# A Novel MIR System Based on Improved Melody Contour Definition

Peng Li
School of Education Technology
Beijing Normal University
Beijing, China
e-mail: lpfai@mail.bnu.edu.cn

MingQuan Zhou
College of Information Science and Technology
Beijing Normal University
Beijing, China
e-mail: mqzhou@bnu.edu.cn

Xuesong Wang
School of Education Technology
Beijing Normal University
Beijing, China
e-mail: Hanson_w75@tom.com

Nansha Li
College of Information Science and Technology
Beijing Normal University
Beijing, China
e-mail: linansha_823@163.com

*Abstract*— **This paper proposes a method of retrieval a music information by a voice or whistling clip. The method is based on the content of the music file. Firstly, extract the pitch as a kind of musical feature form the digital clip file, and then fix and smooth the discrete data. Bring forward a novel melody contour representation method to depict the musical feature properly. At last, adopt an improved LCS algorithm to matching the feature strings stored in the database. The proposed retrieval method is experimentally evaluated. The evaluation result shows that the improved method could retrieval the object more accurate.**

***Keywords—Music Information Retrieval (MIR); pitch; melody contour; Longest Common String(LCS); Query by Voice***

## I. INTRODUCTION

Music Information Retrieval (MIR) is a field that was concerned by increasing people. The first paper on music information retrieval was completed by Michael Kassle dates back to the mid -1960s. Kassle and his colleagues were the earliest researches, and for many years thereafter, very little were done. But now, with the explosive expansion of digital music the field is getting more attention than before. However, in a long time, music information retrieval was based on textual metadata such as title, composer, singer or lyric. These various metadata-based schemes for music retrieval have suffered from many problems including extensive human labor, incomplete knowledge and personal bias.

At the same time, content-based retrieval technology is developing as hot topics, such as content-based image retrieval, content-based audio/video retrieval.

Compared with traditional keyword-based music retrieval, content-based music retrieval provides more flexibility and expressiveness. Users may express queries by existing recordings, or produce an audio clip in a real-time. Query-by-Voice, such as humming, singing, whistling and et al is one of the popular content-based retrieval methods.

Since the content-based music retrieval attracted widespread attention in the world, researchers has been made significant progress. Bainbridge, Nevill Manning, Witten Smith and McNab accomplished a paper on music information retrieval won the best paper award at the Digital Libraries '99 conference, and almost every Digital Libraries, Computer Music or Multimedia conference has had one or more papers on music retrieval and/or digital music libraries.

Content-based music retrieval is usually based on a set of extracted music features such as pitch, duration, and rhythm. For using this information, we are able to get note onset/offset information from the segmentation information of silent or ignorable frames and fundamental frequency.One common approach or developing content-based music retrieval is to represent music melody as a string using a set of characters.

For Example, Seungmin Rho(2004) and his colleagues use three characters for representing the pitch contour: U(up),D(down),S(same). In order to find similar melody strings from melody source, information retrieval techniques, especially string matching methods are widely used.

In this paper, we propose a music retrieval framework, mainly include 3 parts. The first step is pitch extraction, when we got a music clip produced by voice; we are able to acquire the pitch variation feature. After some necessary processing, the second step is to define a melody representation method, in other words, we must find a strategy to convert those pitch value to strings, for we can search the database with the string matching algorithms.

Although researchers engaged in different implementation of the program around the MIR, most of them contain those important steps as figure 1.

978-0-7695-3556-2/08 $25.00 © 2008 IEEE
DOI 10.1109/MMIT.2008.65

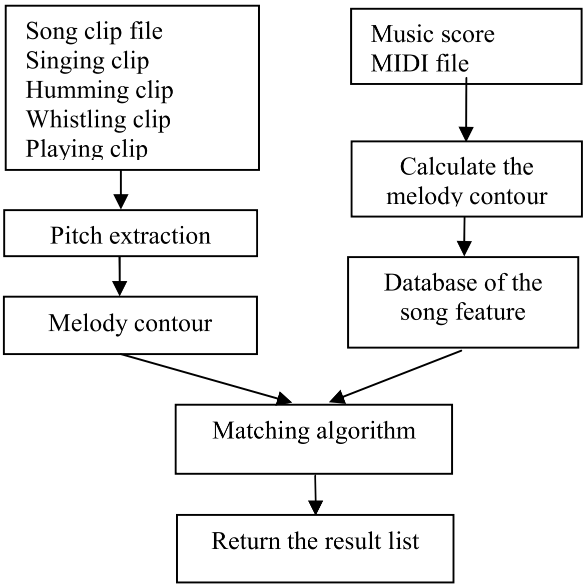


Figure 1. The common framework of MIR.

