Query by Humming Based on the Hierarchical Matching Algorithm

Menglu Li
College of Information Engineering
Communication University of China, CUC
Beijing, China

Zhijun Zhao
College of Information Engineering
Communication University of China, CUC
Beijing, China

Ping Shi
College of Information Engineering
Communication University of China, CUC
Beijing, China

***Abstract*—Along with the continuous development of science and technology, the multimedia retrieval is not only confined to the text retrieval. Content-based music retrieval technology has been widely used in modern society. The query by humming method based on the matching of user's humming fragments and music segments is a research hotspot in recent years. The establishment and the matching based on a large database is the research focuses of query by humming. This paper proposed a query by humming method based on the hierarchical matching algorithm. The hierarchical matching between humming fragments and melody features database by music phrases effectively improved the accuracy of query by humming based on a large database.**

Keywords-Query by humming; large database; hierarchical matching;

I. INTRODUCTION

With the rapid development of science and technology, the amount of multimedia information is growing exponentially. In addition to the visual media, the most important media of multimedia is the voice media, which accounting for 20 percent of the total amount of media information [1]. The text retrieval is a common music retrieval method, such as Baidu, Yahoo and Google. Those search engines are all using the retrieval method based on the text. Text retrieval has some inherent defects. For example, the text cannot express some content features of music, such as tone, rhythm and melody features. For the users, without the understanding of the text characteristic of music, it is hard to adopt the method of text retrieval to get correct results. At the same time, with the rapid growing of the amount of music data, text annotation is becoming time-consuming and laborious. In order to make up for the inadequacy of text retrieval, content-based music retrieval technology rises in response to the proper time and conditions. Content-based music retrieval technology is the retrieval method using content features such as melody and rhythm [2]. QBH (Query by Humming) is a rapid and convenient retrieval method among content-based music retrieval methods. In the 1995 ACM multimedia conference, Ghias et al at the University of Southampton introduced the research achievements of QBH and developed a QBH system. On the basis of Ghias et al, McNab and others researched about the extraction of music rhythm information, improving the success rate of retrieval system. Many domestic research institutes and colleges have done some research in QBH, such as Tsinghua University, Northwestern University, South China University of Technology, etc.

In this paper, the melody features of MIDI music were extracted to build a large MIDI melody feature database, and a hierarchical matching algorithm was proposed based on the large database. Firstly, MIDI files were automatically divided into phrases. Then the characteristic parameters, such as the mean pitch, the variance of the pitch and the phrase length, were extracted by phrases. According to these parameters and recorded humming fragments, MIDI files were executed preliminary coarse screening. Finally humming fragments and MIDI melody feature files were calculated the similarity by the hierarchical matching algorithm which combines EMD (Earth Mover's Distance) and DTW (Dynamic Time Wrapping), and according to the similarity, the results were returned in turn. The experimental results showed that the hierarchical matching algorithm effectively improved the retrieval efficiency under ensuring the condition of the accuracy.

II. THE HIERARCHICAL MATCHING ALGORITHM BASED ON THE LARGE MIDI DATABASE

In this paper, MIDI melody features were extracted and MIDI files were automatically divided into phrases to build large MIDI music melody characteristic library. Recorded WAV format humming fragments were also executed the melody feature extraction and the automatic phrase segmentation with the same algorithm as the MIDI phrase segmentation. WAV and MIDI files are hierarchically matched by phrases, so as to improve the retrieval efficiency of the large database.

A. The Establishment of the Large MIDI Melody Feature Database

MIDI is a kind of music standard format wildly used in the arrangement field. Music is recorded by digital control signal of notes. A complete MIDI file contains 16 music channels, and the channel which could express music melody information more completely is defined as main channel. What transmitted in MIDI files are time, position, intensity, duration, vibrato and other digital information of notes, which are uniformly defined as MIDI message. The rhythm information of music can be obtained via iterating through MIDI files and reading all kinds of information. Therefore in most cases QBH systems adopt MIDI format to establish music melody feature database and match the humming

978-1-4673-8126-0 /15/$31.00 ©2015 IEEE

fragments. The establishment of a large MIDI melody feature database consists three parts: the extraction of the MIDI main melody channel, the extraction of melody feature vectors and the automatic segmentation of MIDI phrases.

