Query by Humming Systems Using Melody Matching Model Based on the Genetic Algorithm

Jing Qin [1,2]
1.School of Computer Science and Technology, Dalian University of Technology, Dalian, China, 116024
2.College of Information Engineering, Dalian University, Dalian, China, 116622
Email: jqins@yahoo.com.cn

Hongfei Lin Xinyue Liu
School of Computer Science and Technology, Dalian University of Technology, Dalian, China, 116024
Email: hflin@dlut.edu.cn, xyliu@dlut.edu.cn

***Abstract*—Query by humming (QBH) refers to music information retrieval systems where short audio clips of singing or humming act as queries. Melody is considered as the most important feature in the queries and the songs. This paper proposes a QBH system using melody matching model based on the genetic algorithm and improving the ranking result by local sensitive hashing algorithm. An approximate template of the query music is constructed by melody contour aligning algorithm based on GA, which is used to align and correct the input pitch template. The validity of the algorithm is presented by the prototype of QBH system and effects of the algorithm are also shown by the experiment results.**

***Index Terms*—music information retrieval, melody contour representation, audio system**

I. Introduction

Traditional music retrieval approaches, based on keyword and textual metadata, face serious challenges. If the user is familiar with the name of the song or other significant information to describe the music, retrieval is straightforward. However, if one does not know the title, singer, alternative retrieval methods are necessary. Content-based music information retrieval (MIR) is gaining widespread attention and can be very helpful, since it forsakes the need of keyword. Often, it consists of a form of query-by-example, such as through singing, humming or playing a sample of the piece, as a query to the database.

From the mid-1990s, there is lots of different field research based on the music information retrieval research. In 1995, Asif Ghias[1] first proposed the method based on the Query by Humming, the method is based on up and down of the melody. From then on, most approaches use pitch sequence to represent the query. For retrieval, dynamic programming[2], dynamic time warping (DTW)[3] and hidden Markov models [4]were used. Recently locality sensitive hashing (LSH) [5][6]approaches are used in this field and gets an impressive result. Since 2005, a number of QBH systems have been evaluated in Music Information Retrieval Evaluation eXchange (MIREX)[7][8].

The major challenges for QBH systems include i) queries vary from different people, how to extract audio features which precisely represent music content, ii) how to describe the musical features, and iii) which method to be used for feature matching.

There are three levels of musical feature, physical features, acoustic features and perceptual features. The physical features express audio content on the format of flow media. The acoustic feature mainly includes time and frequency domain features, such as pitch frequency (or fundamental frequence,F0), short-time energy, zero crossing rate, LPC coefficient and MFCC coefficient, etc. They are the most expressive feature of the audio, and are usually used for different phases of speech recognition. Perceptual features reflect people's feelings such as pitch, rhythm, intensity, timbre, etc. And, the perceptual features can usually be extracted by the physical prosperities. Meanwhile, they are used to recognize and judge the music content, such as the emotion that a song wants to express.

Music is a time-dependent sequence of discrete notes, while we still feel that it is a complete entity. Gestalt Theory (GT) [9] is the framework about the psychological phenomena, the mental process and the psychological application. And this theory proofs that human perception form has a hidden rule which is proximal, similar, and continuous, the rule reveals which pattern organization will be perceived on the condition of stimulus feature. Melody is the main perceptive feature of a song, and in 1986, Dowling [10] proofed that melody satisfy the proximity, similarity and continuity of Gestalt theory. Melody contour could be got by pitch tracking approach, and be used to represent a song. Thus, a melody representation model is the main point in QBH system.

Manuscript received December 28, 2010; revised March 1, 2011; accepted March 28, 2011.

Corresponding author: Jing Qin.

© 2011 ACADEMY PUBLISHER

doi:10.4304/jsw.6.12.2416-2420

In the previous work, [11] proposed a model based on the melody of the standard pitch template and humming-input pitch template, normalized two templates and got the results though matching. In this paper, the melody model was improved and refined, and the matching algorithm was better to match the final search results than previous work.

The human voice frequency ranges from 50 to 3200HZ, while music pitch usually ranges from 16 to7000HZ (the equivalent of notes C2～a5), there are some differences between them. The humming fundamental frequency is often several times less than the standard one. Furthermore, because each person is not in the same pitch range, it could cause variation. If the normalization is simply used to the input template and standard template, much detail information of the pitch contours neglected, it would lead to the distortion result.

