Classical Music Retrieval Based on Accumulated Path Similarity in AAC Compression domain

Yung-Ting Chung
Central University
Jhongli, Taiwan
+886-3-4227151 #57969
ytchung@vaplab.ce.ncu.edu.tw

Tai-Ming Chang
Central University
Jhongli, Taiwan
+886-3-4227151 #57969
tmchang@vaplab.ce.ncu.edu.tw

Pao-Chi Chang
Central University
Jhongli, Taiwan
+886-3-4227151 #34461
pcchang@ce.ncu.edu.tw

ABSTRACT
This work focuses on classical music cover song retrieval in AAC compression domain. In our proposed system, the modified discrete cosine transform coefficients (MDCT) are directly used to represent 12-dimensional chroma feature without a fully decoding process, which can save about 70% decoding complexity. The MDCT coefficients are processed for enhancing chord characteristics and the dot-product operation is used to calculate the chroma similarity matrix. Finally, the similarity score between two songs is evaluated by counting the similarity values along optimal path in the chroma similarity matrix. The proposed system can reach 97% of precision and save over 90% matching time compared with traditional approach operated in the waveform domain.

Categories and Subject Descriptors
H.3.4 [**Information Storage and Retrieval**]: Systems and Software

General Terms
Algorithms, Performance, Design, Experimentation.

Keywords
Classical music, cover song, AAC, compression domain, content-based music retrieval

1. INTRODUCTION
With the rapid development of Internet and multimedia compression techniques, people can easily download or share multimedia data through networks. Therefore, how to efficiently retrieve query from a huge multimedia database becomes an important issue in the present research. The most commonly used method of search engines is through textual labels. However, the label created by people may be ambiguous or even with errors. This problem in retrieving classical music occurs more often than pop music. Accordingly, content-based music retrieval (CBMR) was proposed to extract the feature from music content as the representation to overcome the errors in labeling from human being. In recent years, CBMR has become a high-profile research area in multimedia applications. Ellis *et al.* extracted beat-synchronous chroma features and used cross-correlation to compute similarity between songs [1]. Serra *et al.* used an enhanced version of chroma features, and utilized dynamic programming local alignment algorithms to measure the similarity scores [2]. Kim used note temporal changes (delta chroma feature) and evaluated covariance matrix to measure similarity of songs [3]. Chuan extracted an enhanced chromagram and used Bayes classifier to calculate similarity [4]. Bertin-Mahieux *et al.* used a fingerprinting-inspired model to conquer large cover song dataset recognition [5]. However, the preceding related works all extracted features from WAV format files. As a matter of fact, most digital audios transmitting on Internet have already been compressed. Using above-mentioned retrieval methods, a fully decoding process and time-frequency analysis before feature extraction are required as shown in Fig.1. On the contrary, the most useful information is preserved in compressed audio files. It is a reasonable approach that the compressed files can be partially decoded to extract features directly. Hence, in this paper, we propose a system architecture that retrieves classical music cover song in Advanced Audio Coding (AAC) coded files.

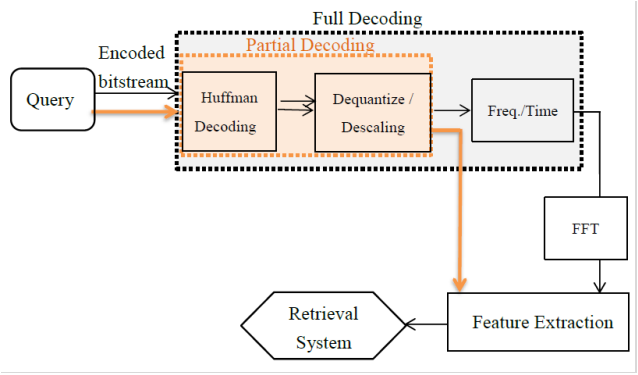


Figure 1. Music information retrieval system

The rest of this paper is organized as follows. The detail of proposed method is described in Section 2. In Section 3, experimental results are presented. Finally, the conclusions are stated in Section 4.