## II. Pitch Extraction

As is shown in Figure 1, the first step is to extract the pitch from the music clip file. There are many techniques to extract pitch from a voice query. We will introduce some traditional methods widely used and then, to explain a practical and efficient method we adopted.

### A. Ttraditional methods

There are several typical acoustic signals processing techniques for music information retrieval. In general, methods for detecting pitches can be divided roughly into two categories: time-domain based and frequency-domain based.

Zero Crossing Rate (ZCR) is a popular method in the time-domain. The basic idea is that ZCR gives information about the spectral content waveform cross zero per unit time (David, 2003). Recently, ZCR appeared in a different form such as VZCR (variance of ZCR) or SZCR (smoothing ZCR) (Huang and Hansen, 2006). They are more suitable for dealing with audio files than ZCR.

In the frequency-domain, Fast Fourier Transformation (FFT) is one of the most popular methods. This method is based on the property that every waveform can be divided into simple sine waves. But, a low spectrum rate for longer window may increase the frequency resolution while decreasing the time resolution. Another problem is that the frequency bins of the standard FFT are linearly spaced, while musical pitches are better mapped on a logarithmic scale. So, Forberg (1998) used an alternative frequency transformation such as constant Q transform spectrums which are computed from tracked parts.

Ryynanen and Klapuri (2006) proposed a singing transcription system based on the HMM-based notes event modeling. The system performed note segmentation and labeling and also applied multiple-F0 estimation method (Klapuri, 2005) for calculating the fundamental frequency.

### B. Our methods

Auto Correlation Function (ACF) is an effective method in the time-domain, it based on the cross correlation function. While a cross correlation function measures the similarity between two waveforms along the time interval, ACF can compare one waveform with itself.

ACF is of the most commonly used features function related to short-time signal analysis. Musical signal $s(m)$ was separated by a $N$ length window, after that, we got a number of windows $S_n(m)$, define the ACF of $S_n(m)$ is:

$$R_n(k) = \sum_{m=0}^{N-k-1} S_n(m) S_n(m+k)$$

In this function, $k = (-N+1) \sim (N-1)$ .Since the ACF of signal will reach the peak when the pitch period occurs, the pitch period can be detected by the peak location.

In our research, we used the following methods to implementation.

The first step towards the audio file obtained from the record. The file is divided into frames of 256 samples, with 50% overlap at the two adjacent frames. If the energy of an individual frame is below a predefined threshold, the whole frame is marked as silence frame and is ignored for further processing. After silence reduction, the audio frames are hamming windowed. Then, the pitch value will be calculated in every window using ACF.

The discrete feature data we get above could be seen as the value of the pitch reflecting query clip, however, any detection algorithm may obtain the inaccuracy value, so the pitch values curve need to smoothing for the abnormal values.

Linear smoothing is adopting a sliding window to deal with:

$$y(n) = \sum_{m=-L}^{L} x(n-m) \cdot w(m)$$

In this function, the Variable $\{w(m) |\ m = -L, -L+1, \cdots, 0, 1, 2, \cdots, L\}$ is the sliding window with $(2L+1)$ points, and:

$$\sum_{m=-L}^{L} w(m) = 1$$

After those procedures, the features are combined to form the feature vector to represent the individual audio file. We will explain how to obtain the well melody contour from pitches in next section.

## III. Melody Contour Representation

The object of melody contour representation is to find well rules to convert a set of discrete feature data to a set of strings.

So far, many researches proposed lot of methods to reach the target, a reasonable solution is use UDR string parsed from pitch information to represent music. However, Seungmin Rho and Eenjun Hwang (2006) thought, there are several restrictions in using the UDR string. First, current UDR string cannot describe sudden pitch transitions. They believe classifying intervals into five extended types could relieve this: up, up a lot, repeat, down and down a lot.

Also, They proposed an alternative representation scheme for time contour. Similar to the way a pitch contour is described by the relative pitch transition, a time contour can be described by the duration between notes. Time in music is classified in one of three ways: R for a repetition of the previous time, L for longer than the previous time, or S for shorter than the previous time.

Actually, the above-mentioned method is still too vague, In other words, to judge pitch up or up a lot just rely on an experience threshold, but we need to confront hundreds of query clips sung by various people, therefore, we propose a new method help to judge "up" and "up a lot" or "L" and "S".

