CONTENT-BASED RETRIEVAL OF MP3 SONGS FOR ONE SINGER USING QUANTIZATION TREE INDEXING AND MELODY-LINE TRACKING METHOD

Tsung-Han Tsai and Jui-Hung Hung

Department of Electrical Engineering
National Central University, Chungli, Taiwan, ROC
E-mail : {han,remix@dsp.ee.ncu.edu.tw}

ABSTRACT

In this paper, the content-based retrieval with audio examples based on MP3 digital music archive for one singer is considered. In our proposed method, the Sub-Band Coefficients (SBC) in a MP3 frame is used for feature extraction. Both Quantization-Tree indexing (QT) approach and the Melody-Line Tracking (MLT) approach are proposed for indexing MP3 objects. Finally, a two-stage comparison method is used to measure the similarity between MP3 objects. We also employ the measurement of the similarity between query sample and database items for a singer. Evaluations on a content-based MP3 retrieval system are performed. Experimental results show that our proposed approach can achieve good performance and high accuracy.

1. INTRODUCTION

As there is a growing amount of MP3 music object available on the Internet nowadays, the problems related to content-based audio retrieval are getting more attention recently. As many research works were done on the content-based retrieval of image and video data, less attention was received to the content-based retrieval of audio data. Traditionally, the audio objects are treated as a long sequence described by texts. Except for some descriptive characteristics such as song name, file format, or sampling rate, the audio object are treated as an opaque large object in which its semantics is omitted. Although many significant feelings can be aware of human beings when listen to an audio object, it is difficult to extract these feelings from the raw data of the audio object.

The automatic segmentation, indexing and retrieval of audiovisual data have become a significant technique in this field. Many applications such as audio segments delivering for professional media production, storage and retrieval of multimedia databases on educational application and audio-on-demand system are very welcome [1][2].

Most of the multimedia indexing approaches store and retrieve multimedia data based on keyword annotations. There are many restrictions on these keyword-based methods, especially in the content-based multimedia databases. First, it is difficult to describe the content of multimedia object by human languages. Second, manual annotation of text phrases is prohibitively laborious for a large database in terms of time and efforts. Third, since it is difficult to describe the same multimedia object with a complete set of keywords, users have different interests in the same multimedia object. Finally, even if all relevant object characteristics are annotated, problem may arise due to the use of different indexing languages or vocabularies by different users.

In this paper we focus on the solution and implementation of a content-based retrieval system. Our realization of the database is based on singer with his/her all published songs. Since nowadays MP3 songs are the very popular audio media we can access, our realization is focus on this compressed domain with some challenges. We utilize content-based retrieval [3][4][5] by audio example to access multimedia database in an efficient and practical way.

2. THE PROCEDURE OF MP3 DECODING

In this section, we describe the procedure of MP3 Decoding. For content-based audio retrieval systems, the features should be extracted from the compressed domain raw data. So we analyze the MP3 decoder. In MP3 decoder, the basic unit of an MP3 bitstream is a frame. An MP3 bitstream consists of 1152 samples. A typical MP3 music object is sampled at 44.1 kHz. Therefore, there are 0.02612 seconds per frame. An MP3 bitstream is unpacked and dequantized, frame by frame, into MDCT (modified discrete cosine transform) coefficients. The 576 MDCT coefficients are then mapped into the polyphase filter coefficients (32 subbands) using inverse MDCT. Finally, these polyphase filter coefficients are synthesized into PCM audio. Both the MDCT coefficients and the polyphase filter coefficients can be used to compute the MP3 frames.

According to the time consumption analysis of the total functions in the MP3 decoder shown in Fig. 1, we observed that the function of IMDCT and Synthesis Poly-phase Filter Bank are more time consuming. Thus, we extract the MDCT coefficients in the position which is between the function of Alias-Reduction and IMDCT in order to reduce

the whole processing time. This not only reduces the time consumption but also achieve fine retrieval accuracy.

