MUSICAL KEY ESTIMATION OF AUDIO SIGNAL BASED ON HIDDEN MARKOV MODELING OF CHROMA VECTORS

Geoffroy Peeters

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

ABSTRACT

In this paper, we propose a system for the automatic estimation of the key of a music track using hidden Markov models. The front-end of the system performs transient/noise reduction, estimation of the tuning and then represents the track as a succession of chroma vectors over time. The characteristics of the Major and minor modes are learned by training two hidden Markov models on a labeled database. 24 hidden Markov models corresponding to the various keys are then derived from the two trained models. The estimation of the key of a music track is then obtained by computing the likelihood of its chroma sequence given each HMM. The system is evaluated positively using a database of European baroque, classical and romantic music. We compare the results with the ones obtained using a cognitive-based approach. We also compare the chroma-key profiles learned from the database to the cognitive-based ones.

1. INTRODUCTION

Considering its numerous applications (search/ query music databases, playlists generation or automatic accompaniment), automatic estimation of musical key (key-note and mode) or of chord progression over time of a music track has received much attention in the recent years. Because symbolic transcriptions of music tracks are not always available, and because automatic transcription algorithms (audio to symbolic) are still limited and costly, many systems attempts to extract the key or chord progression directly from the audio signal. Most existing algorithms therefore start by a front-end which converts the signal frames to the frequency domain (FFT or CQT [1]) and then map it to the chroma domain [2] (or Pitch Class Profile [3]). Chroma/ PCP vectors represent the intensities of the twelve semi-tones of the pitch classes over time. Algorithms then try to find the key or chord progression that best explains the succession of extracted chroma vectors. In order to estimate the key, Chew [4] proposes the Spiral Array Model/ Center of Effect Generator. Most other authors [5], [6], [7], [8] use theoretical chroma/ PCP profiles corresponding to the various keys. These profiles are derived from the probe tone experiment of Krumhansl & Schmukler [9] or from the modified version proposed by Temperley [10]. These experiments aimed at describing the perceptual importance of each semi-tone in a key. The result is a pitch distribution profile for each key. These profiles are then converted by the authors to key-chroma profiles.

Polyphonic audio signal: However, when trying to estimate the key from a music audio signal, a major difference with these experiments is the work with polyphonic signal (several notes played at the same time) and with audio signal (not only the frequencies of the pitch notes are observed but also all their harmonics; therefore high values exist in the chroma vector at the fifth, third, ... intervals of the pitch notes). This problem has been addressed by Gomez [5]. She proposes to take into account the contribution of the harmonics of a note (by using a theoretical spectral envelope) and the polyphony (by considering the three main triads in each key) during the creation of the key-chroma profiles. This problem has also been addressed by Izmirli [6] who estimates directly the contribution of the harmonics of a note by measuring it with a database of piano notes.

Key estimation: The systems compute chroma/ PCP vectors on a frame-based. The systems then try to find the most likely key that explains the overall set of frames. An assumption is often made about the existence of the key in the beginning of the track. Therefore, only the first part (first 20s of the first movement) of the track is considered. Several approaches are taken to go from the frame level to the global key. Gomez [5] finds the key by choosing the key-chroma profile which has the highest correlation coefficient with a global average chroma vector. Izmirli [6] computes at each time a cumulated chroma-vector (average from the beginning of the track) $\overline{c}(t)$; he then searches at each time the key-chroma profile which has the highest correlation with $\overline{c}(t)$, assigns to this key a score equal to the distance between the 1^{st} and 2^{nd} maximum correlation; and finally chose the key which has the maximum score over the first 20s.

When implementing such a key estimation system, one soon notes that the choice of the parameters (choice between Krumhansl or Temperley profile, number of considered harmonics, number of considered triads, ...) and the choice of the key estimation algorithm (type of distance, type of score, considered track duration) influence a lot the recognition rate of key. In fact many assumptions are made in these systems (fixed spectral envelope of a note, fixed polyphony, no modulation of the key) that do not necessarily correspond to the reality of the audio signal corresponding to a music track in a given key.

