한국ITS학회논문지
제6권, 제2호 (2007년 8월)
pp.95~106

A New Tempo Feature Extraction Based on Modulation Spectrum Analysis for Music Information Retrieval Tasks

Hyoung-Gook Kim
(김형국)

요 약

본 논문은 음악 정보검색에 사용되는 효과적인 템포 특징 추출방식을 제안한다. 제안된 템포 정보는 협소 밴드상의 일시적인 변조 성분에 의해 형성된다. 이러한 변조 성분은 시간 축 상의 음악 신호로부터 스펙트럼을 구한 후, 각 스펙트럼 성분에 대한 주파수 영역 분석을 통해 획득된 변조 스펙트럼으로 구성된다. 실제 구현에 있어서는 MP3 음악파일로부터 부분 디코딩에 의해 출력된 변형된 이산 코사인 변환 계수에 퓨리에 변환을 취하여 변조스펙트럼을 구하였다. 획득된 변조 스펙트럼의 진폭으로부터 고속으로 추출된 음악 템포 특징값은 다양한 음악 정보 검색에 적용되었다. 음악 무드 및 장르 분류에서는 로그 변조 주파수 계수를 적용하여 분류 성능을 개선시켰으며, 적응 변조 스펙트럼에서 유도된 비트 벡터는 오디오 핑거프린팅에 적용되어 잡음환경 하에서도 검색 성능을 크게 향상시켰다.

Abstract

This paper proposes an effective tempo feature extraction method for music information retrieval. The tempo information is modeled by the narrow-band temporal modulation components, which are decomposed into a modulation spectrum via joint frequency analysis. In implementation, the tempo feature is directly extracted from the modified discrete cosine transform coefficients, which is the output of partial MP3(MPEG 1 Layer 3) decoder. Then, different features are extracted from the amplitudes of modulation spectrum and applied to different music information retrieval tasks. The logarithmic scale modulation frequency coefficients are employed in automatic music emotion classification and music genre classification. The classification precision in both systems is improved significantly. The bit vectors derived from adaptive modulation spectrum is used in audio fingerprinting task. That is proved to be able to achieve high robustness in this application. The experimental results in these tasks validate the effectiveness of the proposed tempo feature.

Key Words : Tempo feature extraction, music information retrieval, modulation spectrum, joint frequency analysis

Ⅰ. Introduction

Advances in information technology, such as digital libraries, the World Wide Web, and peer-to-peer information systems, are producing an ever-growing volume of music data. The value of a large music collection is limited by how efficiently a user can explore it. Portable audio

* 주저자 : 광운대학교 전파공학과 교수
논문접수일 : 2007년 6월 28일

players can store over 1,000 songs and online music shops offer more than 1 million tracks. Furthermore, a number of new services are emerging which give users nearly unlimited access to music.

Therefore, new tools are necessary to deal with this abundance of music. Of particular interest are tools which can give recommendations, create playlists, and organize music collections.

Toward this end, content-based Music Information Retrieval(MIR) tools provide a unified interface for searching and retrieving music related data. A list of relevant topics can be found online. The topics include, among many others, music genre, emotion, and artist/song name of unknown music files. For example, if a user wants to enjoy some pop songs which can make him feel calm, the music retrieval system will list all songs with pop genre, and calm emotion. Via audio fingerprinting technique, an original piece of music can be identified by its severely distorted clip.

With the technology of music genre/emotion classification employed in the MIR system, each song can be labeled as its genre or the emotion that it brings to the listeners. The automatic music genre classification organizes music songs into genre hierarchy by exploring spectral and temporal features of audio signals. Automatic music emotion classification attempts to cluster the music archives by song's emotion, which expresses the relation between music audio signals and their influence on listeners' emotion.

Audio fingerprinting refers to the method which only extracts highly compressed discriminative information and excludes other complex and complete descriptive information from the audio signals for the purpose of music identification. One basic issue of the music identification concerns the robust feature, because the successful applications require that the method must be able to support recordings and retrieval anywhere and in any situation. So the feature should not only be able to represent an audio clip uniquely for eliminating false positive matching, but also it should be resistant against signal distortions cause by amplifying, filtering, acoustic transmission, channel distortion, severe background noise, audio encoding, and decoding, etc.

