ISSN : 0974 - 7435 Volume 10 Issue 23

2014

BioTechnology

An Indian Journal

FULL PAPER

BTAIJ, 10(23), 2014 [14515-14519]

Research of retrieval about digital music information basing on its content

Li Hao[1], Yang Jie[2], Di Jun-An[3,*]

[1]Straits Institute, Minjiang University, Fuzhou, Fujian, 350121, (CHINA)

[2]School of Information Engineering, Wuhan University of Technology, WuHan, HuBei, 430070,

[3]Department of Electronic Information Engineering, Minjiang University, Fuzhou, Fujian, 350121, (CHINA)

E-mail: eastlihao@163.com, E-mail: citybill@163.com, E-mail:dmieorg@163.com

ABSTRACT

Content-based music retrieval is one of the most activity research fields.Traditional retrieval assumption object data are formed by MIDI format, but the assumption is unsatisfied with practical request. Most of music is stored in wave format. In this paper, we present an approach to deal with WAV music files and design a query-by-humming system. The experiment result demonstrates this approach is effectively.

KEYWORDS

Query-by-humming system; Digital music; Content.

INTRODUCTION

With the rapid development of multimedia technology, toady people are likely to use computers to store and manage multimedia information. However, the existing information retrieval technology is still unable to meet people's demand for huge amounts of information effectively. In the past, information was mostly stored in relational databases in discrete form, queried and retrieved in structured query language (SQL). However, multimedia data are continuous, diversified and huge. At present, the management of multimedia database is generally done manually, classified and retrieved, based on text description. Although text description can be applied to some multimedia data, manual operation is too laborious and time-consuming. For music description, it is highly subjective, inaccurate and misleading. The goal of content-based technology is just to solve this problem. It can be divided into classification and query, i.e., to use the inherent features of music for automatic classification, instead of manual text description. We can retrieve music by humming. Previous music retrieval methods assumed that the object of processing was MIDI music data. But it was difficult to satisfy this assumed condition in actual use. WAV music data are more common. How to retrieve WAV songs is a problem to be solved in this article. By comparing the features of hummed songs and original songs, we make a similarity calculation based on improved Dynamic Time-Warping (DTW) and get corresponding results.

RELEVANT STUDIES

A common way to study the classification of music is to extract statistical data from sound signals and then use these data to classify them. In content-based Music Query, humming retrieval is one of the main research directions. But the vast majority of current studies tend to adopt score-based storage form (MIDI format). Few have paid attention to the original waveform data (WAVE). At present, some large-scale projects in this aspect include:

QBH System by University of Southampton, UK. In recent years, almost all of the research reports on music content retrieval have quoted the article by University of Southampton, UK released on ACM Multimedia 1995 [1]. They also developed a system called QBH (Query by Humming). This system allowed users to search music databases by humming to the microphone. They figured out the distribution diagram of fundamental frequency of sound wave input by means of autocorrelation and used U (the current pitch was higher than the previous), D (the current pitch was lower than the previous), R (the current pitch was equal to the previous) to represent the correlation between the current pitch and previous/next syllable and further form character strings. However, the defect of this system was that no complete set of note-cutting process had been developed. When using QBH, users needed to cut notes by themselves. But the researchers' work did take an important step in music retrieval by means of intuitive singing. Rodger J.McNab from the University of Waikato and New Zealand Digital Library cooperated to develop a system called MT (Melody Transcript) [2][3]. They used Gold - Rabiner Algorithm to find the fundamental frequency distribution of sound wave input and then convert them into standard notes.

Next, they combined MT with Digital Library, to develop a MELDEX system [4], to enable users to search music database directly by humming to the microphone. The accuracy was about 77% - 89%. Nonetheless, MELDEX cannot cut notes correctly, so when users were humming, they had to leave out a short pause or insert more "ticks" between notes by themselves. Therefore, it was still rather inconvenient and unnatural for non-professionals to use.

STRUCTURE OF THE MUSIC RETRIEVAL SYSTEM

The structure of the music query system can be represented in the following figure:

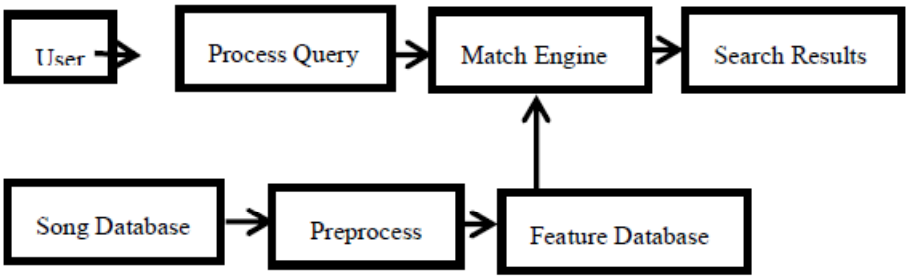

Figure 1 The Composition of Music Query System

At preprocessing, two features, melody and rhythm (generally expressed as symbol sequences similar to musical notation), are extracted from music database (perhaps a variety of media formats) and saved as structured feature database, for later query.

