Log-scale Modulation Frequency Coefficient: A Tempo Feature for Music Emotion Classification

Yuan-Yuan Shi, Xuan Zhu, Hyoung-Gook Kim, Ki-Wan Eom, Ji-Yeun Kim

Samsung Advanced Institute of Technology
{yy.shi, xuan.zhu, hyounggook.kim, kiwan.eom, jiyeun.kim}@samsung.com

Abstract. This paper proposes a tempo feature extraction method. The tempo information is modeled by the narrow-band, low-pass temporal modulation component, which is decomposed into a modulation spectrum via joint frequency analysis. In implementation, the modulation spectrum is directly estimated from the modified discrete cosine transform coefficients, which are output of partial MP3 (MPEG 1 Layer 3) decoder. Then the log-scale modulation frequency coefficients are extracted from the amplitude of modulation spectrum. The tempo feature is employed in automatic music emotion classification. The accuracy is improved with several hybrid classification methods based on posterior fusion. The experimental results confirm the effectiveness of the presented tempo feature and the hybrid classification approach.

1 Introduction

In recent years, interests in the technologies of automatic music information retrieval have grown up considerably. New approaches to facilitate retrieval, organization and management of digital music libraries are presented. Among them, tasks like audio fingerprinting retrieval, music classification, query by humming, music similarity query, and singer clustering gain much attention. Content-based descriptors are the fundamentals to these tasks. The descriptors are proposed to describe energy distribution, pitch, harmonic structure, timbre, melody, tempo, rhythm, or other aspects of music. For example, Tzanetakis *et.al.* [1] propose to represent the music content in three feature sets, that is, timbral texture, rhythmic content and pitch content. Frequently music content description is achieved by combining the three factors. And 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. In musical terminology, tempo is the beat rate of music and corresponds to the music speed perceived by human. One related measure is the number of beats per minute (bpm). Much endeavor has been put to estimate bpm from music signal [3][4]. Besides the tempo estimation and tracking methods, other approaches, which encode the tempo information in a compact form rather than estimate the bpm quantitatively, are proposed. Typical methods are the beat spectrum [2], the beat histogram [1], and the periodicity distribution [5][6]. The beat spectrum [2] 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 [1], 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 [5], 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. [6] gives another representation of signal periodicities, called as Inter-Onset Interval Histogram. All these methods try to represent tempo info by a distribution of modulation strength at various periodicities.

In this paper we propose a computationally efficient tempo feature extraction method based on the long-term modulation spectrum analysis. It is designed particularly to fulfill the requirements of embedded systems. The feature is applied to the task of music emotion classification, where the western style pop music is mainly considered because of its relatively simple concept of tempo.

In Section 2 the motivation behind the presented feature and the detailed feature extraction method are introduced. Section 3 describes applying the proposed feature to music emotion classification. The experimental result illustrates the effectiveness of the feature. Finally, the conclusion is drawn in Section 4.

2 Tempo Feature Extraction

On the assumption that a non-stationary signal is modeled by the product of a narrow bandwidth, low-pass modulating component and a high-pass carrier, the signal can be decomposed into acoustic frequencies and modulation frequencies via a transform, which gives the joint frequency representation $\Theta_x(\omega,\eta)$ of the signal *x,* where ω is the acoustic frequency and η is the modulation frequency. Such a transform typically involves two steps: firstly, a spectrogram is estimated to represent the time-frequency content $\Theta_x(t,\omega)$ of the signal; secondly, another transform, *e.g.* Fourier, is applied along the time dimension of the spectrogram to estimate the joint frequency representation $\Theta_x(\omega,\eta)$. Here we name the Fourier-based joint frequency transform as *modulation spectrum analysis* after Tyagi *et al.*[7].

Several different methods have been proposed to extract the time-varying information from the modulation frequency components. For example, Tyagi *et al.*[7] develop the mel-cepstrum modulation spectrum to extract the long term dynamic variations of speech signal. Peeters *et al.*[8] depict the long-term dynamic information of music by performing a Fourier transform on the mel-band filtered signal. Sukittanon *et al.*[9] use the continuous wavelet transform to mimic the constant-*Q* effect of human perception on modulation frequency.

