes C.

In our system, we first map the values of the Fourier transform to a *semi-tone pitch spectrum*, smooth the corresponding channels over time and then map the results to the *semi-tone pitch class spectrum* (chroma spectrum).

Semi-tone pitch spectrum: The mapping function between the frequencies f_k of the Fourier transform and the semi-tone pitch scale n (expressed in a midi-note scale) is defined as:

$$n(f_k) = 12\log_2\left(\frac{f_k}{440}\right) + 69 \quad n \in \mathbb{R}^+ \tag{3}$$

The computation of the semi-tone pitch spectrum is made using a set of filters $H_{n'}$ centered on the semi-tone pitch frequencies $n' \in [43, 44, \ldots, 95]$ (corresponding to the notes G2 to B6 or the frequencies 98Hz to 1975 Hz). In order to increase the "pitch resolution", we define a factor $R \in \mathbb{N}^+$ which fixes the number of filters used to represent one semi-tone. The center of the filters are now defined by $n' \in [43, 43 + \frac{1}{R}, 43 + \frac{2}{R}, \ldots, 95]$. Each filter is defined by the function

$$H_{n'}(f_k) = \frac{1}{2}\tanh\left(\pi(1-2x)\right) + \frac{1}{2} \tag{4}$$

where x is the relative distance between the center of the filter n' and the frequencies of the Fourier transform: $x = R\,|n' - n(f_k)|$. The filters are equally spaced and symmetric in the logarithmic semi-tone pitch scale, extend from $n'-1$ to $n'+1$ with a maximum value at n'.

The values of the semi-tone pitch spectrum $N(n')$ are then obtained by multiplying the Fourier transform values $A(f_k)$ by the set of filters $H_{n'}$:

$$N(n') = \sum_{f_k} H_{n'}(f_k)A(f_k) \tag{5}$$

Smoothing: The semi-tone pitch spectrum $N(n')$ is computed for each frame t. The output signal of each filter $N(n', t)$ is then smoothed over time using median filtering. This provides a reduction of transients and noise in these signals. Also, for the rest of the process, only the filters centered on the exact semi-tone pitches are considered (i.e. among the R filters representing one semi-tone, we only consider the middle one; for example if $R = 3$, we only keep n'=69 but not n'=68.666 and n'=69.333). We can do this because the tuning is guaranteed to be 440 Hz. This process also allows a reduction of the influence of noise in the computation of the chroma spectrum.

Semi-tone pitch class spectrum (chroma spectrum): The mapping function between the semi-tone pitches n and the semi-tone pitch classes (chroma) c is defined as $c(n) = mod\,(n, 12)$. The mapping to the 12-chroma scale vector $C(l)$ (pitch classes) is achieved by adding the equivalent pitch classes

$$C(l) = \sum_{n' \text{ so that } c(n')=l} N(n') \quad l \in [0, 12[\tag{6}$$

Parameters: The analysis is performed using Short Time Fourier Transform with a window of type Blackman, length 371.5ms and 50% overlap. Because of frequency resolution limits (the frequency distance between adjacent semi-tone pitches becomes small in low frequency), we only consider frequencies above 100Hz. The upper limit is set to 2000Hz. The value of R is set to 3.

Choice of the amplitude scale: The choice of $A(f_k)$ in (5) plays a crucial role. In section 3, we will compare the use of the DFT and HPS for $A(f_k)$. We will also test for each case, the influence of the scale of $A(f_k)$: amplitude scale, energy scale but also a sone-converted value scale. The sone values are obtained in a similar way as proposed in [9]:

$$\begin{aligned} A_s(f_k) &= 2^{\frac{1}{10}(A_{db}(f_k)-40)} \text{ if } A_{db}(f_k) > 40 \\ &= 1/40 A_{db}(f_k)^{2.642} \text{ else} \end{aligned} \tag{7}$$

where $A_{db}(f_k) = 10\log_{10}(A(f_k)10^{96/20})$ is the spectrum after scaling to the maximum signal resolution and conversion to decibel scale.

2.4. Key estimation from the chroma-vectors

The key is estimated from the succession of chroma-vectors over time $\underline{C}(t)$. This can be done in several ways.

