Music Signal Spotting Retrieval by a Humming Query

Hiroki Hashiguchi†, Takuichi Nishimura†* , Junko Takita‡, J. Xin Zhang‡, and Ryuichi Oka†

† **Real World Computing Partnership**
1-6-1 Takezono Tsukuba-shi, Ibaraki 305-0032, JAPAN

‡ **Mathematical Systems Inc.**
2-4-3 Shinjuku, Shinjuku-ku, Tokyo, 160-0022, JAPAN

‡ **Mediadrive Corp.**
3-195 Tsukuba, Kumagaya-shi, Saitama, 360-0037, JAPAN

ABSTRACT

This paper presents a method to retrieve similar intervals of feature sequence of music acoustic data by using a query sequence composed of intervals extracted from a humming or singing wave data. The method is called "Model driven path Continuous Dynamic Programming (mp-CDP)" which is a modification of Continuous Dynamic Programming (CDP) commonly used for spotting retrieval. The key feature of mp-CDP is to determine variations of the local path of CDP in advance by a reference sequence pattern which is a model to be matched. The paper confirms the usefulness of the proposed method based on retrieval experiments for 20 popular music tracks in WAV format.

1. INTRODUCTION

Interesting researches on the use of multimedia databases and their retrieval methods have spread throughout the world, thanks to the ready availability of high-performance, low-cost personal computers. Such databases include image, audio and video data in addition to traditional text and numerical data. The retrieval system requires the convention of a human interface, high-speed processing and accurate searching. Oka [5], a co-author of this paper, proposed Continuous Dynamic Programming (CDP) in order to create a spotting retrieval system in the field of voice recognition.

In this paper we describe a melodic matching method for musical acoustic data by a humming query. This method is an extension of CDP and is called "Model driven path Continuous Dynamic Programming (mp-CDP)". Another aim is to develop a retrieval system for music tracks by humming.

Several previous works confirmed that a music retrieval system actually realized. Ghias et al. [1], Kageyama and Takashima [3] and Sonoda et al. [6] used an alphabet of three possible relationships between pitches ('U', 'D', and 'S'), representing the situations where a note is above, below or the same as the previous note. Kosugi et al. [4] developed a music retrieval system by humming based on beats instead of notes and reported that the retrieval time is at most one second for over 10,000 songs. Their goal was that they used MIDI data or human humming as a database, so it is much easier to extract from melodies such data than from musical audio signals.

The key feature of this paper is to determine melody likeness from musical audio signals and to compare the melody likeness in a music database with the pitch sequence of a humming query.

This paper is divided into five sections. In the next section we outline the retrieval system for a humming query. Section 3 deals with mp-CDP which is central to this paper. In Section 4 we experimentally validate mp-CDP based on retrieval experiments for 20 popular music data in WAV format. Finally, Section 5 summarises conclusions and directions of future work.

2. MUSIC SIGNAL RETRIEVAL SYSTEM BY HUMMING

As shown in Figure 1, we organize the retrieval system into the following principal procedures: extracting from humming/music as the feature vector, matching the mutual features, and playing the detected parts from a music database. In Figure 1 the notation x represents a music note and t represents time or a frame of FFT.

* Presently he is working for National Institute for Advanced Industrial Science and Technology (AIST), Cyber Assist Research Center

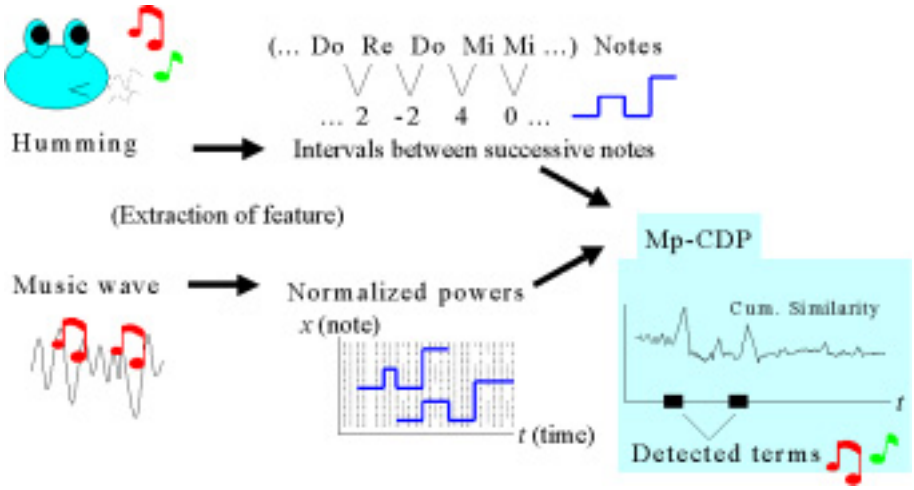


Figure 1: Outline of a music signal retrieval system using a humming query

In order to acquire the pitch sequence extracted from humming, we carry out FFT with a harmonic process. Each pitch transposes the corresponding note. In a similar manner to the above process, we acquire some melody likeness in music based on the powers analyzed by FFT.

