ICASSP 2003 (The 2003 IEEE International Conference on Acoustics, Speech, and Signal Processing) Proceedings, pp.V-437-440, April 2003.

A CHORUS-SECTION DETECTING METHOD FOR MUSICAL AUDIO SIGNALS

Masataka Goto

"Information and Human Activity," PRESTO, Japan Science and Technology Corporation (JST). /
National Institute of Advanced Industrial Science and Technology (AIST).
1-1-1 Umezono, Tsukuba, Ibaraki 305-8568, JAPAN. < m.goto@aist.go.jp >

ABSTRACT

This paper describes a method for obtaining a list of chorus (refrain) sections in compact-disc recordings of popular music. The detection of chorus sections is essential for the computational modeling of music understanding and is useful in various applications, such as automatic chorus-preview functions in music browsers or retrieval systems. Most previous methods detected as a chorus a repeated section of a given length and had difficulty in identifying both ends of a chorus section and in dealing with modulations (key changes). By analyzing relationships between various repeated sections, our method called *RefraiD* can detect all the chorus sections in a song and estimate both ends of each section. It can also detect modulated chorus sections by introducing a similarity that enables modulated repetition to be judged correctly. Experimental results with a popular-music database show that this method detects the correct chorus sections in 80 of 100 songs.

1. INTRODUCTION

Chorus (refrain) sections of popular music are the most representative and prominent thematic sections in the music structure of a song, and human listeners can easily understand where the chorus sections are because these sections are most repeated and memorable portions of a song. Their automatic detection is essential for building a computational model that can understand musical audio signals in a human-like fashion, and is useful in various practical applications. In music browsers or music retrieval systems, it enables a user to quickly preview a chorus section as an "audio thumbnail" in order to find a desired song. It can also increase the efficiency and precision of music retrieval systems by enabling them to match a query with only the chorus sections.

Most previous chorus-detection methods [1, 2, 3] obtained a single segment from several chorus sections by detecting a repeated section of a given length as the most representative of a song. Logan and Chu [1] developed a method using clustering techniques and Hidden Markov Models to categorize short segments (1 sec) in terms of their acoustic features and then regarded the most frequent category as a chorus. Bartsch and Wakefield [2] developed a method that calculated the similarity between acoustic features of beat-length segments obtained by beat tracking and found the given-length segment with the highest similarity averaged over its segment. Cooper and Foote [3] developed a method that calculated the similarity between acoustic features of short frames (100 ms) and found the given-length segment with the highest similarity between it and the whole song. None of the previous methods, however, addressed the problem of detecting all the chorus sections in a song. They also assumed that the output segment length is given and did not identify both ends of a chorus section. While chorus sections are sometimes modulated (the key is changed) during their repetition in a song, the previous methods were not able to deal with modulated repetition.

This paper describes a method, called *RefraiD* (Refrain Detecting Method), that exhaustively detects all the chorus sections appearing in a song. It can obtain a list of the start and end points of every chorus section in real-world audio signals and can detect modulated chorus sections. Furthermore, because it detects chorus sections by analyzing various repeated sections in a song, it can generate an intermediate-result list of repeated sections that usually reflect the music structure of the song; for example, the repetition of the structure like the verse A, verse B, and chorus is often found in the list.

The following sections describe the problems dealt with, specify the RefraiD method in detail, and show experimental results indicating that the method is robust enough to detect the correct chorus sections in 80 of 100 songs of a popular-music database.

2. CHORUS-SECTION DETECTING PROBLEM

Given an audio signal of a song, we want to obtain a list of all the chorus sections without using any prior information about the spectral characteristics of chorus sections. Because the chorus sections are usually the most repeated sections in popular music, the basic idea behind dealing with this problem is to find various groups of repeated sections and then output the group that appears most frequently. It is, however, generally difficult to find the repeated sections automatically because they do not completely match each other. The main issues can be summarized as follows:

[Problem 1] Acoustic features and similarity measure

Whether or not one section is the repetition of another must be judged on the basis of the similarity between acoustic features of those sections. The simple power spectrum or MFCC features are not powerful tools for this judgment because they are liable to change considerably when arrangements of accompaniments or melody lines are changed after repetition.

