AUTOMATIC CHORD RECOGNITION BASED ON PROBABILISTIC INTEGRATION OF CHORD TRANSITION AND BASS PITCH ESTIMATION

Kouhei Sumi,[†] Katsutoshi Itoyama,[†] Kazuyoshi Yoshii,[‡]
Kazunori Komatani,[†] Tetsuya Ogata,[†] and Hiroshi G. Okuno[†]

[†]Dept. of Intelligence Science and Technology
Graduate School of Informatics, Kyoto University
Sakyo-ku, Kyoto 606-8501 Japan
{ksumi, itoyama, komatani, ogata, okuno}@kuis.kyoto-u.ac.jp

[‡]National Institute of Advanced Industrial
Science and Technology (AIST)
Tsukuba, Ibaraki 305-8568 Japan
k.yoshii@aist.go.jp

ABSTRACT

This paper presents a method that identifies musical chords in polyphonic musical signals. As musical chords mainly represent the harmony of music and are related to other musical elements such as melody and rhythm, the performance of chord recognition should improve if this interrelationship is taken into consideration. Nevertheless, this interrelationship has not been utilized in the literature as far as the authors are aware. In this paper, bass lines are utilized as clues for improving chord recognition because they can be regarded as an element of the melody. A probabilistic framework is devised to uniformly integrate bass lines extracted by using bass pitch estimation into a hypothesis-search-based chord recognition. To prune the hypothesis space of the search, the hypothesis reliability is defined as the weighted sum of three reliabilities: the likelihood of Gaussian Mixture Models for the observed features, the joint probability of chord and bass pitch, and the chord transition N-gram probability. Experimental results show that our method recognized the chord sequences of 150 songs in twelve Beatles albums; the average frame-rate accuracy of the results was 73.4%.

Keyword: chord recognition, bass line, hypothesis search, probabilistic integration

1 INTRODUCTION

In recent years, automatic recognition of musical elements such as melody, harmony, and rhythm (Figure 1) from polyphonic musical signals has become a subject of great interest. The spread of high-capacity portable digital audio players and online music distribution has allowed a diverse user base to store a large number of musical pieces on these players. Information on the content of musical pieces such as their musical structure, mood and genre can be used together with text-based information to make music information retrieval (MIR) more efficient and effective. Manual annotation requires an immense amount of effort, and maintaining a consistent level of quality is not easy. Thus, techniques for extracting musical elements are essential for obtaining content-based information from musical signals.

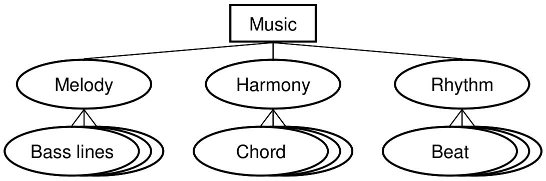

Figure 1. Musical elements

A key principle in analyzing musical signals is that musical elements are related to each other. Because composers exploit the interrelationship among musical elements, this interrelationship should be considered when analyzing the elements as well. Most studies in the literature have dealt with these elements independently.

This paper exploits the interrelationship between chord sequences and bass lines to improve the performance of chord recognition. The chord sequence is regarded as an element of the harmony, while bass lines are regarded as an element of the melody. The chord sequence consists of a chord symbol sequence and chord boundary sequence. As the chord sequence may represent the mood of music, it can be used to calculate the similarity in mood between musical pieces. This similarity is important in MIR and music recommendation. On the other hand, the bass line represents a melody in the bass register; thus, it leads the chord progression.

A recent approach adopted recently by many researchers for automated description of the chord sequence is the use of Hidden Markov Models (HMMs). Several methods have been suggested to explore the analogy between speech recognition and chord recognition and to consider the temporal connection of chords [1, 2, 3]. Sheh *et al.* [1] proposed a method that uses the extended Pitch Class Profile (PCP) [4] as a feature vector. They used an HMM that had one state per chord with a large set of classes (147 chord types). However, they were not able to obtain good enough results for recognition. Bello *et al.* [2] used chroma features and an HMM; they improved accuracy by incorporating musical knowledge into the model. Lee *et al.* [3] built key-specific models for automatic chord transcription. They used a 6-dimensional feature vector, called Tonal Centroid that is based on Tonnetz [5]. Higher accuracies were obtained by limiting the number of chord types that could be recognized.

