A method of Query-by-Humming System for Polyphonic Audio

Junzhi Du

1-585-622-8036
dj910906@gmail.com

ABSTRACT

Content-based multimedia retrieval focuses on the intrinsic characteristics of the target. Query-by-Humming is one of its applications that make users retrieve their objects based on the intrinsic characteristics of music-melody. Most related works are aimed to retrieve music from symbolic music data by humming query. However, the applications of retrieving music in the form of polyphonic raw radio would be more useful and necessary ninth real world since this format is more common in the music consumer market. The focus of this project is to implement a QBH system that extracts melody features and represents humming query by the probability characteristic of note occurrence. Then, DP matching method is adopted to measure the similarity between humming and music data. The paper presents developed algorithms in melody feature extraction and DP matching, along with experimental results of the system.

1. INTRODUCTION

In the field of music information retrieval, one problem that has attracted particular interest is to find similar objects form a dataset to a query object. The domain that this project focuses on is the application of Query-by-Humming system, whose goal is to search a music database for the K most similar songs to a humming query in melody. It happens a lot when we remember that we heard a song before with a segment of melody from the song in our heads but cannot recall any identity information of it, like name, artist, album and released year. One straightforward solution is to hum the melody segment of song that you could remember and make this humming segment as a query and search in a large music database to find the object song we want. With a QBH system, it is possible to find the object song from a list of the K most similar songs.

Currently, there have been several technologies to accomplish this function with promising results in terms of retrieval accuracy and speed. However, most of these systems solve the problem on the level of symbolic format(i.e. MIDI files)[1][2]. Most of these technologies contain an representation of humming and music database melodies with single note at the same time with pitch and duration of one note and a DTW or other DP matching methods to find the k most similar songs. Many works haven shown remarkable performance.

However, symbolic music format is not the most common format for music storing in actual consumer market. Nowadays most music songs are decoded in polyphonic audio files, like MPEG. One way to solve this problem is to convert the polyphonic music into MIDI files, as symbolic format. There are two possible methods: one is converting song manually, which requires musicians' hard work and would consume plenty of time to extract the voice or other object information from a raw audio; the other one is to use converting techniques to solve the problem[3][4][5], which comes with a lack of accuracy in the situation of large sound mixture.

In this project, another method that tries to avoid the problem is implemented. This algorithm is trying to solve this problem by extracting melody information on mid-level instead of signal level. In mid-level melody information is represented by sequences of vectors whose elements describe the estimated strength of notes in polyphonic audio signal, expressing the probability of each note actually being contained in a fragment of a song as a melody element. This approach focuses on the method to extract and represent melody information in order to enable the matching of human query and the music data accurately, from polyphonic audio, and how to measure the (dis)similarity between melody representations of humming and music data.

This projects presents a algorithm of extracting and representing the melody information and an approach of matching features from music database and humming query in the mid-level. Experimental results are also shown in the paper.

2. SYSTEM OVERVIEW

The QBH system contains three main modules, which are music database feature extraction module, humming feature extraction module and (dis)similarity measuring module as shown in Figure 1. The music database feature extraction module contains extracted sequences of vectors whose elements indicate the probability or strength of occurring of notes in a polyphonic audio frame or segments in music database. The humming feature extraction module extracts melody feature, also represented by sequences of vectors, from the humming query acquired via humming samples stored in a database or sung by microphone. The note segmentation process groups the adjacent audio frames with seemingly identical notes. The similarity measuring module provides a matching