For these reasons, content-based descriptors from music signals are the fundamentals to these MIR tasks. The descriptors are proposed to describe energy distribution, pitch, harmonic structure, timbre, melody, tempo, rhythm or other aspects of music. Frequently, they are designed for specified applications to emphasize different musical elements.

In this paper we focus on extracting the tempo information from the acoustic signal. It aims at representing the tempo information in a compact form that is feasible for the tasks of music classification and retrieval.

In musical terminology, tempo is the beat rate of music and corresponds to the music speed perceived by human. It is measured by the number of beats per minute(bpm). Much endeavor has been put to estimate the tempo from music signal. Alonso et al. [1,2] report the highest tempo extraction precision, 96%(at least one tempo correct), 55.71%(two tempos correct), 25%(at least one phase correct) and 5%(two phases correct) in MIREX'05. It seems that estimating music tempo accurately is very difficult.

Unlike the tempo estimation and tracking method in [2,3], other approaches are proposed, which encode the tempo information in a compact form. Typical methods are the beat spectrum [4], the beat histogram [5] and the periodicity distribution [6-8]. The beat spectrum [4] is estimated by summing the diagonal of similarity matrix. Peaks in the beat spectrum correspond to repetitions in the audio signal. In the beat histogram method [5], the enhanced autocorrelation function of the energy envelop is calculated and its peaks are detected. The first three peaks in the appropriate range for beat detection are accumulated into a histogram. In [7], the periodicity distribution is in the form of a 2-dimensional histogram, which counts for the periodicities with different tempos and different strength levels over time. And each periodicity is computed via a comb filter. [8] gives another representation of signal periodicities, called as Inter-Onset Interval Histogram (IOIH). All these methods try to encode the tempo information in a spectrum whose coefficients present the strength at the periodicities.

In this paper we propose a computationally efficient tempo feature extraction method based on modulation spectrum estimation(Section II). Two kinds of features are derived. They are designed particularly to fulfill the requirements of different applications, and applied to three tasks, including music emotion classification(Section III), music genre classification(Section IV), and audio fingerprinting(Section V). The feature effectiveness is verified in these tasks.

Ⅱ. Tempo Feature Extraction

For most signals, such as music, short-time spectral estimates change with time. If a signal is observed for a long period of time and the spectrum changes periodically, then the signal can be modeled as a cyclostationary signal. The modulation frequency analysis is to characterize the periodic time-varying behavior of the audio signal.

Based on the assumption that tempo is modeled by the modulation components of narrow-band temporal modulated signal, the modulation signal can be decomposed in acoustic frequencies and modulation frequencies via modulation spectrum estimation. Therefore, the tempo information is able to be represented by the modulation spectrum.

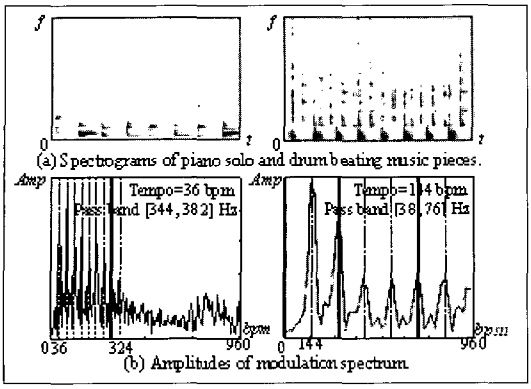
Similarly, Tyagi et al. [9] develop the mel-cepstrum modulation spectrum to extract the long term dynamic variations of speech signal. Peeters et al. [10] depict the long-term dynamic information of music by performing a Fourier transform on the mel-band filtered signal. Sukittanon et al. [11] use the continuous wavelet transform to mimic the constant-Q effect of human perception on modulation frequency.

In this paper the modulation spectrum estimation is introduced to extract tempo feature. Fourier transform, performed on modified discrete cosine transform (MDCT) coefficients, is still inherited for its computational efficiency and easy implementation in embedded systems.

1. Tempo Characterization

The tempo decomposition effect on monophonic pieces is illustrated in Figure 1.

