PHASE-CHANGE BASED TUNING FOR AUTOMATIC CHORD RECOGNITION

Maksim Khadkevich

FBK-irst, Universitá degli studi di Trento,
Via Sommarive, 14 - Povo - 38050
Trento, Italy
khadkevich@fbk.eu

Maurizio Omologo

Fondazione Bruno Kessler - irst
Via Sommarive,18 - Povo - 38050
Trento, Italy
omologo@fbk.eu

ABSTRACT

This paper focuses on automatic extraction of acoustic chord sequences from a piece of music. Firstly, the evaluation of a set of different windowing methods for Discrete Fourier Transform is investigated in terms of their efficiency. Then, a new tuning solution is introduced, based on a method developed in the past for phase vocoder. Pitch class profile vectors, that represent harmonic information, are extracted from the given audio signal. The resulting chord sequence is obtained by running a Viterbi decoder on trained hidden Markov models. We performed several experiments using the proposed technique. Results obtained on 175 manually-labeled songs provided an accuracy that is comparable to the state of the art.

1. INTRODUCTION

In the past few decades, since scientific and technological progress has allowed us to store a great deal of multimedia information, a keen interest in the chord sequence extraction from a piece of music has been emerging. Automatic analysis of digital music signals has attracted attention of many researchers, establishing and evolving Music Information Retrieval (MIR) community. Chord recognition is a part of the large research field of computer audition which deals with all kinds of information extraction from audio signals. Harmonic structure can be described in terms of chord sequences. A chord can be introduced as a number of notes sounding simultaneously, or in a certain order between two time instants, known as chord boundaries. Therefore the task of chord transcription includes chord type classification and precise boundary detection.

Chromagram has been the most successfully used feature for the chord recognition task. It consists of a sequence of chroma vectors. Each chroma vector, also called Pitch Class Profile (PCP), describes harmonic content of a given frame. Each component of the vector represents the energy of one pitch class. Since a chord consists of a number of tones and can be uniquely determined by their positions, chroma vector can be effectively used for the chord representation.

Fujishima was the first one who used the chroma feature in the music computing tasks [1]. He proposed a real-time chord recognition system, describing extraction of 12-dimensional chroma vectors from the Discrete Fourier Transform (DFT) of the audio signal and introducing numerical pattern matching method using built-in chord-type templates to determine the most likely root and chord type. His system showed promising results on pieces of music containing a single instrument.

Sheh and Ellis proposed a statistical learning method for chord recognition [2]. The Expectation-Maximization (EM) algorithm was used to train hidden Markov models, meanwhile chords were treated as hidden states. Their approach involves statistical information about chord progressions – state transitions are identical to chord transitions.

Lee et al. [3] described a chord recognition system that used symbolic data, taken from MIDI[1] files, to train hidden Markov models. This allowed them to avoid a time consuming task of human annotation of chord names and boundaries. The advantage of their system is the possibility of concurrent estimation of key and chord progression, which is achieved by means of building 24 key-dependent HMMs.

Papadopoulos and Peeters [4] presented a new method for chord recognition, which simultaneously estimates chord progression and downbeats from an audio file. They paid a lot of attention to possible interaction of the metrical information and the harmonic information of a piece of music.

On the stage of feature extraction for chord recognition and key estimation of a piece of music a lot of attention has been paid to tuning issues [5, 6, 7]. The necessity of tuning appears when audio was recorded from instruments that were not properly tuned in terms of semitone scale. They can be well-tuned relatively to each other, but, for example, "A4" note is played not at conventional 440 Hz but at 447Hz. This mis-tuning can lead to worse feature extraction and as a result less efficient or incorrect classification. Several approaches to circumvent the problem have been developed.

Harte and Sandler [5] suggested using 36 dimensional chroma vectors. The frequency resolution in this case is one-third of a semitone. After the peak-picking stage and computing a histogram of chroma peaks over the entire piece of music they find mistuning deviation. And prior to calculating 12-bin conventional chromagram they accurately locate boundaries between semitones. The resulting 12-bin semitone-quantized chromagram is then compared with a set of predefined chord templates.

Peeters [7, 8] tested a set of candidate tunings, i.e. the quarter-tone below and the quarter-tone above "A4" note. For each possible tuning the amount of energy in the spectrum is estimated. After defining the global tuning center, the signal is resampled so that it becomes tuned to 440Hz.