For these reasons, in this paper we study a system that models the keys using a set of hidden Markov models trained directly on the chroma representation. This choice allows us to avoid making the above-mentioned assumptions.

The paper is organized as follows: in section 2.1, we present the front-end of our system which extracts the chroma representation over time of a music track; in section 2.2, we present our key estimation system based on hidden Markov modeling; in section 3, we evaluate this system with a database of 300 tracks and compare the results with the ones obtained using a cognitive-based approach (combining [5] and [6] approaches). We then compare the key-chroma profiles learned by the HMMs to the cognitive-based ones.

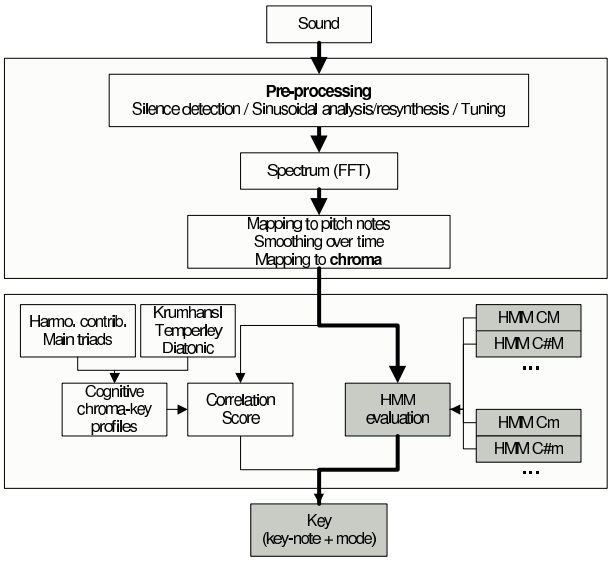

Figure 1: *Key estimation systems: based on cognitive models (thin arrows) and on hidden Markov models (thick arrows).*

2. PROPOSED METHOD

Our key estimation system is represented in Figure 1. In the following we detail the various steps of it.

2.1. Front-end: signal observation

2.1.1. Pre-processing

We first apply a set of pre-processing algorithms to the audio signal. In order to save computation time, the signal is first converted to mono and down-sampled to 11.025 Hz.

A *silence detectors* (based on loudness and spectral flatness measure) is applied in order to detect the actual beginning of the music in the audio signal.

A simple *sinusoidal analysis/re-synthesis* (spectrum peak-picking and short-term partial tracking) is applied in order to reduce transient and noise influence in the measures.

As in [11], the *tuning of the track* is then estimated. This is necessary because the instruments used during the recording may have used another tuning than 440 Hz and because possible transcoding of the audio media may have changed its tuning. We suppose the tuning constant over the track duration. We test a set of tunings between 427 Hz and 452 Hz (the quarter-tones below and above A4). For each tuning t and for each frame m, we compute a "modeling error" defined as the ratio between the energy of the spectrum explained by the current tuning (sum of the energy at the frequencies f_t corresponding to the semi-tones pitches based on the tuning t) and the total energy of the spectrum:

$$\epsilon(t,m) = 1 - \sum_n A(f_{t,n},m) / \sum_f (A(f,m) \qquad (1)$$

where A denotes the amplitude of the Fourier transform and $f_{t,n}$ are the semi-tones pitches based on the tuning t:

$$f_{t,n} = t \cdot 2^{(n-69)/12} \quad n \in [43,44,\ldots,95] \quad t \in [427,452] \quad (2)$$

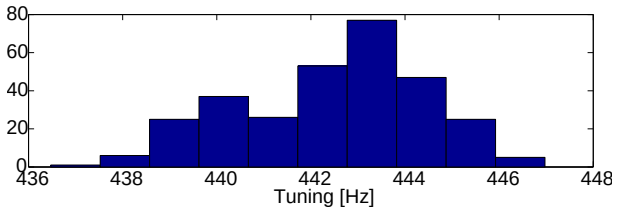

Figure 2: *Histogram of the tunings estimated on the evaluation database (see section 3.1).*

The tuning is then chosen as the value t which minimizes the modeling error over time. In Figure 2, we present the histogram of the tunings estimated on the 300 tracks database we will used in our experiments (see section 3.1). The signal is then re-sampled (using a polyphase filter implementation) in order to bring the tuning back to 440 Hz. The rest of the system can now be based on a tuning of 440 Hz.