Yoshioka *et al.* pointed out that chord symbols affect chord boundary recognition and vice versa [6]. They developed a method that concurrently recognizes chord symbols and boundaries by using a hypothesis search that recognizes the chord sequence and key.

While previous studies have treated only the features of chords, we focus on the interrelationship among musical elements and integrate information about bass lines into chord recognition in a probabilistic framework. The framework enables us to deal with multiple musical elements uniformly and integrate information obtained from statistical analyses of real music.

This paper is organized as follows: Section 2 describes our motivation for developing an automatic chord recognition system, the issues in involved in doing so, and our solution. Section 3 explains our method in concrete terms. Section 4 reports the experimental results and describes the effectiveness of our method. Our conclusions are discussed in Section 5.

2 CHORD RECOGNITION USING BASS PITCH ESTIMATION

2.1 Motivation

A bass line is a series of tonal linear events in the bass register. We focus on it specifically because it is strongly related to the chord sequence. Bass sounds are the most predominant tones in the low frequency region, and bass lines have the following important properties:

- They are comprised of the bass register of the musical chords.
- They lead the chord sequence.
- They can be played with a single tone and have a pitch that is relatively easy to estimate.

We improve chord recognition by exploiting the above properties. Information about bass lines is obtained through bass pitch estimation.

To process multiple musical elements simultaneously, we use a probabilistic framework that enables us to deal with them uniformly because each evaluation value is scaled from 0 to 1. In addition, with statistical training based on probability theory, it is possible to apply information obtained from the analysis of real music to the recognition. Thus, the framework has both scalability and flexibility.

2.2 Issues

We use the hypothesis-search-based (HSB) method proposed by Yoshioka *et al.* to recognize chord symbols and chord boundaries simultaneously. We chose this method over the HMM-based one because it expressly solves the mutual dependency problem between chord symbols and chord boundaries. Furthermore, the HSB method makes it easier to probabilistically integrate various musical elements. That is, we are able to integrate bass pitch estimation into the chord recognition. In this paper, Yoshioka's HSB method is called the baseline method. However, two issues remain in using it to calculate the evaluation value of the hypothesis.

2.2.1 *Usage of bass pitch estimation*

Although information about bass sounds is used in the baseline method, it is not used in a probabilistic framework. When the predominant single tone estimated is different from the harmonic tones of the chord, penalties are imposed on the certainty based on bass sounds. Errors in estimating single tones also tend to produce errors in chord recognition.

2.2.2 *Non-probabilistic certainties*

The certainties based on musical elements for the evaluation function are not probabilistic in the baseline method. When the observation distribution of chroma vectors [7] is approximated with a single Gaussian, the Mahalanobis generalized distance between acoustic features is used as the certainty. Another certainty based on chord progression patterns is defined as a penalty. The criterion for applying this penalty is related to progression patterns that appear several times. However, as the scales of the values was inconsistent, it becomes difficult to integrate multiple elements. Additionally, optimizing the weighting factors of each value takes a lot of time and effort.

2.3 Our Solution

To resolve the above issues, we use bass pitch probability (BPP) and define hypothesis reliability by using a probabilistic function.

2.3.1 *Bass Pitch Probability*

We utilize BPP as information about bass lines to reduce the effect of bass pitch estimation errors on chord recognition. BPP can be estimated using a method called PreFEst [8]. Because BPP is uniform in non-bass-sound frames, it does not significantly affect chord recognition in these frames.

