HMM-BASED APPROACH FOR AUTOMATIC CHORD DETECTION USING REFINED ACOUSTIC FEATURES

Yushi Ueda, Yuki Uchiyama, Takuya Nishimoto, Nobutaka Ono and Shigeki Sagayama

Graduate School of Information Science and Technology, The University of Tokyo
7-3-1, Hongo, Bunkyo-ku, Tokyo, 113-8656, Japan
{ueda, uchiyama, nishi, onono, sagayama}@hil.t.u-tokyo.ac.jp

ABSTRACT

We discuss an HMM-based method for detecting the chord sequence from musical acoustic signals using percussion-suppressed, Fourier-transformed chroma and delta-chroma features. To reduce the interference often caused by percussive sounds in popular music, we use Harmonic/Percussive Sound Separation (HPSS) technique to suppress percussive sounds and to emphasize harmonic sound components. We also use the Fourier transform of chroma to approximately diagonalize the covariance matrix of feature parameters so as to reduce the number of model parameters without degrading performance. It is shown that HMM with the new features yields higher recognition rates (the best in MIREX 2008 audio chord detection task) than that with conventional features.

Index Terms— Audio chord detection, HMM, HPSS, diagonalization, dynamic features

1. INTRODUCTION

Automatic chord detection is one of the important issues in music analysis with many possible applications such as music information retrieval, music identification and automatic music transcription. The problem is to estimate chords from the observed notes in symbolic or acoustic form. Chord progression or harmony which is one of the most important elements of Western tonal music, plays a dominant role in determining the music structure and mood. When we listen to a music, even without knowing the individual notes in the music, we can hear the harmony. Thus, chord progression of music can be an aid in the task of automatic music transcription. Numerous related works have been done in this field from artificial intelligence approaches earlier and more probabilistic approaches recently.

In 1999, we formulated this problem with Hidden Markov Models (HMM) for representing hidden chord progression behind the notes given as melodies [1, 2]. Though the input notes were in the symbolic form and the main purpose was automatic harmonization to the given melodies, it is essentially the first use of HMM for chord estimation. In 2003, Sheh *et al.* [3] combined HMM and pitch class profiles (PCP) (later, referred to chroma vectors) proposed by Fujishima [4] for chord detection from audio signal inputs. Since then, numerous methods have been proposed using HMM and chroma vectors [5, 6, 7].

2. PROBABILISTIC MODEL OF CHORD PROGRESSION

2.1. Modeling the music production process

In this section, we review HMM-based chord estimation both for symbolic [1, 2] and audio [3] inputs. Most of Western tonal music are composed according to harmony theory which is an established discipline and taught in music schools as one of the most important subjects containing typical topics such as triads, seventh chords, harmonic progression, tonality, inversion, non-harmonic tones, modulation, etc. Harmony theory can be compared to the grammar of natural languages.

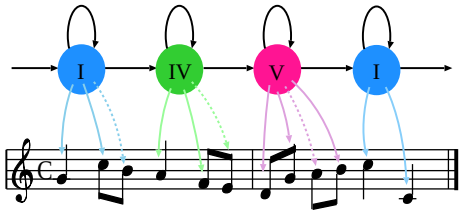


Fig. 1. Melody generation from an underlying chord progression

Most polyphonic music can be considered as being produced from underlying chord progression from which notes in the melody and accompaniment are generated (as illustrated in Fig. 1). Since there are common chord transitions as well as rare ones in real music, chord transitions can be formulated with the n-gram probabilities, being bigram probabilities in the simplest case. Notes in the melody and accompaniment are generated from the chord probabilistically.

2.2. Hidden Markov Model of polyphonic music

The problem of chord detection can be regarded as an inverse problem of this process, that is, to estimate the hidden chord sequence from a melody.

Given a symbolic sequence of music notes such as in the MIDI data format, chord detection would be simple if all observed notes were harmonic tones without any omitted notes or non-harmonic tones. In real situations, music pieces often contain incomplete chords and non-harmonic tones. While a well-trained person identifies these in the framework of harmony theory and understands the chord sequence, common listeners can 'feel' the chord without identifying all notes in music.