The detail melody representation strategies are as follows:

- Step1: draw the pitch change curve.
- Step2: Statistic the pitch data, to find the stable value of each a pronunciation syllable.
- Step4: Statistic all the pitch range and pitch Continue length, to produce a reasonable threshold.
- Step3: define 3 strings to depict the pitch curve, the first is U,S,D to depict the pitch is up, same or down since last pitch; the second is A,B,C to represent the length of same pitch; the third is X,Y,Z to stand for the Changes value in the pitch range

Those 3 strings are able to reflect the melody contour more completely than others.

## IV. MATCHING METHODS

After the melody representation has been completed, the song feature is ready to match with the ones in database. Since we convert the melody contour to strings, we need a quick and efficient string matching algorithm.

String matching algorithm should calculate the similarity between the input query and records in database. Some efficient methods have been developed based on statistic model or feature space, such as HMM, KNN, DTW, Boyer–Moore algorithm and GMM. In general, algorithms for string matching can be divided into two kinds: exact and approximate matching. Although the exact matching is faster than approximate, we can not chose it for the query string is not exact actually. Approximate matching could tolerance the inaccuracy in query string; consequently, it is the suitable method.

There are two matching algorithms could be implemented to matching approximately in prototype system. Dynamic Programming Algorithm: This has been popular in the field of approximate string matching. Since melody contours are represented as character strings, dynamic programming was applied to melody comparison and has become a standard technique in music information retrieval.

When melodies are viewed as strings, one of the popular measures of similarity is the number or cost of editing operations that must be performed to make the strings identical. The minimum cost is called "edit distance." The most common editing operations for melody comparisons are inserting a note (insertion), deleting a note (deletion) and replacing a note (replacement).

Those three basic operations establish the foundation of dynamic programming algorithms applied to melody comparison.

Another effective matching algorithm is Longest Common Subsequence (LCS) Algorithm. This algorithm involves establishing a recurrence for the cost of an optimal solution.

Both of the two algorithms are available in our prototype system, but they cannot be used directly. The reason is that, the melody contour strings in the database include a whole song; it is much longer than the query strings.

In order to handle the problem, we design a feature storage format; each song was depicted with series of strings separated by "#" according to the sentences. After that, we proposed a Local Longest Common Subsequence (LLCS) Algorithm.

The Specific operations are as follows, also seen in figure 2:

- Step1: read the feature strings of the first song form the database, and stored to a buffer.
- Step2: get the first sentence which end with "#" ,and calculate the LLCS with the query string.
- Step3: record the length of the LLCS
- Step4: repeat the comparison till the song finished.
- Step5: save the longest length of the LLCS to the database according to the song.
- Step6: repeat the calculate operation to all the songs in the database.
- Step7: sort all the songs according to the value of the length, return the song list.

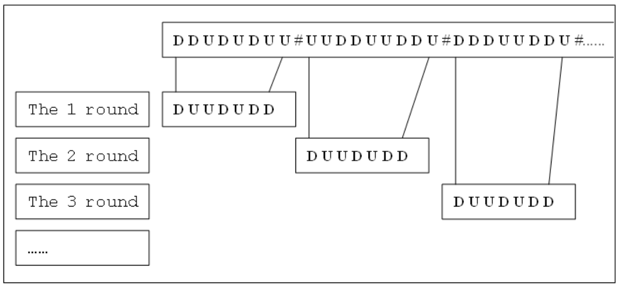
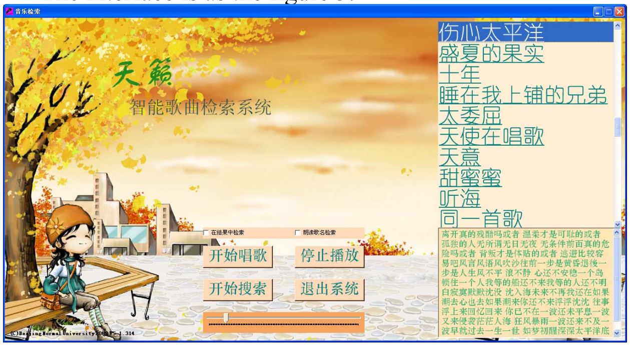


Figure 2. The LLCS algorithm

## V. IMPLEMENTATION

We developed a music retrieval system by voice, using the algorithms described in chapter II to IV. The initial prototype system allowed searching of a database of 200 pop songs, The system allows users to humming, singing, or whistling at the front of their microphones, and then the digital clips well be analysis by the pitch extract part, after the melody representation and search, a sing list will be return to the users, the order is according the similarity between the clip and the songs stored in the database.

The interface is as the figure 3.