2.3.2 *Probabilistic design of hypothesis reliability*

It is necessary to reformulate hypothesis reliability in order to use BPP in the reliability calculation. Like the baseline

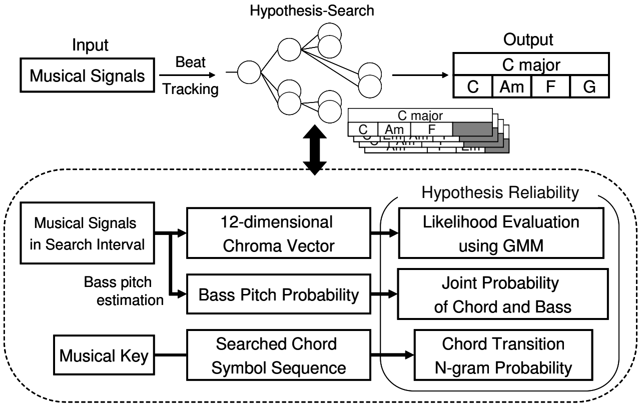

Figure 2. System Overview

method, we use acoustic features and chord progression patterns. However, we define the reliabilities based on these features by using a probabilistic framework. The values of the acoustic features are based on the likelihood of Gaussian Mixture Models (GMMs) for 12-dimensional chroma vectors. The values for the progression patterns are based on the transition probability obtained from statistical N-gram models. This reformulation enables three reliabilities to be integrated: those based on acoustic features, on BPP, and on transition probability.

3 SYSTEM

By applying information about BPP when calculating hypothesis reliability, we integrate the baseline method and bass pitch estimation in our chord recognition system. The three reliabilities based on the three elements are formulated probabilistically so that the system deals with them uniformly.

Figure 2 shows an overview of our automatic chord recognition system. First, the beat tracking method from [9] is used to detect the eighth note level beat times of an input musical piece. Second, the hypotheses are expanded over this beat time, and hypothesis reliability is calculated based on these three reliabilities. Third, a beam-search method using hypothesis reliability as the cost function is used to prune the expanded hypotheses. These operations are repeated until the end of the input signal. Finally, we obtain a tuple comprising the chord symbol sequence, chord boundary sequence, and key from the hypothesis that has the highest reliability.

3.1 Specification of Chord Recognition System

We define automatic chord recognition as a process of automatically obtaining a chord symbol sequence, a chord boundary sequence, and a key. These elements are defined as follows:

- **Chord symbol sequence**
  $\mathbf{c} = [c_1 \cdots c_M], c_i \in \mathbb{C} \equiv \mathbb{R} \times \mathbb{T}$
  The system uses 48 classes (major, minor, diminished, and sus4 for each of the 12 roots). Triads, sevenths, and so on are included as subclasses of these larger classes. we focused on discriminating the larger classes. For MIR applications we believe the larger classes would be sufficient to capture the characteristics or moods of accompaniments of musical pieces.
- **Chord boundary sequence**
  $\mathbf{t} = [t_0 \cdots t_M], t_i \in \mathbb{N}$
  where M is the number of chord symbols,
  t_i denotes the boundary time of c_i and c_{i+1},
  t_0 denotes the beginning time of the input signal,
  and t_M denotes the end time of the input signal.
- **Key**
  $k, k \in \mathbb{K} \equiv \mathbb{R} \times \mathbb{M}$

$\mathbb{R}$, $\mathbb{M}$, $\mathbb{T}$ are defined as follows:

$$\begin{aligned}\mathbb{R} &\equiv \{\mathrm{C}, \mathrm{C\#}, \cdots, \mathrm{B}\}, \quad \mathbb{M} \equiv \{\mathrm{Major}, \mathrm{Minor}\},\\ \mathbb{T} &\equiv \{\mathrm{Major}, \mathrm{Minor}, \mathrm{Diminished}, \mathrm{Sus4}\}\end{aligned}$$

We also assume the tempo stays constant, the beat is a common measure (four-four time), and the key does not modulate.