Mauch et al. [6] used a quite similar approach: after computing 36-bin chromagram they pick one of three possible sets of 12-bin chromagram, relying on the maximum energy inside candidate bins (e. g. {1, 4, 7... 34 }).

The above-mentioned tuning approaches are similar in sense that they all utilize the information taken just from energy spectrum. However, to apply a tuning technique one can start from dif-

[1] http://www.midi.org

ferent energy spectrum representations, based, for instance, on different window type and size that are chosen. In order to perform a more accurate and precise mis-tuning estimation the proposed paper firstly investigates different possible windowing to apply DFT; secondly, it introduces a tuning technique that concurrently analyses magnitude and phase spectrum. The rest of the paper is organized as follows: section 2 describes the front-end processing and the resulting feature extraction procedure. In section 3 the here adopted HMM-based classification engine is briefly outlined. The experimental results and conclusion are then given in section 4 and section 5, respectively.

2. FRONT-END PROCESSING

2.1. Feature extraction

On the stage of front-end processing the signal is downsampled to 11025 Hz and converted to the frequency domain by a DFT using a windowing function. We consider the range of frequencies between 100 Hz and 2 kHz, mainly because in this range the energy of the harmonic frequencies is stronger than non-harmonic frequencies of the semitones. We use a conventional 12-dimensional Pitch Class Profile (PCP) vector as acoustic feature set, which represents the energies of the 12 semitone pitch classes. A sequence of PCP vectors is known as chromagram. Chromagram is computed in several steps. At first, energy spectrum is calculated using DFT. Then the obtained spectrum is mapped to chroma domain, as shown in (1).

$$n(f_k) = 12\log_2\left(\frac{f_k}{f_{ref}}\right) + 69, n \in \Re^+ \quad (1)$$

where f_{ref} denotes the reference frequency of "A4" tone, while f_k and n are the frequencies of Fourier transform and the semitone bin scale index, respectively. In order to reduce transients and noise, similarly to Peeters [8] and Mauch et al. [6], smoothing over time using median filtering is applied. After filtering semitone bins are mapped to pitch classes, as shown in (2)

$$c(n) = \mathrm{mod}(n, 12) \quad (2)$$

As a result, a sequence of 12-dimensional PCP vectors is obtained.

2.2. Tuning

In order to circumvent the problem of audio recording mis-tuning, a technique that was formerly developed for phase vocoder [9] is here used. The proposed method allows for very precise and accurate frequency estimation of each sinusoid by performing the analysis of the degree of phase change. The block diagram of the proposed estimation scheme is depicted in figure 1. The basic principle is to compute a second Fourier transform of the same signal, windowed by the same function shifted by D samples. Let $x[n]$ be a sequence of samples of the analyzed signal that contains some fundamental and harmonic components. Discrete Fourier Transform (DFT) is performed on the signal weighted by window function $w[n]$ as follows:

$$X_w[t_0, k] = \sum_{n=0}^{N-1} w[n]x[n + t_0]e^{-2\pi jnk/N} \quad (3)$$

where k and N denote a bin number and the window size respectively. A peak extraction algorithm is applied to the obtained