1) The extraction of the MIDI main melody channel: In the study of content-based digital music retrieval methods, complex sound music files are commonly used in the database. However, a retrieval system with the complex sound music data index needs a huge computation and the complex matching algorithm. MIDI files contain 16 logical channels. When a MIDI file is played, 16 channels will be played at the same time. Melody characteristics are stored in the main melody channel; therefore the main melody channel needs to be detected.

This paper adopts the method proposed by Zhao Fang, Wu Yadong, which executed the MIDI main channel extraction based on the channel characteristic parameters [3]. MIDI characteristic quantities of 16 channels are extracted, namely the channel name (F1), the channel number (F2), the balance of the sound track (X1), the average strength (X2), the master volume (X3), the articulatory time (X4), and the articulatory area (X5). The channels which are not the MIDI main channel are eliminated according F1 and F2. For example, F1 of the main channel usually adopts MELODIES, VOCAL, SING, SOLO, LEAD and VOICE as keywords, and the accompany channels often use ACCU, DRUM, BASS, PERCUSSION, COMPANION and BACK as the channel name. The No. 10 channel is generally retained for percussion instrument, so we can define No.10 channel as the accompany channel.

The characteristic parameters of each channel are comprehensively used to describe the possibility of the main melody channel. We define the score function Y (k) of the kth channel as follows:

$$\mathrm{Y(k)} = \sum_{m=1}^{5} \alpha m * \mathrm{Xm(k)} \qquad k = 1,2 \ldots \ldots 16 \qquad (1)$$

In formula (1), αm is the weight of Xm(k), Xm(k) is the channel characteristic parameters. According to the possibility calculated by (1), we can detect several possible main melody channels. We keep the top 3 channels to execute the extraction of melody feature vectors and automatic segmentation of MIDI phrases.

2) The extraction of melody feature vectors: In this paper pitch and duration values were adopted to constitute a two-dimensional sequence as melody feature sequence to complete the matching. Melody feature sequence can be formulated as follows, where V is the feature point sequence of the whole notes in music files, n is the total number of notes. For a single feature point v, Pitch is the value of the pitch, and Time is the description of note duration. Melody feature sequence can accurately express the melody information of a MIDI file.

$$\mathrm{V} = \{\mathrm{v1}, \mathrm{v2}, \mathrm{v3} \ldots \mathrm{vn}\}$$
$$\mathrm{v} = < Pitch, Time > \qquad (2)$$

Because of the randomness of humming fragments, the reference tone of user's humming is not always matching the actual reference tone in music. Individual tone may be inaccurate, and humming speed also varies with each individual. So in this paper, absolute sequence is converted to relative sequence to complete the matching. The relative pitch sequence is concluded via the following note's pitch minus the previous note's pitch. The relative duration sequence was concluded via dividing the following note's duration by the previous note's duration. Relative two-dimensional melody feature sequences are deposited in the database to prepare for the matching.

3) The automatic segmentation of MIDI phrases: The automatic segmentation of music phrase is an indispensable link for multiple music phrases retrieval, which can effectively improve the efficiency and accuracy of retrieval. It was proposed that according to the distribution of note's duration, the appropriate threshold could be determined to estimate if the note is the end of phrase or not [4]. The mass music feature database can be established automatically. Experiment results showed that, like MIDI note's duration, time intervals between notes also have a very significant regularity. In this paper a method based on the note's duration and time intervals was presented for the automatic phrase segmentation.

According to the rule that it can be commonly seen that the notes at the end of music phrases have long durations and long intervals, in this paper a method of multi-step phrase segmentation was presented, which comprehensively considered note's duration and interval.