For music signal, the periodicity in the signal causing non-zero terms on the modulation frequencies can result from the beat or tempo. The beat information is usually represented as the high value terms in $\Theta_x(\omega,\eta)$ at low modulation frequency η, ranging from 0.5 to 5 Hz, which corresponds to 30 to 300 bpm. Therefore, with appropriate manipulation the tempo information can be extracted from the modulation spectrum.

In this paper the modulation spectrum analysis based on the discrete Fourier transform performed on modified discrete cosine transform (MDCT) coefficients is proposed for its computational efficiency and easy implementation in embedded systems. MDCT is widely employed in MP3, AC-3, Ogg Vorbis and AAC for audio compression. In MP3 MDCT is applied to the output of a 32-band polyphase quadrature filter bank. It subdivides the sub-band output in frequency into 576 sub-bands to provide better spectral resolution. Thus, the input signal is decomposed in its spectral components represented in MDCT. It is used here to approximate the time-frequency representation $\Theta_x(t,\omega)$. And MDCT along the time dimension forms the MDCT sub-band signal.

2.1 Feature Extraction Algorithm

This section describes the detailed steps of tempo feature extraction. For a standard MP3 file with the properties of 44100Hz, stereo and 128kbps, the feature is extracted as follows:

Step 1. Partially decode the MP3 file to narrow-band-pass filtered samples $\{s^n(t)\}$, $n \in [1,576]$. n is the index of sub-bands. 576 sub-bands can be obtained on the intermediate level of decoding process. The bandwidth of each sub-band is about 38Hz.

Step 2. Full wave rectification followed by a low-pass filtering is a typical envelope detection method. Here, a one-pole filter with α=0.96875 is adopted to obtain a smoothed envelope.

$$\bar{s}^n(t) = \left|s^n(t)\right| \tag{1}$$

$$x^n(t) = (1-\alpha)\bar{s}^n(t) + \alpha x^n(t-1), \qquad 0.95 \le \alpha < 1 \tag{2}$$

Step 3. The deviation between two time-adjacent filtered samples is employed for emphasizing signal variation.

$$\hat{x}^n(t) = x^n(t) - x^n(t-1) \tag{3}$$

Step 4. Long-term Fast Fourier Transform (FFT) is applied on hamming windowed deviation signals. The analysis window is 13.4 seconds (512 samples) and the window shift is 1 second (38 samples). For the i^{th} frame, the k^{th} sample is

$$y_i^n(k) = \hat{x}^n(38 * i + k), \qquad 0 \le k < 512 \tag{4}$$

The result of FFT is described as in Equation (5), where $w(k)$ is the weight of hamming window.

$$Y_i^n(m) = \sum_{k=0}^{512} w(k) \cdot y_i^n(k) \cdot e^{-j2\pi km/512}, \qquad 0 \le m < 512 \tag{5}$$

Step 5. The power of modulation spectrum is smoothed by a log-scale triangular filter-band $\{H(p,m)|1\leq p<12,\ 0\leq m<256\}$, where the p^{th} filter is given by

$$H(p,m)=\begin{cases} 0, & m<f(p-1) \\ \dfrac{2(m-f(p-1))}{(f(p+1)-f(p-1))(f(p)-f(p-1))}, & f(p-1)\leq m\leq f(p) \\ \dfrac{2(f(p+1)-m)}{(f(p+1)-f(p-1))(f(p+1)-f(p))}, & f(p)\leq m\leq f(p+1) \\ 0, & m>f(p+1) \end{cases} \tag{6}$$

and $\{f(p)|0\leq p<13\}$ is the center frequency of each filter, which increases logarithmically. Then, the 12-order vector is calculated as the log-energies at the output of filter-bank:

$$c_i^n(p)=\ln\left(\sum_{m=0}^{256}\left|Y_i^n(m)\right|^2 H(p,m)\right),\qquad 1\leq p\leq 12 \tag{7}$$

It is named as Log-scale Modulation Frequency Coefficient. For each frame, LMFCs on the lowest 5 sub-bands (n=1~5) are extracted to constitute a 60-dimensional feature vector.