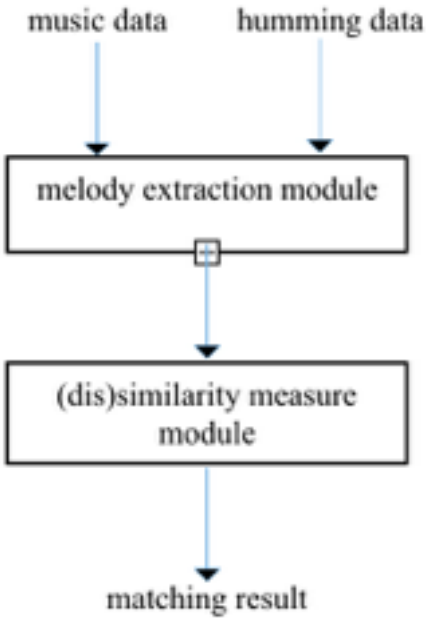

Figure 1. Overview of QBH System

algorithm which outputs the matching results in order of similarity for the music database and humming feature. In this module, when comparing the two representation of feature, the differences of the overall length between music database and humming features and the local variance influenced by voice are considered.

According to [6], other modules such as music filtering

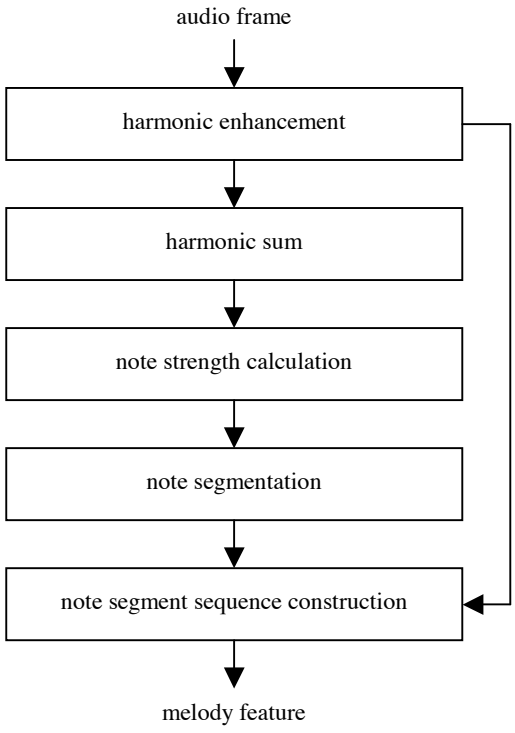

Figure 2. Melody feature extraction process

module and melody part detection module can be added to complete the system and a better performance in accuracy and speed.

3. MELODY FEATURE EXTRACTION

In mid-level, the extracted melody feature of both polyphonic audio and humming are represented as a set of the note strengths of audio frames or segment, instead of a definite music note. These note are supposed to be chosen as the most audible notes among all the simultaneous notes that generated from multiple sound sources. When adjacent audio frames are considered to hold the identical notes, they would be combined together as a segment. As shown in Figure 2, the process is consist of five steps.

3.1. Harmonic Enhancement

Polyphonic music is usually composed of various sound sources, such as vocal, piano, guitar, percussion, etc. Most tasks of polyphonic music analysis contain a step of source separation and note recognition, which are both believed to be hard-to-solve problems. However, QBH system application could try to avoid these problems. This thought is based on the fact that only those parts of music that are easy for human to be recognized and memorized can be hummed. The first step of melody extraction is aimed to extract those predominant pitches from all the notes generated by different sources. Musical sounds are constructed by harmonics whose positions in frequency and amplitudes are supposed to be specific. For most audible pitches, their harmonic structure should mask all other partials of harmonics. This makes the other harmonics much less audible. Consequently, when a signal source generates the most large amplitude signal tends to mask other signals nearby in frequency domain. Harmonic enhancement process extracts the harmonics that outstandingly mask other surrounding signals and make themselves peaks compared with there surroundings. It can be described as equation (1).

$$E_t^{EP}(k) = \sum_{i=-W}^{W} A\left(E_t(k) - E_t(k+i)\right), 0 \le k < N \quad \text{- (1)}$$

where $A(x) = x, \forall x \ge 0$ and $A(x) = 0, \forall x < 0$

In equation(1), N is the FFT index range, $E_t^{EP}(k)$ represents the degree of the predominance of the harmonic in the frequency in the frequency index k, considering the spectral amplitudes $E_t(k)$ of surrounding signals within the frequency range W in time t. $E_t^{EP}(k)$ gets large when the spectral amplitude of harmonic in the index k reaches a peak and no adjacent peaks within the given window W. This process eliminates the small peaks and white noise and emphasizes prominent peaks.