In Kawakami's formulation [1, 2] of chords, Hidden Markov Model (HMM) is applied to chord sequence modeling for chord detection given the melodies. The hidden (unseen) chord progression was modeled as Markovian state transitions and melody was considered as the stochastic output from the hidden states. Chord detection is a problem of estimating the hidden chord sequence from a time series of features. This kind of problem was solved by using HMM in the framework of Maximum A Posteriori (MAP) estimation with a wide variety of HMM configurations including ergodic HMM, vocabulary formulation of chord subsequences, key-dependent HMMs for finding the key and detection of key modulation.

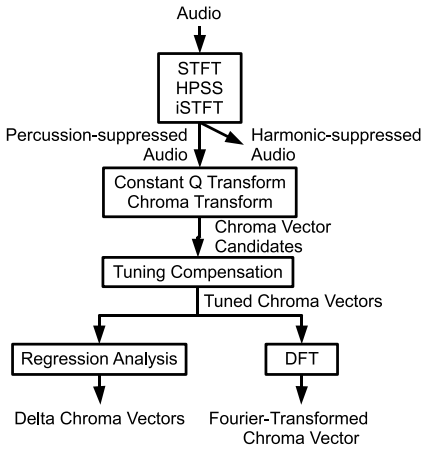


Fig. 2. Flowchart of the proposed feature extraction

In the basic formulation, a chord period is modeled as one state and chord sequences are modeled as ergodic transitions of these states. Each state stochastically emits feature vectors. Let $S = (s_0, s_1, \cdots, s_T)$ be a hidden chord sequence and $X = (x_0, x_1, \cdots, x_T)$ be the output feature sequence. Detecting chord sequence $\hat{S}$ can then be formulated as equation (1), using the Bayes' theorem

$$\hat{S} = \underset{S}{\operatorname{argmax}}\, P(S|X) = \underset{S}{\operatorname{argmax}}\, P(X|S)P(S). \tag{1}$$

Chord transitions are approximated by assuming the first order Markov property in which a state only depends on the previous state and does not depend on any other past states or outputs. The outputs are chroma vectors which are emitted stochastically from each state. In this formulation, the maximum probability chord sequence can be formulated as

$$\hat{S} \simeq \underset{S}{\operatorname{argmax}}\, P(x_0|s_0)P(s_0)\prod_{t=1}^{T} P(x_t|s_t)P(s_t|s_{t-1}) \tag{2}$$

and calculated efficiently by the Viterbi algorithm. If the ground truth data exist, the transition and emission probabilities can be optimized by the Baum-Welch algorithm. It is also possible to use prior musical knowledge to model the chord transition probabilities [5, 7] such as C major is more likely to appear after G major than C# major.

3. FEATURE EXTRACTION FOR AUDIO INPUT

3.1. Feature extraction procedure

The HMM-based chord recognition framework reviewed in the previous section can be simply applied to audio input by replacing input MIDI pitch with acoustic features representing the energy profile of possible pitches [4, 3]. The problems here are that most music recordings, especially popular music contain percussive sounds and harmonic overtones which make it difficult to estimate which notes are performed. Furthermore, even if every performed note was known, i.e., in the case of symbolic input data, we could not directly obtain the chords performed due to different voicings of chords, omission of chord tones and insertion of non-chord tones. In addition, there may be tuning differences from recording to recording.

In the following subsections, we describe our approach to deal with the problems. Fig. 2 shows the flowchart of the proposed feature extraction.

3.2. Suppression of percussive sounds

To suppress percussive sound, we can use the harmonic/percussive sound separation (HPSS) technique [9]. Harmonic sounds usually have stable pitches, which are concentrated in certain frequency bins and have a relatively smooth time envelope. Percussive sounds on the other hand have no harmonic structure, instead they have smooth frequency envelopes and are concentrated in a short time. The HPSS is based on these differences between harmonic sounds and percussive sounds, and therefore does not require templates or prior knowledge of timbre.