In this paper, we employ a template with standard fundamental frequency range to approximate the input humming template and be matched with the standard template instead. A melody contour alignment algorithm based on GA is suggested, which might be a linear shift to input templates, and seems reserve the detail information rather than using normalization directly. LSH NN search is employed for templates indexing and matching, the final ranked list seems be improve. The system framework is shown in Fig. 1.

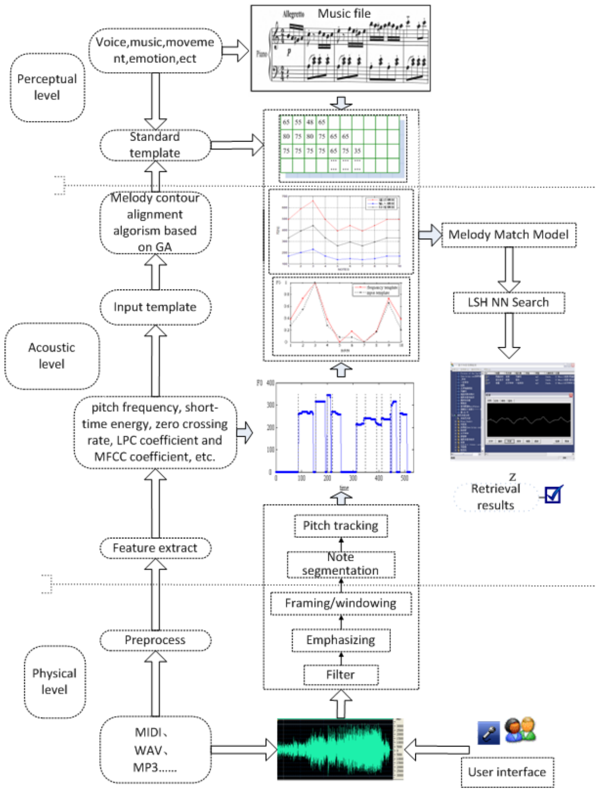

Figure 1. The framework of the proposed QBH system

II. Constructing Melody Match Model

A. Definition

Genetic algorithm [12] as a global optimization algorithm in parallel and global searching capability owns to prominent characteristics. This algorithm does not require the prior knowledge, but can obtain the optimal solution. Therefore, the application using genetic algorithm gain the most similar matching template between the standard pitch contour and the input template, we finally achieved the goal of the template translation.

Melody alignment problem can be defined as follows: Let $P=\{p_1,p_2,\ldots,\ p_i,\ldots,\ p_n\}$ be input template, where p_i represents the certain note, n is the number of notes. We can achieve the approximate template $Q=\{q_1,q_2,\ldots,\ q_i,\ldots,\ q_n\}$ through scaling P template into the range of standard F0 template. In the sequel, we can use similarity measure based on cosine-norm. The goal using cosine-norm is for finding the Q template victor which has the shortest distance with P template.

B. Genetic representation

According to the melody alignment problem, a population of decimal numeral strings, which encoded candidate approximate template with variable length, evolved toward better solutions. A chromosome represented an approximate template, which length varied by the input template. In numbered musical notation, standard pitches distribute into three octaves with 21 notes, the relationship between notes and standard pitches is shown in Table 1.

TABLE1 MAPPING TABLE OF NOTES AND PITCHES

syllable names	Do	Re	Mi	Fa	So	La	Si
F0 (three octave, Hz)	130	146	164	174	196	220	247
	261	293	330	349	392	440	494
	523	587	659	698	784	880	988

Every bit in the population strings was a F0 number in Table1. An input template with a length of *n*, was approximated by a chromosome, which was a decimal numeral strings. An optimized chromosome represented the approximate standard pitch template for an input melody.

For an input template, $P=\{p_1,p_2,\ldots,\ p_i,\ldots,\ p_n\}$, the encoding solution is like that, g_i is a chromosome bit, $i=1,\ldots.,n$, $g_i\in\{x\mid x\ is\ \text{standard}\ F0\ \}$, g_i was cording to q_i in approximate template

© 2011 ACADEMY PUBLISHER

$Q=\{q_1,q_2,\ldots,\ q_i,\ldots,\ q_n\}$, chromosome representation is shown in Table 2. The evolution usually starts from a population of randomly generated individuals, g_i could be one of the standard F0. In each generation, the fitness of every individual in the population was evaluated and modified to form a new population Q. The new population was then used in the next iteration of the algorithm. The algorithm terminates when a satisfactory fitness level has been reached for the best solution, then P and Q are the most closely templates. Because of decimal numeral encoding, the decoding process is simple.