3.2 Formulating of Hypothesis Reliability

Denoting the observed feature sequence over frames $\tau = (\tau_s, \tau_e)$ as X_τ, we can probabilistically define hypothesis reliability Rel_τ as follows:

$$Rel_\tau = p(c_\tau | X_\tau) \quad (X_\tau = [x_{\tau_s} \cdots x_{\tau_e}]) \tag{1}$$

The BPP, β_f^τ, during a duration τ is defined from the frame-by-frame BPP, $w^{(t)}(f)$, as follows:

$$\beta_f^\tau = \sum_{i=\tau_s}^{\tau_e} w^{(i)}(f)/\alpha = \{w^{(\tau_s)}(f) + \cdots + w^{(\tau_e)}(f)\}/\alpha \tag{2}$$

where f denotes the frequency of the bass pitch and α is a normalization constant. Hypothesis reliability integrating BPP on the duration τ is defined as follows:

$$Rel_\tau = p(c_\tau | X_\tau) = \sum_f p(c_\tau, \beta_f^\tau | X_\tau) \tag{3}$$

This hypothesis reliability is converted with Bayes' theorem into another form as follows:

$$\sum_f p(c_\tau, \beta_f^\tau | X_\tau) \propto \sum_f p(X_\tau | c_\tau, \beta_f^\tau) p(c_\tau, \beta_f^\tau) \tag{4}$$

$$= \sum_f p(X_\tau | c_\tau, \beta_f^\tau) p(c_\tau | \beta_f^\tau) p(\beta_f^\tau) \tag{5}$$

We use 12-dimensional chroma vectors as the observation features. Since the vectors only depend on the chord symbol, we set the following expression.

$$p(X_\tau | c_\tau, \beta_f^\tau) = p(X_\tau | c_\tau) \tag{6}$$

Thus, hypothesis reliability over τ becomes as follows:

$$Rel_\tau = p(X_\tau|c_\tau) \sum_f p(c_\tau|\beta_f^\tau)p(\beta_f^\tau) \qquad (7)$$

where $p(X_\tau|c_\tau)$ denotes the reliability based on acoustic features, and $\sum_f p(c_\tau|\beta_f^\tau)p(\beta_f^\tau)$ denotes the reliability based on BPP.

The key k is independent of chord boundaries. With the conditional probability of a chord symbol sequence given a key, overall hypothesis reliability Rel_{all} is defined as follows:

$$\begin{aligned} Rel_{\text{all}} &= p(\mathbf{c}|k) \prod_\tau Rel_\tau & (8) \\ &= p(\mathbf{c}|k) \prod_\tau p(X_\tau|c_\tau) \sum_f p(c_\tau|\beta_f^\tau)p(\beta_f^\tau) & (9) \end{aligned}$$

where $p(\mathbf{c}|k)$ denotes the reliability based on transition probability of chords.

3.3 Reliability based on Acoustic Features

We use 12-dimensional chroma vectors as acoustic features; these vectors approximately represent the intensities of the 12-semitone pitch classes. As the chord symbols are identified by the variety of tones, it is essential for chord recognition.

Because we focus on four chord types, major, minor, diminished and sus4, we use four M-mixture GMMs (Maj-GMM, Min-GMM, Dim-GMM, and Sus4-GMM). The parameters of each GMM, λ_t, are trained on chroma vectors calculated at the frame level. Note that chords of different roots are normalized by rotating chroma vectors. This normalization reduces the number of GMMs to four and effectively increases the number of training samples. The EM algorithm is used to determine the mean, covariance matrix, and weight for each Gaussian.

After chroma vectors from the input signals are rotated to adapt to the 12 chords having different root tones, we calculate the log likelihood between them and the 4 GMMs. The likelihood is equal to $p(X_\tau|c_\tau)$. Thus, the reliability divided by the number of frames in the hypothesis, g_c is defined as follows:

$$g_c = \log p(X_\tau|c_\tau) = \log p(X_r|\lambda_t), \qquad (10)$$

where r denotes the number of rotating chroma vector indexes and t denotes the number of GMM types.

3.4 Reliability based on Bass Pitch Probability

BPP is obtained by bass pitch estimation, and it is used to represent the degree of each pitch of bass sounds. Since the bass lines determine the chord sequence, they should be simultaneously analyzed for recognizing the chord sequence.