The mp-CDP matching method finds successive notes in music that have similar intervals and tempos to the ones of the humming query

3. MODEL DRIVEN PATH CDP (MP-CDP)

3.1 Overview

A new method called "mp-CDP " finds paths having the approximately maximum similarity accumulated with local paths in a 3-dimensional lattice, while the ordinal CDP walks in 2-dimensional lattice. The mp-CDP differ from the ordinal CDP as follows:

- The database as an input is represented by music note and time sequences in a 2-dimensional lattice, and
- A humming query is represented by a scalar vector whose element is an interval of music notes.

Figure 2 shows the conceptual diagram of mp-CDP . The horizontal axis represents notes and the vertical axis represents the frames for query in Figure 2–(a). Being able to move parallel in the (x, τ)-plane permits any transposition of a humming query.

Figure 2–(b) shows the projection of the note axis x onto (t, τ)-plane and permits time elasticity from half to the twice of humming duration in the same manner as ordinary CDP. A small difference in tempo between humming and music is shown to be acceptable in Figure 2–(b). Figure 2–(c) shows the integration of the permission for transposition (a) and the time elasticity (b) in 3-dimensional space. When $\tau = T$, Figure 2–(d) shows the variety of cumulative similarity $S(x, T, t)$. The mp-CDP method detects the path for which $S(x, T, t)$ is maximum.

3.2 Formula

Let f_b[Hz] denote the lowest frequency of melody. We compute the melody likeness for the frequency $2^{x/12} f_b$ [Hz] through the FFT analysis and we write the sequence of the likeness of size W as $I_t(x)$ $(t = 1, \ldots, W,$ $x = 1, \ldots, N,$ $0 \leq I_t(x) \leq 1)$. Here $2^{N/12} f_b$[Hz] represents the highest frequency which compose melody. Here the melody likeness $I_t(x)$ is based on the power spectrum in consideration of harmonic structure of acoustic signals [2].

In a similar manner, we obtain the sequence of the degree of belief for pitch $R_\tau(x)$ $(\tau = 1, \ldots, T,$ $x = 1, \ldots, N,$ $0 \leq R_\tau(x))$ from a humming query. This belief $R_\tau(x)$ is based on the power spectrum in consideration of harmonic structure; see Hashiguchi et al. [2] for details. Therefore, if $R_\tau(x)$ is very small, say $R_\tau(x) < \epsilon$ for a suitable threshold ϵ, the corresponding frame (τ) must be a silent period. We derive the note number on each frame as

$$x_\tau = \arg \max_{1 \leq x \leq N} R_\tau(x)$$

and briefly write its power as

$$R'(\tau) = R_\tau(x_\tau).$$

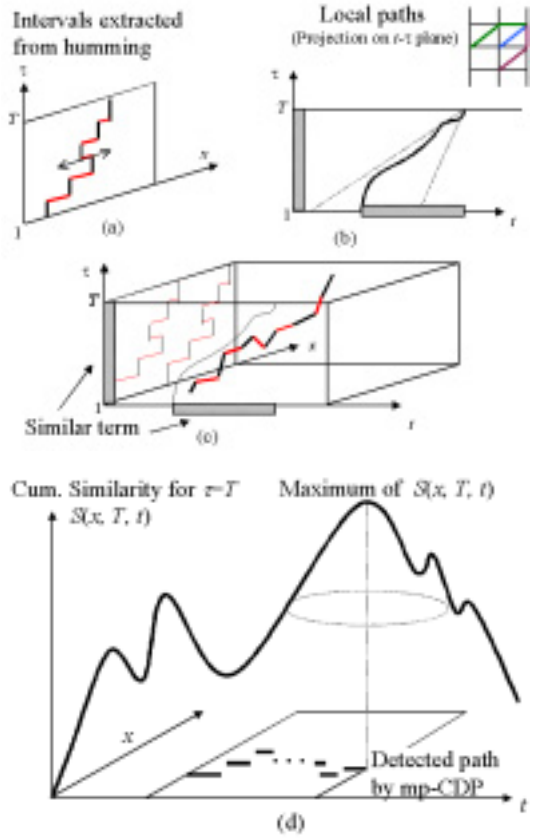


Figure 2: Conceptual diagram of mp-CDP

Without loss of generality, the powers $R'(\tau)$ for the first and the last frames ($\tau = 1$ and T) are assumed to be greater than ϵ.