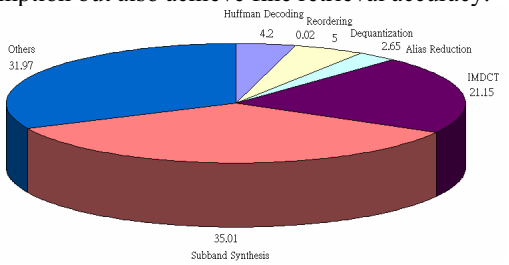

Fig. 1 The diagram of time consumption in MP3 decoder

3. PROPOSED APPROACH

In the preceding section, the fundamental concept and the general architecture for content-based retrieval of audio example based on MP3 songs are introduced. As shown in Fig. 2, the entire system can roughly be divided into two parts. The right part is the client part which is the interface provided for users operation, and the left part is the server part performed the database construction.

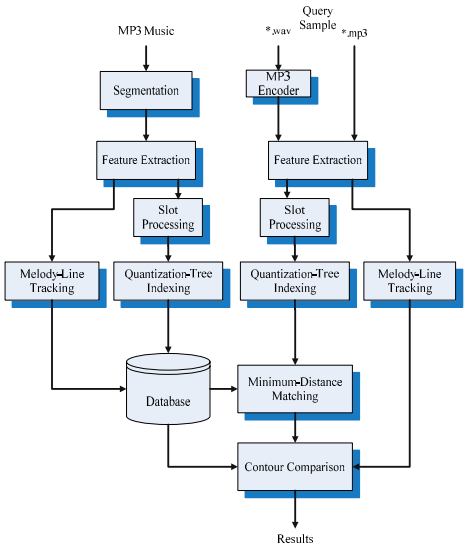

Fig. 2 The architecture of proposed retrieval system.

3.1 FEATURE EXTRACTION

Generally, audio features can be classified into two types (physical features and perceptual features). Physical features refer to mathematical measurements computed from sound wave, such as energy function, spectrum, fundamental frequency, etc. Perceptual features are subjective terms which related to perceptual sounds by human beings, including loudness, brightness, pitch, and so on.

In our paper, we choose the spectrum energy as physical features. We extract the Sub-Sand Coefficients (SBC) right after the processing of Alias-Reduction and before the IMDCT operation.

3.2 SLOT PROCESSING

Here, we discuss two requirements for content-based retrieval of music data. One is **"Unrestricted Search"** (User can search from any position in music objects). Another requirement is the **"input fault tolerance"**. (The input of user does not coincide with the patterns in music objects of database).

According to the first requirement described above, users can pick a segment from a song arbitrarily. This exists a variable length problem (The length of query sample is different from the database items) in each query process. We use the slot processing method to solve this problem. We segment the MP3 music of database and variable length query sample into several slots.

TABLE I shows the analysis of slot length and accuracy. We observe that the smaller value of slot length causes high accuracy. On the contrary, the bigger value of slot length causes low accuracy. Nevertheless, the smaller value of slot length generates numerous data which unrelated to query sample. It is a tradeoff between accuracy and data size. Therefore, we choose 40 frames as a slot for high accuracy according to the experiments. If one slot is shorter than 40 frames, it will be truncated.

TABLE I. The analysis of slot length and retrieval accuracy.

Length	Top-1	Top-3	Top-5	Top-10	Top-20
10	87%	93%	100%	100%	100%
20	87%	93%	93%	100%	100%
30	73%	80%	93%	100%	100%
40	73%	80%	86%	93%	100%
50	67%	73%	80%	96%	100%
60	40%	47%	53%	67%	100%

(Length: The value of slot length)
(Top-n: The order of song in retrieval results)

3.3 INDEX CONSTRUCTION

3.3.1 Quantization Tree Indexing

Assume that the sequence of slots of query sample and database items are $\{q_1, q_2,...,q_N\}$ and $\{s_1, s_2,...,s_N\}$, respectively. We exploited a tree-structured quantizer to convert slot sequences into indices. Analysis and statistics of all the extracted frame values are made to obtain the distribution. The slot sequences $\{q_1, q_2,...,q_N\}$ or $\{s_1, s_2,...,s_N\}$, where N is the slot length, will be put into tree-structured quantizer. Then, the distribution of all the frame values is divided into totally 29 regions. The frame value of a slot sequence will be classified through these specified regions one by one and then be assigned into a specified bin as shown in Fig. 3.