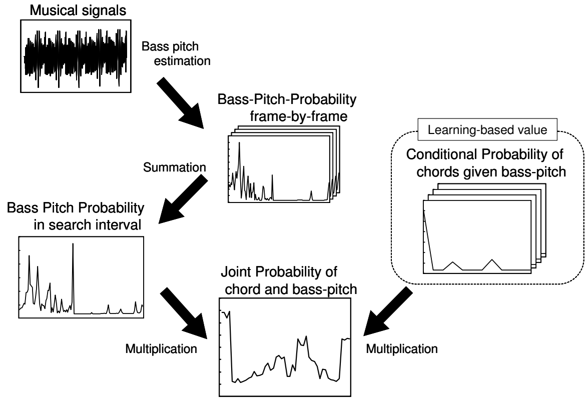
Musical signals
Bass pitch estimation
Bass-Pitch-Probability frame-by-frame
Learning-based value
Conditional Probability of chords given bass-pitch
Summation
Bass Pitch Probability in search interval
Joint Probability of chord and bass-pitch
Multiplication
Multiplication

Figure 3. Bass-Pitch Processing

Figure 3 shows an overview of the process for calculating the reliability based on BPP. The prior probability of bass sounds $p(\beta_f^\tau)$ is defined by using BPP $w^{(t)}(f)$ as follows:

$$p(\beta_f^\tau) = w^{(\tau)}(f) = \sum_{j=\tau_s}^{\tau_e} w^{(j)}(f) \qquad (11)$$

On the other hand, the conditional probability of chord c_i given bass pitch β_f^τ is obtained from real music by using correct chord labels and the results of PreFEst for the particular duration. We statistically calculate the frequency of appearance of each bass pitch for each chord.

As the reliability based on BPP is the log joint probability of the chord and bass pitch, the reliability g_b is defined in terms of the BPP $w^{(\tau)}(f)$ and the conditional probability $p(c_i|\beta_f^\tau)$.

$$g_b = \log(\sum_f p(c_i|\beta_f^\tau)w^{(\tau)}(f)) \qquad (12)$$

3.5 Reliability based on Transition Probability

Music theory indicates that the genre and the artist usually determine the chord progression patterns for a given musical piece. The progression patterns are obtained from the key and scale degree. We probabilistically approximate the frequency of a chord symbol appearing, thus reducing the ambiguity of chord symbols.

We use two 2-gram models, one for major keys and one for minor keys. They are obtained from real music in advance. In the learning stage, we obtain the 2-gram probabilities from common chord symbol sequences consisting of the key and correct chord labels. We calculate the 2-gram probability $p(c_i|c_i^{(-1)})$ from the number of progression patterns and use smoothing to handle progression patterns not appearing in the training samples.

We estimate the transition 2-gram probability on a log scale by using the hypothesis's key. The reliability g_p is defined as follows:

$$g_p = \log p(c_i|k) = \log p(c_i|c_i^{(-1)}) \qquad (13)$$

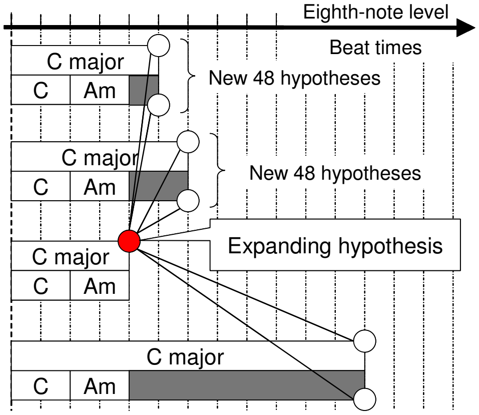

Figure 4. Hypothesis Expansion

3.6 Integrating three reliabilities

Because hypothesis reliability on a log scale is the weighted sum of the three log reliabilities described above, Rel is defined as follows:

$$Rel = w_c \times g_c + w_b \times g_b + w_p \times g_p \tag{14}$$

where w_c, w_b, and w_p are weight constants.

3.7 Updating the Hypothesis

3.7.1 *Hypothesis Expansion*

Figure 4 shows the hypothesis expansion process. We consider the minimum size of boundary intervals to be eighth-note level beat times, which is the same as the baseline method. For each unit between one and eight beats, 48 hypotheses are generated for 48 chord symbols.