2.4.1. Key estimation based on cognitive models

Key-chroma profiles: This approach is close to the one proposed in [5]. The key profiles are created by extending the Temperley-Diatonic pitch-distribution-profiles [4] to the polyphonic case (several pitches) by considering the contribution of the three main triads (tonic, dominant and subdominant) in each key. The key profiles can be further extended to the audio case (harmonics of each pitch) by considering a contribution of the h harmonic of each pitch with an amplitude of 0.6^{h-1}. In [5], the first H=4 harmonics are considered. In section 3, we will test the simple polyphonic model (H=1) and the extended-to-audio model (H=4). In both cases, the result is a 12-dimensions key-chroma profile for each of the 24 keys: $\underline{\hat{C}}_i$ $i \in [1, 24]$.

Key decision method: In the following we note $\underline{C}(t)$ the 12-dimension chroma-vector extracted from the signal at time t. [5] proposes to estimate the global key of a piece as the key-chroma profile $\underline{C}_i$ which has the highest correlation with an averaged over time chroma-vector: $\max\{\underline{\hat{C}}_i \cdot \mu(\underline{C}(t))\}$ We have found better results using the maximum of the average correlation between key-chroma profiles and instantaneous chroma-vectors: $\max\{\mu(\underline{\hat{C}}_i \cdot \underline{C}(t))\}$. We call this method *MeanInstCorrel*. We also test the decision method proposed in [6]. At each time t, we estimate the $\underline{C}_i$ that has the highest correlation with a cumulated-over-time chroma-vector [5]. We attribute a score to this key proportional to the distance between its correlation value and the correlation value of the second most likely key. This score acts as a reliability coefficient. The final key is chosen as the one with the maximum score cumulated over time. We call this method *ScoreCorrelCumul*.

[4] During our experiment, we have found better results using the Temperley-Diatonic profiles proposed in [6] than the Krumhansl ones.

[5] At time t, the cumulated-over-time chroma-vector is computed by averaging the chroma-vector $\underline{C}(\tau)$ since the beginning of the track: $1/t\sum_{\tau=0}^{t}\underline{C}(\tau)$.

2.4.2. Key estimation based on HMM

In [12], we have proposed a method for key estimation based on training a set of hidden Markov models on the chroma representation corresponding to the various keys. No a priori musical knowledge at all is introduced in this method. The characteristics (signal and musical) of the keys are learned directly from the training. In this case the chroma-vectors were derived from the DFT expressed in sone scale. 24 models corresponding to each possible key need to be trained. However, because the number of instances in our database strongly differs among the 24 keys, training directly the HMMs on the set of items belonging to a specific key could lead to over-fitting (learning the track characteristics instead of the key characteristics). We therefore start by training only two models, a Major and minor mode model, and then map the two trained models to the various possible keys. For this the chroma-vectors of all the tracks are mapped to a key-note of C (by using circular permutation of chroma-vectors). Two HMMs are trained corresponding to C Major and C minor. The training is done using the Baum-Welsh algorithm. The parameters of the two models are then used to construct the 24 HMMs corresponding to the various keys. This is done by circular permutation of the mean vectors and covariance matrices of the state observation probability. For a song with unknown key, we evaluate the log-likelihood (using the forward algorithm) that its chroma-vector sequence has been produced by each of the 24 HMMs. The model with the maximum log-likelihood gives the key. In a 10-fold cross-validation, we have obtained the best results using the following configuration: 3 hidden states / 3 Gaussians per state (GMM) with diagonal matrices.

3. Evaluation

So far, we have proposed several alternatives for each stage of the key estimation system. We would like now to compare the effect of each of them on the global recognition rate. We would especially like to test the influence of: • the periodicity observation (DFT or HPS), • the scale used to represent the value before the mapping to chroma (amplitude, energy or sone scale), • the value of H in the key-chroma profiles (H=1 means ignoring the harmonic contribution, H=4 means considering the first four harmonics), • the key decision method (MeanInstCorrel or ScoreCorrelCumul). We also test the performances of the HMM-based approach [12] applied directly to the chroma-vectors derived from the DFT in sone scale.

For each track of the database, only the first 20s are analyzed. We therefore make the underlying hypothesis that the main key is used in the beginning of the track but not necessarily right at the beginning of the track (as it is often the case in romantic music).

3.1. Test set

A database of 302 European baroque, classical and romantic music extracts have been created. This includes pieces by Bach (48), Corelli (12), Handel (16), Telleman (17), Vivaldi(6), Beethoven (33), Haydn (23), Mozart (33), Brahms (32), Chopin (29), Dvorak (18), Schubert (23), Schuman (7). The pieces are for solo keyboard (piano, harpichord), chamber and orchestra music. It should be noted that no opera or choir music was considered in the present study. The distribution of the test set is indicated in Table 1. As in [6], the database was derived from the NAXOS web radio service. The ground-truth key (key-note and mode) was derived from the title of the piece. Only the first movement of each piece, supposed to correspond to the provided key, was used. Note that we had to manually correct the annotation of part of the baroque pieces, since they were based on a tuning of A4=415Hz.