The left column is (a) the spectrogram and (b) the amplitude of modulation spectrum extracted from an excerpt of piano solo. And the right column is the ones obtained from a piece of drum beating. Their tempo rates are captured precisely by the modulation spectrum estimated in the



(a) Spectograms of piano solo and drum beating music pieces.

(b) Amplitudes of modulation spectrum

Figure 1. Modulation spectrum estimation on MDCT

acoustic frequency range of the target instruments.

Another experiment is carried out on the drum beat annotation data made by Tanghe et al. [12]. There are 49 real-world polyphonic music fragments (each in length of 30 seconds) whose drum events are annotated by experienced drummers or percussionists.

The data are processed as follows:

1) The drum events are stored in MIDI format. They are synthesized to waveform samples (note pitch in 36~77 Hz).
2) The waveform is encoded into MP3 file (44100 Hz sampling rate, stereo, 16 bits, 128 kbps).
3) The 2nd MDCT sub-band signal is extracted by the partial MP3 decoder. Its frequency range, 38~76 Hz, has covered the pitch range of drum events.
4) The modulation spectrum is extracted. About 18 frames can be estimated from each 30-second fragment.
5) The preview of the polyphonic music fragment, which is encoded in very low quality and 5~11 seconds in length, is also transformed to 128 kbps MP3 file.
6) The modulation spectrum is only estimated from the 2nd MDCT sub-band signal. Only one frame can be extracted from each polyphonic preview.

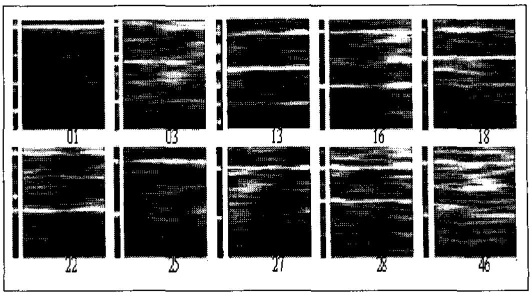


Figure 2. Modulation spectrum amplitudes of synthesized drum events and polyphonic music: Number is the track number given in [12]. The vertical bar depicts the amplitude of modulation spectrum estimated from the polyphonic music. The square map depicts the amplitudes of modulation spectrums estimated from the synthesized drum events. On the square map, the x-axis is time in 30 seconds; the y-axis is modulation frequency from 0.5~5 Hz.

The amplitudes of the modulation spectrums are shown in Figure 2.

Ten fragments are illustrated here. Their tempo rates are 250 bpm. As reference, the only bright horizontal bar in the square map of track 01 just locates at 4.2 Hz, indicating the beat rate clearly. It can be seen that the modulation spectrums estimated from the drum events and from their polyphonic music pieces have peaks at the same frequencies. So the tempo rate can be captured in the modulation spectrum, although the harmonics of beat rate is extracted more likely from the polyphonic pieces.

2. Feature Extraction

In this section, we describe a novel music tempo feature extraction based on modulation spectrum estimation. For computational efficiency the modulation spectrum is obtained by applying Fourier transform on MDCT that are partially decoded results of MP3 files. The tempo feature extraction follows the following steps:

1) The music signal is narrow band-pass filtered. An efficient implementation on MP3 data is to use MDCT coefficients along time as the sub-band signal.
2) The MDCT sub-band signal is low-pass filtered by half-hamming window convolution. Its cut-off frequency is 10 Hz.
3) The deviation operation between two adjacent low pass filtered samples is performed for emphasizing signal variant.
4) The modulation spectrum is obtained by doing a long-term Fourier transform on the deviation signal. The analysis window is 12 seconds and the analysis shift is 1 second.
5) The amplitude of modulation spectrum is smoothed by a log-scale triangular filterbank. The smoothed coefficients are named as log-scale modulation frequency coefficients(LMFC). 12-order LMFC is extracted frame by frame in each sub-band.
6) The tempo feature is composed of the LMFCs in the lowest 5 MDCT sub-bands, covering 0~200 Hz approximately in case of 44100 Hz sampling rate. It focuses on the onsets of percussion and bass instruments.
7) A Karhunen Loeve (KL) transform is applied to the tempo feature to remove the linear dependency among dimensions and reduce dimensionality.