3.7.2 *Hypothesis Pruning*

We use a beam search to prune the expanded hypotheses; this prevents the number of hypotheses from expanding exponentially. In the beam search, hypotheses are pruned using a beam width BS. After the pruning, the BS hypotheses with higher reliability than the rest of the hypotheses are expanded. Furthermore, by delaying the expansion of hypotheses until it comes time to evaluate them, we can apply pruning to newly generated hypotheses. As a result, by decreasing wasteful expansions of hypotheses we can reduce both the amount of computation and memory required.

3.7.3 *Update of Hypothesis Reliability*

When a hypothesis is expanded, we need to update hypothesis reliability. To treat hypotheses with different numbers of intervals fairly, the reliability of a new hypothesis Rel_{new} is defined as follows:

$$Rel_{new} = \frac{Rel_{prev}}{N_h} + Rel_{next}, \tag{15}$$

where Rel_{prev} denotes the reliability of the previously expanded hypothesis, Rel_{next} denotes the hypothesis reliability of the newly expanded interval, and N_h denotes the number of previously expanded intervals.

Table 1. Parameter values

$BS = 25$	$M = 16$	
$w_c = 1.0$	$w_b = 2.0$	$w_p = 0.3$

Table 2. Results of 5-fold cross validation
[1]:Baseline Method, [2]:Ac, [3]:Ac+Ba, [4]:Ac+Pt, **[5]:Our Method**

Groups	[1]	[2]	[3]	[4]	**[5]**
1st	65.1	62.9	71.7	68.3	**78.3**
2nd	62.7	62.6	70.6	65.7	**74.9**
3rd	57.7	60.4	66.8	61.2	**69.5**
4th	61.0	61.3	70.2	64.0	**72.7**
5th	61.6	60.9	69.2	64.8	**71.4**
Total	61.6	61.6	69.7	64.8	**73.4**

4 EXPERIMENTAL RESULTS

We tested our system on one-minute excerpts from 150 songs in 12 Beatles albums (a total of 180 songs), which have the properties described in Section 3.1. These songs were separated into 5 groups at random for 5-fold cross validation. For training the parameters of the GMMs, we used chroma vectors calculated from audio signals of 120 of these songs and audio signals of each chord played on a MIDI tone generator. These songs were also used to train the conditional probability of chords given the bass pitch (Section 3.4). We utilized 150 songs (137 in the major key, 13 in the minor key) as training data for the 2-gram models. As the correct chord labels, we used ground-truth annotations of these Beatles albums made by C. A. Harte [10]. The implementation for experiments used the parameters listed in Table 1.

To evaluate the effectiveness of our method, we compared the frame-rate accuracies of the following five methods of computing the hypothesis reliability:

1. Baseline method
2. Using only acoustic features
3. Using acoustic features and BBP
4. Using acoustic features and transition probability
5. Our method (three elements)

The results are listed in Table 2. With our system, the average accuracy for the 150 songs was 73.4%. Compared with using only acoustic features, the method using both acoustic features and bass pitch probability improved the recognition rate by 8.1 points. Furthermore, the method using acoustic features, BBP, and transition probability improved the recognition rate by 11.8 points. In addition, our system's accuracy was higher than that of the baseline method. This is because the probabilistic integration enabled us to

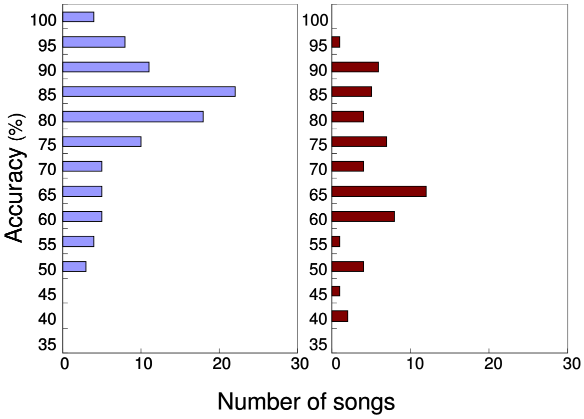

Figure 5. Accuracy Histograms.(Left) Histogram for the songs with correct key. (Right) Histogram for the songs with incorrect key.

utilize information about bass lines as a clue in chord recognition. Thus, the results prove both the importance of considering the interrelationship between chord sequence and bass lines and the effectiveness of the probabilistic integration of these elements.