After completing all the classification of the frame values in a slot, the frame values which assigned into the same bin will be summed up. The sum in each bin will be gathered to form a histogram. Finally, the histogram in Fig. 4 can be

regarded as a vector which has 29 dimensions and we call this vector as a QT index.

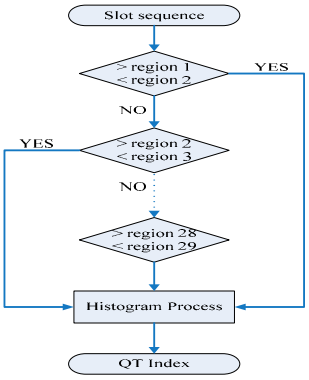

Fig. 3 The flow chart of Quantization Tree Indexing.

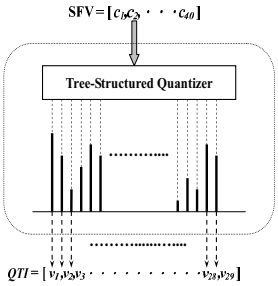

Fig. 4 The Diagram of Tree-Structure Quantizer.

(SFV: Frame value of subband coefficients)
(QTI: Quantization tree indexing vector)

3.3.2 Melody-Line Tracking

We consider the practicality of the retrieval system. The retrieval system should adopt variable types of query formats. Therefore, we employ MP3 encoder to re-encode the query sample into MP3 format in the beginning of the retrieval process. However, there is a problem after re-encode process. The feature values extracted from MP3 bit-stream will be different from that in database items and these feature values have variation after quantization tree indexing.

We use the Melody-Line Tracking (MLT) approach to solve the problem described above. MLT tolerates the deviation due to the re-encoding of query sample. The spectral energy of a song is utilized to constitute the melody contour of music, and then we use this melody contour to measure the similarity between query example and database items.

The melody contour is composed of several tone lines. Each of tone line represents a spectral energy of a music segment. First, we utilize the extracted feature values to obtain the spectral energy. Every 12 frames are divided into one tone length, and feature values of each tone length are summed up to obtain tone energy. Finally, a sequence with feature values is transformed into several tone lines. By the way, we compare the current tone energy with its preceding one. If the current tone energy is smaller than its preceding one, it is noted as "D", otherwise, it will be noted as "U". Thus, the sequence is converted into a string with two letters (U, D). For example, D, U, U, U, D, U, U notation.

In the process of the tone length selection, we observe that the accuracy of retrieval system is affected by the value of tone length. If the value of tone length is too large, a lot of similar retrieval results are obtained and along with high accuracy. On the other hand, if the value of tone length is too small, there might be almost no results founded. In order to determine an acceptable value, we evaluate the precision rate by (1) in TABLE II with different tone length. Then, the tone length can be derived by multiplying the accuracy rate of top-1 in TABLE II with the precision rate by (2). Finally, we choose 12 for tone length value because it has the maximum value.

$$\text{Precision rate} = \frac{\text{No. of retrieved reference that are relevant}}{\text{No. of reference that are retrieved}} \quad (1)$$

$$\text{Tone length} = \max\{\text{accuracy rate}|_{\text{top-1}} * \text{precision rate}\} \quad (2)$$

TABLE II. The decision of tone length.

Length	5	12	15	20	25	40
Precision rate	50%	50%	33%	25%	14%	9%
Top-1 Accuracy rare	20%	67%	67%	73%	80%	87%

4. SIMILARITY MEASUREMENT

Basically, two matching stages are used in similarity measurement. The flow chart is shown is Fig. 5. The first stage is minimum-distance matching and the second stage is contour comparison.

4.1 Minimum-Distance Matching

In this stage, the QTI are used to calculate the distance between the audio segment of query sample and database items. The similarity is dependent on the computed distance. Furthermore, there is a clustering property where the values of QTI are clustered together in a few dimensions. Based on this phenomenon, we propose a weighted measurement approach to weight several dimension in specified area. The following steps are performed:

Step1: Find out the maximum value of dimension in QTI.
Step2: Center with the maximum value in step1 and extend 3 dimensions in both left side and right side.
Step3: After step 2, 7 dimensions are assigned to weight as an emphasized area.