[Problem 2] Repetition-judgment criterion

The appropriate criterion of the similarity for judging repetition depends on the song. For a song containing many-repeated accompaniment phrases, for example, only a section with very high similarity should be considered the chorus-section repetition. For a song containing a chorus section with accompaniments changed after repetition, on the other hand, a section with somewhat lower similarity can be considered the chorus-section repetition. Because a criterion hand-tuned for a few songs is not robust for a large song set, the criterion must be adjusted automatically for each song.

[Problem 3] Estimating both ends of repeated sections

It is necessary to estimate the start and end points of repeated sections by analyzing relationships between various repeated sections. For a song whose structure is (A B C B C C), for example, the repetition of (B C) would be obtained by a simple repetition search. In this case, both ends of this C can be estimated by using the information of the repetition of the last C.

[Problem 4] Detecting the modulated repetition

Although the modulated repetition is difficult to detect because acoustic features are not similar after a modulation, this detection is important because the repetition of chorus sections (especially around the end of song) sometimes includes the modulation.

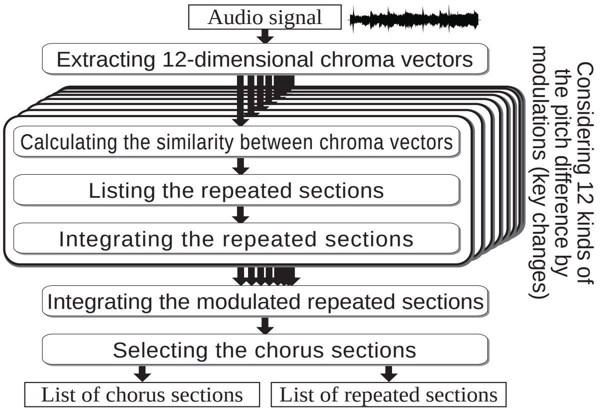
Audio signal
Extracting 12-dimensional chroma vectors
Calculating the similarity between chroma vectors
Listing the repeated sections
Integrating the repeated sections
Considering 12 kinds of the pitch difference by modulations (key changes)
Integrating the modulated repeated sections
Selecting the chorus sections
List of chorus sections
List of repeated sections

Fig. 1. Overview of the chorus-section detecting method *RefraiD*.

3. CHORUS-SECTION DETECTING METHOD: REFRAID

Figure 1 shows an overview of the RefraiD method. It first extracts a 12-dimensional feature vector called a *chroma vector*, which is robust for small changes of accompaniments, from each frame of an input audio signal and calculates the similarity between these vectors *(solution to Problem 1)*. Each element of the chroma vector corresponds to one of the 12 pitch classes and is the sum of power at frequencies of its pitch class over six octaves. The method then lists pairs of repeated sections by using an adaptive repetition-judgment criterion that is configured by an automatic threshold selection method based on a discriminant criterion [4] *(solution to Problem 2)*. To organize common repeated sections into groups and to identify both ends of each section, it integrates those pairs by analyzing their relationships over the whole song *(solution to Problem 3)*. Because each element of a chroma vector corresponds to a different pitch class, a before-modulation chroma vector is close to the after-modulation chorus vector whose elements are shifted (exchanged) by the pitch difference of the key change. By considering twelve kinds of shifts (pitch differences), it then calculates 12 sets of the similarity between non-shifted and shifted chroma vectors, lists pairs of repeated sections from those sets, and integrates all of them *(solution to Problem 4)*. Finally, it evaluates the possibility of being chorus sections for each group and outputs the repeated sections with the highest possibility as well as other groups of repeated sections.