2. THE PROPOSED CLASSICAL MUSIC COVERSONG RETRIEVAL SYSTEM
The proposed system architecture is illustrated in Fig. 2. First, AAC audio file is decoded partially to get MDCT coefficients, which are pre-processed by filtering suitable ranges of frequency and magnitude. Then the MDCT coefficients are refined to 12-dimensional chroma feature. In the matching step, the dot-product

ICIMCS'14, July 10–12, 2014, Xiamen, Fujian, China.

operation is used to generate the chroma similarity matrix and the dynamic time warping (DTW) operation calculates the similarity score. Finally, the similarity weighted mean between the original and cover song is evaluated along the optimal similarity accumulated path.

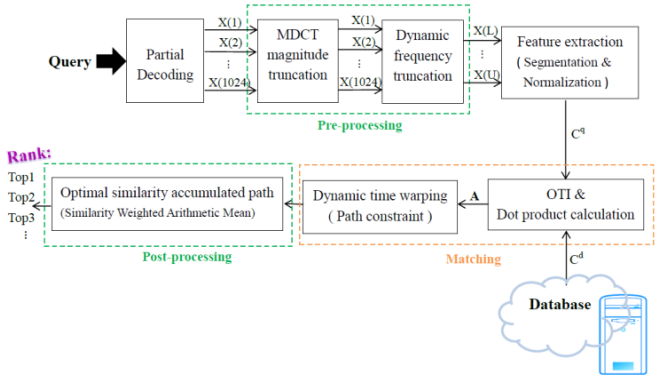


Figure 2. Diagram of proposed system architecture

2.1 Pre-processing

In the AAC audio encoding process, the time-frequency analysis tool is the modified discrete cosine transform (MDCT) [6], which is defined as follows,

$$X(k) = 2\sum_{n=0}^{N-1} x(n)cos\left(\frac{2\pi}{N}\left(n+\frac{N}{4}+\frac{1}{2}\right)\left(k-\frac{1}{2}\right)\right) \quad (1)$$

$$for\ 1 \le k \le \frac{N}{2}$$

where x is windowed input sequence, n is sample index, k is spectral coefficient index, and N is the length of the transform window.

2.1.1 MDCT magnitude truncation

To raise the resolution of chroma histogram and reduce the interference of noise and low magnitude feature in matching result, this system normalizes and truncates the MDCT coefficients magnitude for each frame as defined in (2) and (3).

$$X_{normal}(k,l) = \frac{X(k,l)}{\sqrt{\sum_{k=1}^{\frac{N}{2}} X(k,l)^2}} \quad (2)$$

$$\text{if } X_{normal}(k,l) \le T_{mag},\ then\ X(k,l) = 0 \quad (3)$$

where l represents frame index. In (3), the magnitude of normalized MDCT coefficient below a threshold T_{mag} would be truncated.

2.1.2 Dynamic frequency truncation

Semitone frequency in low pitch class needs relatively high frequency resolution, which may cause many low MDCT coefficients index to map into wrong chroma bins. Considering the sampling rate of songs in our database is 44.1 kHz, the system chooses 260Hz as the lower bound. In addition, the harmonics of pitch also cause a wrong mapping problem at high multiples of pitch. The wrong mapping starts from the third harmonic frequency. Hence the system truncates high frequency dynamically by observing the whole song frequency energy distribution [7]. If the energy average distributes in 260~2 KHz (Octave 4-6), i.e., the energy in high frequency range is significant, the system treats the frequency beyond 2 KHz as harmonics and chooses 2kHz as upper bound. Otherwise, 1 kHz (Octave 4-5) is chosen as the upper bound.

2.2 Feature Extraction

In feature extraction step, we calculate each frequency component of the pre-processed MDCT coefficient and determine which chroma bin b it belongs to [8]. The chroma bin is calculated as defined in (4), where f_k is center frequency of MDCT coefficient, $B(b)$ is the set of MDCT coefficients which belongs to chroma bin b, and the f_0 is set to be 16.352Hz in our experiment.

$$B(b) = \left\{k \middle| mod\left(round\left(12\,log_2\left(\frac{f_k}{f_0}\right), 12\right) + 1\right) = b\right\} \quad (4)$$

Chroma features record the energy intensity associated with each of the 12 semitones and the energy in the same notes is folded together. Therefore, chroma feature matrix H can be obtained by (5):

$$H(b,l) = \sum_{k \in B(b)} X(k,l) \quad (5)$$

For reducing matching time, a number of frames are merged into a segment as (6)

$$C(b,m) = \frac{1}{P}\sum_{j=(m-1)\times P+1}^{m\times P} H(b,j) \quad (6)$$

where P is segment size, m is segment index, and C is segmentation chroma feature matrix. Empirically, the segment size is determined by how many frames a segment contained in around one second of time.