Let $W_{i,j}$ be a power spectrogram of the input audio signal, where i and j represent indices of frequency and time bins. To separate the spectrogram $W_{i,j}$ into a harmonic spectrogram $H_{i,j}$ and a percussive spectrogram $P_{i,j}$, differences between the adjacent bins along time and frequency are evaluated by

$$\Omega_P = \tfrac{1}{2\sigma_P^2}\sum_{i,j}(\sqrt{P_{i-1,j}} - \sqrt{P_{i,j}})^2 \tag{3}$$

$$\Omega_H = \tfrac{1}{2\sigma_H^2}\sum_{i,j}(\sqrt{H_{i,j-1}} - \sqrt{H_{i,j}})^2, \tag{4}$$

where σ_P and σ_H are parameters to control the degree of time and frequency smoothness.

Let $m_{P_{i,j}}$ and $m_{H_{i,j}}$ be the time-frequency masks which decompose the original spectrogram $W_{i,j}$ into $P_{i,j}$ and $H_{i,j}$. We measure the distance between $m_{P_{i,j}}W_{i,j}$, $m_{H_{i,j}}W_{i,j}$ and $P_{i,j}$, $H_{i,j}$ with I-divergence. The optimal $m_{P_{i,j}}$ and $m_{H_{i,j}}$ are found by minimizing

$$\begin{aligned} J(H,P) &= \sum_{i,j} m_{P_{i,j}}W_{i,j}\log\left(\frac{m_{P_{i,j}}W_{i,j}}{P_{i,j}}\right) \\ &+ \sum_{i,j} m_{H_{i,j}}W_{i,j}\log\left(\frac{m_{H_{i,j}}W_{i,j}}{H_{i,j}}\right) \\ &- \sum_{i,j}(W_{i,j} - P_{i,j} - H_{i,j}) + \Omega_P + \Omega_H, \end{aligned} \tag{5}$$

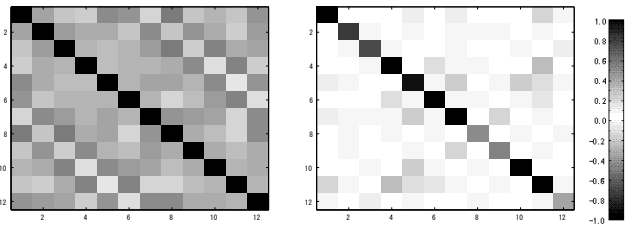
with an EM-like algorithm efficiently.

3.3. Chroma vectors

In this subsection we review PCPs [4] or chroma vectors often used in key or chord detection, since they are effective features for audio signals. They are time series vectors corresponding to the chromatic notes and are based on the assumption that the difference between octaves can be ignored when identifying chords. This assumption works well and we can cope with different voicings of chords as well as reduce the dimensions of the features. There are various methods to calculate a chromagram, among which we use the constant Q transform [10] in order to increase the frequency resolution at low frequencies and obtain frequency resolution in log frequency scale. Let f_k be the center frequency of the kth frequency bin, then f_k is calculated by $f_k = f_{min}2^{\frac{k}{12}}$. The center frequencies of the bins corresponds to the equal-tempered pitch. Using the constant Q transform representation $H_{CQ}(k,\tau)$, where k and τ represent indices of note and time bins, a chromagram $C(l,t)$ is calculated by summing $H_{CQ}(k,\tau)$ over octaves,

$$C(l,\tau) = \log\left(\sum_{k\equiv l(\mathrm{mod}12)} |H_{CQ}(k,\tau)|^2\right), \tag{6}$$

where l represent pitch class indices. We take the logarithm here, since the power distribution of a chromagram usually lean toward small values. By taking the logarithm, the distribution approaches a Gaussian, and the approximation of the output probability of HMM to Gaussian fits well.



(a) chroma (b) FT of chroma

Fig. 3. Covariance matrices – the covariance matrices of chroma vectors are not diagonal but almost circulant, and are therefore diagonalized by the Fourier Transform.