Firstly initial segmentation of phrase is carried out according to note's duration, as shown in (3), where Ta is the duration value, N is the total number of notes in a music file, k is the coefficient.

$$\mathrm{C} = \frac{k}{N} * \sum_{i=1}^{N} Ta(i) \qquad (3)$$

After setting the appropriate coefficient k, we can get the phrase segmentation threshold C. The value of coefficient k plays a decisive role for the phrase segmentation. A small k value can cause that a complete phrase is wrongly separated into multiple phrases, and a large k value may result in that two consecutive phrase failed to be disconnected [4]. In this paper 100 MIDI files were experimented，and the results showed that the k value of 1.5 could achieve the first step segmentation accurately with a little influence on the second step segmentation.

After the first step of phrase segmentation, notes are preliminarily classified into two groups: notes at the end of phrases or not. Each note is added a variable Isinternal(i) with the value of true or false to estimate whether it is at the end of phrases. The time intervals are extracted, which express the time lag between a note and the following one, and the average intervals of each note group are calculated, as shown in (4) and (5).

$$\mathrm{avei} = \frac{1}{N1} * \sum_{i=1}^{N1} Ti(i) \qquad (4)$$

$$\mathrm{aven} = \frac{1}{N2} * \sum_{i=1}^{N2} Ti(i) \qquad (5)$$

In (4) and (5), avei is the average of ending notes and aven is the average of not-ending notes; N1, N2 are the number of notes; Ti is the time interval.

By traversing time interval value of notes and comparing the two averages, we can finish the second step music phrase

segmentation. After got the two averages, we orderly compare the time intervals and the two averages, as shown in (6).

$$\text{Isinternal(i)} = \begin{cases} true & (Ti(i) > avei) \\ false & (Ti(i) < aven) \end{cases} \quad (6)$$

The notes' Isinternal(i) have been preliminary assigned after the first step segmentation. The not-ending notes are compared with the average avei. If Ti(i) is greater than avei, the note will be classified into the ending group. While the ending notes are compared with aven, if the Ti(i) is lesser than aven, the note will be classified into the not-ending group.

The misjudged notes in the first step will be amended after the second step segmentation. The two-dimensional vectors of music melody features are placed into database, which will contribute to the similarity calculation and matching.

B. Melody Feature Extraction of WAV Format Fragments

For recorded WAV format humming fragments, the melody features are extracted which are composed of note's pitches and durations, and the absolute melody feature sequence needs to be transformed into the relative sequence (the pitch interval and the duration ratio).The melody feature extraction process of WAV format consists following steps: framing and preprocessing, the detection of mute frames, filtering, fundamental frequency extraction, the representation of melody characteristics and phrase segmentation.

1) Framing and Preprocessing: In this paper, humming fragments are WAV format waveform files. Firstly the audio was processed by framing and preprocessing. In general, frame shift is 1/2 of the frame length. The window length is not less than two pitch period, and the detection result will be more accurate with the longer window length, with more computational burden. In general, the window length is 20ms-40ms.

2) The Detection of Mute Frames: Short-time energy describes the signal strength of humming fragments, which can be applied to detect the mute frames. While the short-time energy of a frame in humming signals is lower than the predetermined threshold, the short time frame is the mute frame. For signal x (m), the short-time energy is defined as follows, where w (m) is the window function.

$$E_n = \sum_{m=-\infty}^{+\infty} [x(m) \cdot w(n-m)]^2 \quad (7)$$

3) The Fundamental Frequency Extraction: For the WAV format music, the information of notes and melody can be obtained by detecting the fundamental frequency. Note's relative pitch and duration value is placed into a two-dimensional vector and relative accurate matching results can be got.

The algorithms of fundamental frequency detection include the autocorrelation function, the average magnitude difference function, cepstrum analysis, etc. In this paper, autocorrelation function was used to extract the fundamental frequency. Autocorrelation function is defined as:

$$R(v) = \sum_{n=-\infty}^{n=\infty} x(n)x(n+v) \quad (8)$$

Each frame of the music signal is transformed by (8). Because each frame is taken tens or hundreds of milliseconds to be analyzed, and there has not been much change of the fundamental frequency information of each frame. Therefore, extract the autocorrelation period of each frame and carry on the smoothing processing, and we can get the fundamental frequency information.