TABLE2 CHROMOSOME REPRESENTATION

q_1	q_2	……	q_i	……	q_n
g_1	g_2	……	g_i	……	g_n

TABLE3 MELODY CONTOUR ALIGNMENT ALGORISM

Algorithm 1: *Melody contour alignment algorism*

```
Input: A humming template $P$;
       The max generation number $MAXGEN$;
Output: A approximate template $Q$;
Description:
Choose the initial population of individuals $B$, randomly produce a number of chromosome;
Let $S$ be a chromosome, $i$ be a counter of generations;
Evaluate the fitness $F(s)=\cos(P,Q)$ of each individual in population $B$;
Repeat
    Select the best-fit individuals for reproduction;
    Breed new individuals through crossover and mutation operations to give birth to offspring;
    Evaluate the individual fitness of new individuals;
    Replace least-fit population with new individuals;
until $i < MAXGEN$;
return $S$.
```

C. Fitness function

Let the input template be $P=\{p_1,p_2,\ldots,\ p_i,\ldots,\ p_n\}$, the approximate template be $Q=\{q_1,q_2,\ldots,\ q_i,\ldots,\ q_n\}$ through using genetic operation. Each approximate template similarity can be calculated as follow

$$Sim(P,Q)=\cos(P,Q)$$
$$=\sum_{i=1}^{n}\frac{w_{ip}}{\sqrt{\sum_{k=1}^{n}w_{kp}{}^{2}}}\times\frac{w_{iq}}{\sqrt{\sum_{k=1}^{n}w_{kq}{}^{2}}} \qquad (1)$$

Through genetic algorithm, the optimal chromosome should own to the maximal fitness function

$$F(s)=Sim(P,Q)=\cos(P,Q). \qquad (2)$$

Therefore, cosine-norm measure was used as fitness function, and achieved the optimal approximate template.

D. Melody contour alignment algorism

Process of the melody contour alignment algorism based on GA is described in Table 3.

III. SIMILARITY MATCHING

Euclidean distance is the most widely used similarity measure for time series similarity research. For the input pitch template P and the standard pitch template S, the Euclidean distance is given as (3)

$$D_E(P,S)=\sqrt{\sum_{i=1}^{m}(w_{pi}-w_{si})}\ . \qquad (3)$$

There are lots of advantages of Euclidean distance. For the instance, it can easily calculate and understand, it satisfies the triangle inequality, and supports the multi-dimensional space index.

TABLE4 LSH NN SEARCH ALGORITHM

Algorithm 1: LSH NN search algorithm

```
Input: An approximate template $Q$;
       A window $\omega$;
       Database melody $s$;
Output: The best similarity song list $L$;
Description:
For $i=1\ldots ni$ do
    divide $s(i)$ into several melody fragments $mf(j), j=1\ldots mj$;
    Normalize these melody fragments ;
end for
Index all fragments by LSH;
Search approximately $k$ nearest neighbors in the tables of LSH;
The candidate melodies are ranked according to the entire database melody occurrence number in $k$ nearest neighbors;
Let $L$ be a list of top $n$ candidate songs;
return $L$.
```

An input audio clip is a part of a song, so we need to get a similarity between a subsequences and a whole sequence which represent pitch information of an entire song. Melody fragment was employed to diverse a song into many subsequences, and constructed an index using locality sensitive hashing by [5].The benefits of LSH against Euclidean distance is that it can get a sublinear time complexity. In our method, we employed the approach above, and made some changes to fit the

© 2011 ACADEMY PUBLISHER

melody templates match, the algorithm is described in Table 4.

IV. Experiments and Analysis

A. Music Database

The music database used in this study consisted of 100 pop songs. There are several melody fragments in a song. The input signal which is got from microphone, was sampled on the format of 11.025KHz/8bit/monc, filtered by the bandpass filter with the close frequency of $f_H = 3400Hz$ and $f_L = 60 - 100Hz$, emphasized by a first-order digital filer $H(Z) = 1 - \mu^{-1}$, where μ is 0.98. Hamming window is applied to the signal, which is framed with the window size of 128 and with the overlap of 64.

