International Journal of Information, Control and Computer Sciences
ISSN: 2517-9942
Vol:3, No:2, 2009

Effective Digital Music Retrieval System through Content-based Features

Bokyung Sung, Kwanghyo Koo, Jungsoo Kim, Myung-Bum Jung, Jinman Kwon, Ilju Ko

***Abstract*—**In this paper, we propose effective system for digital music retrieval. We divided proposed system into Client and Server. Client part consists of pre-processing and Content-based feature extraction stages. In pre-processing stage, we minimized Time code Gap that is occurred among same music contents. As content-based feature, first-order differentiated MFCC were used. These presented approximately envelop of music feature sequences. Server part included Music Server and Music Matching stage. Extracted features from 1,000 digital music files were stored in Music Server. In Music Matching stage, we found retrieval result through similarity measure by DTW. In experiment, we used 450 queries. These were made by mixing different compression standards and sound qualities from 50 digital music files. Retrieval accurate indicated 97% and retrieval time was average 15ms in every single query. Out experiment proved that proposed system is effective in retrieve digital music and robust at various user environments of web.

***Keywords*—**Music Retrieval, Content-based, Music Feature and Digital Music.

I. Introduction

RECENTLY, various IT devices (Mobile, PMP, Netbook, etc.) and IT infrastructure like high-speed wire/wireless internet has spread rapidly. With these streams, Digital Contents Market has been also more growing. In digital content market, various kinds of contents such like music, Film and e-learning are in circulation. Among them all, Music content is considered as Core content from distribution volume and market size point of view. While music content demands are increasing, it has been occurred demands about not only typical content distribution but also novelty services. According to these requirements, technology demands have been also requested for service development.

Now, requested technology demands are ‘Music search with music’ and ‘User resource recognition’. ‘Music search with music’ is very important technology for developing and diversification of typical retrieval interface. ‘User resource recognition’ is essential for Service application development, content contribution management and statistical work. These requested technologies are able to be built through content-based music retrieval system.

In this paper, we propose effective digital music retrieval system through extracted content-based features from music. Proposed system is divided into two parts like ‘Client Part’ and ‘Server Part’.

‘ Client Part’ is divided into Input stage for retrieval request and Result Output stage. First, Input stage is to input music file for retrieving music or music related contents. Even inputted music files by user are same contents, there is a strong probability that those have different digitalize standard. In this case, Hashing Key retrieval method is able to carry out high rates retrieval error. When people convert own digital music, they use various digitalize standards like In-coding, Sampling, Quantization, Bit-rate and etc. If one music file is changed in different digitalize standard, discrete values for representing waveform are able to be transformed. Content-based feature extracting with music files pre-processing can carry out music features with keeping consistency. We use First-order differentiated MFCC [1]-[4] as content-based feature. Through first-order differentiation computation, we could remove MFCC feature error when volume is changing. Second, Result output stage is to display user retrieval result. In this stage, music contents or music related contents are provided.

Serve part is divided into ‘Query String construct’, ‘Music Matching/Result’ and ‘Music Server’. First, Music Server consists of two Databases. One is Music-ContentsDB that Music contents and music related contents are stored. The other is FeatureDB that content-based features from all music content are stored. FeatureDB is used to retrieve works. Music-ContentsDB is used to output requested content from DB. Second, ‘Music Matching/Result’ searches same digital music between Inputted music and FeatureDB with similarity criterion. Similarity is measured by DTW [5]-[9]. DTW presents robustness at minute time axis error.

The structure of this paper is as follows. In section2, we present related work in content-based retrieval fields. In section 3, we propose structure and detail methods of our effective retrieval system through Content-based Feature. In section 4, we describe our experiments and results. In section 5, we provide some conclusions and suggest future works.