3.4. Tuning compensation

In real audio, tuning pitch may differ from recording to recording and ignoring this difference blurs the chromagram. This is because chromagrams assume that center frequencies of the filterbank match with the performed pitches. We can assume that chroma vectors tuned closer to the correct tuning of the recording have larger energy than those tuned farther because energy distributions of the performed pitches fit the filterbank. So, one way to deal with the problem is to choose the chroma vector with largest energy from n tuning frequency candidates which are placed equally every $100/n$ cents, *i.e.*, $f_{min} = f_0$, $f_0 2^{\pm 1/12n}$, $f_0 2^{\pm 2/12n}$, $\cdots$, $f_0 2^{\pm (n-1)/24n}$. This is similar to the tuning algorithm proposed by Mauch *et. al.*, where they used 3 candidates [8]. We assume that the tuning of the recording does not change over time. Then the maximum-energy chroma vectors $C_{\hat{j}}$ can be obtained by summing C_j over chroma and time bins and comparing them as

$$\hat{j} = \underset{j}{\operatorname{argmax}} \sum_{\tau} \sum_{l=0}^{11} C_j(l,\tau) \quad , \quad j = 0, \cdots, n-1, \qquad (7)$$

where C_j represents the chroma vectors tuned in jth frequency candidate.

3.5. The Fourier transform of chroma vectors

In general, the bins of a chroma vector are not independent of each other. The covariance matrix Σ of 180 songs of the Beatles are shown in Fig. 3 (a). Non-diagonal elements are non-zero obviously. Musical sounds usually contain harmonic overtones, for example when a single C note is played energy will also be present in the chroma bins of its overtones G, E and Bb. Also there are co-occurrences of pitches, as notes are often played together in polyphonic music.

Now we consider the assumptions that each note of the input signals has the same harmonic structure and the amount of occurrence of the same intervals (*e.g.* C-G and D-A) is the same. Though there are various harmonic structures and the amount of occurrence of the same intervals differ among recordings, we can consider these assumptions approximately hold as a whole. Therefore the covariance matrix Σ becomes circulant matrix as Fig. 3 (a). A circulant matrix is diagonalized by the DFT matrix F independent of its values [12].

$$F_{ij} = \begin{cases} \dfrac{\cos(2\pi ij/12)}{\sqrt{\sum_j \cos(2\pi ij/12)^2}} & \text{if} \quad i = 0,1,2,3,4,5,6 \\ \dfrac{\sin(2\pi(i-6)j/12)}{\sqrt{\sum_j \sin(2\pi(i-6)j/12)^2}} & \text{if} \quad i = 7,8,9,10,11 \end{cases} \qquad (8)$$

Table 1. Chord categories

original chord	treated as
C maj, C aug, C sus4, C maj + tension note	C maj
C min, C dim, C min + tension note	C min

That is, the covariance matrix $F\Sigma F^T$ of the Fourier Transform of chromagram $C_{FT} = FC$ is almost a diagonal matrix as shown in Fig. 3 (b). F_{ij} is the (i,j) component of the DFT matrix F.

Using the Fourier transform of chroma vectors as features we can deal with harmonic overtones without increasing the number of parameters and the model become more robust. When the dimension of chroma vector is 12 and the output probability distribution is single Gaussian with a full covariance, the number of parameters is 90. However, when the covariance is diagonal the number of parameter is 24.

The Fourier transforms of chroma vectors are similar to the tonal centroid [13]. The tonal centroid is the second, third and fifth coefficients of the Fourier transform of chromagram. They correspond to minor third, major third and fifth which are the most important intervals in chords. Lee and Slaney [6] use the tonal centroid to audio chord detection and the effectivity was confirmed. Though the Fourier transform of a chromagram has higher dimensionality than the tonal centroid, it loses no information on the assumption that each note in the input signals has the same harmonic structure, and therefore can identify all chord that a chromagram can identify.

3.6. Dynamic features

In speech recognition, dynamic features first proposed by Sagayama *et al.* in 1979 [14] such as delta cepstrums or delta MFCCs are often used together with static features and known to significantly improve recognition performance [15]. Similarly, dynamic features of chroma vectors (delta chroma vectors) can be used in chord detection. We can assume that by using delta chroma vectors, we can obtain higher accuracy on chord boundaries since delta chroma vectors have large values on sound changes.