Ⅲ. Music Emotion Classification

Music is perceived historically and pervasively as an important carrier of human emotion. There is solid empirical evidence from psychological research that listeners often strongly agree about what type of emotion is expressed in a particular music piece [13]. The topic here is trying to derive the emotion expressed in the music solely relying on the audio data [14,15].

1. Feature Extraction

Intensity, timbre and tempo features are extracted from the MP3 music data. Intensity is represented by the average scale factor(scf) of each MP3 frame. Also its delta value(dscf) is calculated as Equation (1).

$$dscf(i) = \sum_{j=-2}^{2} j \times scf(i+j) \qquad (1)$$

The timbre features presented in [14] is adopted here. They include centroid, bandwidth, roll-off frequency, and spectrum flux of the amplitude spectrum. Also the peaks, valleys and arithmetic averages on the 7 logarithmic sub-bands are used. We substitute the MDCT coefficients for the amplitude spectrum. Thus, 25 values extracted from MDCT and the scale factor features of each MP3 frame are used as the segmental features. The 12×5 dimensional LMFC coefficients are used as the tempo features.

2. Classification of Music Emotion

Figure 3 gives the system structure of emotion classification. It is constituted by three parts: feature extraction, hierarchical classifier and classification rule. Features are extracted and inputted to the classifiers with multiple layers.

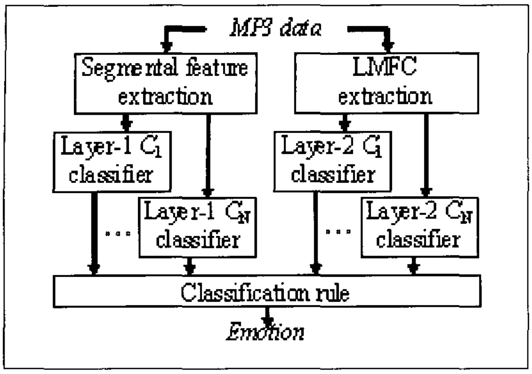


Figure 3. System structure of music emotion classification

Each single classifier is trained by AdaBoost [16]. Finally the music emotion is outputted according to the classification rule.

The final classifier includes two layers, each responsible for one type of features. At each layer, there are numbers of pairwise classifiers, each of which is trained for distinguishing class C from its anti-class $\tilde{C}$. The pairwise classifier is composed of a Karhunen Loeve (KL) transform and a Gaussian mixture model(GMM). The dimensionality of the KL transform matrix and the number of the Gaussian mixtures are determined by AdaBoost [16]. This structure facilitates the implementation of AdaBoost algorithm, which is a pairwise training method and has the advantages of high classification precision, optimal feature selection and model parameter adjusting.

In the recognition stage, each pairwise classifier classifies an input feature frame to positive (C) or negative ($\tilde{C}$). The classification rule determines the emotion of the music piece according to the ratio between the positive frames and the negative frames:

$$I = \mathrm{ar\,gmax}\left\{ \alpha \frac{N_{1,C_j}}{N_{1,C_j} + \widetilde{N_{1,C_j}}} + (1-\alpha)\frac{N_{2,C_j}}{N_{2,C_j} + \widetilde{N_{2,C_j}}} \right\} \quad (2)$$

In Equation (2) $N_{i,C_j}$is the number of positive frames of class j in layer i, and $\widetilde{N_{i,C_j}}$ is the number of negative frames of class j in layer i. N is the number of emotion classes. I is the emotion output. In our implementation, α is equal to 0.7.

3. Database for Music Emotion Classification

Four emotion classes, including calm, sad, pleasant, and excited, are selected, because (1) they are relatively consistent and widely accepted emotions in music, (2) it is easy to get the ground truth data of these categories, and (3) they distribute at the four corners in Russell's dimensional map [17].