B. Sung is with the Department of Media, Soongsil University, Seoul, Korea (e-mail : ivsinger@ssu.ac.kr)

K. Koo is with the Department of Media, Soongsil University, Seoul, Korea (e-mail : khkoo@mog.kr)

J. Kim is with the Department of Media, Soongsil University, Seoul, Korea (e-mail : dotline@ssu.ac.kr)

M. Jung is with the Department of Media, Soongsil University, Seoul, Korea (e-mail : nzin@ssu.ac.kr)

J. Kwon is with the Department of Media, Soongsil University, Seoul, Korea (e-mail : lovekpo@gmail.com)

I. Ko is with the Department of Media, Soongsil University, Seoul, Korea (e-mail : andy@ssu.ac.kr)

II. Related Work

Content-based Music Retrieval methods are able to be branched mainly two parts. One is a method to use MIDI and Tab information. The other is a method to use extracted information through music spectral analysis.

For retrieving musical information, it is simple and easy to use MIDI [10]-[12] & Tag information [13], [14]. MIDI has musical information like tempo and chord. Tag information consists of objective music information like composer, song title and release date. In spite of that, it is hard to get good retrieval result when you retrieve digital music through using MIDI and Tag. You are be able to retrieve only music that MIDI and Tag information is stored.

Retrieval method to use extracted information from music waveform stays basically DSP (Digital Signal Processing) Technology. Unlike method to use MIDI, DSP extracts musical information from music waveform through various extracting algorithm. That is, it is possible to apply all kind of digitalized music file (MP3, OGG, WMA and etc.) that have spectral information. By applying DSP technology, Pitch [15], Timbre [16], Harmony [17] and etc are extracted. Then those are used as elements of retrieval.

Currently, Melody-based retrieval [18], [19] and Humming based retrieval are active research topic in content-based retrieval fields.

Melody-based retrieval is to index melody of all music then to retrieve. Melody is categorized information as mid-level information. In Classical Music that record performing sounds of some instruments, to apply melody-based retrieve is very easy. On the contrary, it is occurred high rate error when melody-based retrieve is applied to pop style music. This is a weakness example of Melody-based retrieval.

Humming-based retrieval is to understand pitch of human voice and retrieve digital music through understand information. An example of this is MIDOMI[20]. Humming-based retrieval is able to be applying to Mobile Device, Auto Karaoke system and various entertainment fields. All of that, it has limitation that retrieval result accurate depend on accurate of user humming.

III. Effective Digital Music Retrieval

A. Structure of Propose System

In Fig.1, we can see the overall structure of Effective Digital Music Retrieval System. Flow of proposed system is as follows. User inputs own digital music file. Inputted music takes Pro-processing and Feature extraction stages in Applet of Client. Extracted features are sent to Server. Then that is reconstructed as Query. Retrieval Query is used to find same content that satisfy similarity criterion in FeatureDB. FeatureDB consists of extracted Content-based features from all of music contents. After this stage, we can get Matched music information or related contents as retrieval result. These are sent to Applet of Client then displayed to user through browser.

In digital contents retrieval fields, one of the main problems is that many digitalize standard exists in inputted digital contents. In real contribution market, large volume music have different digitalize standard even if those are same music. For this reason, it has no fixed digitalized standard of inputted music. For solving this problem, pre-processing stage is very important. Pre-processing is carried out before feature extracting. It has two stage sequences. First stage is to transform different standard into same standard. Second one is to align time axis.

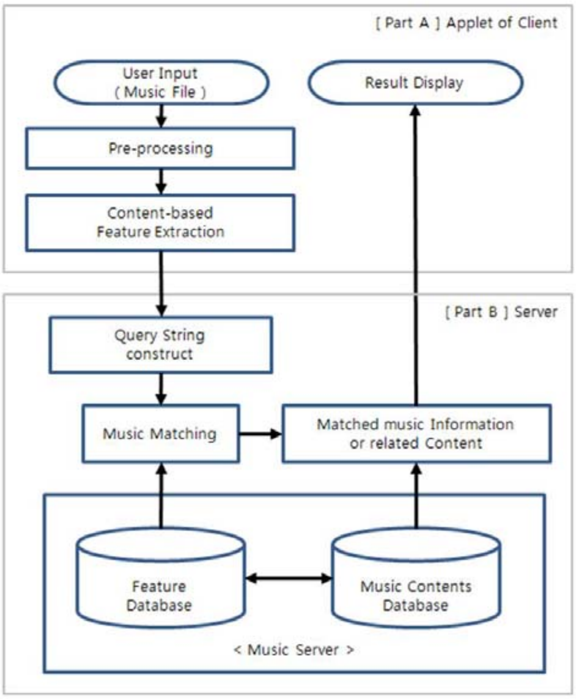


Fig. 1 This is a proposed structure for effective digital music retrieval through content-based features. It divides into two parts. Part A presents Applet of Client. Part B shows the structure of Server.