We compared our results with those obtained from other systems proposed by Bello [2] and Lee [3]. They used two Beatles albums (*Please Please Me*, *Beatles for Sale*) as the test data set. Our system had an average accuracy of 77.5%. Although both Bello's and Lee's system used different training data from those that we used, their systems had 75.0% and 74.4% accuracy, respectively.

Upon investigating the accuracy distribution for all songs, we fount that the accuracy histogram for all songs is polarized with two peaks split at approximately 70%. Figure 5 shows two accuracy histograms, one for songs where the key was estimated correctly and another where it was incorrect. As these histograms describe the polarization, it is clearly important to correctly estimate the key for chord recognition. To improve key estimation, we plan to develop a method that searches the hypotheses by using not only a forward search but also backtracking.

5 CONCLUSION

We presented a chord recognition system that takes into account the interrelationship among musical elements. Specifically, we focus on bass lines and integrate hypothesis-search-based chord recognition and bass pitch estimation in a probabilistic framework. To evaluate hypotheses, our system calculates the hypothesis reliability, which is designed by the probabilistic integration of three reliabilities, based on acoustic features, bass pitch probability, and chord transition probability. The experimental results showed that our system had a 73.4% frame-rate accuracy of chord recognition in 150 songs. They also showed an increase in accuracy when the three reliabilities were integrated compared with the baseline method and a method using only acoustic features. This shows that to recognize musical elements (not only musical chords but also other elements), it is important to consider the interrelationship among musical elements and to integrate them probabilistically. To obtain more information about how to recognize chord sequences more effectively, we will design a way to integrate other musical elements such as rhythm.

6 ACKNOWLEDGEMENTS

This research was partially supported by the Ministry of Education, Culture, Sports, Science and Technology (MEXT), the Global Century of Excellence (GCOE) Program, and the CrestMuse Project of the Japan Science and Technology Agency (JST). The authors thank Takuya Yoshioka for his valuable discussion and comments, C. A. Harte for providing us ground-truth annotations of the Beatles albums, and Graham Neubig for his careful reading of our English manuscript.

7 REFERENCES

[1] A. Sheh and D. P. W. Ellis, "Chord Segmentation and Recognition Using EM-Trained Hidden Markov Models," *Proc. ISMIR*, pp. 183-189, 2003.

[2] J. P. Bello and J. Pickens, "A Robust Mid-level Representation for Harmonic Content in Music Signals," *Proc. ISMIR*, pp. 304-311, 2005.

[3] K. Lee and M. Slaney, "A Unified System for Chord Transcription and Key Extraction Using Hidden Markov Models," *Proc. ISMIR*, pp. 245-250, 2007.

[4] T. Fujishima, "Realtime Chord Recognition of Musical Sound: a System Using Common Lisp Music," *Proc. ICMC*, pp. 464-467, 1999.

[5] C. A. Harte, M. B. Sandler, and M. Gasser, "Detecting Harmonic Change in Musical Audio," *Proc. Audio and Music Computing for Multimedia Workshop*, pp. 21-26, 2006.

[6] T. Yoshioka, T. Kitahara, K. Komatani, T. Ogata, and H. G. Okuno, "Automatic Chord Transcription with Concurrent Recognition of Chord Symbols and Boundaries," *Proc. ISMIR*, pp. 100-105, 2004.

[7] M. Goto, "A Chorus-Section Detecting Method for Musical Audio Signals," *Proc. ICASSP*, **V**, pp. 437-440, 2003.

[8] M. Goto, "A Real-time Music-scene-description System: Predominant-F0 Estimation for Detecting Melody and Bass Lines in Real-world Audio Signals," *Speech Comm.*, **43**:4, pp. 311-329, 2004.

[9] M. Goto, "An Audio-based Real-time Beat Tracking System for Music With or Without Drum-sounds," *Journal. of New Music Research*, **30**:2, pp. 159–171, 2001.

[10] C. A. Harte, et al., "Symbolic Representation of Musical Chords: A Proposed Syntax for Text Annotations," *Proc. ISMIR*, pp. 66-71, 2005.