B. Results of GA

The results of GA seem that it is effective in getting approximate templates. For an example, a chinese song named “Tian mi mi”, the music score of the first sentence is like “3563121253”, the input pitch template is shown in Fig. 2.

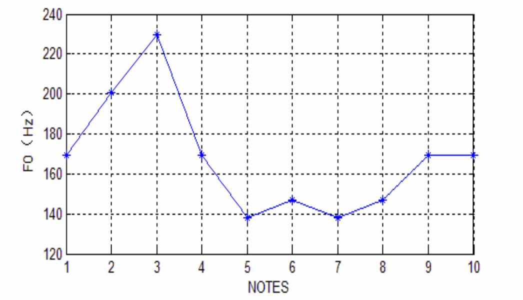

Figure 2. An input template before alignment

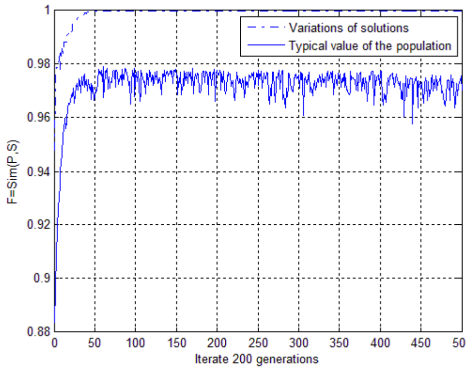

Figure 3. Variations of solutions and typical value of the population size

The original input template would be the input of GA. According to the variable length encoding, the number of notes was another input parameter. In the example, the length of chromosome was 10, population size was 40, generation gap was 0.9, the maximum number of generations was 200, and Fitness-based Reinsertion was used. The result is shown in Fig. 3. The template got a constringency trend after 20 generations, and was steady around 40 computations. The final similarity of the input template and approximate template was 0.9999.

The standard template, input template and approximate template were similar in contour and amplitude, as shown in Fig. 4, even without normalization.

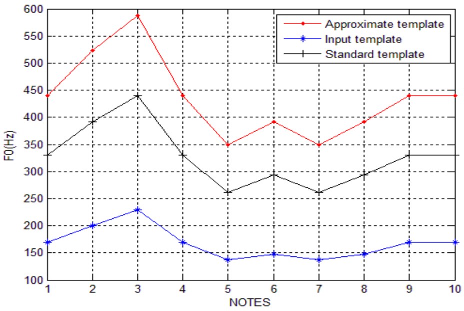

Figure 4. Templates output in melody contour alignment algorism

After normalization, the approximate template coincided with the standard template, which is shown in Fig. 5; it indicates that the difference between each input was eliminated by GA melody contour alignment algorism. It solved a main problem in QBH system as we mentioned above, the best similarity of the approximate template and standard template was a great premise for the matching result.

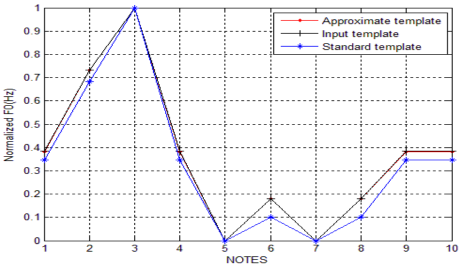

Figure 5. Normalized templates

To test the efficiency of GA, 40 input melodies were involved. The results show that 85% of them got better rank list, the input templates were corrected and the match precision was improved.

C. QBH system search results

Table5 presents the result of 20 piano or other orchestral instruments played audio clips ordered by similarity, under the condition of the great quality input could get Top-3 hit rate of 65%. The process suggested that Top-X hit rate is changed by the number of songs in database, when melody fragments was less, Top-X hit rate was higher. Although Top-3 hit rate was lower, it could make the match melody on the front position, and the similarities of the front melody were close. It indicates that the standard template and input template are similar, but there are many fragments with high similarity in database, they interfere with the results.

© 2011 ACADEMY PUBLISHER

TABLE5 INSTRUMENTS MELODY RETRIEVAL RESULTS

Ordinal	1	2-3	4-10	10-20	20-30
Number of melodies	8	5	1	3	1
Top-X hit rate (%)	40	65	70	85	90

To test the humming queries, we randomly analyzed 20 melody fragments from different people which include 10 girls and boys, the results ordered by similarity, as shown in Table 4, the Top-10 hit rate of 50% is lower than instrument fragments, it suggested that the input quality of instrument is better than human, but the system still useful.