A sequence of intervals is given by the difference of successive notes. Now consider the frame τ and its nearest frame $\tau' < \tau$ whose power is greater than ϵ,

$$\tau' = \max_{\tau^* < \tau} \{\tau^* \mid R'(\tau^*) > \epsilon\}$$

we define the interval $q(\tau)$ at τ ($\tau = 1, \ldots, T$) as

$$q(\tau) = \begin{cases} x_1, & \text{if } \tau = 1; \\ x_\tau - x_{\tau'}, & \text{if } \tau \geq 2 \text{ and } R'(\tau) > \epsilon; \\ 0, & \text{otherwise} \end{cases}$$

If $R'(\tau) > \epsilon$, the interval $q(\tau)$ means the difference between the current note and the previous one at the nearest frame τ' with $R'(\tau') > \epsilon$. Then we define the local similarity $d(x, \tau, t)$, and the cumulative similarity $S(x, \tau, t)$ as the following equations (1) $\sim$ (3):

$$d(x, \tau, t) = \begin{cases} I_t(x), & \text{if } R'(\tau) > \epsilon \\ 0, & \text{otherwise} \end{cases} \quad (1)$$

$$S(x, 1, t) = 3d(x, 1, t); \quad (2)$$

$$S(x, \tau, t) = \max \begin{cases} S(x - q(\tau), \tau - 1, t - 1) \\ \quad + 2d(x, \tau, t - 1) + d(x, \tau, t); \\ S(x - q(\tau), \tau - 1, t - 1) \\ \quad + 3d(x, \tau, t); \\ S(x - q(\tau) - q(\tau - 1), \tau - 2, t - 1) \\ \quad + 3d(x - q(\tau), \tau - 1, t) + 3d(x, \tau, t); \end{cases} \quad (3)$$

where the boundary conditions for $d(x, \tau, t)$ and $S(x, \tau, t)$ satisfy

$$d(x, \tau, t) = S(x, \tau, t) = 0. \ (x \leq 0 \text{ or } \tau \leq 0 \text{ or } t \leq 0).$$

The value $S(x, \tau, T)$ when the frame τ arrives at T is divided by the constant $3T_v$ so that it can be normalized from 0 to 1 where T_v is the number of non-silences in a humming query. The final result of mp-CDP is a collection of (x, t)'s which attain the approximately maximum value of $S(x, T, t)$, *i.e.,* for a given threshold α ($0 \leq \alpha \leq 1$) the following result is obtained;

$$\left\{ (x, t) \mid \frac{1}{3T_v} S(x, T, t) > \alpha \right. \quad (4)$$
$$\left. \text{and } S(x, T, t) \text{ is a local maximum value} \right\}.$$

4. RETRIEVAL EXPERIMENT

4.1 Music database and humming queries

We prepared 20 WAV files in 16 kHz sampling and monaural recording as the music database to test the mp-CDP matching method. This database includes 10 Japanese pops, 8 children's songs, an animation song and a Japanese *enka*. There are 4 male and 16 female vocal artists in the database. On the other hand, 3 males and 2 females sang a part of each song in the database for about 20 seconds. Hence, the total number of queries of 20 second duration was 300. To investigate the effect of the length of queries, we prepared shortened queries of 5 and 10 seconds from the 20-second originals. We adopted FFT analysis (2048 samples for each frame, 64 msec frame interval) for extracting features from both the music signal database and the humming signal data.