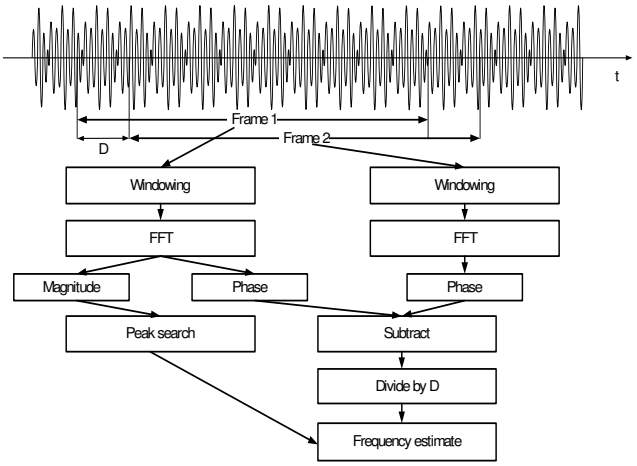

Figure 1: Block diagram of precise frequency estimates

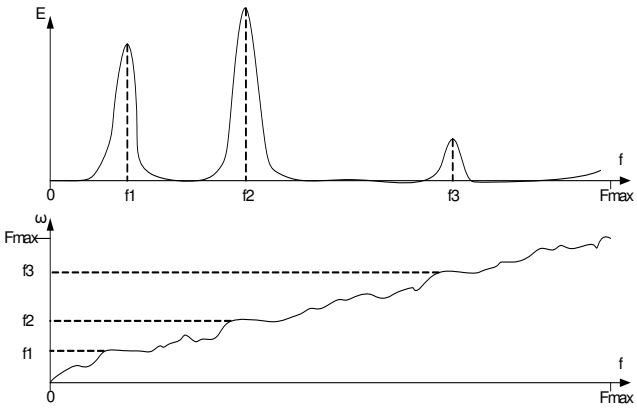

Figure 2: Magnitude and Phase-change spectrum

magnitude spectrum, which results in a list of possible candidates. The main problem of accurate frequency detection based just on the magnitude information is that the main lobe of low frequency harmonics is wider than the spectral resolution (and sometimes than a semitone distance). In such cases the energy of a harmonic component is distributed between adjacent bins, which represents an obstacle in the way of an accurate frequency estimation.

To cope with the above-mentioned problem, a second DFT is applied on the signal weighted by the same window, shifted by D samples, from which the difference of the two given phases divided by the time interval of D samples is calculated as follows:

$$\omega(D, N, t_0) = \frac{\arg X_w[t_0 + D, k] - \arg X_w[t_0, k]}{D} \quad (4)$$

The time interval D is chosen so that the phase change for the maximum frequency is less than 2π. Thus we can ignore the phase-wrapping effect. Analyzing the obtained spectra in terms of phase-change allows for determining frequencies of harmonic components in a more accurate way, since all the adjacent bins containing the energy of a single harmonic have the same degree of phase change (see fig. 2).

Now information obtained from peak-search algorithm is combined with phase-change spectrum in order to provide the fi-

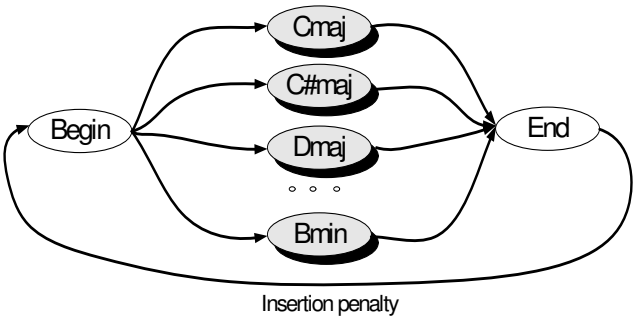

Figure 3: Connection schema of trained models for decoding.

nal estimation. Positions of all possible candidates are checked in terms of the flatness of the corresponding frequency intervals in the phase-change spectrum.

A set of detected harmonics is compared to the table of nominal frequencies. Mean value and standard deviation of closest log-distance (based on a semitone metric) to the nearest nominal frequency are calculated in order to determine the mis-tuning and the consequent consistency of the estimate. Once this procedure has been applied, a new value is assigned to the reference frequency, which is subsequently used for feature extraction. For example, frequency of "A4" is set to 443Hz and frequencies of all the other notes are determined according to equally tempered intervals.

3. HIDDEN MARKOV MODELS

In this work a quite standard application of hidden Markov models (HMM) is addressed, for which a chroma vector represents the basic feature vector to be observed in order to build and use related acoustic models. As opposed to many existing approaches [4, 2, 3], where a chord is represented as a hidden state in one ergodic HMM, a separate model is here created for each chord. Observation vector probabilities in each state can be characterized by a number of Gaussians in 12 dimensions, described by a mean vector and a covariance matrix. It is assumed that feature vector components are uncorrelated one each other, so the covariance matrix has diagonal form. For each observation we use a mixture of 512 12-dimensional Gaussians. Songs from the training set are segmented according to the ground-truth labels so that each segment represents one chord. Chromagrams extracted from these segments are used for training, which is based on the application of the Baum-Welch algorithm. Before running the recognition task, feature vectors are extracted from the test data. There is no preliminary segmentation as done on the training data for which a chroma vector sequence is extracted for each chord segment; only one chromagram is obtained for the whole test song. The trained chord HMMs are connected as shown in figure 3. The Viterbi algorithm is then applied to the test data by using the resulting connected trained model in order to estimate the most likely chord sequence for each song. The full schema of chord recognition system is depicted in figure 4.