The query processing is mainly to process the humming input of the user who hums into the same features as in the database and match them.

After converting the original music data and input into feature sequences, we can use string matching or text searching for query. The role of matching search engines is to compare the features of converted input sequences with those in preprocessed feature database, find potential matched result and output the results according to the degree of similarity matching.

Relevant studies mainly use MIDI songs as the object of processing. Since MIDI is a kind of score-based storage form, it is more systematic. WAV data that have not been preprocessed are easy to process. But in actual situations, WAV files are more widely used. The contents and expressions that they contain far surpass the content that MIDI files are able to describe. How to describe the features of WAV files has become an urgent issue. This article uses low-level MFCC parameters as feature description information. MFCC features are widely used in the research fields of voice recognition, audio classification and retrieval. This article not only considers MFCC features, but also short-time energy function and short-time zero-crossing rate, for the purpose of feature description.

SEARCHING ALGORITHM

The searching algorithm used here is improved DTW matching algorithm. Template matching is one of the most commonly used similarity calculation methods in multidimensional pattern recognition system. In the process of training, through feature extraction and the compression of feature dimensions, as well as clustering or other methods, each pattern generates one or a few templates. At recognition stage, a similarity calculation is made on the eigenvectors of recognition pattern and various templates and then judge which class it belongs to. In music retrieval, template matching can also be used to make a similarity calculation. But just the same as the typical application in voice recognition, there are also time alignment problems with feature dimensions, some particular circumstances that do not exist in the matching calculation for pattern recognition. For humming retrieval, some people hum quickly, while some hum slowly. Besides, the length of each note in one humming is unlikely to be completely consistent with the original music. Therefore, when matching, if we just conduct linear time warping on eigenvector sequences, the phonemes in them may be misaligned. In this case, a certain kind of nonlinear time warping template matching algorithm should be used, which has been widely used in voice recognition.

DTW is to adopt a dynamic planning method to disintegrate a complex global optimization problem into several local optimization problems and make decisions step by step.

We assume that the eigenvector sequence of reference template is $A=\{a_1, a_2,\ldots, a_i\}$, input the eigenvector sequence as $B=\{b_1, b_2,\ldots, b_j\}$, $i \neq j$. DTW Algorithm is just to find an optimal time warping function, to map the timeline i of tested template non-linearly to the timeline j of reference template and minimize the total cumulative amount of distortion.

Define the minimum cumulative distortion function $g(i, j)$ as the cumulative matching distance of optimal path among all possible paths by the end of the matching point (i, j).

Below, the specific steps of DTW Algorithm will be introduced:

(1) To initialize: define $i(1)=j(1)=1$, $g(1,1)=2d(a_1,b_1)$

$$g(i,j)=\begin{cases} 0 & (i,j)\square \mathrm{Re}\, g, \\ huge & (i,j)\square \mathrm{Re}\, g, \\ \text{where Re}g \text{ is a constrained parallelogram.} & \end{cases}$$

(2) To calculate cumulative distance recursively:

$$g(i,j)=\min\{g(i-1,j)+d(a_{i-1},b_j)\bullet W_n(1),\ g(i-1,j-1)+ d(a_i,b_j)\bullet W_n(2),\ g(i,j-1)+d(a_i,b_j)\bullet W_n(3)\}$$

$i=2, 3,\ldots, I$, $j=2,3,\ldots,J$; $(i,j)\square \mathrm{Re}\, g$

(3) To backtrack all matching pairs: backtrack from (I, J) straight to (1, 1), according to the optimal local path derived from the previous step. This backtracking process is indispensable for the evaluation of average template or clustering center, but the recognition process is not always necessary.

QUERY-BY-HUMMING SYSTEM

This experiment system has collected a total of 84 pop songs as retrieval materials, including 13 English songs, 71 Chinese songs, 14 female singers, 32 male singers and also part of movie and television songs. The original files are all MP3 format. The sampling rate is 11025Hz, mono and 8 bit WAV files. The schedule time for users' humming is 10 seconds. These are sufficient to meet the requirements. The recording sampling rate is 11025 Hz, mono and 8 bit WAV files. The method to preprocess and frame the humming input is the same as in genre classification experiment. Late, the endpoint is tested, to extract effective parts from the input:

Calculate the energy of each frame, $E_n=\frac{1}{N}\sum_m[S(m)w(n-m)]^2$, where $S(m)$ is the input signal, $w(m)$ is the rectangular window of the width N, whence the average energy $E=\sum_i E_i/m$, with $g=c*E$ (with c as constant) as silent period or noise. When extracting features, we only calculate non-silent periods.

To extract MFCC features from the corresponding segments of the hummed song and the original song respectively, we can see from the corresponding envelope that they are indeed similar. Through experiment, it is found that for 12-dimensional MFCC, only low-dimensional coefficient has high similarity with zero-crossing rate.