During the labelling procedure, listeners are asked to describe that the music piece is supposed to indicate one of the emotions or none of them. 3 females and 3 males (Korean and Chinese) in the ages of 20~35 attend the labelling work. 695 homogeneous songs are labelled from hundreds of western classical, waltz, march, jazz, electronic, popular and rock, with the average length of 3minutes. Among them, 286 pieces (68/calm, 59/excited, 85/pleasant, 74/sad) are labelledby the 6 persons consistently and are used as trainingdata; 409 pieces (100/calm, 100/excited, 107/pleasant, 102/sad) are labelled by 3 Korean and used as testing data.

4. Experimental Results for Music Emotion Classification

To compare the performance of different features, the classification only based on the segmental features or the tempo features is separately carried out as comparison. Table 1 shows their results.

Table 1. Comparative result of emotion classification

Precision	Calm	Excited	Pleasant	Sad
Segmental	88%	87%	91%	77%
Tempo	72%	81%	74%	70%
Integrated	96%	90%	97%	76%

Table 2. Confusion matrix of emotion classification

Emotion types	Calm	Excited	Pleasant	Sad
Calm	96	0	2	2
Excited	0	90	10	0
Pleasant	1	0	104	2
Sad	24	0	0	78

Table 2 gives the confusion matrix of the integrated features.

We have observed that most of the errors take place in the categories with similar tempo mode. Listening test reveals that the excited error songs usually sound like pleasant rock. They are noisier than the pleasant pop, but not as heavy and tense as the mental. Also it is difficult to distinguish the emotion of sad from calm, because some of them are similar vocal singing with slow tempo accompaniment. The main difference is that the voice of calm songs is smooth and relaxed, whereas the voice of sad songs is sometime trembling. It is even suspected that several pieces are labelled as sad based on lyrics. Clearly, the difference of singing style and the lyrics content can not be captured by the current features.

Although the experiment data is very limited, the result validates the effectiveness of the tempo feature. Especially by integrating it appropriately with other features, the classification performs promisingly.

IV. Music Genre Classification

Similar to the emotion classification, the genre classification also belongs to statistical pattern recognition domain. Significant approaches are presented in [5]. Further improvements can be found in [18-20].

Here we also apply the tempo feature to the task of genre classification. Music songs are divided into five genre categories, which include classical, pop & rock, jazz & blues, hip-hop and mental & punk. The dataset is composed of1,000 songs, equally 200 songs in each category. The length of each song is no less than 30 seconds. These songs are collected from different ways, e.g. compact disks, internet, etc., and part of the training data in MIREX'04 genre classification evaluation are also included.

The system is the same as that of emotion classification. Its performance is evaluated on the genre data using 20-fold cross-validation. The results are shown in Table 3.

It can be seen that the segmental feature is much effective than the tempo feature. The best classification precision is 86.8% by employing segmental features, while the optimal precision with tempo featuresis only 77.0%. That is probably because the songs with different genres are usually played by different instruments and segmental features can capture the characteristics of instruments well. On the contrary, the songs

Table 3. Comparative results of genre classification

Precision	The number of mixtures of GMM					
	4	8	12	16	20	24
Segmental	83.0%	85.1%	85.5%	86.8%	86.5%	84.7%
Tempo	73.0%	75.0%	77.0%	76.0%	77.0%	76.0%

Table 4. Confusion matrix of genre classification

Genre types	CL	PR	JB	HH	MP
CL	188	8	4	0	0
PR	0	188	0	4	8
JB	4	4	180	4	8
HH	0	20	0	173	7
MP	0	7	0	0	193

even in the same genre always have diverse tempos, so the boundaries of tempos between music genres are very fuzzy. But for the songs, which are played by similar instruments and hardly distinguished solely depending on segmental features, introducing tempo features is a good choice. Table 4 gives the confusion matrix of the 1000 songs. It precision is 92.2%.

In Table 4, CL, PR, JB, HH, MP respectively represent classical, pop & rock, jazz & blues, hip-hop, mental & punk. The confusion matrix proves that the boundaries between different music genres are very indistinct. Pop & rock, hip-hop, mental & punk form a confusion set; classical music can be incorrectly recognized as pop & rock or jazz & blues; and jazz & blues is the most easily confused set, which can be wrongly classified into any other category.