Step4: The following equation (3) is exploited to compute the distance between query sample and database items, where d_n is the distance between query sample and database items, P_n, and Q_n is two distinct QTI, W_i is the weight function.

$$d_n = (P_n, Q_n) = \sum_{i=0}^{28} W_i \mid P_{ni} - Q_{ni} \mid \quad (3)$$

Depending on these distances, the similarity of the retrieved music among the database is arranged. The weighted measurement approach can enhance the precision rate in the retrieval process.

4.2 Contour Comparison

In this stage, as shown in Fig. 5, after the minimum-distance matching, the contour comparison is followed to perform the second stage comparison. The melody-line information of query and database items are used to determine how similar they are. As described in preceding subsection, when input is not in MP3 format, it has to be re-encoded to MP3. This will cause a deviation and result in low retrieval accuracy. The second stage matching of contour comparison is used to improve the whole retrieval accuracy after similarity measurement.

As finishing the measurement of similarity, the searching results will be arranged according to the similarity degree and listed as the song titles to user.

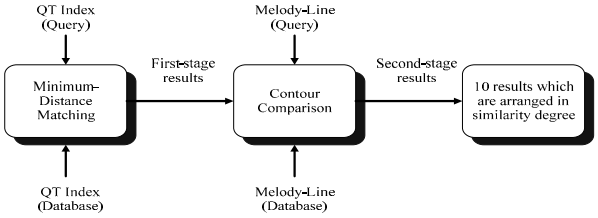

Fig. 5 Two stage similarity measurement scheme.

5. EXPERIMENTAL RESULTS

In order to examine the robustness of our proposed algorithm, we choose 176 MP3 songs of one singer as our database. All these MP3 songs are singing by one female singer (*Mariah Carey*). The experiment is performed in the form of query-by-example. We segment three arbitrary length pieces from a song as query examples to perform retrieval. We implement a GUI interface of content-based MP3 retrieval system on PC as shown in Fig. 6. For the results some of *Mariah Carey's* songs such as "Shake it off", "Say something" and so on are also retrieved with the top-1 ranking. TABLE III shows all the possible ranking, the corresponding number of songs and the accumulated accuracy rate, respectively. The top-n which means the correct song is listed in the n-th position of results. From this total 176-song database, at least within the rank of top-7 all songs can be acquired successfully.

Fig. 6 The GUI interface of proposed content-based MP3 retrieval system

TABLE III. The ranking of one female singer (*Mariah Carey*)

Top-1	Top-2	Top-3	Top-4	Top-5	Top-6	Top-7
10	14	41	65	30	15	1
5.6%	13.6%	36.9%	73.8%	90.9%	99.4%	100%

6. CONCLUSIONS

In this paper, the content-based retrieval of audio example based on MP3 digital music archive for one singer is considered. In our proposed approach, the sub-band coefficients in a MP3 frame are used. These values are extracted from the MP3 decoder to compute the MP3 features for indexing the MP3 objects. A quantization-tree and melody-line tracking method are also proposed for indexing MP3 objects. Experiment results show that the accuracy at least top-4 can achieve about 74%, and an accuracy of about 90% can be achieved within top-5.

7. REFERENCES

[1] E. Wold, T. Blum, D. Keislar, and J. Wheaton,"Content-Based Classification, Search, and Retrieval of Audio", IEEE Multimedia, Vol. 3, No. 3, pp. 27-36, Fall 1996.

[2] T. Zhang and C. C. Jay Kuo, "Content-Based Audio Classification and Retrieval for Audiovisual Data Parsing", Springer, December 2000.

[3] Faisal I. Bashir, S. Khanvilkar, D. Schonfeld, and A. Khokhar, "Multimedia Systems: Content-Based Indexing and Retrieval Sec. 4, Chapter 6 in 'The Electrical Engineering Handbook', Academic Press, October 2004.

[4] I. Karydis, A. Nanopoulos, Apostolos N. Papadopoulos, and Y. Manolopoulos, "Audio Indexing for Efficient Music Information Retrieval", The Proceedings of the 11th International Multimedia Modelling Conference, pp. 22 – 29, January 2005.

[5] C. Yang, "Peer-to-Peer Architecture for Content-Based Music Retrieval on Acoustic Data", The Proceedings of the 12th international conference on World Wide Web, Budapest, Hungary, pp. 376 – 383, May 2003.