Table 1. Distribution of the test set.

	Keyboard	Chamber	Orchestra	
Baroque	61	37	6	104
Classical	42	N/A	47	89
Romantic	46	10	53	109
	149	47	106	

Table 2. MIREX score (MI), recognition rate of key (KE), key-note (KN) and mode (MO) for various configurations.

		MeanInstCorrel				ScoreCorrelCumul			
		MIREX	Correct key	Correct key-note	Correct mode	MIREX	Correct key	Correct key-note	Correct mode
DFT ampl	H=1	86,1	79,8	82,8	91,4	86,7	81,8	84,4	91,1
	H=4	**88,4**	83,4	86,4	92,1	**87,9**	**83,8**	86,1	91,7
DFT ener	H=1	84,9	78,5	80,8	90,7	80,5	73,2	75,2	86,4
	H=4	85,1	78,8	80,8	91,1	79,7	71,9	74,2	86,1
DFT sone	H=1	84,6	76,5	79,8	90,7	86,6	81,8	85,4	90,7
	H=4	**88**	82,5	84,8	**92,7**	**87,6**	**83,8**	**87,1**	**92,1**
HPS ampl	H=1	86,4	80,5	83,1	91,4	84,3	79,5	82,8	88,1
	H=4	86	80,1	82,8	91,4	81,9	75,5	80,1	86,8
HPS ener	H=1	84,6	77,8	80,8	90,4	82,7	76,5	79,8	87,4
	H=4	81,6	73,2	76,8	88,4	76,2	67,5	71,9	82,5
HPS sone	H=1	85,9	80,5	83,1	90,4	**89,1**	**84,8**	**87,7**	**93**
	H=4	87,8	83,1	85,4	92,1	88,2	84,1	87,1	92,1
HMM DFT sone						85,5	81	87,4	88

3.2. Results

In Table 2, we indicate the recognition rate of key (KE), key-note alone (KN), and mode alone (MO). We also indicates the score used for the MIREX-2005 key estimation contest (MI) [6]. According to this evaluation the best recognition rate of key (KE), as well as the highest MIREX score (MI) are obtained using the HPS using a sone scale and the ScoreCorrelCumul decision method (MI=89.1%, KE=84.8%). We also see that changing only one of the processes (scale, H or decision method) can change drastically the results. What conclusion can we draw from this evaluation?

Concerning the choice of a specific decision method: there is no clear trend in the results, the highest value of MI is not necessarily obtained with the same decision method as the highest value of KE.

[6] This score uses the following weights: - 1 for correct key estimation, - 0.5 for perfect fifth relationship between estimated and ground-truth key, - 0.3 if detection of relative major/minor key, - 0.2 if detection of parallel major/minor key.

Table 3. Recognition rate of key by music genre and instrumentation type (HPS, sone scale, ScoreCorrelCumul).

	Keyboard	Chamber	Orchestral	
Baroque	89,8	94,6	100	92,1
Classical	96,2	N/A	93	94,5
Romantic	85,4	92	76,8	81,8
	90,3	94	85,3	

Concerning the choice of a scale: the choice of the energy scale systematically decrease both MI and KE. The amplitude and sone scale give very close results in the case of the DFT, but the sone scale surpasses the amplitude in the case of the HPS.

Concerning the value of H: in the case of the DFT, H=4 allows increasing MI and KE (this can be understand by the fact that the DFT does not remove the higher harmonics contribution therefore it is necessary to include it in the key-chroma profiles), while in the case of the HPS choosing H=4 decreases the results (for the opposed reason).

Concerning the choice of the periodicity observation (DFT or HPS): in the case of the sone scale, the HPS surpasses the DFT, however this is not the case for the amplitude scale.

Why are the results better in sone scale for the HPS ?: This can be explained considering Figure 3 where we see that the HPS allows to emphasize the existing pitch frequencies but does not provides an accurate estimation of their amplitude. Because the sone scale performs a compression of $A(f_k)$ it provides a reduction in the amplitude discrepancy.