3.1. Extracting acoustic features

The 12-dimensional *chroma vector* $\vec{v}(t)$ is extracted from the power spectrum, $\Psi_p(f,t)$ at the log-scale frequency f at time t, calculated by using the short-time Fourier transform (STFT). Each element of $\vec{v}(t)$ corresponds to a pitch class c ($c = 1, 2, \ldots 12$) in the equal temperament and is represented as $v_c(t)$:

$$v_c(t) = \sum_{h=\mathrm{Oct_L}}^{\mathrm{Oct_H}} \int_{-\infty}^{\infty} BPF_{c,h}(f)\ \Psi_p(f,t)\ df. \qquad (1)$$

The $BPF_{c,h}(f)$ is a bandpass filter that passes at the log-scale frequency $F_{c,h}$ (in cents[1]) of pitch class c in octave position h[2]

$$F_{c,h} = 1200h + 100(c-1) \qquad (2)$$

and is defined using a Hanning window as follows:

$$BPF_{c,h}(f) = \frac{1}{2}\left(1 - \cos\frac{2\pi(f-(F_{c,h}-100))}{200}\right). \qquad (3)$$

This filter is applied to octaves from $\mathrm{Oct_L}$ to $\mathrm{Oct_H}$.

In the current implementation, the input signal is digitized at 16 bit / 16 kHz, and then the STFT with a 4096-sample Hanning window is calculated by using the Fast Fourier Transform (FFT). Since the FFT frame is shifted by 1280 samples, the discrete time step (1 frame shift) is 80 ms. The $\mathrm{Oct_L}$ and $\mathrm{Oct_H}$, the octave range for the summation of Equation (1), are respectively 3 and 8. This covers six octaves (130 Hz to 8 kHz).

There are several advantages to the chroma vector.[3] Because it captures the overall harmony (pitch-class distribution), it can be similar even if accompaniments or melody lines are changed in some degree after repetition. In fact, we have confirmed that the chroma vector is effective for identifying chord names. The chroma vector also enables the modulated repetition to be detected as described in Section 3.5.

3.2. Calculating the similarity

The similarity $r(t,l)$ between the chroma vectors $\vec{v}(t)$ and $\vec{v}(t-l)$ is defined as

$$r(t,l) = 1 - \frac{\left|\frac{\vec{v}(t)}{\max_c v_c(t)} - \frac{\vec{v}(t-l)}{\max_c v_c(t-l)}\right|}{\sqrt{12}}, \qquad (4)$$

where l ($0 \le l \le t$) is the lag. Since the denominator $\sqrt{12}$ is the length of the diagonal line of the 12-dimensional hypercube with edge length 1, $r(t,l)$ satisfies $0 \le r(t,l) \le 1$.

3.3. Listing the repeated sections

Pairs of repeated sections are obtained from $r(t,l)$. Considering that $r(t,l)$ is drawn within the right-angled isosceles triangle in the two-dimensional time-lag space as shown in Figure 2, the method finds line segments that are parallel to the horizontal time axis and indicate consecutive regions with high $r(t,l)$. When the section between the time $T1$ and $T2$ is denoted $[T1, T2]$, each line segment between the points $(T1, L1)$ and $(T2, L1)$ is represented as $(t = [T1, T2], l = L1)$ and means that the section $[T1, T2]$ is similar to (i.e., is the repetition of) the section $[T1 - L1, T2 - L1]$. In other words, a line segment indicates a repeated-section pair.

To find $(t = [T1, T2], l = L1)$ in $r(t,l)$, the possibility of containing line segments at the lag l, $R_{all}(t,l)$, is evaluated at the current time t (e.g., at the end of song) as follows (Figure 2):

$$R_{all}(t,l) = \int_l^t \frac{r(\tau,l)}{t-l}\ d\tau. \qquad (5)$$

Before this calculation, $r(t,l)$ is normalized by subtracting the mean of $r(t,l)$ in the adjacent area while removing noises. The method then picks up high peaks above a threshold Th_R of $R_{all}(t,l)$ for searching the line segments, after smoothing $R_{all}(t,l)$ by using a moving average filter. Because the threshold Th_R is closely related to the repetition-judgment criterion that should be adjusted for each song, we use an automatic threshold selection method based on a discriminant criterion [4]. When dichotomizing the peak heights into two classes by a threshold, the optimal threshold is obtained by maximizing the discriminant criterion measure that is defined by the following between-class variance:

$$\sigma_B^2 = \omega_1\omega_2(\mu_1 - \mu_2)^2, \qquad (6)$$

where ω_1 and ω_2 are the probabilities of class occurrence (the number of peaks in each class / the total number of peaks), and μ_1 and μ_2 are the mean of peak heights in each class.