4) The Representation of Melody Characteristics: After extracting the fundamental frequency information in each frame, fundamental frequency information needs to be transformed into note information according to the comparison table of frequency and notes.

According to the principle of proximity, the fundamental frequency was mapped into note numbers of MIDI format, which were considered as the absolute pitch sequences, and the durations of relatively same frequency were calculated, which were recognized as the absolute duration sequences. Absolute melody feature sequences were converted to relative melody feature sequences. Relative pitch difference and relative duration ratio were saved into two-dimensional sequences to prepare for the matching.

5) The Phrase Segmentation: For WAV format humming fragments, the music is segmented into phrases using the same method as used in MIDI format. According to the note's duration and interval, humming fragments can be divided into phrases by the multistep segmentation. Music files are converted to two-dimensional relative sequences composed of the pitch difference and the duration ratio, and the end of phrases are labeled. Then similarity calculation and the matching can be carried out with the melody feature database.

C. The Hierarchical Matching Algorithm

Music retrieval technology based on humming has some means to dispose the humming information, the main purpose of which is to extract the music melody characteristics. Through the similarity matching process between the inquiry sequence and the feature sequences in database, the system can get the retrieval target [5]. This paper adopted the mean value and the variance of pitch and the length of phrases to execute preliminarily coarse screening, and carry out the similarity calculation of feature vectors combining EMD (Earth Mover's Distance) and DTW (Dynamic Time Warping). Finally the system sorts on the basis of similarity, and returns the most similar music fragments.

1) The Preliminarily Coarse Screening: In this paper, the mean and variance of pitch are calculated in phrases, and compared with the humming fragments to complete the first step coarse screening. Then we complete the second step coarse screening through the phrase length. On the basis of similarity, the system will return the most similar music fragments.

Firstly in the MIDI melody features database, the mean and variance of pitch is calculated in phrases, and the highest and lowest mean value and the corresponding phrase index are stored in the melody feature database. We respectively define the highest and lowest mean values are Epmax, Epmin, and the variances of corresponding two phrases are Dpmax, Dpmin.

Recorded humming fragments are calculated the mean value and the variance of pitch, which are recorded as Eh and Dh. Eh and Dh are compared in turn with the Epmax and Epmin. If the statistical magnitudes in a MIDI file meet the conditions of (9), the MIDI file will be excluded from similar files, and the system will continue to determine the next MIDI file. If the MIDI file does not meet the conditions, the system will execute the second step coarse screening for this file. Experimental results show that when coefficient k1 takes 1.5 and coefficient k2 takes 0.8, the system has a good screening effect.

$$\begin{cases} Eh > Epmax \\ Dh > k1 * Dpmax \ \vee Dh < k2 * Dpmax \end{cases}$$
$$\begin{cases} Eh < Epmin \\ Dh > k1 * Dpmin \ \vee Dh < k2 * Dpmin \end{cases} \quad (9)$$

For MIDI files incongruent with (9), system will execute the second step of coarse screening. The phrase length of humming fragments is compared to the phrase length in MIDI file. If the difference values of continuous several phrases are within a certain threshold value (three notes in this paper), the system will carry out the two step similarity calculation.

The similarity calculation algorithm in this paper combines the EMD and DTW.

2) The Rough Matching Based on EMD Algorithm: EMD (Earth Mover's Distance) is a kind of distance measurement, which is used to measure the distance of two distributions. We define P is m factories from P1 to Pm, and there are goods that weights Wi at the factory Pi. Q is n storages from Q1 to Qn, and the storage capacity of Qi is Wj. In order to carry all of the goods from P to Q efficiently, we define dij is the distance from Pi to Qj, and the fij is the weight of carried goods, so the workload of a transport is dij*fij. The workload will be more with the higher value of dij or fij. We can get the minimum total workload as follows:

$$W = min\left(\sum_{i=1}^{m}\sum_{j=1}^{n} d_{ij}f_{ij}\right) \quad (10)$$

The earth mover's dance is defined as (11), where the constraint condition of the most optimal solution fij is shown in (11) [6]. We can get a most optimal solution to obtain the minimum of EMD.