2.1.2. Chroma representation

Shepard [12] 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:

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

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

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

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 multipying the Fourier transform values $A(f_k)$ by the set of filters $H_{n'}$:

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

Smoothing: The semi-tone pitch spectrum $N(n')$ is computed for each frame m. The output signal of each filter $N(n', m)$ 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 now 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

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

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 100 Hz. The upper limit is set to 2000 Hz. The variable $A(f_k)$ in (5) can represent either amplitude, energy, log-amplitude or sone-converted values of the DFT. The results given in the following were obtained using the sone-converted values, which has given the best results in our case. The computation of the sone-converted values is similar to the one used in [13]. The value of R is set to 3.

2.2. Key estimation

In the following we will compare the estimation of key based on key-chroma profiles derived from Krumhansl/ Temperley experiments with a system based on hidden Markov models trained directly on chroma representations. We first start by presenting the key-chroma profile method we will use in the experiment then we present our system.

2.2.1. Key estimation based on cognitive models

We have tested several systems based on cognitive models. The best results were obtained using a combination of [5] and [6].

Creation of key-chroma profiles: We use an approach similar to Gomez [5]: the key profiles are created by extending Krumhansl & Schmukler (Temperley or Diatonic) pitch distribution profile to the polyphonic (several pitches) and audio (several harmonics for each pitch) cases. For each key, we consider the three main triads in this key: the tonic, dominant and sub-dominant triads (for example in C Major: C-E-G, G-B-D, F-A-C). The chroma vector corresponding to each single note of a specific triad is computed by adding the contribution of its harmonics h. The harmonic h is given a contribution of 0.6^{h-1}. Only the first 4 harmonics are considered. For a specific triad, the chroma vectors corresponding to the three notes are added. Finally for a specific key, the key-chroma vector is computed by adding the three triad-chroma vectors. Each triad-chroma vector is weighted by the value of the Krumhansl's (Temperley or Diatonic) profile at the position corresponding to the position of the root of the triad in the key (for example 6 for the F-A-C triad in C Major). The result is a 12 dimensions chroma profile vector for each of the 24 keys: $\underline{C}_i$ $i \in [1, 24]$. In the following, we will use the pitch distribution profile proposed in [6] (combination of Temperley and Diatonic profile) which has given the best results in our case.

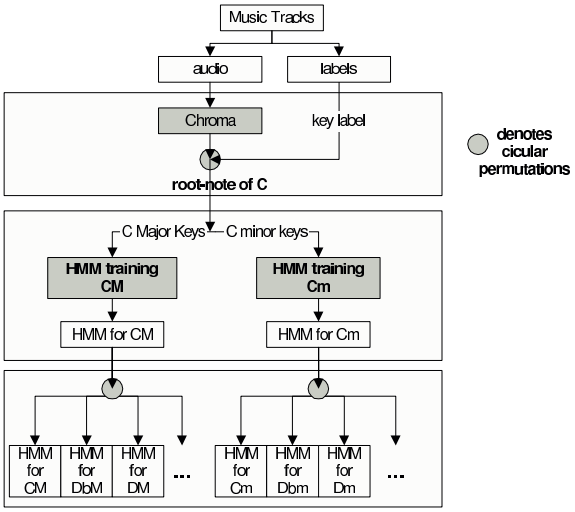

Figure 3: *HMM training for key estimation.*

Estimation of key: The most likely key of the track is estimated using an approach similar to Izmirli [6]. The chroma vectors $\underline{c}(t)$ are extracted on a frame basis. At each time t, we estimate the key $\underline{C}_i$ that has the highest correlation with a cumulated-over-time chroma-vector[1]. 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 decision is chosen as the key with the maximum score cumulated over time. Only the first 20 seconds of the tracks are considered.