4. EXPERIMENTS

In order to evaluate the proposed system we used songs from 12 Beatles albums; the total duration of the collection is 7 hours 44 minutes. Database annotations, kindly provided by C. A. Harte [10], were used as ground-truth. We consider 24 different chord

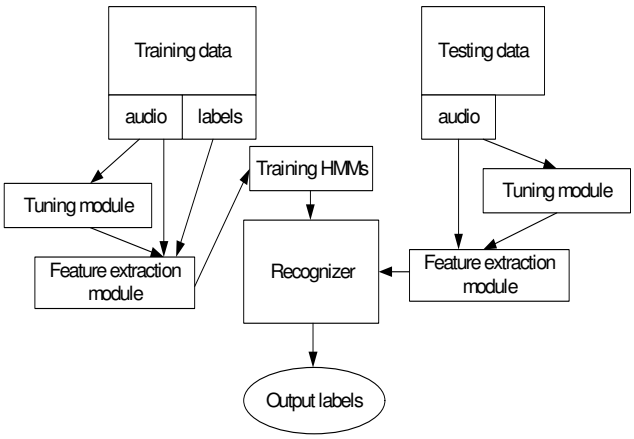

Figure 4: Chord recognition system.

types (major and minor for each of 12 roots) that can be distinguished by the system, while 7th, min7, maj7, minmaj7, min6, maj6, 9, maj9, min9 chords are merged to their root triads. Suspended augmented and diminished chords are discarded from the evaluation task, since the percentage of their duration results to be 2.71% out of the whole material. In order to prevent from lack of training data (some chord types can appear only few times in the training corpus) only two models are trained: C-major and C-minor. For this purpose, all chroma vectors obtained from labeled segments are mapped to the C-root using circular permutation. Then, mean vectors and covariance matrices are estimated for the two models. All the other models can be obtained by a circular permutation procedure. For evaluation, a recognition rate measure similar to the "recall" one was used, which in the given case corresponds to the total duration of correctly classified chords divided by the total duration of chords, as reported in the following:

$$rec.rate = \frac{|recognized\_chords| \cap |ground - truth\_chords|}{|ground - truth\_chords|} \quad (5)$$

The evaluation is performed frame by frame. As done under the MIREX 2008 competition[2], the "precision" measure is not used since a contiguous sequence of chords is assumed, i.e. each time unit features a chord label. Another important characteristic denoting the quality of the transcribed chord labels is the "fragmentation" measure, which is defined as a relative number of chord labels [6].

It is worth noting that varying the chord insertion penalty allows for obtaining output labels with different degree of fragmentation. The recognition accuracy as a function of insertion penalty for Hamming window is displayed in figure 5. For each window size, there is an optimal value of insertion penalty, which produces labels with a fragmentation rate very close to the ground-truth.

In order to find the best windowing function, a set of tests were carried out involving window lengths of 1024(92.8 ms), 2048(185.7 ms), 4096(371.5 ms), 8192(743.0 ms), for Blackman, Hamming and Hanning window types (with 50% overlapping and the optimal insertion penalty). The results are reported in table 1.

The highest performance (69.00 %) was achieved with Hamming window of length 2048 samples, while other window types showed results that are very close to this value. Window length of 2048 samples appeared to be a reasonable trade-off between the

[2]http://www.music-ir.org/mirex/2008/index.php/Main_Page

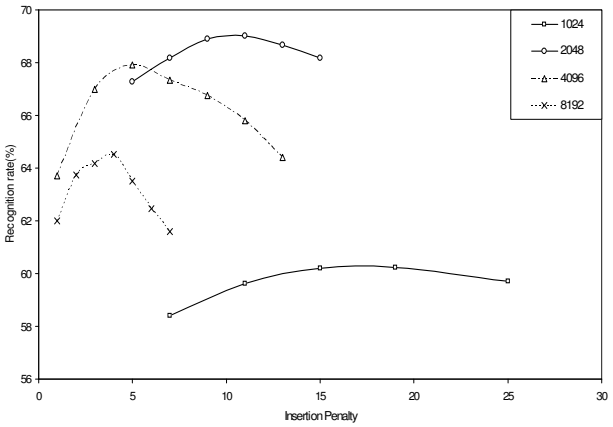

Figure 5: Recognition rate as a function of insertion penalty.