Content-based Features is fundamentally extracted through MFCC. MFCC has been popularly used in most voice, music and audio analysis and recognition fields because of good performance. None the worse, MFCC has a weak to depend on normalize status of waveform. If normalize status is different, MFCC feature values are presented in different value area. For solving these kinds of problems, we use differentiated MFCC as content-based feature.

Criterion of Music Matching is Similarity. It is measured by DTW. Music is a media that include time attribute. Result of similarity measuring depends on time axis change rates between two digital music files. DTW algorithm measures similarity while consider dynamically changing time axis. For this reason, DTW is suitable to measure similarity of media with time attribute.

B. Applet of Client

Applet of Client is divided into Music Input, Pre-processing and Feature extraction and Result Display.

Music Input is that user input own digital music file to retrieve through Applet. Inputted music files by user are generally compressed music files into various compressed standard. Sampling and Bit rate also are presented in various standards. If music files of same content have different digitalize standard, time code and discrete value of waveform are little bit different. That is, inputted music file and original music file have value difference if inputted music has numbers of converted history.

Pre-processing means blocked sequences of Clipping, Decoding, Silence removing and Framing. Time code gap occurs whenever music file is converted into different digitalize standard. Therefore, pre-processing is requested to minimize time code gap

First, clipping stage is to slice some part from whole music and make it as a clip. We don't need whole part of inputted music for identification recognizing. We therefore just take some part as a clip. In this paper, we just use a clip that is sliced from fixed time area. For making clip, we use 0~25sec part of all inputted music.

Second, Decoding stage is to transform clips that have different digitalize standard into criterion standard. Until don't weigh with retrieval result, clips are down sampled to minimum criterion standard. Minimum criterion standard is 44100Hz sampling rate, 16bit, Mono and PCM.

Third, Silence Removing [21], [22] stage is to remove no sound area in clip signal. Clip is front part of original music. Clip therefore has around 1~2sec silence sound. Silence sound is inputted while digital file is converted. In case of CD Ripping, Silence sound length depends on user setting. In case of reconverting, silence sound length is able to be changed. For this reason, it doesn't guarantee that same music files have same time code because of alterability of silence sound length. In this stage, Silence sound area is judged and removed by ZCR(Zero Crossing Rate)[23].

Fourth, framing stage is to segment music clip into frame by decided Hop size and Frame size. Minimum length of sound recognition ability of human is 20ms. We therefore decide 20ms as a frame length. Hop size is length of overlapping area. Hop occurs when clip is divided into frame with overlapping. Data loss occurs when music features are extracting. Hop is essential important to minimize this. In this paper, we set Hop size to 10ms half length of frame length.

Feature extracting means blocked sequences block of MFCC feature extracting, Grouping and Enveloping.

First, MFCC features are extracted from pre-processed frames. MFCC features are extracted into numbers of decided output channel. In this paper, output channel was fixed up 13. 13 features are 0~12 orders data. Fundamentally, MFCC derived expression is as follow Formula (1).

$$y_t^{(m)}(k) = \sum_{m=1}^{M} \log |Y_t| \cos(k(m - \frac{1}{2})\frac{\Pi}{M}),\ k = 0,\ldots,L \qquad (1)$$

Second, extracted MFCC features from frame sequences form a group. If we extract 13 features in every frame and store all of them in DB, Column numbers of DB Table will increase by geometric progression. This is an inefficient method from efficient retrieval system construct of view. Frames form a group into length that have no an effect on retrieval result. Then, it sums up all features of some groups. This is used as a Key value of that group. In this paper, one group is decided to sum of 50 frames and we extract Key value of group from that.