4.2 Search rate for music database and queries

The search rate, which depends on a threshold $\alpha (0 \leq \alpha \leq 1)$, is defined as the average of precision rate N_C/N_D and recall rate N_C/N_T, where N_C, N_D and N_T represent the number of similar terms to the humming query in detected terms, the number of detected terms by mp-CDP and the number of similar terms to the humming, respectively. The detected term by mp-CDP is correct if the following overlapping rate:

$$\text{overlapping rate} = \frac{\text{similar term} \cap \text{detected term}}{\text{similar term} \cup \text{detected term}} \quad (5)$$

Table 1: Average of search rates (%)

Query duration	5s	10s	20s
Average search rate	55.3%	73.9%	89.0%

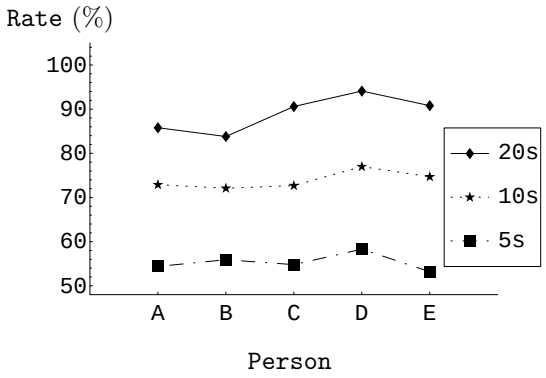


Figure 3: Average of search rate of each person for three different query durations

is greater than 0.5. This means the overlapping terms between similar and detected ones has 70-% intersection when the lengths of both terms are mutually the same. Since the search rate depends on the threshold α, "the search rate for a humming query" is defined as the maximum value running over all α.

4.3 Experimental results

Table 1 shows the average of the search rate for 5 persons× 20 songs = 100 queries for the query duration of 20 seconds. Furthermore, Table 1 also shows the average for the query duration of 5 seconds and 10 seconds.

Figure 3 shows the average of the search rate for 20 songs for each person and Figure 5 shows the average of the search rate for 5 persons for each song. Persons A, B and C are male, and the others are female.

The experimental results are summarized below.

- From the results of Table 1 the longer the duration of a query, the higher the average. If the duration of a query is short, there will be many similar terms to the query in the music database.
- From Figures. 4 and 5, the average for songs 8 and 14 is lower than the others on every persons. Comparing disirable terms with undisirable detected terms in originals by playing, we confirmed that these terms are similar to each other.
- For song 3 (female vocal), the search rate for the query 20s of Male A is 58 %, while that of Female D is 100 %. The sequences of notes for staff, Male A and Female D are shown in Figure 6–(a), (b) and (c). This humming of Female D is probably better than that of Male A. This result shows that the similarity in (3) as well as the search rate reflect the goodness of humming.

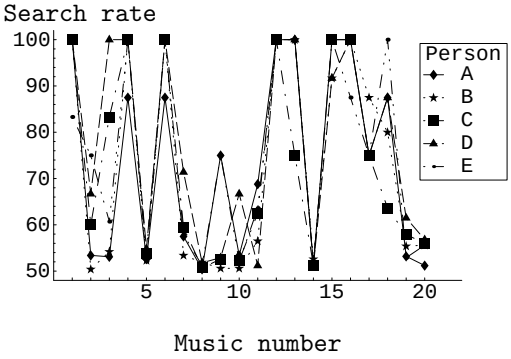


Figure 4: Search rate of each son for five persons (Query duration 10 s).

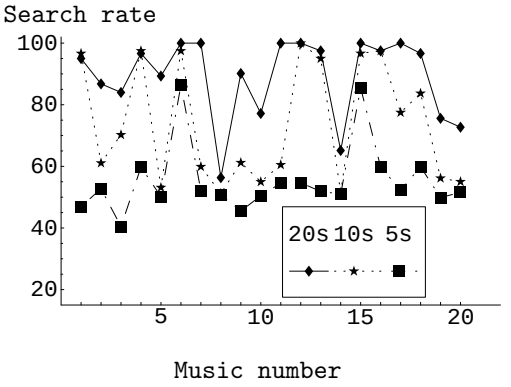


Figure 5: Average of search rate of each song for three different query durations

5. CONCLUSIONS

We have proposed a methodology for matching pieces of music accoding to the similarity of melody likeness and the pitch of humming. The mp-CDP matching method is effective for queries of about 5 to 20 seconds and a music database containing clear, melodic songs.

The computing time for retrieval is directly proportional to the duration of the humming query. The computing time requried, for each 5 second query was, about 1/40 of the database length on the current fastest machine (OS: Windows 2000, CPU: Pentium IV 1.5

GHz). For 10 or 20 queries, the retrival time took about 1/20 or 1/10 of the music database length. Reduction of retrieval time is a feature task. One possible solution for the reduction is to project the 3-dimmensional searching space to 2-dimensional space and to compress the similarity terms in the database.