The last row of Table 2, indicates the recognition rate obtained using the HMM-based approach. This result has been obtained using a ten-fold cross-validation. It is interesting to consider that this method without any introduction of musical knowledge achieves quiet reasonable results (the KN value is very close to our winning algorithm).

To conclude we indicate in Table 3, the MI score of the winning algorithm by music genre and instrumentation type. This table emphasizes the fact that the results strongly depend on the considered music genre. The lowest recognition rate is obtained for the romantic period (81.8%). Part of the tracks of this period (Brahms, Schuman) actually contains mainly a neighboring key in the first 20s.

4. Conclusion and Future Work

In this paper, we have presented a system for the automatic estimation of key based on chroma representation. The main contribution of this paper is the Harmonic Peak Subtraction function expressed in a sone scale to be used as front-end for spectral representation of the signal. We have tested various way of estimating the key from the succession of chroma-vector over time including a proposed HMM-based approach. In an evaluation using a database of 302 baroque, classical, romantic music tracks, the best results (89.1% MIREX score) were obtained using our HPS function in sone scale with Gomez polyphonic key-chroma profiles and Izmirli score-based key decision method. It is however worth mentioning that the results obtained during the evaluation strongly depends on the music period. The more harmonically complex romantic music has a lower recognition rate than the baroque and classical music. This could indicate the limitation of such a straightforward approach for key estimation.

On the signal side, future works will concentrate on improving the amplitude associated to the peaks of the HPS function. We would also like to test the performance of a multi-pitch detection algorithm ([7][18]) mapped to the chroma domain in order to know the limits of the chroma-based approach. On the music analysis side, the key decision method should certainly be improved in order to take into account potential modulation over time, this was in fact the prime reason for testing the HMM approach.

Acknowledgments

Part of this work was conducted in the context of the European I.S.T. project Semantic HIFI [16] [7] . To my father.

References

[1] J. Brown. Calculation of a constant Q spectral transform. *JASA*, 89(1):425–434, 1991.

[2] C.-H. Chuan and E. Chew. Fuzzy analysis in pitch class determination for polyphonic audio key finding. In *ISMIR*, pages 296–303, London, UK, 2005.

[3] M. Cremer and C. Derboven. A system for harmonic analysis of polyphonic music. In *AES 25th Int. Conf*, pages 115–120, London, UK, 2004.

[4] T. Fujishima. Realtime chord recognition of musical sound: a system using common lisp music. In *ICMC*, pages 464–467, Bejing, China, 1999.

[5] E. Gomez. Tonal description of polyphonic audio for music content processing. *INFORMS Journal on Computing, Special Cluster on Computation in Music*, 18(3), 2006.

[6] O. Izmirli. Template based key finding from audio. In *ICMC*, pages 211–214, Barcelona, Spain, 2005.

[7] A. Klapuri. A perceptually motivated multiple-f0 estimation method. In *WASPAA*, New Paltz, New York, 2005.

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

[9] E. Pampalk, S. Dixon, and G. Widmer. Exploring music collections by browsing different views. In *ISMIR*, pages 201–208, Baltimore, USA, 2003.

[10] S. Pauws. Musical key extraction from audio. In *ISMIR*, pages 96–99, Barcelona, Spain, 2004.

[11] G. Peeters. Music pitch representation by periodicity measures based on combined temporal and spectral representations. In *ICASSP*, Toulouse, France, 2006.

[12] G. Peeters. Musical key estimation of audio signal based on HMM modeling of chroma vectors. In *DAFX*, McGill, Montreal, Canada, 2006.

[13] H. Purwins, B. Blankertz, and K. Obermayer. A new method for tracking modulations in tonal music in audio data format. In *Neural Networks, IJCNN, IEEE Computer Society*, 2000.

[14] R. Shepard. Circularity in judgements of relative pitch. *JASA*, 36:2346–2353, 1964.

[15] D. Temperley. What's key for key? the Krumhansl-Schmuckler key finding algorithm reconsidered. *Music Perception.*, 17(1):65–100, 1999.

[16] H. Vinet. The Semantic Hifi project. In *ICMC*, pages 503–506, Barcelona, Spain, 2005.

[17] G. Wakefield. Mathematical representation of joint time-chroma distributions. In *SPIE*, Denver, USA, 1999.

[18] C. Yeh, A. Roebel, and X. Rodet. Multiple fundamental frequency estimation of polyphonic music signals. In *ICASSP*, volume 3, pages 225–228, Philadelphia, PA, USA, 2005.

[7] http://shf.ircam.fr