Music is perceived historically and pervasively as an important carrier of human emotion. There is solid empirical evidence from psychological research that listeners often strongly agree about what type of emotion is expressed in a particular music piece. The topic here is trying to derive the emotion expressed in the music solely relying on the audio data.

V. Audio Fingerprinting

Music retrieval applications based on audio fingerprinting technique have been successfully launched to the commercial market recently. Different appoaches are described in [21~25] and Cano et al. make a rather complete review in [26]. The audio fingerprinting is composed of two distinct phases, such as database generation stage and query stage.

During the database generation stage, the index bit vectors and the fingerprint bit vectors are extracted both based on the modulation spectrum estimation. Each fingerprint bit vector is indexed by four index bit vectors. Then, they are stored in the database by implementing a linear hash table data structure. In the query stage, the bit vectors of the query clip are extracted. Then, two-stage searching methods firstly locates the positions of the fingerprint bits vectors indexed by the index bit vectors; secondly computes the distances between the indexed fingerprint bit vectors and those of the query clip. The music piece with the minimal distance is evaluated whether to be the retrieval result.

1. Audio Fingerprinting Design

The fingerprint extraction process follows the steps below:

1) Partially decoding the MP3 frames to MDCT coefficients;
2) Performing a long-term Fourier transform on the MDCT sub-band signal along the time;
3) Getting the amplitude of the modulation spectrum;
4) Smoothing the amplitude by a low pass filter (a third-order finite impulsive response low pass filter is used here)
5) Quantizing the smoothed amplitude by performing the one bit delta quantization along

the modulation frequency as Equation (3) illustrated

$$fbv(i) = \begin{cases} 1, & LFA_{MS}(i) > LFA_{MS}(i+1) \\ 0, & otherwise \end{cases} \quad (3)$$

where fbv is one frame of fingerprint bit vector, LFA_{MS} is the low pass filtered amplitude, and i indexes modulation frequency.

6) Selecting M×N×T bits in N sub-bands (M bits per band per frame) of adjacent T frames to form the fingerprint block.

The system with optimal parameters is as follows. Briefly, only the 2nd~5th sub-bands, covering 38~191 Hz in case of 44100 Hz sampling rate, are used. Then the modulation spectrum is estimated on each sub-band. Its analysis window is 3.3 seconds long with 104.5 milliseconds shift. Thus, a fingerprint bit vector with 127 bits can be quantized. Four fingerprint bit vectors, respectively extracted from the 4 MDCT sub-bands, compose a long fingerprint bit vector with 512 bits (actually 4 bits are absent). Finally the 8 adjacent 512-bit vectors are assembled into one fingerprint block with 4096 bits. In order to extract a complete fingerprint block, the retrieval clip should be no less than 4 seconds. The fingerprint block with the minimal bit error rate under a verification threshold (typically 0.4) is located in database and the corresponding music can be identified.

2. Robustness Evaluation

1000 MP3 songs are transformed from CD tracks or downloaded from internet, including rock, jazz, popular, hip-hop and folk. Then 100000 clips are excerpted from them, 100 clips from each song averagely. Environmental noise is recorded by a Samsung digital camera Digimax V4 on streets and in cars. Then around 2 hours of the wide band coloured noise are added to the clean clips with different gains, making noise corpus sets with different signal to noise ratios (SNR). The fingerprints of the noise corrupted clips are extracted and searched among the clean fingerprint database.

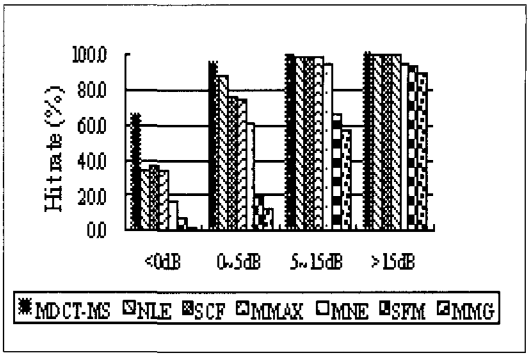


Figure 4. Hit rate of different feature (False positive rate is less than 1%)

The comparison is made among several features, including spectral flatness measure (SFM), spectral crest factor (SCF), maximum of energy(MMAX), geometric average of energy (MMG) and arithmetic average of energy(MNE) calculated in 8 equal-spaced sub-bands. The sub-band values are transformed to bits via one bit delta quantization along time and frequency dimensions simultaneously [25]. Also the fingerprint(DLE) described in [25] is also implemented. Figure 4 gives their results.