3.2. Harmonic Sum

The most significant ingredient factor in music recognition is the harmonicity of music. Harmonic sum could be applied as the second step of melody extraction In this project harmonic sum is utilized to implement pitch extraction. The process of extraction pitch information is shown in equation (2).

$$F_t(p) = \frac{1}{\lfloor N/p \rfloor} \sum_{m=1}^{\lfloor N/p \rfloor} E_t^{EP}(mp) \quad \text{- (2)}$$

In equation(2) , $\lfloor X \rfloor$ is an integer which is not bigger than x and $F_t(p)$ is the average strength of harmonics having the fundamental frequency p in the audio frame at time t. $F_t(p)$ is determined by averaging degree of prominence value obtained by step 1 in equidistance frequency indexes. If predominant harmonics of fundamental frequency p exists, the value $F_t(p)$ becomes large, which represents the possibility of the occurrence of sound of fundamental frequency p is large.

3.3. Note Strength Calculation

The note of a human singing varies commonly without noticing the slight variations of singing. Also, in order to speed up the matching, the dimension of the frequency index generated by STFT needs to be minimized in order to accelerate the matching process by reducing the feature dimension. Hence , it is necessary that frequencies are supposed to be quantized into the frequency bands according to the frequencies of musical notes represented in 12-note scale. The fundamental strength in frequency index is converted to one in musical note index as shown in equation (3) in the third step.

$$NS_t(m) = \frac{\int_{L_m}^{U_m} F_t(p)\, dp}{\left| U_m - L_m \right|}, \quad 0 \le m \le M-1 \text{ -(3)}$$

In equation (3), m is a musical note index and Um and Lm are frequencies of the upper and lower limit for a musical note indexed as m, and M is the total number of musical notes in the musical scale. In our method 512 frequency index is converted to 108 note index (12 notes per octave, 9 octaves).

3.4. Page Numbering, Headers and Footers

In the spectrogram, time frames with similar pitch characteristics are considered as one segment and grouped together in this step. It is based on the learning that human voices occurs with a transient phase before reaching a specific note. This period, which is called onset period here, is generated with very little energy. We could use the enhanced harmonic data in section 3.1 to complete the note segmentation step. When the enhanced harmonics shows a value of local minimum,, the point is considered as a segment boundary point, as defined in equation(4).

$$SB = \{t \mid \min_t(FE(t)), \min(FE(t)) < TH\}$$
$$FE(t) = \frac{1}{N} \sum_{k=0}^{N-1} EP_t^2(k) \quad \text{-(4)}$$

$FE(t)$ is the energy of enhanced harmonics at time frame t and TH is the threshold value which is applied here to avoid selecting too many local minima. Additionally, when the frame energy of the enhanced harmonics remain little for a specific duration, that group of frames is classified as silent segment. A segment between two silent segments is merged with the silent segments constructing a single silent segment, if it is shorter than a certain threshold.

3.5. Construction of Note Segment Sequence

With the audio segment generated in section 3.4, we could construct pitch information of segment by averaging the note strength vector of each segment as shown in equation (5).

In equation (5), Sl is the note strength vector of a segment, C is the number of audio frames included in the segment, and ls and le are the start audio frame and the end audio frame of the note segment respectively, which are obtained by equation (5).

$$S_l(m) = \frac{1}{C} \sum_{t=l_s}^{l_e} NS_t(m), \quad 0 \le m \le M-1 \quad \text{-(5)}$$

Finally, several peaks with large note strength are selected as note candidates.

4. MATCHING

The humming query is very possible to have different lengths with music and the variations of notes from human voice could be hard to be recognized. Also, erroneous note and segment information is also mixed in the pitch information during music extraction.

These problems could be solve with DP matching method. DP matching is used to match two pattern of different length while permitting partial variations and errors[7]. DP array uses note strength vector while shifting vector index as the existence of overall biases in features between music and humming.