$$EMD(P,Q) = \frac{\sum_{i=1}^{m}\sum_{j=1}^{n} d_{ij}f_{ij}^{*}}{\sum_{i=1}^{m}\sum_{j=1}^{n} f_{ij}^{*}}$$
$$\begin{cases} f_{ij} \geq 0 \\ \sum_i f_{ij} = W_j \\ \sum_j f_{ij} = W_i \end{cases} \quad (11)$$

In the use of the EMD algorithm, the system recognizes the melody features of humming fragments as factories, and the MIDI feature database as the storage. Dij is the distance of any two elements of the relative pitch and duration two-dimensional melody feature vector (using the Euclidean distance formula to calculate the distance), and fij is the most optimal solution. The EMD distance calculated by (11) is the distance of two melodies. This paper uses EMD to calculate the distance and exclude the MIDI files with low matching degree.

3) The Fine Matching Based on DTW Algorithm: DTW is an algorithm translating the global optimization problem to the local optimum problem, which is a nonlinear warping technique combining time alignment with calculation of distance measures. Automatic searching for a path using local optimization to acquire the minimum distortion between two feature vectors can avoid the error introduced by different length effectively. For pitch sequences matching by the sentence, DTW algorithm is relatively simple and DTW can solve the problem of the inconsistency in the note amount of humming fragments and that of the actual music. So in this paper DTW algorithm was adopted to make similarity calculation.

For QBH, the melody feature vector of the humming fragment is considered axis X, while the melody feature vector of the music in the database is axis Y. The shortest distance of two vectors was calculated by DTW. The formula is expressed by equation (7), in which i is the index of the axis X, and j is the index of axis Y, D (i, j) is the Euclidean distance between two vectors, and dist (i, j) is the cumulative distance. According to the equation (12) we can choose a best path from the lower-left corner to the upper-right corner, and finally conclude the minimum cumulative distance value.

$$D(i,j) = \min\begin{cases} D(i-1,j-2) \\ D(i-1,j-1) + dist(i,j) \\ D(i-2,j-1) \end{cases} \quad (12)$$

In order to guarantee the accuracy and improve the retrieval efficiency, this paper adopted a double matching model. Firstly EMD algorithm is recognized as the fast matching algorithm based on notes to finish the rough matching. The MIDI files which have low matching degree are excluded. Then the rest of MIDI files in database are executed the fine matching by DTW algorithm. Finally the music files similar to the humming fragment are returned to users.

III. The Experimental Results

In this paper, a database of 200 MIDI files was established, and 10 WAV format humming fragments were recorded to match with the database. Firstly files were divided into phrases, and through the two step coarse screening, MIDI files inconformity with the conditions were excluded. Then the similarity was calculated by the algorithm combined EMD and DTW. Finally the results with high similarity were returned to users. Experiments showed that this method can greatly improve the accuracy and efficiency of a large database.

A. Phrase Segmentation Experiment of MIDI Files

. Because the hierarchical matching retrieval of the large database is based on the phrases, phrase segmentation is the

precondition of accuracy and efficiency of the QBH system. This paper used 40 MIDI files to count the accuracy of phrase segmentation (the number of phrases correctly identified divided by the total number of phrases). The phrase segmentation method proposed in this paper was compared to the existing method proposed by Li Juan et al, which segmented the phrases by distribution of note's duration. The automatic segmentation results of two methods were shown in Table I.

Experimental results showed that by the method of this paper the accuracy increased by 11.6%. The method proposed in this paper effectively improved the accuracy of automatic segmentation, and provided a new thinking of effective similarity calculation and the matching.

TABLE I. THE COMPARISON OF THE ACCURACY OF SEGMENTATION BY TWO METHODS.