Third, Enveloping stage is to calculate change tendency of Key values. MFCC feature has a weak that feature values is able to be changed easily as volume level is changed. According to transformed data in converting process, gap of feature values is occurred even if there are same music files. Fig.2. shows waveform differences when same music files are converted into different digitalize standard and volume level. Human recognize same music when they hear converted music files with different status. Fig.3. show comparing Key values between extracted MFCC feature groups from two music file. We can see difference of feature value even if changing tendency is similar. For protecting these kind of error, we can compute changing tendency through First-order differentiate and use it. One feature presents a moment sound. Changing tendency of features has more information than feature sequences. Therefore, changing tendency of features is more suitable than feature sequences as a music feature with high discrimination.

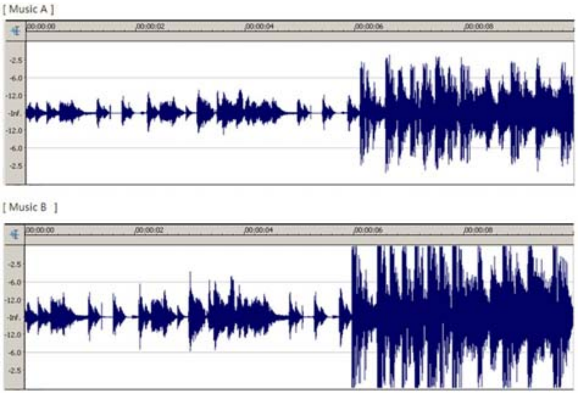


Fig. 2 MusicA and MusicB are same music. Those are converted with different digitalizing standard and different volume level. We can see some points of difference between waveform of two music files

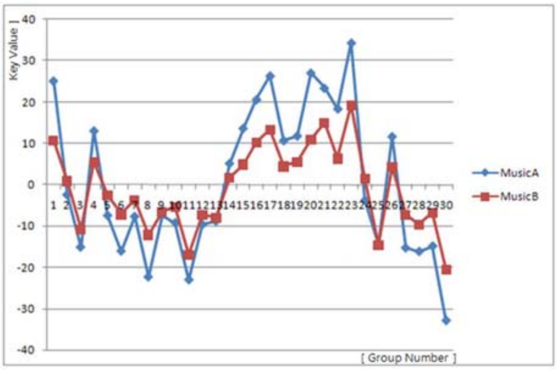


Fig. 3 This is a comparing graph of Key Value of grouped MFCCs of MusicA and Music B. MusicA and MusicB are same music with Fig.2. X axis means grouped MFCC group number, Y axis means Key value of MFCC group

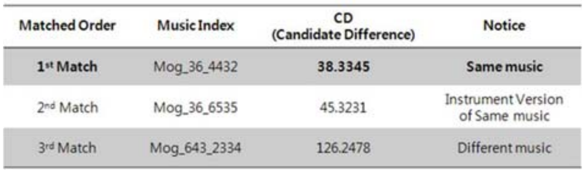
C. Server

In server part, Music Matching stage is core step. Music matching stage consists of candidate music retrieving with Query and similarity measuring by DTW.

First, candidate music retrieving uses Input Query for finding candidate list. Extracted feature data from input music is reconstructed into Query. Candidate list is carried out with difference between Query and each record of FeatureDB. In that process, difference means similarity. Values of each element in candidate list are in threshold limitation range. Formula (2) show derived expression of difference. CD (Candidate Difference) means difference between Query and feature sequences of each record. G means numbers of MFCC Group.

$$CD = \frac{1}{m}\sum_{m=1}^{G}(Query[m] - ServeFeature[m]) \quad (2)$$

Fig.4. show one example of candidate retrieval result. Retrieval result of candidate is three candidates that are ranked top 3 Candidate among difference values.