The line segments are finally searched in the direction of the horizontal time axis on the one-dimensional function $r(\tau, L1)$ ($L1 \le \tau \le t$) at the lag $L1$ of each high peak. After smoothing

[1] Frequency f_{Hz} in hertz is converted to frequency f_{cent} in cents so that there are 100 cents to a tempered semitone and 1200 to an octave: $f_{\mathrm{cent}} = 1200\log_2(f_{\mathrm{Hz}}\ /\ (440 \times 2^{\frac{3}{12}-5}))$.

[2] In the Shepard's helix representation of pitch perception [5], c and h respectively correspond to *chroma* and *height*.

[3] The chroma vector is similar to the chroma spectrum [6] that is used in reference [2], although its formulation is different.

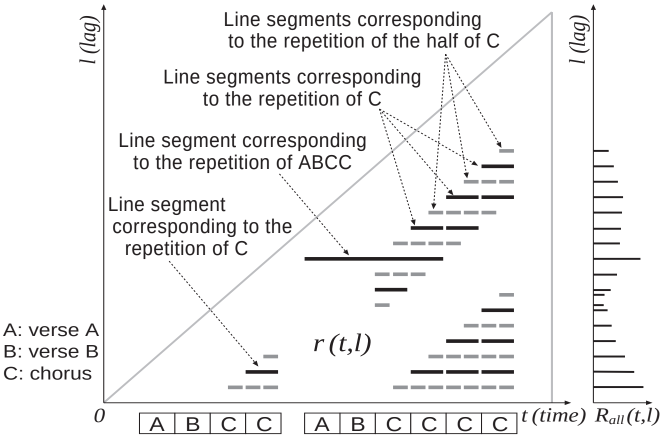
l (lag)
Line segments corresponding to the repetition of the half of C
Line segments corresponding to the repetition of C
Line segment corresponding to the repetition of ABCC
Line segment corresponding to the repetition of C
A: verse A
B: verse B
C: chorus
r (t,l)
0
A B C C A B C C C C
t (time)
l (lag)
R_all(t,l)

Fig. 2. A plot of line segments, the similarity $r(t,l)$, and the possibility $R_{all}(t,l)$ of containing line segments. The similarity $r(t,l)$ is defined in the right-angled isosceles triangle in the lower right-hand corner. The actual $r(t,l)$ is noisy and ambiguous and usually contains many line segments irrelevant to chorus sections.

$r(\tau, L1)$ by using a moving average filter, the method obtains line segments on which the smoothed $r(\tau, L1)$ is above a threshold. This threshold is also adjusted by using the automatic threshold selection method.

3.4. Integrating the repeated sections

Since each line segment indicates just a pair of repeated sections, it is necessary to organize into a group the line segments that have common sections. First, line segments that have almost the same section $[Ts_i, Te_i]$ are organized into a group, which is represented as $\phi_i = ([Ts_i, Te_i], \Gamma_i)$, where $\Gamma_i = \{\gamma_{ij} \mid j = 1, 2, \ldots, M_i\}$ (M_i is the number of line segments) is a set of the segment's lag γ_{ij} — corresponding to the high peaks in $R_{all}(t,l)$ — in this group. A set of these groups is denoted by $\Phi = \{\phi_i \mid i = 1, 2, \ldots, N\}$ (N is the number of all the groups).

By using $r(t,l)$ just within $[Ts_i, Te_i]$ of each group ϕ_i, line segments are then searched again in order to recover some line segments that the process described in Section 3.3 did not find. In Figure 2, for example, we can expect that two line segments corresponding to the repetition of the first and third C and the repetition of the second and fourth C, which overlap with the long line segment corresponding to the repetition of ABCC, are found here even if they were hard to be found in the process described in Section 3.3. For this purpose, starting from

$$R_{[Ts_i, Te_i]}(l) = \int_{Ts_i}^{Te_i} \frac{r(\tau, l)}{Te_i - Ts_i}\, d\tau \qquad (7)$$

instead of $R_{all}(t,l)$, the method performs the almost same peak-picking process described in Section 3.3 and forms a new set Γ_i of the peaks γ_{ij} in $R_{[Ts_i, Te_i]}(l)$. It then removes inappropriate peaks in each Γ_i as follows: it removes too many peaks that are equally spaced, a peak whose line segment has a highly deviated $r(\tau, \gamma_{ij})$ ($Ts_i \leq \tau \leq Te_i$), and a peak that is too close to other peaks and makes sections overlap.