	The method in this paper	The existing method
The highest value	99.9%	99.2%
The lowest value	80.0%	62.3%
The average value	91.1%	79.5%

B. The Hierarchical Matching Experiment

In this paper, a melody feature database of 200 MIDI files was established. 15 single phrase humming fragments and 15 multiple phrases humming fragments were recorded to calculate the similarity with the melody feature database by hierarchical matching algorithm. The algorithm in this paper was compared with the existing methods using EMD and DTW without the hierarchical matching method. The comparison of results was shown in Table II.

TABLE II. THE COMPARISON OF THE ACCURACY OF HIERARCHICAL MATCHING BY THE METHOD IN THIS PAPER AND EXISTING METHODS.

	The method in this paper	The existing method with EMD and DTW	The existing method with DTW
The top five retrieval accuracy (single phrase)	53.3%	46.7%	40.0%
The top ten retrieval accuracy (single phrase)	66.7%	66.7%	53.3%
The top five retrieval accuracy (multiple phrases)	73.3%	66.7%	60.0%
The top ten retrieval accuracy (multiple phrases)	80.0%	73.3%	66.7%

By the experiment results, the hierarchical matching algorithm has obviously improved retrieval accuracy. The retrieval accuracy was improved obviously by new method. For the single phrase, the top five retrieval accuracy was improved by 6.7% (compared to the method with EMD and DTW); the top five and top ten retrieval accuracy were improved by 13.3% (compared to the method with DTW). For multiple phrases, the top five and top ten retrieval accuracy were improved by 6.7% (compared to the method with EMD and DTW); the top five and top ten retrieval accuracy were improved by 13.3% (compared to the method with DTW). Because multiple phrases humming fragments contain more melody characteristic information, in the case of phrase segmentation relatively accurate, the retrieval results are more accurate for multiple phrases than the single phrase. This paper provides a new way of thinking to improve the performance of QBH.

IV. THE CONCLUSION

This paper proposed the hierarchical matching retrieval method based on the large database of QBH. Firstly the melody feature database was segmented into phrases by note's duration and interval. Then the recorded WAV format humming fragments were pretreated and extracted the fundamental frequency and segmented into phrases. Finally the melody features of humming fragments and the database were hierarchical matched. The coarse screening is executed by the mean value and the variance of pitch, and the two step similarity calculation was through the algorithm combining EMD and DTW. The experimental results showed that the method had great improvement for large database retrieval accuracy.

Content-based retrieval as an important aspect of the information retrieval technology research has received more and more attention. But because of the limitation of various technical conditions, audio retrieval technology is relatively backward and not known to the public [7]. At present, audio retrieval technology still exist many unsolved problems, but the overall trend is progressive. Due to the extremely high convenience, there will be great development prospect and public recognition of content-based audio retrieval system.

REFERENCES

[1] X.-F. Wang, The key techniques of content-based music retrieval, The doctoral dissertation of Northwestern Polytechnical University, 2011.

[2] H.-P. Guo, The algorithm research and system implementation of content-based music retrieval, The master degree theses of Northeast Normal University, 2007.

[3] F. Zhao, Y.-D. Wu, J.-K. Su, "The MIDI main melody channel extraction method based on the characteristic parameters," Computer engineering, 2007.

[4] J. Li, M.-Q. Zhou, and P. Li, "The construction of music database by the MIDI features extraction," Computer engineering and application, 2011.

[5] M. Rocamora, P. Cancela, and A. Pardo, "Query by humming: Automatically building the database from music recordings," Pattern Recognition Letters, vol. 36, 2013, pp. 272-280.

[6] H.-B. Ling and K. Okada, "An Efficient Earth Mover's Distance Algorithm for Robust Histogram Comparison," IEEE Transactions On Pattern Analysis And Machine Intelligence, vol. 29, no. 5, 2007, pp. 840-853.

[7] W. Wang, L.-Z. Xu, and Y.-S. Dong, "A Construction Method of Music Features Libraries in QBH," Proc. 2011 7th International Conference on Digital Content, Multimedia Technology and its Applications (IDCTA), 2011.