2.2.2. Key estimation based on hidden Markov models

In the following we propose the use of hidden Markov models [14] trained and evaluated directly on the temporal sequences $\underline{c}(t)$ of chroma vectors.

In comparison with the above-mentioned approach, the advantages are

1. it does not necessitate to make assumptions about the presence of harmonics of the pitch notes in the chroma representation, or assumptions about specific polyphony since they will be inherently learned from the training set;
2. it does not necessitate the choice of a specific pitch distribution profile (Krumhansl, Temperley or Diatonic);
3. it allows to take into account possible modulation of key over time (through the transition probabilities).

Training of key-chroma HMMs: We want to create a specific HMM for each of the 24 possible keys (12 key-notes, 2 modes). However, because the number of instances in our database strongly differs among the 24 keys, training directly the HMMs on the set of

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

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 a minor *mode* model, and then map the two trained models to the various possible *key-notes*. The training process is depicted in Figure 3. It consists in the following steps:

1. map the chroma-vectors of all the tracks of the training set to a root-note of C (by using circular permutation of chroma vectors);
2. train an HMM for the C Major (C minor) key by using all the tracks in C Major mode (C minor mode);
3. construct the HMMs for the other Major (minor) keys by mapping the Major (minor) HMM to the various key-notes (Db, D, Eb, ...). This is done by circular permutation of the mean vectors and covariance matrices of the state observation probability. 24 HMMs are obtained in this way from the two trained HMMs.

The training of the HMMs is made using the Baum-Welsh algorithm. We have tested various HMM configurations (number of states, number of mixtures) [2] [3].

Estimation of key: For a song with unknown key, we evaluate the log-likelihood of its chroma-vector sequence given each of the 24 HMMs. This is done using the forward algorithm. The model giving the maximum log-likelihood determines the key.

3. EVALUATION

3.1. Test set

The evaluation of our system is performed on a database of 302 European baroque, classical and romantic music extracts: 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. No opera or choir music has been considered in the present study. 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.

3.2. Evaluation method

For each track, we extract the chroma vectors of the first 20s as described in section 2.1. We then compare the estimation of the key using the cognitive-based (section 2.2.1) and HMM-based (section 2.2.2) approaches. For the system based on HMMs, we have performed a ten-folds cross-validation (each time the HMMs were trained on 9 folds and evaluated on the remaining one). We indicate the recognition rate of key, key-note alone and mode alone. We also indicate the score used for the MIREX-2005 key estimation contest[4]. This score uses the following weights: - 1 for correct key estimation (CM → CM), - 0.5 for perfect fifth relationship between estimated and ground-truth key (CM → GM), - 0.3 if detection of relative Major/ minor key (CM → Am), - 0.2 if detection of parallel Major/ minor key (CM → Cm).

	MIREX score	Correct key	Correct key-note	Correct mode	5th up	5th down	relative M/m	parallel M/m	Correct + Neighboring
Cognitive-based approach	**88,9**	85,1	88,4	92,7	0,6	5	1	3,3	95
HMM S=3, M=1 diag	84,6	80,4	86,2	86,8	0,9	2,7	3,9	5,8	93,7
HMM S=6, M=1 diag	81,3	76,1	85,3	83,2	1,5	3,3	3	9,1	93
HMM S=12, M=1 diag	84,2	79,8	88	85,6	2,4	1,5	2,7	8,2	94,6
HMM S=3, M=3 diag	**85,5**	81	87,4	88	2,1	2,4	3	6,4	94,9
HMM S=6, M=3 diag	84,2	79,5	87,7	85,3	1,8	2,1	3,6	8,2	95,2
HMM S=12, M=3 diag	85,5	80,7	86,2	87,4	2,1	2,7	3,9	5,5	94,9
HMM S=12, M=3 full	62,7	52,5	66	59	2,1	1,5	16,5	13,4	86