Finally, by using the lag γ_{ij} corresponding to each peak of Γ_i, the method searches for a group whose section is $[Ts_i - \gamma_{ij}, Te_i - \gamma_{ij}]$ (i.e., is shared by the current group Γ_i) and integrates it with Γ_i if it is found. They are integrated by adding all the peaks of the found group to Γ_i after adjusting the lag values (peak positions); the found group is then removed. In addition, if there is a group that has a peak indicating the section $[Ts_i - \gamma_{ij}, Te_i - \gamma_{ij}]$, it too is integrated.

3.5. Detecting the modulated repetition

The processes described above do not deal with the modulation (key change), but they can easily be extended to it. A modulation can be represented by the pitch difference of its key change, tr $(0, 1, \ldots, 11)$, which denotes the number of tempered semitones. For example, $tr = 9$ means the modulation of nine semitones upward or the modulation of three semitones downward.

One of the advantages of the 12-dimensional chroma vector $\vec{v}(t)$ is that tr of the modulation can naturally correspond to the amount by which its 12 elements are shifted. When $\vec{v}(t)$ is the chroma vector of a performance and $\vec{v}(t)'$ is the chroma vector of the performance that is modulated by tr semitones upward from the original performance, they satisfy

$$\vec{v}(t) \doteqdot S^{tr}\vec{v}(t)', \qquad (8)$$

where S is a shift matrix defined by

$$S = \begin{pmatrix} 0 & 1 & 0 & \cdots & \cdots & 0 \\ 0 & 0 & 1 & 0 & \cdots & 0 \\ \vdots & & \ddots & \ddots & \ddots & \vdots \\ \vdots & & & \ddots & \ddots & \vdots \\ 0 & \cdots & \cdots & 0 & 1 & 0 \\ 0 & \cdots & \cdots & \cdots & 0 & 1 \\ 1 & 0 & \cdots & \cdots & \cdots & 0 \end{pmatrix}. \qquad (9)$$

To detect the modulated repetition, we can define 12 kinds of extended similarity for each tr as follows:

$$r_{tr}(t,l) = 1 - \frac{\left| \frac{S^{tr}\vec{v}(t)}{\max_c v_c(t)} - \frac{\vec{v}(t-l)}{\max_c v_c(t-l)} \right|}{\sqrt{12}}. \qquad (10)$$

Starting from each $r_{tr}(t,l)$, the processes of listing and integrating the repeated sections are performed as described in Sections 3.3 and 3.4, except that the threshold adjusted at $tr = 0$ is used for the processes at $tr \neq 0$. After these processes, 12 sets of line-segment groups are obtained for 12 kinds of tr. To organize non-modulated and modulated repeated sections into the same groups, the method integrates several groups across all the sets if they share the same section.

Hereafter, we use $\Phi = \{\phi_i | \phi_i = ([Ts_i, Te_i], \Gamma_i)\}$ to denote the groups of line segments obtained from all the tr. By unfolding each line segment of γ_{ij} to the pair of repeated sections indicated by it, we can obtain

$$\Lambda_i = \{([Ps_{ij}, Pe_{ij}], \lambda_{ij}) \mid j = 1, 2, \ldots, M_i + 1\}, \qquad (11)$$

where $[Ps_{ij}, Pe_{ij}] = [Ts_i - \gamma_{ij}, Te_i - \gamma_{ij}]$ represents one of the unfolded repeated section, and λ_{ij} is its possibility of being chorus sections. The λ_{ij} is defined as the mean of the similarity $r_{tr}(t,l)$ on the corresponding line segment. For $j = M_i + 1$, we define $[Ps_{ij}, Pe_{ij}]$ and λ_{ij} as follows: $[Ps_{ij}, Pe_{ij}] = [Ts_i, Te_i]$ and $\lambda_{ij} = \max_{k=1}^{M_i} \lambda_{ik}$. The modulated sections are labeled with their tr for reference.