Finally, the chroma feature matrix is shown in Fig. 3. The horizontal direction expresses segment length in the entire piece, and the vertical direction expresses the energy of each semitone.

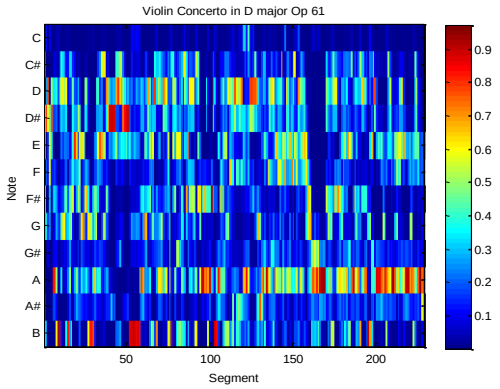


Figure 3. Chroma feature histogram

2.3 Matching

In the matching stage, we first utilize dot-product calculation to compute chroma similarity matrix of the similarity between query and reference songs. Then the dynamic time warping is used to calculate similarity score. The details of chroma similarity matrix and dynamic time warping are described in the following subsections.

2.3.1 Chroma similarity matrix

Because the query and reference song may have different keys, the query is transposed to the same key as the reference. The system utilizes the optimal transposed index (OTI) to calculate the key distance between two songs [2]. The OTI function is defined in (9).

$$k = arg \max_{0 \le J \le 11} \{mean(C^d) \cdot circshift(mean(C^q), J)\} \quad (9)$$

where "$\cdot$" indicates a dot product and *circshift()* is a function that rotates the vector h^q with J positions. After estimating the key distance by OTI, we transpose the chroma feature matrix of the query into the same key as in reference song.

$$C'_q(b,m) = C_q(mod((b+k)/12), m) \quad (10)$$

After the transposition, the query and the reference song are in the same key. And then, the chroma similarity matrix *A* is calculated as (11):

$$A = C'_q \cdot C_d^T \quad (11)$$

The chroma feature matrix is normalized so that all entries in chroma similarity matrix are between 0 and 1. The query/reference song of chroma similarity matrix is shown in Fig. 4. The more the entry close to 1, the more the similarity between two segments is.

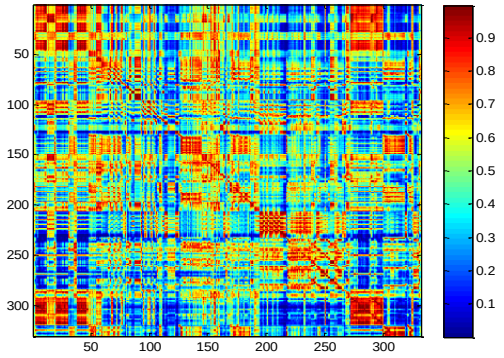


Figure 4. Chroma Similarity Matrix

2.3.2 Dynamic time warping (path constraint):

The chroma similarity matrix is used as input for DTW algorithm to calculate similarity grade. First, we initialize a (*M*+1)×(*E*+1) matrix *G* in which entries in the first row and column are zero and the second row and column are equal to the entries in first row and column of matrix A. Then, we calculate other entries in matrix *G* through a recursive formula as defined in (13). In order to reduce DTW algorithm computational time, a path constraint is given to DTW computational region which is obtained empirically and defined in (14).

$$G(i,j) = \begin{cases} D & \in Y(i,j) \quad for\ i = 3,\dots,M+1 \\ skip & otherwise \quad j = 3,\dots,E+1 \end{cases} \quad (12)$$

$$D = max \begin{cases} G(i-1,j-1) + A(i-1,j-1) \\ G(i-1,j-2) + A(i-1,j-1) \\ G(i-2,j-1) + A(i-1,j-1) \end{cases} \quad (13)$$

$$Y = \left\{ (i,j) \middle| 4Mi - 3Ej > -\frac{3}{5}ME \cap 3Mi - 4Ej < \frac{3}{5}ME \right\} \quad (14)$$

The output of DTW algorithm is shown in Fig. 5.