Matched Order	Music Index	CD (Candidate Difference)	Notice
1st Match	Mog_36_4432	38.3345	Same music
2nd Match	Mog_36_6535	45.3231	Instrument Version of Same music
3rd Match	Mog_643_2334	126.2478	Different music

Fig. 4 This is an example graph of retrieving result of retrieving candidate group. Results are three matched music. Query music is one Korean song(Singer: Wonder girls, Song title: Nbody). In this case, 1st match is same music, 2nd match is instrument version of same music and 3rd match is different music

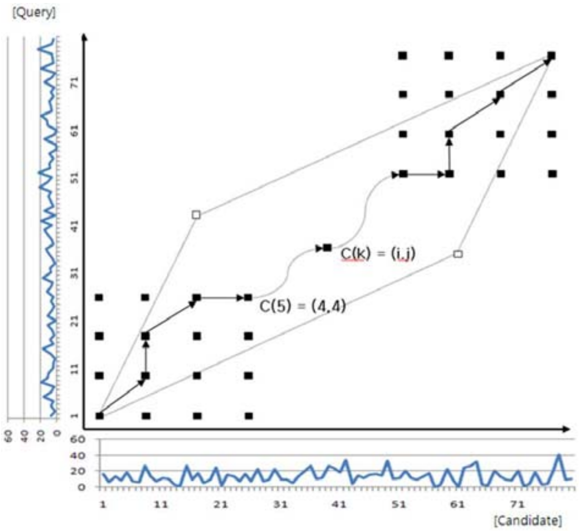


Fig. 5 This is an example of similarity measuring between Query and one of Candidates. Similarity measuring is carried out by using DTW

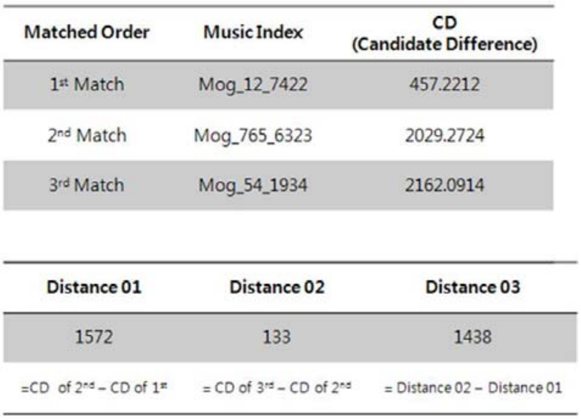
Second, Similarity measuring is to choose final result after measure similarity between Query and each candidate by DTW. DTW is similarity measuring algorithm. This is robust at time axis changing. For this reason, it has been popularly used in audio recognition fields. This find dynamically the best warping route in two data with different length then compare them. Fig.5. shows one example of the result of applying DTW to compare Query and Candidate. Formula (3) is the computation formula of DTW. In D(A,b), 'A' and 'B' are two music inputs. Function D computes the distance difference.

$$D(A,B) = \frac{\min}{F}\left[\frac{\sum_{k=1}^{k} d(c(k)w(k))}{\sum_{k=1}^{k} w(k)}\right] \quad (3)$$

IV. Experimental & Result

For experimental, the music server contained 1,000 songs in random order from Korean, Pop and classical music. All of them were sampled at 44100Hz sampling, Mono and PCM and quantized 16tit. Then features were extracted from all of them through noticed method in section 3.

The input music was 50 songs. For keeping experimentation autonomy, the input music was collected outside even if those were also present in the music server. 450 Query were made by mixing different compression standards and sound qualities from these 50 songs. The nine standard specifications used MP3 (64kbps, 128kbps, 320kbps), OGG (96kbps, 128kbps, 350kbps), WMA (64kbps, 128kbps, 160kbps).



Matched Order	Music Index	CD (Candidate Difference)
1st Match	Mog_12_7422	457.2212
2nd Match	Mog_765_6323	2029.2724
3rd Match	Mog_54_1934	2162.0914

Distance 01	Distance 02	Distance 03
1572	133	1438
=CD of 2nd – CD of 1st	= CD of 3rd – CD of 2nd	= Distance 02 – Distance 01

Fig. 6 This is a result table of retrieved candidate group with distance values among all of server music

Fig.6 is one example of candidate retrieval result. We can rely on candidate retrieval result through Distance01 value is relatively more big than Distance02 one. Fig.7 is a comparing graph of Query, 1st Match, 2nd Match and 3rd Match. This graph shows that Query and 1st Match is almost a unit. It also shows that 2nd Match is not a unit with Query in some parts and 3rd Match is different with query in many parts.