伤心太平洋
盛夏的果实
十年
睡在我上铺的兄弟
太委屈
天使在唱歌
天意
甜蜜蜜
听海
同一首歌

离开真的残酷吗或者 温柔才是可耻的或者
孤独的人无所谓无日无夜 无条件前面真的危
险吗或者 背叛才是体贴的或者 逃避比较容
易吧风言风语风吹沙往前一步是黄昏退後一
步是人生风不平 浪不静 心还不安稳一个岛
锁住一个人我等的船还不来我等的人还不明
白寂寞默默沈没 沈入海未来不再我还在如果
潮去心也去如果潮来你还不来浮浮沈沈 往事
浮上来回忆回来 你已不在一波还未平息一波
又来侵袭茫茫人海 狂风暴雨一波还来不及一
波早就过去一生一世 如梦初醒深深太平洋底

Figure 3. The user interface

The evaluation procedure include 3 query types, 5 seconds (s), 10s and 15s input, we statistics the result form to aspects, the object song appeared in top3 songs, or top 10 songs, the results shows in figure 4, and the improved method could retrieval the object more accurate when the recording time is property.

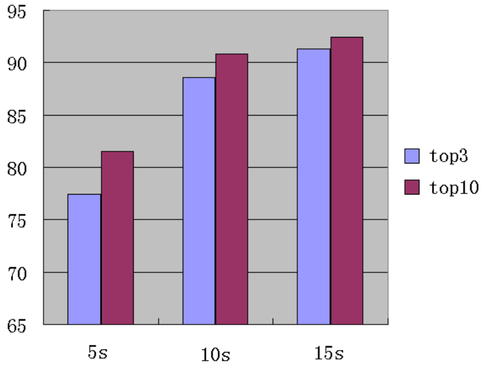


Figure 4. The evaluation result chart

If only pick the pitch various information, they are the first and the third melody contour, our system's precision is up to 47.2%, higher than 24.2% on the system developed by Seungmin.

## VI. CONCLUSIONS

In this paper, we propose a music information retrieval framework and a prototype system including some effective strategy. A query music signal is captured by a microphone in real world. Then a robust feature extraction method called ACF is implemented to get the pitch information form the raw digital clip. The proposed system has been tested with well performance because of the accurate melody contour representation and appropriate string matching algorithm. Experimental comparisons for music retrieval with other systems are presented and it demonstrates the superiority of LLCS method in terms of the retrieval accuracy. Future work will involve the development of new features and further analysis of retrieval system for speech information.

## ACKNOWLEDGMENT

The research is sponsored by the National Natural Science Foundation of China (No. 60673100).

## REFERENCES

[1] Bainbridge, D., and Bell, T. "The challenge of optical music recognition". Computers and the Humanities, 35(2) 2001, pp.95–121.

[2] Bainbridge, D., Nevill-Manning, C., Witten, I., Smith, L., & McNab, R. "Towards a digital library of popular music". In The 4th ACM conference on digital libraries. Berkeley, California1999. pp. 161–169

[3] David Bainbridge, Michael Dewsnip and Ian H. Witten. "Searching digital music libraries", Information Processing and Management 41 (2005) pp.41–56

[4] David Gerhard. "Pitch Extraction and Fundamental Frequency:History and Current Techniques". Technical Report TR-CS 2003-06.

[5] Huang, R., Hansen, J.H.L. "Advanced in unsupervised audio classification and segmentation for the broadcast news and NGSW Corpora". IEEE Trans. on Audio, Speech and Language Processing 14(3) 2006, pp.907–919.

[6] Johan Forberg. "Automatic conversion of sound to the MIDIformat".TMH-QPSR 1-2/1998.

[7] Klapuri, Anssi P. "A perceptually motivated multiple-f0 estimation method". 2005 IEEE workshop on applications of signal processing to audio and acoustics, pp.291–294.

[8] Liu, Z., Huang, Q. "Content-based indexing and retrieval-byexample in audio". ICME 2, 2000.pp.877–880.

[9] Matti Ryynanen, Anssi Klapuri. "Transcription of the singing melody in polyphonic music", ISMIR 2006.

[10] Seungmin Rho and Eenjun Hwang, "FMF: Query adaptive melody retrieval system,"The Journal of Systems and Software. 79 (2006) pp.43–56

[11] Seungmin Rho, Byeong-jun Han b, Eenjun Hwang b and Minkoo Kim , " MUSEMBLE: A novel music retrieval system with automatic voice query transcription and reformulation," The Journal of Systems and Software. 81 (2008) pp.1065−1080