Delta chroma vectors robust against noise can be obtained from weighted regression analysis of chroma vector sequences. Let weight $w(k)$ be an even function such as triangular function, the delta chroma vector of time τ can be calculated using δ samples before and ahead as

$$\Delta C(l,\tau) = \frac{\sum_{k=-\delta}^{\delta} k \cdot w(k) C(l,\tau+k)}{\sum_{k=-\delta}^{\delta} k^2 w(k)}. \qquad (9)$$

3.7. Number of hidden states per chord

Another issue is the appropriate number of hidden states per chord. Within a duration of a same chord, the beginning and ending parts may have different spectral characteristics due to percussive sounds often synchronizing strong beats. Thus, more than one hidden state per chord can better model the spectral observation. It has to be studied experimentally as simply increasing the states may fall the performance of HMM-based chord detection due to the training data sparseness

Table 2. Recognition rates using HPSS and FT chroma

Features	chroma		FT of chroma	
covariance	diagonal	full	diagonal	full
without suppression	50.94%	74.24%	73.77%	74.24%
with suppression	74.19%	78.48%	78.46%	78.48%

Table 3. Recognition rates using delta chroma

covariance		number of states		
FT chroma	Δ chroma	1	2	3
full	–	78.48%	77.65%	78.21%
diagonal	–	78.46%	78.09%	78.38%
diagonal	diagonal	77.96%	78.62%	78.15%
diagonal	full	79.01%	79.53%	78.98%
full	full	79.06%	79.70%	79.81%

4. EXPERIMENTAL EVALUATION

4.1. Experimental conditions

We conducted experiments to evaluate the effectiveness of the proposed methods. We used 180 songs in 12 albums of The Beatles already used in previous works [7, 6] and at the MIREX2008 [11], so that we were able to compare the results.

Each song was mixed to monaural and down-sampled to 11025 Hz to reduce time for computation. We used 60 channels of constant Q filterbanks ranging from 55.0 Hz (A1) to 1661.2 Hz (G#6). We had 5 tuning candidates ($n = 5$). Delta chroma vectors were calculated from 7 sample points ($\delta = 3$), sampled every 33.3 ms.

The ground-truth chord annotations were provided by Harte *et al.* [16]. We sampled this data at 10 Hz to train parameters and evaluate the recognition rate. In this experiment, chord categories were narrowed down to major and minor triads. Since the annotations contain other chords as well, we remap them to major or minor triads, as shown in Table 1. The chords were treated as major or minor according to their third tone being major/minor except suspended 4 chords treated as major chord. There was also a category for a non-chord for the periods that cannot be assigned as a chord, such as silence or speaking.

We used a three-fold cross-validation. Four albums were used as test data and remaining eight albums were used as training data. The method was evaluated by the recognition rate that was the number of frames correctly recognized divided by the total number of frames.

4.2. Experimental results

First, we evaluated the effectiveness of percussion-suppressed and Fourier-transformed chroma vectors. The results are shown in Table 2. Suppression of percussive sounds yielded higher recognition performance in all cases. In the case of full covariance, error rate was reduced by 16.46 %. With full covariance matrices, both chroma vectors and Fourier-transformed chroma vectors gave exactly the same results, since no information was lost. With diagonal covariance matrices, however, by using the Fourier transform of chroma vectors, recognition rate almost did not change with full covariance while using conventional chroma vectors decreased the recognition rate. This was because the Fourier transform of chroma vectors almost diagonalized the covariance matrix and gave accurate approximation of using diagonal covariance.

Second, we evaluated the use of delta chroma vectors. The results are shown in Table 3. The best recognition rate was obtained when the covariance of both Fourier-transformed chroma vectors and Fourier-transformed delta chroma vectors were full and number of states per chord was 3. The error rate was reduced by 21.62 % compared with chroma vectors with full covariance and one state per chord. We believe the recognition rate increased because precise modeling of HMM became possible using dynamic features.