3.6. Selecting the chorus sections

After evaluating the total possibility ν_i of being chorus sections for each group (ϕ_i, Λ_i), the group m that maximizes ν_i is selected as the chorus sections: $m = \operatorname{argmax}_i\ \nu_i$. The total possibility ν_i is a sum of λ_{ij} weighted by the length of the section and is defined by

$$\nu_i = \left(\sum_{j=1}^{M_i+1} \lambda_{ij} \right) \log \frac{Te_i - Ts_i}{\mathrm{D_{len}}}, \qquad (12)$$

where $\mathrm{D_{len}}$ is a constant (1.4 sec). Before calculating ν_i, the possibility λ_{ij} of each repeated section is adjusted according to the following three assumptions (heuristics):

[Assumption 1] The length of the chorus section has an appropriate range (in the current implementation, 7.7 to 40 sec). If the length is out of the range, λ_{ij} is set to 0.

[Assumption 2] When there is a repeated section that is long enough to be likely to correspond to the long-term repetition like the verse A, verse B, and chorus, the chorus section is likely to

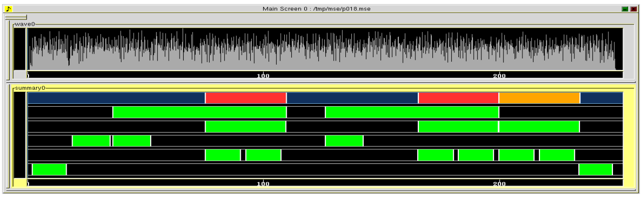

Fig. 3. The detected chorus sections of the database song RWC-MDB-P-2001 No. 18. The horizontal axis is the time (sec) corresponding to the whole song. The window above shows the power. The top line in the window below shows the list of the detected chorus sections, which were correct for this song and the last of which was modulated. The bottom five lines show the list of various repeated sections.

be near its end. If there is a repeated section $[Ps_{ij}, Pe_{ij}]$ whose end is close to the end of another long repeated section (longer then 50 sec), its λ_{ij} is doubled.

[Assumption 3] Because a chorus section tends to have two half-length repeated sub-sections within its section, a section that has those sub-sections is likely to be the chorus section. If there is a repeated section $[Ps_{ij}, Pe_{ij}]$ that has those sub-sections in another group, half of the mean of the possibility of those two sub-sections is added to its λ_{ij}.

These assumptions fit a large class of popular music.

4. EXPERIMENTAL RESULTS

The RefraiD method has been implemented in a real-time system that takes a musical audio signal as input and outputs the list of the detected chorus sections. Along the real-time audio input, the system displays visualized lists of chorus sections and other repeated sections, which are obtained by using just the past input and are considered most probable every moment.[4] A chorus-section viewer that shows those visualized lists as shown in Figure 3 and enables a user to play back a selected section has also been developed.

The system was tested on 100 songs of the popular-music database *"RWC Music Database: Popular Music"* (RWC-MDB-P-2001 No. 1−100) [7], which is an original database available to researchers around the world. These 100 songs were originally composed, arranged, performed, and recorded in a way that reflected the complexity and diversity of real-world music. We compared the system output with the correct chorus sections that were hand-labeled by using a music-structure labeling editor. The degree of matching between the detected and correct chorus sections was evaluated by using the F-measure [8], which is the harmonic mean of the recall rate (R) and the precision rate (P):

$$\text{F-measure} = \frac{2RP}{R+P} \tag{13}$$

$$R = \frac{\text{sum of the length of chorus sections detected correctly}}{\text{sum of the length of correct chorus sections}} \tag{14}$$

$$P = \frac{\text{sum of the length of chorus sections detected correctly}}{\text{sum of the length of chorus sections detected}}. \tag{15}$$

The system output of a song was judged to be correct if its F-measure is more than 0.75 under the condition that estimated *tr* of modulations must be correct.