It is clear that although all the features achieve high hit rate when the SNR is above 15dB, only the MDCT modulation spectrum based fingerprint gains high hit rate under the severe noise conditions. Even when the SNR is under 0dB its hit rate is still above 60%. The fingerprint's hit rates are 100.0%, 99.8% and 95.4% respectively, when the SNR is above 15dB, 5~15dB and 0~5dB. That proves its advantage of robustness.

3. Discussion

The experiment verifies the high robustness of the fingerprint to environmental noise on a small corpus. Also it is very resistant to severe channel distortion. But we find out two drawbacks of this feature.

The biggest problem is that the variation of MDCT coefficients, caused by different encoding parameter and time mismatch of encoded frame. That severely influences the fingerprint match precision. It can be solved by extracting the fingerprint from the sub-band synthesized values.

The other drawback is that most of the search errors come from the classical music, especially when the query clip is a piece of chord instrument playing without clear onsets. In this case, the amplitude values of modulation spectrum are small fluctuations and uninformative. Quantizing such values to derivative bit results in a random bit map and the bit error rate is near to be 0.5. So the query clip can not be identified. On the contrary the fingerprint works quite well on rock and rhythmic popular songs. So we believe that combining this feature with other spectral representations can improve the performance further.

Another issue is the fast searching method. [27] describes the solution in details. The system is based on the modulation spectrum estimated from the synthesized sub-band signal. Then a two-stage bit vector searching implements for the fast retrieval. Its hash index is extracted from the logarithmic scale modulation frequency coefficients. The system is evaluated on the database with 3000 MP3 music files. It achieves the query precisions of 99.67%, 97.00% and 92.52%, when the query SNR is above 15dB, 5~15dB, and 0~5dB.

VI. Conclusion

A tempo feature extraction method for the music information retrieval is proposed in this paper. The tempo information is modeled by the narrow-band temporal modulation components. It is decomposed into a modulation spectrum via joint frequency analysis. Furthermore, the modulation spectrum coefficients are transformed to more compact forms that are implemented to three MIR tasks in this paper.

In practice, the modulation spectrum is estimated from the MDCT coefficients embedded in a MP3 decoding process. The logarithmic scale modulation frequency coefficients are proposed and applied to music emotion classification and music genre classification. The bit vectors derived from the modulation spectrum is employed in the audio fingerprinting task. That validates the high robustness of the feature further.

While achieving promising results in these applications, the proposed tempo feature also reveals the drawbacks in the audio fingerprinting task. Therefore, how to combine the propose tempo feature with timbre features in a variety of MIR tasks should be further explored in the future.

References

[1] M. Alonso, B. David, and G. Richard, "Tempo

extraction for audio recordings," *MIREX'05*, September 2005.

[2] M. Alonso, B. David, and G. Richard, "Tempo and beat estimation of musical signals," *ISMIR*, pp. 158-163, October 2004.

[3] E. Scheirer, "Tempo and beat analysis of acoustic music signals," *J. Acoust. Soc. Am.*, vol. 103, no. 1, pp. 588-601, January 1998.

[4] J. Foote, "The beat spectrum: A new approach to rhythm analysis," *ICME*, pp. 881-884, August 2001.

[5] G. Tzanetakis and P. Cook, "Automatic musical genre classification of audio signals," *IEEE Trans. Speech and Audio Processing,* vol. 10, no. 5, pp. 293-302, July 2002.

[6] F. Gouyon, S. Dixon, E. Pampalk, and G. Widmer, "Evaluating rhythmic descriptors for musical genre classification," *Proc. AES 25th Int. Conf.*, pp. 196-204 , June 2004.

[7] E. Pampalk, S. Dixon, and G. Widmer, "Exploring music collections by browsing different views," *ISMIR*, pp. 201-208, October 2003.