The algorithm using percussion-suppressed chroma vectors performed best in the task of Audio Chord Detection in MIREX 2008 [11]. The use of delta chroma vectors improved this performance.

5. CONCLUSION

In this paper, we first reviewed modeling polyphonic music based on harmony theory with Hidden Markov Model and then proposed effective features for automatic chord detection. Percussion-suppressed chroma vectors obtained through HPSS reduced error rate by 16.46 %. The recognition rate of the Fourier transform of chroma vectors with a diagonal covariance matrix was almost equivalent to the ones with full covariance matrix. This means we were able to reduce the number of parameters without degrading the chord detection performance. Delta chroma vectors reduced error rate by 21.62 %.

In the future we plan to extend the model to estimate both key and chord and also use percussive component obtained by HPSS to improve chord boundary. For such extension that the degree of freedom increase, using the Fourier-transformed chroma vector and approximating the covariance matrix as diagonal, *i.e.*, decreasing the number of parameters, may outperform the one with full covariance matrix.

6. ACKNOWLEDGMENT

This research was partly supported by CrestMuse Project under JST.

7. REFERENCES

[1] T. Kawakami, M. Nakai, H. Shimodaira and S. Sagayama, "Harmonization for melody using HMM," in *Proc. JHES*, F-61, p. 361, 1999. (in Japanese)

[2] T. Kawakami, M. Nakai, H. Shimodaira and S. Sagayama, "Hidden Markov Model Applied to Automatic Harmonization of Given Melodies," in *Technical Report of IPSJ*, 1999-MUS-034, pp. 59–66, 2000, and in *Technical Report of IEICE*, SP99-156, pp. 25–32, 2000. (in Japanese)

[3] A. Sheh and D. P. W. Ellis, "Chord segmentation and recognition using EM-trained hidden Markov models," in *Proc. ISMIR*, pp. 183–189, 2003.

[4] T. Fujishima, "Real-time chord recognition of musical sound: A system using common lisp music," in *Proc. ICMC*, pp. 464–467, 1999.

[5] J. P. Bello and J. Pickens, "A robust mid-level representation for harmonic content in music signal," in *Proc. ISMIR*, pp. 304–311, 2005.

[6] K. Lee and M. Slaney, "Acoustic chord transcription and key extraction from audio using key-dependent HMMs trained on synthesized audio," *IEEE Trans. on Audio Speech and Language Processing*, vol. 16, no. 2, pp. 291–301, 2008.

[7] H. Papadopoulos and G. Peeters, "Large-scale study of chord estimation algorithms based on chroma representation and HMM," in *Proc. CBMI*, pp. 53–60, 2007.

[8] M. Mauch and S. Dixon, "A discrete mixture model for chord labelling," in *Proc. ISMIR*, pp. 45–50, 2008.

[9] N. Ono, K. Miyamoto, J. L. Roux, H. Kameoka and S. Sagayama, "Separation of a monaural audio signal into harmonic/percussive components by complementary diffusion on spectrogram," in *Proc. EUSIPCO*, 2008.

[10] J. Brown, "Calculation of a Constant Q Spectral Transform," *Journal of the Acoustical Society of America*, vol. 89, no. 1, pp. 425–434, 1991.

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

[12] G. Golub and C. Van Loan, *Matrix Computations*, Johns Hopkins University Press, 1996.

[13] C. Harte, M. Sandler, and M. Gasser, "Detecting harmonic change in musical audio," in *Proc. Audio Music Computer Multimedia Work Shop*, pp. 21–26, 2006.

[14] S. Sagayama and F. Itakura, "On individuality in a dynamic measure of speech," in *Proc. Spring ASJ Meeting*, 3-2-7, pp. 589–590, June 1979. (in Japanese)

[15] L. Rabiner and B.-H. Juang, *Fundamentals of Speech Recognition*, Upper Saddle River, NJ: Prentice Hall, 1993.

[16] C. Harte, M. Sandler, S. Abdallah and E. Gómez, "Symbolic representation of musical chords: A proposed syntax for text annotations," in *Proc. ISMIR*, pp. 66–71, 2005.