Table 1: *Recognition rate of key for the cognitive-based and HMM-based approach*

3.3. Results

The results are indicate into Table 1. We have tested various configurations of the HMMs: number of emitting states (S=3, 6, 12), number of Gaussian distributions for each state (M=1, 3). Each Gaussian is described by its 12-dimensions mean vector $\mu_{s,m}$ and its covariance matrix $\Sigma_{s,m}$. We have considered independence between chroma values, so that $\Sigma_{s,m}$ are diagonal matrices.

Compared to the cognitive-based approach (88.9% MIREX score), the HMM-based approach leads in all cases to a lower recognition rate (maximum of 85.5% MIREX score). The confusion with the 5th up, relative and parallel Major/minor of the key are larger than the ones obtained with the cognitive-based approach. This can be explained partly by the fact that - the tracks used for the training of the models do not only contain the main key but also neighboring keys - part of the tracks (especially for the romantic period: Brahms, Schuman) start in a neighboring key. The last column of the table indicates the number of "not so bad recognition" (correct recognition + recognition of neighboring keys). This number is very similar for both cognitive and HMM-based approaches (95%). This means that the number of gross errors is similar in both cases. It is important to note that no a priori musical knowledge has been introduced in the HMM-based approach.

Comparing the various configurations of the HMMs, we see that increasing the number of Gaussian distributions for each state (M) slightly increases the recognition rate, but increasing the number of emitting states (S) does not influence significantly the results. In the last row, we indicate the results obtained when considering dependence between the chroma values (full covariance matrix) for the case S=12/ M=3. This decreases significantly the results.

[2] Note that HMM has already been used in the context of chord progression estimation. The system proposed in [15] recognizes the chord progression by decoding a single HMM in which each state represents a specific chord. In our case, a specific key is represented by a specific HMM and the meaning of its states remains unspecified. Our system recognizes the key by finding the most likely HMM over 24 HMMs.

[3] Also, the idea of chroma rotation has been used in the context of chord progression estimation. [15] uses it after the training of its single HMM in order to average the parameters of the state observation probabilities (the chord models). Our process is different since we use it before the training (in order to train only two HMMs) and after the training (in order to construct the 24 HMMs). We do not perform averaging.

[4] http://www.music-ir.org/ mirex2005/ index.php/ Audio_and_Symbolic_Key_Finding

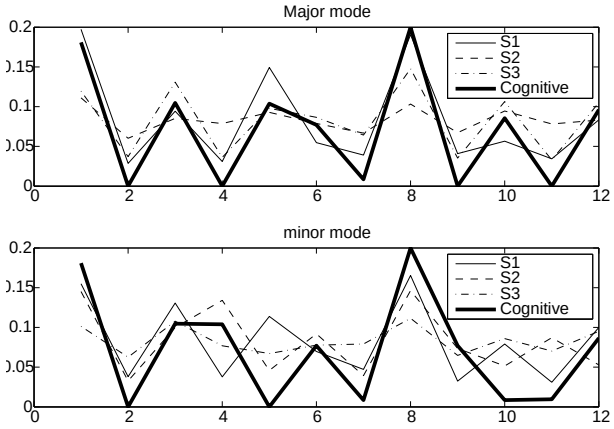

Figure 4: *Comparison between learned and cognitive-based key-chroma profiles for the Major and minor mode*

3.3.1. Comparison between HMM-based and cognitive-based key-chroma profiles

In Figure 4, we indicate the 12-dimensions mean vector of the emitting states of the Major and minor mode HMMs. For this, we have trained the two HMMs on the whole database in a S=3/ M=1 configuration. We compare these vectors with the two cognitive key-chroma profiles as computed according to section 2.2.1. For the Major mode, state 1 (S1) and 3 (S3) of the HMM are pretty close to the cognitive key-chroma profile. For the minor mode, the closest state of the HMM is S2 but it does not agree with the cognitive profile on the importance of the 10^{th} and 11^{th} pitch classes (A and Bb in C minor). This could correspond to the presence in our database of other minor modes than the harmonic one which is the one used in the Temperley/ Diatonic profile.