Table 1: *System performance obtained with different windowing configurations.*

	1024	2048	4096	8192
Blackman	57.05	68.92	68.67	64.36
Hamming	60.24	69.00	67.91	64.18
Hanning	59.76	68.51	68.40	63.63

stationarity of the analysed frame of signal and frequency resolution. Taking the best configuration from the above-described experiments (Hamming window of length 2048 samples) the system performance was conducted by including the tuning procedure. Different window delays D were explored in terms of recognition rate. The results are given in the table 2. By increasing the delay D, a very small increase in accuracy can be noticed, which can be due to a different uncertainty in frequency that is obtained for the given window length [11]. Besides this aspect, applying the tuning procedure leads to a higher recognition rate. In order to estimate

Table 2: *Recognition rate obtained using the tuning procedure*

delay (samples)	accuracy
1	71.37
2	71.42
4	71.41
10	71.52
12	70.60
15	69.06

the increase of performance introduced by the tuning procedure, a 3-fold cross-validation was accomplished on the given data set. The results are shown in table 3, which show that about 2.5% and 1% improvements are obtained on the reduced and on the whole data sets, respectively.

5. CONCLUSION

In this paper, the results of a set of chord recognition experiments have been outlined which are based on exploring different windowing solutions as well as on the adoption of a tuning procedure to make this task less dependent on possible instrument mis-tuning effects. A new approach for tuning has been introduced, based on concurrent analyzing magnitude and phase-change spectrum, that can be used in high-accuracy feature extraction for chord recognition and key identification systems. The experimental results showed a very interesting performance in a 3-fold cross-validation conducted on a commonly used database of Beatles songs, for which an average recognition rate of 68.28% has been obtained.

Table 3: *Evaluation results on the reduced and on the complete test data set.*

data	baseline		with tuning	
	rec.rate	frag.	rec.rate	frag.
fold1	69.00%	0.80	71.52%	0.81
fold1, fold2, fold3	67.47%	0.84	68.28%	0.84

6. REFERENCES

[1] Takuya Fujishima, "Realtime chord recognition of musical sound: A system using common lisp music," in *Proceedings of the International Computer Music Conference*, Beijing, 1999.

[2] A. Sheh and D. P. Ellis, "Chord segmentation and recognition using em-trained hidden markov models," in *Proc. 4th International Conference on Music Information Retrieval*, 2003.

[3] K. Lee and M. Slaney, "Acoustic chord transcription and key extraction from audio using key-dependent hmms trained on synthesized audio," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 16, no. 2, february 2008.

[4] H. Papadopoulos and G. Peeters, "Simultaneous estimation of chord progression and downbeats from an audio file," in *Proc. ICASSP*, 2008.

[5] Christopher A. Harte and Mark B. Sandler, "Automatic chord identification using a quantized chromagram," in *Proceedings of the Audio Engineering Society*, Spain, 2005.

[6] Matthias Mauch and Simon Dixon, "A discrete mixture model for chord labelling," in *Proceedings of the 2008 ISMIR Conference*, Philadelphia, 2008.

[7] G. Peeters, "Musical key estimation of audio signal based on hmm modeling of chroma vectors," in *Proceedings of DAFX*, McGill, Montreal, Canada, 2006.

[8] G. Peeters, "Chroma-based estimation of musical key from audio-signal analysis," in *Proceedings of the 2006 ISMIR Conference*, Victoria, Canada, 2006.

[9] J. L. Flanagan and R. M. Golden, "Phase vocoder," *Bell Syst. Tech.J.*, vol. 45, pp. 1493–1509, november 1966.

[10] C. Harte, M. Sandler, S. Abdallah, and E. Gómez, "Symbolic representation of musical chords: A proposed syntax for text annotations," in *Proceedings of the 2005 ISMIR Conference*, 2005.

[11] M. S. Puckette and J. C. Brown, "Accuracy of frequency estimate using the phase vocoder," *IEEE Trans. Speech Audio Process*, vol. 6, no. 2, pp. 166–176, march 1998.