2.4 Post-processing

In post-processing step, the similarity score is evaluated from matrix *G*. Different to [2], we count the values which distribute on the optimal accumulated path and sum up the quantity of each scale as the similarity score. In Fig. 5, the max energy point represents a starting point to find its accumulated path, as the black line shown in Fig. 6(a) (c). Each input query finds an optimal accumulated path with each reference song in the database. Then we collect statistics of the entries in chroma similarity matrix *G* on optimal accumulated path. In Fig. 6(b) and 6(d), the more the histogram distribution is on the right side, the more similar between two segments is.

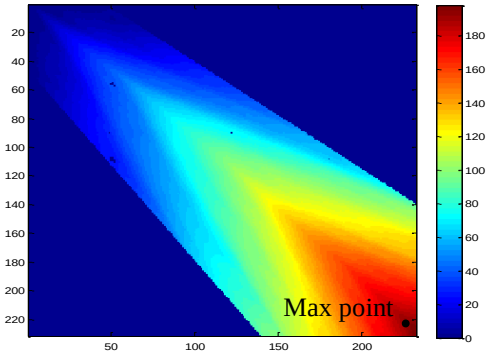
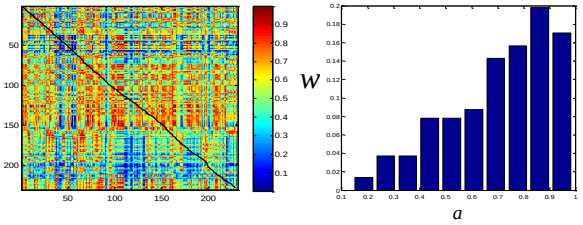
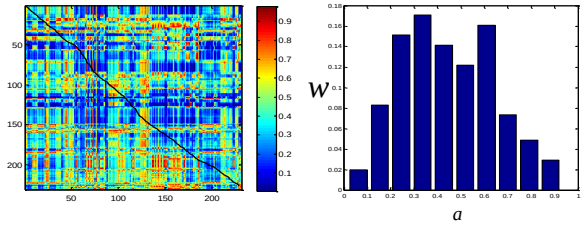


Figure 5. DTW algorithm output



(a) (b)



(c) (d)

Figure 6. (a) Original/Cover chroma similarity matrix
(b) Histogram of (a)
(c) Original/Non-cover chroma similarity matrix
(d) Histogram of (c)

Next, we calculate the weighted arithmetic mean of the entries histogram distribution as final similarity score:

$$W = \sum_i w(i)a(i) \quad (15)$$

$$Rank_1 = arg \max_{d \in database} (W_{q,d}) \quad (16)$$

where a denotes the scale from 0 to 1 with a space 0.1, w is the proposition of a. Finally the retrieving result, Top-N ranks, will be returned.

3. EXPERIMENTAL RESULTS

The experiments of the proposed system are discussed in two parts. First, the decoding computational complexity is analyzed.

Second, the performance of our proposed system is evaluated and compared with the related work.

3.1 Computational Complexity Analysis

The computational complexity of AAC decoder is analyzed in [9]. Fig. 7 summarizes the complexity analysis results of the AAC decoder. In our proposed system, it partially decodes MDCT coefficients from AAC files without fully decoding process, and it could save about 70% computational complexity.

3.2 System Performance Evaluation

Our testing database consists of 985 classical music songs which are 124 queries and 861 covers. The sampling rate of all songs is 44.1 kHz. The number of covers per song is ranging from 2 to 7. In experiments, we evaluated the retrieving accuracy of the proposed system and compared with [1] which needs fully decoding process for the encoded files as in traditional retrieval systems.