ACKNOWLEDGMENT

The authors are very grateful to Dr. Junichi Shimada, a managing director of RWCP, who has not only given them constant advice on this research, but has also provided a superb research environment. We wish to thank Mr. Hironubu Takahashi, Mr. Yasuhiro Mori of RWCP, Mrs. Shinobu Ogi, Mr. Michinao Mizuno of Mathematical System Inc. and Mrs. Megumi Nishimura for their assistance in collecting humming query data.

REFERENCES

[1] A. Ghias, J. Logan, D. Chamberlin and B. C. Smith, Query by humming – Musical Information retrieval in an audio database. In *ACM Multimedia 95–Electronic Proceedings*, 1995.

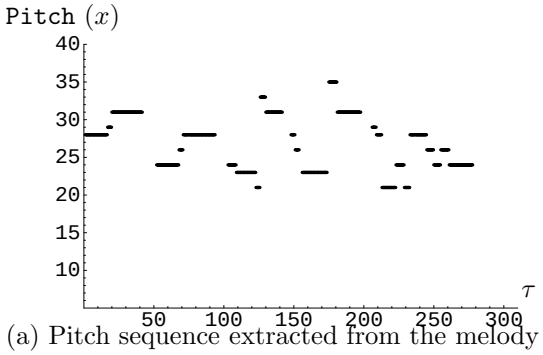
[2] H. Hashiguchi, T. Nishimura, J. X. Zhang, H. Takahashi and R. Oka. Model driven path CDP for music retrieval with humming query (in Japanese). *Technical Report of IEICE*, PRMU-2000-66 (2000-09), pp. 35–40, 2000.

[3] T. Kageyama and Y. Takashima. A melody retrieval method with hummed melody (in Japanese). *Transactions of the Institute of Electronics, Information and Communication Engineers* (D-II), Vol. J77-D-II, No. 8, pp. 1543–1551, 1994.

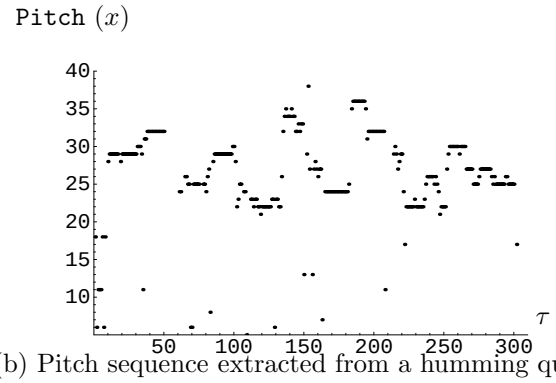
[4] N. Kosugi, Y. Nishihara, T. Sakata, M. Yamamuro and K. Kushima, "A practical Query-by-Humming system for a large music database", *ACM Multimedia 2000*, pp 333–342, 2000.

[5] R. Oka. Continuous word recognition with Continuous DP (in Japanese), Report of the Acoustic Society of Japan, S78-20, pp. 145–152, 1978.

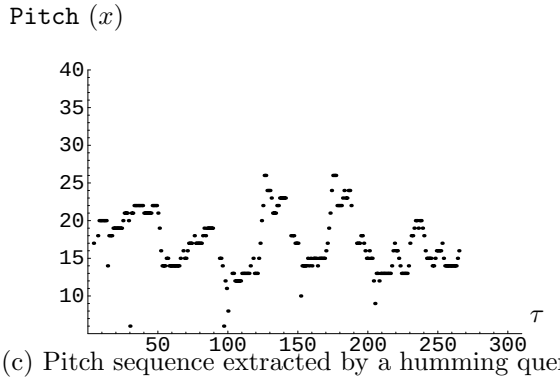
[6] T. Sonoda, M. Goto and Y. Muraoka. A WWW-based melody retrieval system (in Japanese). *Transactions of the Institute of Electronics, Information and Communication Engineers* (D-II), Vol. J82-D-II, No. 4, pp. 721–731, 1999.



(a) Pitch sequence extracted from the melody of music 3



(b) Pitch sequence extracted from a humming query indicating a high search rate.



(c) Pitch sequence extracted by a humming query indicating a low search rate.

Figure 6: Comparison of a music note sequences and extracted pitch sequences indicating high/low search rate.