TABLE6 HUMMING MELODY RETRIEVAL RESULTS

Ordinal	1	2-3	4-10	10-20	20-30
Number of melodies	2	2	6	3	4
Top-X hit rate (%)	10	0	50	65	85

V. CONCLUSIONS

This paper presented an efficient melody alignment algorithm for solving pitch the differences between different people. We obtained our similarity match algorithm by adapting the general LSH method of [5] to gain a sublinear time complexity. Based on the melody match model, the search process is more fuzzy and effective. After all, the size of database is indeed very small, and we hope that there is room for some improvement. Much work needs to be done to solve interference from high similarity of music fragments in large database.

REFERENCES

[1] Asif Ghias, Jonathan Logan, David Chamberlin, Brian C. Smith, “Query by humming-musical information retrieval in an audio database”, In: Proc of the third ACM international conference on Multimedia .Jan. (1995) 231-236

[2] J.-S. R. Jang, C.-L. Hsu, and H.-R. Lee, “Continuous HMM and its enhancement for singing/humming query retrieval”, in Proc. 6th International Conference on Music Information Retrieval, 2005.

[3] Y. Ohishi, M. Goto, K. Itou, and K. Takeda., “A stochastic representation of the dynamics of sung melody,” in Proc. ISMIR, 2007, pp. 371–372.

[4] J.-S. Roger Jang and Hong-Ru Lee, "A General Framework of Progressive Filtering and Its Application to Query by Singing/Humming", IEEE Transactions on Audio, Speech, and Language Processing, No. 2, Vol. 16, PP. 350-358, Feb 2008.

[5] M. Ryyn¨anen and A. Klapuri, “Query by humming of midi and audio using locality sensitive hashing,” in IEEE International Conference on Acoustics, Speech, and Signal Processing, Las Vegas, Nevada, USA,Apr. 2008, pp. 2249–2252.

[6] M. A. Casey, R. Veltkamp, M. Goto, M. Leman, C. Rhodes, and M. Slaney, “Content-Based Music Information Retrieval: Current Directions and Future Challenges”, Proceedings of IEEE, Vol. 96, No. 4, April 2008.

[7] F. Wiering R. Typke and R. C. Veltkamp, “Mirex symbolic melodic similarity and query by singing/humming,” in Intl. Music Information Retrieval Systems Evaluation Laboratory (IMIRSEL), 2006.

[8] A. Ito M. Suzuki, T.J. S. Downie, D. Bryd, T. Crawford, “Ten Years of ISMIR: Reflections on Challenges and Opportunities”, Keynote talk, Kobe, ISMIR 2010.

[9] Nevis, E. (2000) Introduction, in Gestalt therapy: Perspectives and Applications. Edwin Nevis (ed.). Cambridge, MA: Gestalt Press. pp. 3.

[10] W.J. Dowling, “Scale and Contour: Two Components of a Theory of Memory for Melodies”, Psychological Review, lxxxv (1978), 341–54

[11] Jing Qin, Xing-ce Wang, Ming-quan Zhou, Xin-yu Liu, “A novel MIR approach based on dynamic thresholds segmentation and weighted synthesis matching”, in IET Conference on Wireless, Mobile and Sensor Networks 2007 (CCWMSN07), pp.1017–1020

[12] Mitchell, Melanie, (1996), An Introduction to Genetic Algorithms, MIT Press, Cambridge, MA.

Jing Qin received the M.S degree in Computer Science and technology from Beijing Normal University in 2007. She is currently working toward the Ph.D. degree in the School of Computer Science and Technology, Dalian University of Technology, Dalian, China. Her research interests include multimedia information retrieval, music signal processing and machine learning.

Hongfei Lin received the Ph.D degree from Northeastern University, China. He is a professor in the School of Computer Science and Technology, Dalian University of Technology, Dalian, China. His professional interests lie in the broad area of information retrieval, web mining and machine learning, affective computing.

Xinyue Liu is currently working toward the Ph.D. degree in the School of Computer Science and Technology, Dalian University of Technology, Dalian, China. Her research interests include multimedia information retrieval, web mining and machine learning.