[8] F. Gouyon, P. Herrera, and P. Cano, "Pulse-dependent analyses of percussive music," *Proc. AES 22nd Int. Conf. Virtual, Synthetic and Entertainment Audio*, pp. 396-401, June 2002.

[9] V. Tyagi, I. McCowan, H. Misra, and H. Bourland, "Mel-Cepstrum Modulation Spectrum (MCMS) features for robust ASR," *IEEE Workshop on Automatic Speech Recognition and Understanding*, August 2003.

[10] G. Peeters, A. L. Burthe, and X. Rodet, "Toward automatic music audio summary generation from signal analysis," *ISMIR*, pp. 94-100 , October 2002.

[11] S. Sukittanon, L. E. Atlas, and J. W. Pitton, "Modulation-scale analysis for content identification," *IEEE Trans. Signal Processing*, vol. 52, no. 10, pp. 3023-3035, October 2003.

[12] K. Tanghe, "Collecting ground truth annotations for drum detection in polyphonic music," *ISMIR*, pp. 50-57 , October 2005.

[13] P. N. Juslin and J. A. Sloboda, "*Music and emotion: Theory and research*," Oxford Univ. Press, 2001.

[14] D. Liu, L. Lu, and H. J. Zhang, "Automatic mood detection from acoustic music data," *ISMIR*, November 2003.

[15] Y. Z. Feng, Y. T. Zhuang, and Y. H. Pan, "Music information retrieval by detecting mood via computational media aesthetics," *Proc. IEEE/WIC International Conf. on Web Intelligence*, pp. 281-282, October 2003.

[16] R. E. Schapire, "A brief introduction to boosting," *Proc. Int. Joint Conf. on Artificial Intelligence*, pp. 1401-1406, August 1999.

[17] B. L. Feldman and J. A. Russell, "Independence and bipolarity in the structure of affect," *J. Personality and Social Psychology*, vol. 74, pp.967-984, August 1998.

[18] T. Li and G. Tzanetakis, "Factors in automatic musical genre classification of audio signals," *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, pp. 143-146, October 2003.

[19] M. McKinney and J. Breebaart, "Features for audio and music classification," *ISMIR*, pp. 151-158, October 2003.

[20] K. West and S. Cox, "Features and Classifiers for the Automatic Classification of Musical Audio Signals," *ISMIR*, pp. 531-536, October 2004.

〔21〕 J. Herre, O.Hellmuch, and M. Cremer, "Scalable robust audio fingerprinting using MPEG-7 content description," *IEEE Workshop on Multimedia Signal Processing*, pp. 165-168, December 2002.

〔22〕 R. Lancini, F. Mapelli, and R. Pezzano, "Audio content identification by using perceptual hashing," *ICME*, pp. 739-742, June 2004.

〔23〕 C. Burges, J. C. Platt, and S. Jana, "Distortion discriminant analysis for audio fingerprinting," *IEEE Trans. Speech and Audio Processing*, vol. 11, no. 3, pp. 165-174, March 2003.

〔24〕 A. Ribbrock and F. Kurth, "A full-text retrieval approach to content-based audio identification," *IEEE Workshop on Multimedia Signal Processing*, pp. 194-197, December 2002.

〔25〕 J. Haitsma and T. Kalker, "A highly robust audio fingerprinting system," *ISMIR*, pp. 14-17, October 2002.

〔26〕 P. Cano, E. Batlle, T. Kalker, and J. Haitsma, "A review of algorithms for audio fingerprinting," *Int. Workshop on Multimedia Signal Processing*, pp. 169–173, December 2002.

〔27〕 Y. Y. Shi, X. Zhu, H. G. Kim, and K. W. Eom, "A robust music retrieval system," 120th AES Convention, May 2006.

저자소개

Kim, Hyoung-Gook (김 형 국)

2007년 3월~Assistant Professor, Kwangwoon University, Korea

2005년 ~2007년 2월 : Project Leader, Samsung Advanced Institut of Technology

2002년 ~2005년 : Adjunct (Assistant) Professor, Technical University Berlin, Germany

1999년 ~2002년 : Senior Researcher, Cortologic AG, Berlin, Germany

1997년 ~1999년 : Researcher, Siemens, Munich, Germany