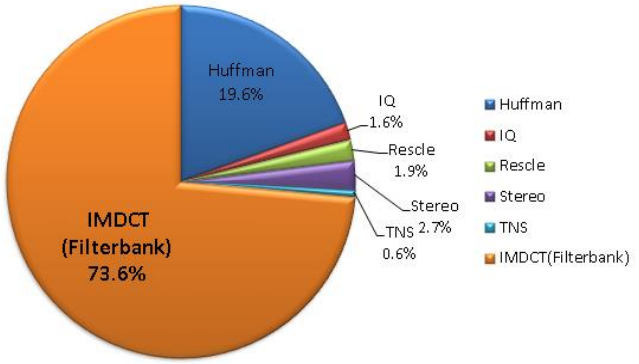


Figure 7. Complexity analysis of AAC decoder

Table 1 shows the system performance. The amount of the correct cover version is recall in Top-N, and the Mean Reciprocal Rank (MRR) reflecting the rank of the first correctly identified cover for each query are presented. As shown in Table 1, the proposed method can retrieve 118 covers in Top-1 and MRR reaches 0.96. In addition, in the matching process, the proposed method can save over 90% time compared with [1]. Secondly, we use all of the songs we collected to evaluate the system performance. The experiment results show that proposed method can reach Precision of 97%, Recall of 0.87, and an F-measure of 0.92. The proposed system can achieve superior performance with much less computational complexity.

Table 1. System performance evaluation

	Ellis system [1]	**Proposed system**
Top-1	65/124	118/124
Top-3	76/124	119/124
Top-5	77/124	119/124
Top-10	82/124	120/124
MRR	0.57	**0.96**
Matching Time(sec.)	513.573	**28.590**
Saving Time		**94%**

4. CONCLUSIONS

This work utilizes simple and low-complexity procedures to enhance the system performance. The system we proposed only partially decodes MDCT spectral coefficients without full decoding. The pre-processing that skips low energy component and limits the frequency range can reduce the computational complexity and promote matching results. We utilize dot-product calculation to get chroma similarity matrix and calculate similarity weighted mean by finding the optimal similarity accumulated path. Since classical music is much more rigorous than popular music, it provides less freedom for cover songs. Hence the path constraint in dynamic time warping eliminates a lot of inefficient searches. As a result, our system can save over 90% matching time and reach Precision of 97%.

5. REFERENCES

[1] Ellis, D. P. W. and Poliner, G. E. 2007. Identifying 'Cover Songs' with Chroma Features and Dynamic Programming Beat Tracking. In *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing* (Honolulu, Hawaii, U.S.A., April 15-20, 2007).

[2] Serra, J. and Gomez, E. 2008. Audio cover song identification based on tonal sequence alignment. In *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing* (Las Vegas, Nevada, U.S.A., March 30- April 4, 2008).

[3] Kim, S., Unal, E., and Narayanan, S. 2008. Music fingerprint extraction for classical music cover song identification. In *Proc. Int. Conf. on Multimedia and Expo*. (Hannover, Germany, June 23-26, 2008).

[4] Chuan, X. 2012. Cover song identification using an enhanced chroma over a binary classifier based similarity measurement framework. In *Proc. Int. Conf. on Systems and Informatics (ICSAI)* (Las Vegas, Nevada, U.S.A., May 19- 20, 2012).

[5] Bertin-Mahieux T. and Ellis, D. P. W. 2011. Large-scale cover song recognition using hashed chroma landmark. In *Proc. IEEE Int. Conf. on Applications of Signal Processing to Audio and Acoustics (WASPAA)* (New Paltz, NY, Oct. 19-20, 2011).

[6] International Organization for Standardization 1997. Information Technology - Generic coding of moving pictures and associated audio information - Part 7: Advanced Audio Coding (AAC). *ISO/IEC 13818-7*.

[7] Muller, M., Ellis, D. P. W., Klapuri, A., and Richard, G. 2011. Signal Processing for Music Analysis. *IEEE Journal of Selected Topics in Signal Processing*, 5, 6 (Oct. 2011), 1088-1110.

[8] Ravelli, E., Richard, G., and Daudet, L. 2010. Audio Signal Representations for Indexing in the Transform Domain. *IEEE Transactions on Audio, Speech, and Language Processing*, 18, 3 (Mar. 2010), 434-446.

[9] Tsai, T. H. and Liu, C. 2007. A Configurable Common Filterbank Processor for Multi-Standard Audio Decoder. *IEICE Transactions on Fundamentals of Electronics, Communications and Computer Sciences*. 90, 9 (Sep. 2007), 1913-1923.