If the input signal is in other formats rather than MP3, for example, in PCM format, such a process can be carried out only substituting a narrow band-pass filter bank designation to Step 1.

2.2 Tempo Decomposition

The tempo decomposition effect on monophonic music pieces has been illustrated in [10]. Here, an experiment is carried out on polyphonic music pieces. There are 49 real-world polyphonic fragments (each in length of 30 seconds) whose drum events are annotated by experienced drummers or percussionists. The data are made by Tanghe *et al.*[11]. They are manipulated 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 with the properties of 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.

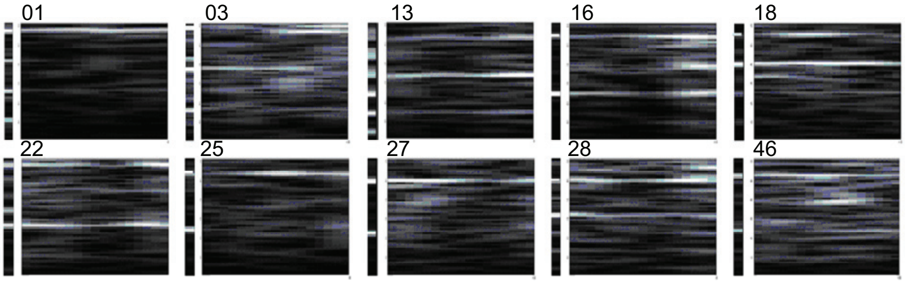

Fig. 1. Modulation spectrum amplitudes of synthesized drum events and polyphonic music. The number above each square map is the track number given in [11]. The vertical bar depicts the amplitude of modulation spectrum estimated from the polyphonic music. The square map depicts the amplitudes of modulation 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 1. Ten fragments are illustrated here. Their tempo rates are 250 bpm. As reference, the bright horizontal bar in the square map of track 01 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. That means the tempo can be captured in the modulation spectrum, although the harmonics of beat rate are also enclosed.

It must be pointed out that there are some limitations to LMFC feature. One constraint is caused by extracting LMFC from only the lowest 5 MDCT sub-bands. Hence, the presented tempo feature mainly indicates the onset periodicity information of drum and other percussion instruments. Another constraint is that the tempo feature is computed from the periodicities at different meters as a whole. So, it is better to apply it to the music with relatively simple tempo and rhythm pattern. That limits the applications of LMFC feature mainly to the style of Western pop music.

3 Music Emotion Classification

LMFC is firstly applied to an automatic music emotion classification task to evaluate its performance for music content-based classification. Music is perceived historically and pervasively as an important carrier of human emotion. There is solid empirical evidence from psychological research that listeners from the same culture often strongly agree about what type of emotion is expressed in a particular piece [12]. Similar work can be found in [13][14]. [13] uses intensity, timbre, and rhythm features to recognize the four emotional states of exuberance, anxious, contentment and depression. The rhythm information is described by the strength, stability, and ratio of onsets. The system obtains from 76.6% to 94.5% precision for different emotion categories. [14] recognizes the four emotions of happiness, sadness, anger and fear by only using three features, that is, mean and variance of the silence ratio and the beat rate estimated by a beat-tracking algorithm. The average precision is 67% and each category's is from 25% to 86%.

3.1 Features

There are three kinds of features, intensity, timbre, and tempo, adopted in the music emotion classification system. All of them are extracted from the MP3 music data.

Intensity is represented by the averaged scale factor (*scf*) of each MP3 frame. Also its delta value (*dscf*) is calculated as in Equation (8) based on the regression method. This delta computation is more resistant to the abrupt noise disturbance than the differentiator with unit sample delay.

$$dscf(i) = \sum_{j=-2}^{2} j \times scf(i+j) \tag{8}$$

The timbre features presented in [13] are adopted here. They include spectral centroid, spectral bandwidth, 95% roll-off frequency and spectral flux of the amplitude spectrum. Also the peaks, valleys, and arithmetic averages of the seven logarithmic sub-bands are used. Here we substitute the MDCT coefficients for the amplitude spectrum. Thus, the 25 values extracted from the MDCT coefficients form the timbre feature set.