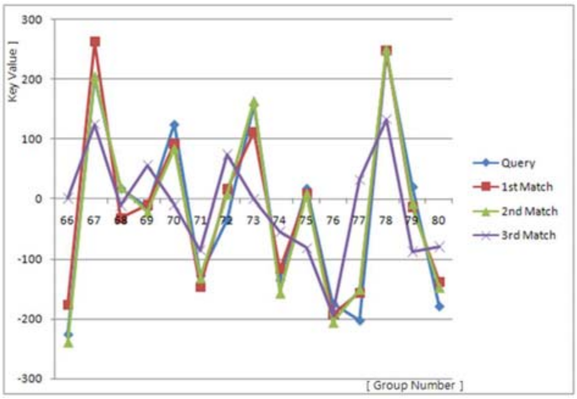


Fig. 7 This is a result graph about comparing value sequences of Query, 1st match, 2nd match and 3rd match. X axis means index number of first-order differentiated Key Values of MFCC Group and Y axis means Key Value of each index.

In our retrieval experiment, we achieved 97% retrieval success rate and 3% retrieval missing rate in total 450 quires. Pre-processing and feature extracting time of 1 music file was average 1.21Sec. Retrieval speed about 1 Query was average 15ms. Average retrieval speed linearly increase as follow growing of server volume. The more server volume increases, the more number of retrieval individual increases. The cause of 3% retrieval missing rate was Time Code gap. Time code gap was occurred because accurate of Silence Removing was not perfect.

V. Conclusion

In this paper, we proposed Content-based retrieval system for effective digital music retrieval. As a content-based feature of proposed system, we used First-order differentiated Key Value of Grouped MFCC. Retrieval query was made up extracted features from input music. This was used to find Candidate list. DTW was used to measure similarity between query of input music and candidate music. Music Matching was consists of Candidate extracting and Similarity measuring. In proposed system, we can get effective result on accurate and speed point of view.

In future work, we suggest 3 topics for development of proposed system. First topic is extra study about retrieval algorithm. We will research a stable retrieval algorithm without reference to server volume increasing. Second one is study about finding content-based feature with good discrimination in similar music group. Third one is to research and apply novel algorithm for sophisticated starting point detecting.

Acknowledgment

This work was supported by the Brain Korea 21 Project of Media department of Soongsil University in 2008.

References

[1] Z. Jun, S. Kwong, W. Gang, and Q. Hong, “Using Mel-Frequency Cepstral Coefficients in Missing Data Technique,” *EURASIP Journal on Applied Signal Processing*, Vol.3, 2004, pp. 340-346.

[2] M. Xu, NC. Maddage, C. Xu, M. Kankanhalli, and Q.Tian, “Creating audio keywords for event detection in soccer video,” *Proc. International Conference Multimedia and Expo ICME’03*, Baltimore, USA, 2003, pp. 281-284.

[3] B.J. Shannon, and K.K. Paliwal, “A Comparative study of filter bank spacing for speech recognition,” *Proc. International Microelctronic engineering research conference*, Brisbane, AUSTRALIA, 2003, pp. 1-3.

[4] Ziyou Xiong, R. Radhakrishnana, A. Divakaran, and T.S. Huang, “Comparing MFCC and MPEG-7 Audio features for feature extraction, maximum likelihood HMM and entropic priop HMM for sports audio classification,” *Proc. International Conference Multimedia and Expo ICME’03*, Baltimore, USA, 2003, pp. 397-400.

[5] J.C. Brown, A. Hodgins-Davis, and P.J.O. Miller, “Classification of vocalizations of killer whales using dynamic time warping,” *The Journal of the Accoustical Society of America*, Vol.119, 2006, pp. EL34-EL40.

[6] C.A. Ratanamahatana, and E. Keogh, “Three myths about dynamic time warping data mining,” *Proc. SIAM International Conference on Data Mining*, Newport Beach, CA, 2005, pp. 506-510.

[7] A.M. Youssef, T.K. Abdel-Galil, E.F. El-Saadany, and M.M.A. Salama, “Disturbance classification utilizing dynamic time warping classifier,” *IEEE Transactions on Power Delivery*, Vol. 19, 2004, pp. 272-278.