4.1. Dissimilarity Calculation using DP Matching

The sequences of music and humming segments containing note information are defined as *R* and *Q*.

$$R = [r_0, r_1, r_2, \ldots, r_i, \ldots, r_{NR-1}]$$
$$Q = [q_0, q_1, q_2, \ldots, q_j, \ldots, q_{NQ-1}]$$

where r_i is the note strength vector of *i*th segment of music and q_j is the note strength vector of *j*th segment of humming. NR and NQ are the number of segments in the music and humming respectively.

To match two sequences *R* and *Q,* a matrix *D size NR* X *NQ* is constructed. An element of the matrix, $d_{ps}(r_i, q_i)$ as given in equation (6), represents the dissimilarity between r_i and q_j when the overall pitch shift is *ps*.

$$d_{ps}(r_i, q_j) = \frac{\sqrt{\sum_{m=0}^{M-1}\left[r_i(m) - q_j(m-ps)\right]^2}}{\sqrt{\sum_{m=0}^{M-1} r_i^{\,2}(m)}\sqrt{\sum_{m=0}^{M-1} q_j^{\,2}(m-ps)}}, 0 \le m, m-ps \le M-1 \quad \text{-(6)}$$

The matching path *Cps* in case of pitch shift *ps* is defined as a set of consecutive vector elements $d_{ps}(r_i, q_i)$ which decides the matching between *R* and *Q*. The *h*th element

$$C_{ps} = c_{ps,1}, c_{ps,2}, c_{ps,3}, \ldots, c_{ps,h}, \ldots c_{ps,H}$$
$$\max(NR, NQ) \le H < NR + NQ + 1 \quad \text{-(7)}$$

of matching path *Cps* is defined as $c_{ps,h}=(i, j)$ and the matching path can be represented as the following equation assuming that the length of the matching path is *H*.

After contracting DP arrays for all pitch shift values, matching path can be selected for the matching cost to be

$$DP_{ps}(C_{ps,\min}) = MIN\left\{\frac{\sqrt{\sum_{h=1}^{H_{ps}} d_{ps}\left(c_{ps,h}\right)}}{H_{ps}}\right\} \quad \text{-(8)}$$

minimized for each pitch shift values.

4.2. Windowing

Avoiding invalid matching patch could reduce much matching time. If the path is too far from the ideal diagonal path, we need to exclude to as a path using windowing method on DP arrays. As shown the dotted line in in figure 3, narrowing the upper and lower dotted lines could reduce the matching time. However, it will cause a lower variation allowed which could lead to a more strict matching area.

4.3. Additions to the Conventional DP Matching

The measure that reflects the amount of how far the matching path is from the ideal diagonal path is given by the following equation.

$$f_{ps}(R,Q) = H_{ps} \Big/ NR + NQ$$

where H_{ps} is the length of the matching path. This measure is applied to the matching cost DP_{ps} as a normalization factor for calculating dissimilarity value between *R* and *Q* in the form of addition or multiplication.

5. EXPERIMENTAION

5.1. Experiment Configuration

It is not that simple to find an appropriate dataset with polyphonic music and humming. However, I am now trying to get a QBH dataset with a modest humming queries and music clips with 12 music clips and humming samples. Table 1 is the configuration used here(NNF: no use,NF1: multiplication,NF2:addition).

Table 1. Configuration of experiment

	NNF	NF1	NF2
W=8	M01	M02	M03
W=4	M04	M05	M06

5.2. Experiment Results

There are various parameters in this algorithm, like the size of neighbors in partial enhancing process and usage of f measure which measures how far the matching path is from the diagonal path.

	Top 1	Top 3	Top 5
M01	8.33	25	41.67
M02	8.33	33.33	41.67
M03	8.33	33.33	41.67
M04	8.33	25	41.67
M05	8.33	33.33	41.67
M06	8.33	33.33	41.67

Table.2 The preliminary results

The results from Table 2 seems not very good. From analysis, we may consider that the scale of database is too small. On the other hand, the result seems to bee too random. I believe there is still much to improve for the ex-

periments, so the report could still be modified. There is a comparison of spectrogram and enhanced harmonic of a music clip in figure 3, which could let us see the effect of harmonic enhancement.