Table 1. Results of evaluating RefraiD: the number of songs whose chorus sections were detected correctly under 4 sets of conditions.

	Condition (enable:○, disable:×)			
Modulation detection	○	×	○	×
Use of assumptions 2 & 3	○	○	×	×
Number of songs (out of 100)	80	74	72	68

The results are listed in Table 1. The method dealt correctly with 80 of 100 songs (the average of the F-measure of the 80 songs was 0.938). The main reasons that it made mistakes were that the number of repetition of chorus sections was not more than that of other sections and that an accompaniment phrase was repeated for most of a song. Among the 100 songs were 10 songs with modulated chorus sections, and the outputs of 9 of them were correct. When the function detecting the modulated repetition was disabled, only 74 songs were dealt with correctly. On the other hand, when assumptions 2 and 3 were not used, the performance fell as shown by the entries in the rightmost two columns of Table 1. There were 22 songs in which accompaniments or melody lines of a repeated chorus section were considerably changed; the outputs of 21 of them were correct and the repeated chorus section itself (i.e., *tr*) was correctly detected in 16 songs.

5. CONCLUSION

We have described the RefraiD method that detects the chorus sections in real-world popular-music audio signals. It basically regards the most repeated sections as the chorus sections. Analysis of the relationships between various repeated sections enables all the chorus sections to be detected with their start and end points. In addition, the introduction of the similarity between non-shifted and shifted chroma vectors makes it possible to detect modulated chorus sections, which previous methods could not detect. Experimental results show that the method is robust enough to detect the correct chorus sections in 80 of 100 songs.

The RefraiD method also has relevance to music summarization methods [9, 10, 11], none of which addressed the problem of detecting all the chorus sections. One of the chorus sections detected by our method can be regarded as a song summary, as could another long repeated section in the intermediate-result list of repeated sections.

Our repetition-based approach was proven effective in popular music. To improve the performance of the method, however, we will need to use prior information about the spectral characteristics of chorus sections. We also plan to experiment with other music genres and extend the method to be widely applicable.

ACKNOWLEDGMENTS: I thank Hideki Asoh for his valuable discussions.

6. REFERENCES

[1] B. Logan and S. Chu, "Music summarization using key phrases," in *Proc. of ICASSP 2000*, pp. II–749–752, 2000.

[2] M. A. Bartsch and G. H. Wakefield, "To catch a chorus: Using chroma-based representations for audio thumbnailing," in *Proc. of WASPAA'01*, pp. 15–18, 2001.

[3] M. Cooper and J. Foote, "Automatic music summarization via similarity analysis," in *Proc. of ISMIR 2002*, pp. 81–85, 2002.

[4] N. Otsu, "A threshold selection method from gray-level histograms," *IEEE Trans. SMC*, vol. SMC-9, no. 1, pp. 62–66, 1979.

[5] R. N. Shepard, "Circularity in judgments of relative pitch," *J. Acoust. Soc. Am.*, vol. 36, no. 12, pp. 2346–2353, 1964.

[6] G. H. Wakefield, "Mathematical representation of joint time-chroma distributions," *SPIE'99*, pp. 637–645, 1999.

[7] M. Goto *et al.*, "RWC music database: Popular, classical, and jazz music databases," in *Proc. of ISMIR 2002*, pp. 287–288, 2002.

[8] C. J. van Rijsbergen, *Information Retrieval*. Butterworths, second ed., 1979.

[9] G. Peeters *et al.*, "Toward automatic music audio summary generation from signal analysis," in *Proc. of ISMIR 2002*, pp. 94–100, 2002.

[10] R. B. Dannenberg and N. Hu, "Pattern discovery techniques for music audio," in *Proc. of ISMIR 2002*, pp. 63–70, 2002.

[11] K. Hirata and S. Matsuda, "Interactive music summarization based on GTTM," in *Proc. of ISMIR 2002*, pp. 86–93, 2002.

[4] Further information, including video clips, is available at the following URL: http://staff.aist.go.jp/m.goto/ICASSP2003/