Below, the processing process of the retrieval system shall be introduced: firstly, to extract and calculate the MFCC1~3 and zero-crossing rate of the original song. For example, the song "I only Care about You" (by Christine Hsu) which has a playtime of 4:30 seconds can be divided into:

$$\frac{(4*60+30)*11025}{128}=23255$$

frames. The 4-dimensional features of MFCC1~3 and zero-crossing rate in each frame are calculated and 23255*4 eigenvectors are obtained. A 10 second humming input, however, only obtains 861 * 4-dimensional eigenvectors. Because of the large number of data, to improve the retrieval speed and efficiency, we adopt the following method: f(n) as the original feature sequence, s(m) as the simplified feature sequence.

```
m=0
for (i=1:10:n)
j=i+10;
if(mean(f(i:j))>mean(f(1:n)))
   s(m)=max(f(i:j))
else
   s(m)=min(f(i:j))
end
m=m+1;
end
```

In this way, a total of 23255*4 eigenvectors in the song "I Only Care about You" can be converted into 2325*4. The humming feature sequence is 86*4. The data size is compressed to 1/10 of the original size. After such processing, the feature sequence can retain and even strengthen the original similarity.

And then, DTW Algorithm is used to match the 86-frame humming feature sequence with the 2325-frame original song. When matching, first of all, label the starting time of each lyric in the original song. For example,

[00:27. 98] Where would I have been,

[00.31.26] if I hadn't met you

[00.34.89] how would things be going

Corresponding information has actually been stored on actual Karaoke tapes. Later, we extract feature sequence and humming sequence to make DTW calculation in corresponding time position in original feature sequences. Different distance values are obtained. When extracting feature sequences, considering that the time spans are unequal, we calculate the distance values of different time spans, such as 0.9/1.0 (equal length) /1.1, etc., of each starting point.

Corresponding to the above examples, we can get the distance value of each starting point:

[00:27.98] (0.9) =5, (1.0) =9, (1.1) =13

[00:31.26] (0.9) =42, (1.0) =28, (1.1) =52

[00:34.89] (0.9) =64, (1.0) = 55, (1.1) =59

From the above, we take the smallest distance [00:27. 98] (0.9) = 5 and believe that the hummed song and the original song are most similar in this position. The time scale is 0.9. Make similar calculation on the 84 songs in the database one by one and get the minimum distance in these 84 songs and then rank the 84 distances from far to near, the following results can be obtained:

Rank	Singer's Name	Song Title
1	Christine Hsu	I Only Care about You
2	Jacky Cheung	Steal Your Heart
3 …	Faye Wong	Fall by the Wayside

PERFORMANCE EVALUATION

The experiment system conducts retrieval tests on over 70 hummings by 6 hummers. When testing, first of all, we provide the repertoire list of the song database to participants, to let them choose songs freely. The statistical results of the test are shown in Figure 2.

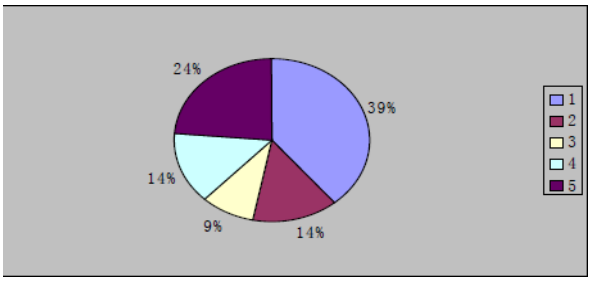

Figure 2 Test Results

In Figure 2, 1: rank=8∼15; 2: rank=2∼5; 3: rank=1; 4: rank>40; 5: rank=16∼40.

These test results are poorer than some methods based on MIDI format and melody features. In Literature [4], the recognition rate of rank=1 can reach 82%, but as what this article deals with is WAV files, moreover, there is no limit to the hummer's humming, (Literature [2] requires that hummers can only hum from the beginning), these results are acceptable. The probability of rank<15 has exceeded 60%. At the same time, the feature of retrieved results constraint has been added to the actual system. Users can search the retrieved results after ranking by means of fuzzy search, according to the author, title, album, lyrics and other text attributes of the song, and thereby assists the screening of results to make them conform to the requirements.

ACKNOWLEDGEMENT

This paper is the subsidy project of the State Key Program of National Natural Science of China (Grant No. 51479159); The Education Department of Fujian province science and technology research project (JB12155); Fujian branch of research project of National Occupation Education Research Institute (GZM13012); Project supported by special fund of central government financial support for the development of local colleges and universities (Min Cai –Jiao-Zhi-2014-50). I would like to express my gratitude here!

REFERENCES

[1] Foote, J."An Overview of Audio Information Retrieval"Institute of Systems Science, National University of Singapore,2013

[2] W.H. Xu, M.Y. Gao and Z.X. Zhang, "Intuitive Singing Input Search Engine". National Tsing Hua University, Taiwan.

[3] A.Ghias, J.Logan, D.Chambetlain, B.C. Smith. "Query by Humming-Musical Information Retrieval in an Audio Database", ACM Multimedia, San Francisco,2012.

[4] Roger J. McNab, Lloyd A. Smith, Melody Transcription for Interactive Applications, Department of Computer Science, University of Waikato, New Zealand

[5] H.S. Chi et al. *Digital Voice Signal Processing*. Electronic Industry Press, 2013.