Additionally, the 12×5 LMFCs are used as the tempo feature set.

3.2 Hybrid Classification

In this section, two approaches, feature fusion and posterior fusion, are explored for integrating the intensity/timbre features and the tempo feature into a hybrid system.

Feature fusion refers to combination in the feature space. The tempo feature is extracted with the same time interval as that of the intensity/timbre features for synchronization. Thus, the synchronized features can form a new feature vector conveniently. Karhunen-Loeve (KL) transform is performed to extract the orthogonal feature vectors. Then, Gaussian Mixture Models (GMM) are trained with the KL transformed feature vectors. The emotion of a testing music piece is estimated as the one that outputs the maximum likelihood.

Contrastively, posterior fusion merges the posterior probabilities with respect to the intensity/timbre features and the posterior probabilities with respect to the tempo feature by machine learning methods. With the assumption of equal prior probabilities of emotions, the posterior probabilities are reduced to likelihood. Then, the information of each music emotion class is learned in the likelihood space. The training procedure is described as follows:

Firstly, KL transform and GMM are applied on the intensity/timbre features and on the tempo feature respectively.

Secondly, the likelihood is normalized to a probability measure by the rule:

$$P_i = \frac{\exp(L(x \mid i))}{\sum_i \exp(L(x \mid i))} \tag{9}$$

Here, $L(x|i)$ is the log likelihood of vector x with respect to emotion hypothesis i.

Thirdly, P_i values of intensity/timbre features and of tempo feature at the same time interval are connected to one vector, referred to as probability vector.

Finally, the probability vector's class distribution is modeled for classification. For a comparison, GMM, Multi-Layer Perceptions (MLP), and Support Vector Machine (SVM) are trained with the probability vectors.

SVM is originally designed for binary classification [15]. Here it is extended to K-class by constructing $K(K$-1)/2 "one-against-one" binary classifiers and combining them with the rule of majority voting. Each binary classifier is trained on probability vectors from class i and class j. In combining the decisions of the $K(K$-1)/2 binary classifiers, the sample is assigned the class when no less than Q classifiers are agreed on the identity, where

$$Q = \begin{cases} \frac{K+1}{2}, & if\ K\ is\ odd \\ \frac{K}{2}, & if\ K\ is\ even \end{cases} \tag{10}$$

In case that two classes have identical votes, we simply select the one with the smaller index. The experiment is done by using LIBSVM [16].

3.3 Data

Four emotion classes, including calm, sad, pleasant, and excited, are selected, because (1) they are relatively consistent and widely accepted music emotions, (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 labeling 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 labeling work. 695 homogeneous pieces are labeled from hundreds of western classical, waltz, march, jazz, electronic, pop, and rock, with the average length of 3 minutes. Among them, 286 pieces (68/calm, 59/excited, 85/pleasant, 74/sad) are labeled by the 6 persons consistently and are used as training data; 409 pieces (100/calm, 100/excited, 107/pleasant, 102/sad) are labeled by 3 Korean and employed as testing data.

3.4 Experimental Results

A preliminary experiment is carried out to compare the performance of different features. So the timbre/intensity features and the tempo feature are used separately. The dimensionality of KL transform and the number of Gaussian mixtures are selected by experiments. Then GMMs are trained on the training data. The emotion of a testing piece is estimated as the one with the maximum likelihood. Table 1 gives the classification precision.

Table 1. Precisions of music emotion classification using different features.

Precision	Calm	Excited	Pleasant	Sad	Average
Intensity/timbre	88%	87%	91%	77%	85.8%
Tempo	72%	81%	74%	70%	74.2%

Obviously the intensity/timbre features outperform the tempo feature by a large margin. But closer inspection of the results shows that some errors are complementary, so it gives rise to the hybrid approach as explained earlier. Table 2 gives the classification precision of each hybrid system.

Table 2. Precisions of music emotion classifiction using different hybrid approaches.