6. CONCLUSION

The project presents a QBH system with polyphonic music data. The three main modules is melody extraction from music and humming and matching the information in mid-level representation. For this goal, we need to emphasize the most audible sound from mixture by harmonic enhancing, calculate the possible notes by harmonic sum, converting amplitude from frequency index to note index, grouping similar frames to segments, and constructing note segment sequence to extract information. We also need to use DP matching to overcome the difference between music and humming features. We could add other modules to improve the accuracy and speed in the future work, or we could try to extract features from a completed song rather than music clips.

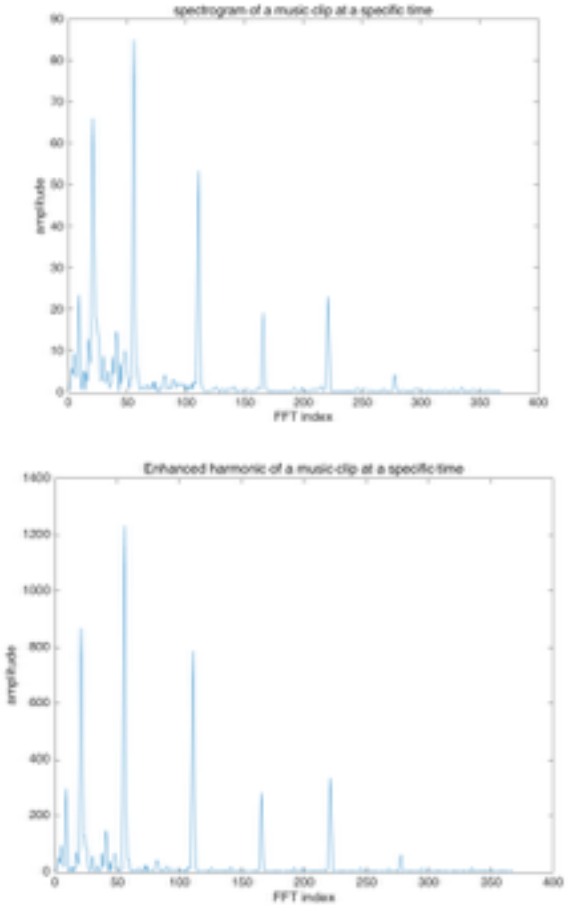

Figure 3. The effect of enhancing harmonic of a music clip in a specific time.

7. REFERENCES

[1] H.Y. Tseng, "Content-based Retrieval for Music Collections," Proc. of 4th ACM conference on Digital Libraries, pp.176~182, California, 1999.

[2] Y. Kim, W. Chai, R. Garcia, B. Vercoe, "Analysis of a Contour-Based Representation for Melody," Proc. International Symposium on Music Information Retrieval, Oct. 2000.

[3] A. Klapuri, “Multipitch Estimation and Sound Separation by the Spectral Smoothness Principle,” IEEE International Conference on Acoustics, Speech and Signal Processing, vol.5, pp.3381~3384, 2001.

[4] T. Miwa, Y. Tadokoro and T. Saito, “Musical Pitch Estimation and Discrimination of Musical Instruments Using Comb Filters for Transcription,” 42nd Midwest Symposium on Circuits and Systems, vol.1, pp.105~108, 2000.

[5] M. Goto, “A Predominant-F0 Estimation Method for CD Recordings: MAP Estimation Using EM Algorithm for Adaptive Tone Models,” IEEE International Conference on Acoustics, Speech and Signal Processing, vol.5, pp.3365~3368, 2001.

[6] Jungmin Song, So Young Bae, Kyoungro Yoon,”Mid-Level Music Melody Representation of Polyphonic Audio for Query-by-Humming System ”

[7]L.Rabiner, B. –H. Juang, *Fundamentals of Speech Recognition,* Prentice-Hall, 1993.