4. CONCLUSIONS

In this paper, we have proposed a system for the automatic estimation of the key of a music track from the analysis of its audio signal. The system is based on a front-end that extracts chroma vectors over time and uses them as observations. The characteristics of the Major and minor modes are learned by training two hidden Markov models on a labeled database. 24 hidden Markov models corresponding to the various keys are then derived from the two trained models. The estimation of the key of a music track is then obtained by computing the likelihood of its chroma sequence given each HMM. The system is evaluated using a database of European baroque, classical and romantic music. The results are compared with the ones obtained using a cognitive-based approach based on extensions of Krumhansl/ Temperley pitch distribution profiles to the audio/ polyphonic case. The results obtained with the HMM approach (85.5%) remain lower than the ones obtained with the cognitive-based approach (88.9%) but the number of gross errors is similar in both cases. This indicates that a system without any a priori musical knowledge can learn the characteristics of the keys from a labeled database. Comparing the chroma-key profiles learned from the database to the cognitive-based ones gives fairly good agreement for the Major mode, but differs at the 10^{th} and 11^{th} pitch classes for the minor mode. Future works will concentrate on improving the chroma representation and on testing the training/evaluation on the whole track duration. We would also like to test this method for modeling the characteristics of composition styles.

5. ACKNOWLEDGEMENTS

Part of this work was conducted in the context of the European I.S.T. project Semantic HIFI [16] [5]. To my father.

6. REFERENCES

[1] J. Brown, "Calculation of a constant q spectral transform," *JASA*, vol. 89, no. 1, pp. 425–434, 1991.

[2] G. Wakefield, "Mathematical representation of joint time-chroma distributions," in *SPIE conference on Advanced Signal Processing Algorithms, Architecture and Implementations*, Denver, Colorado, USA, 1999, pp. 637–645.

[3] T. Fujishima, "Realtime chord recognition of musical sound: a system using common lisp music," in *ICMC*, Bejing, China, 1999, pp. 464–467.

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

[5] E. Gomez, "Tonal description of polyphonic audio for music content processing," *INFORMS Journal on Computing, Special Cluster on Computation in Music*, vol. 18, no. 3, 2006.

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

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

[8] St. Pauws, "Musical key extraction from audio," in *ISMIR*, Barcelona, Spain, 2004, pp. 96–99.

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

[10] D. Temperley, "What's key for key? the krumhansl-schmuckler key finding algorithm reconsidered," *Music Perception.*, vol. 17, no. 1, pp. 65–100, 1999.

[11] Ch. Harte and M. Sandler, "Automatic chord identification using a quantised chromagram," in *AES 118th Convention*, Barcelona, Spain, 2005.

[12] R. Shepard, "Circularity in judgements of relative pitch," *JASA*, vol. 36, pp. 2346–2353, 1964.

[13] E. Pampalk, S. Dixon, and G. Widmer, "Exploring music collections by browsing different views," in *ISMIR*, Baltimore, USA, 2003, pp. 201–208.

[14] L. Rabiner, "A tutorial on hidden markov model and selected applications in speech," *Proccedings of the IEEE*, vol. 77, no. 2, pp. 257–285, 1989.

[15] A. Sheh and D. Ellis, "Chord segmentation and recognition using em-trained hidden markov models," in *ISMIR*, Baltimore, 2003, pp. 183–189.

[16] H. Vinet, "The semantic hifi project," in *ICMC*, Barcelona, Spain, 2005, pp. 503–506.

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