[8] A. Pikrakis, S. Theodoridis, and D. Kamarotos, “Recognition of isolated musical patterns using context dependent dynamic time warping,” *IEEE Speech and Audio Processing*, Vol. 11, 2003, pp. 175-183.

[9] E.J. Keogh, and M.J. Pazzani, “Computer Derivative Dynamic Time Warping,” *Proc. First SIAM International Conference on Data Mining*, Chicago, USA, 2001, pp. 1-11.

[10] S. Velusamy, B. Thoshkahna, and K.R. Ramakrishnan, “A Novel Melody Line Identification Algorithm for Polyphonic MIDI Music,” *Proc. LNCS Advances in Multimedia Modeling*, 2006, pp. 248-257.

[11] Ning Hu, R.B. Dannenberg, and G. Tzanetakis, Polyphonic audio matching and alignment for music retrieval. *Proc. 2003 IEEE Workshop on Application of Signal Processing to Audio and Acoustics*, New York, USA, 2003, pp. 185-188.

[12] R. Clifford, M. christodoulakis, and T. Crawford, “A Fast Randomised Maximal Subset Matching Algorithm for Document-Level Music Retrieval,” *Proc. 7th International Conference on Music Information Retrieval*, Victoria, CA, 2006.

[13] Y.Y. Chung, H.c. Choi, Zhen Zhao, M.A.M. Shukran, Y.S. David, and Fang Chen, “An Efficient tree-based quantization for content based music retrieval system,” *Proc. 2007 annual Conference on International Conference on Computer Engineering and Applications table of contents*, Queensland, AUSTRALIA, 2007, pp. 237-241.

[14] A. Spanias, T. Painter, V. Atti, and J.V. Candy, *Audio Signal Processing and Coding* (The Journal of the Acoustical Society of America, 2007).

[15] G. Peeters, “Music Pitch Representation by Periodicity Measures Based on Combined Temporal and Spectral Representations,” *Proc. 2006 IEEE International Conference on Acoustics, Speech and Signal Processing*, Toulouse, FRANCE, 2006, pp. V53-V56.

[16] Xin Zhang, and W. Ras Zbigniew, “Analysis of Sound Features for Music Timbre Recognition,” *Proc. 2007 International Conference on Multimedia and Ubiquitous Engineering*, Seoul, Korea, 2007.

[17] J.P. Bello, and J. Pickens, “A robust mid-level representation for harmonic content in music signals,” *Proc. International Symposium on Music Information Retrieval 2005*, London, UK, 2005.

[18] M. Marolt, “Melody-based retrieval in audio collections,” *The Journal of the Acoustical Society of America*, Vol. 122, No. 5, 2007.

[19] Xi Shao, N.C. Maddage, Xu Changsheng, and M.S. Kankanhalli, “Automatic music summarization based on music structure analysis,” *Proc. 2005 IEEE International Conf. on Acoustics, Speech, and Signal Processing*, Philadelphia, USA, 2005, pp. 1169-1172.

[20] Yi Yu, J. Stephen Downie, Lei Chen, Vincent Oria, and Kazuki Joe, “Searching musical audio dataset by a batch of multi-variant tracks,” *Proc. The 1st ACM International Conference on Multimedia information retrieval table of contents*, Vancouver, CA, 2008, pp. 121-127.

[21] R. Hu, and R.I. Damper, “Fusion of two classifiers for speaker identification: removing and not removing silence,” *Proc. 8th International Conference on Information Fusion*, Philadelphia, USA, 2005, pp. 429-436.

[22] Xufang Zhao, and D. O’Shaughnessy, “A new hybrid approach for automatic speech signal segmentation using silence signal detection, energy convex hull, and spectral variation,” *Proc. Canadian Conference*

on Electrical and Computer Engineering, Ottawa, CA, 2008, pp. 145-148.

[23] Wenjuan Pan, Yong Yao, Zhijing Liu, and Weiyao Huang, “Audio classification in a weighted SVM,” *Proc. International Symposium on Communications and Information Technologies*, Sydney, AUSTRALIA, 2007, pp. 468-472.