	Feature Fusion	Posterior Fusion		
		GMM	MLP	SVM
Precision	85.8%	90.0%	89.7%	88.3%

Except for the feature fusion method, all posterior approaches outperform the individual baseline. It demonstrates that the integration in likelihood space performs better than in feature space. Another observation is that GMM performs comparable with MLP, while SVM works relatively worse. We think one possible reason is that the classifier combination based on majority voting deteriorates the multi-class classification precision to some extent. So, besides the multi-class precision, the binary precision of each SVM is listed in Table 3. Other approaches can be explored to search a better classifier combination result.

Table 3. Precisions of music emotion classification using binary SVM.

Precision	Calm	Excited	Pleasant	Sad
Calm	—	97.8%	98.1%	83.2%
Excited	—	—	80.9%	86.6%
Pleasant	—	—	—	89.2%

Table 3 also illustrates that the classification between different pair-wise of emotions is unbalanced on precision. This phenomenon is demonstrated more clearly in other hybrid systems. For example, Table 4 gives the confusion matrix of GMM posterior fusion.

Table 4. Confusion matrix of music emotion classification.

#	Calm	Excited	Pleasant	Sad	Precision
Calm	96	0	2	2	96%
Exciting	0	90	10	0	90%
Pleasant	1	0	104	2	97%
Sad	24	0	0	78	76%

Obviously, most of the errors take place in the categories with similar tempo mode, that is, calm versus sad (slow tempo) and excited versus pleasant (fast tempo). The two most typical errors are that exciting songs are misclassified as pleasant, and sad songs are misclassified as calm. Closer inspection of the error pieces shows that those misclassified exciting songs sound like pleasant rock, more noisy than pleasant pop, but not as heavy and tense as black metal. Even worse, to distinguish between sad and calm precisely is difficult in some cases, where they have similar vocal in slow tempo, and accompaniment instruments are also same. One discriminative factor is that the vocal of calm songs is smooth and relaxed, while the vocal of sad songs sounds trembling sometime. It is even suspected that several pieces are labeled as sad based on their sad Korean lyrics. The Chinese listeners cannot distinguish them from calm songs. Clearly the differences of singing style and lyrics content cannot be expressed by current features.

Although the experiment data are very limited, the result still validates the effectiveness of the proposed tempo feature and the hybrid classification approach.

4 Conclusion

A tempo feature extraction method is presented in this paper. The tempo information is modeled by the narrow-band, low-pass temporal modulation components. It is decomposed into a modulation spectrum via joint frequency analysis. In practice, the modulation spectrum is estimated from the MDCT coefficients embedded in a MP3 decoding process. The log-scale modulation frequency coefficients are proposed and applied to music emotion classification. The classification precision is increased by means of integrating the intensity/timbre features and the tempo feature in a posterior fusion scheme.

The proposed tempo feature only focuses on detecting the repetition of onset events of the percussion instruments. But, besides drum etc., there are a variety of other instruments, e.g. guitar, piano, keyboard, and etc., which also indicate rhythmic information. Another limitation is that the tempo feature represents the periodic onsets of different percussion instruments as a whole, although they quite possibly reside at different metrical levels. It is incapable of modeling the rhythmic structure. Therefore, how to involve the complete rhythm pattern and structure with a compact form in the tempo feature is our future work.

5 Acknowledgments

The authors wish to thank Koen Tanghe for allowing us to use the drum annotation data that they developed.

References

1. G. Tzanetakis, and P. Cook, "Automatic Musical Genre Classification of Audio Signals," *IEEE Transactions on Speech and Audio Processing*, vol. 10, no. 5, pp. 293-302, Jul. 2002.
2. J. Foote, "The Beat Spectrum: A New Approach to Rhythm Analysis," *Proceeding of International Conference on Multimedia Expo*, 2001.
3. E. Scheirer, "Tempo and Beat Analysis of Acoustic Music Signals," *Journal of the Acoustical Society of America*, vol. 103, no. 1, pp. 588-601, Jan. 1998.
4. M. Alonso, B. David, and G. Richard, "Tempo and Beat Estimation of Musical Signals," *Proceeding of International Conference on Music Information Retrieval*, 2004.
5. F. Gouyon, S. Dixon, E. Pampalk and G. Widmer, "Evaluating Rhythmic Descriptors for Musical Genre Classification," *Proceeding of AES 25th International Conference*, 2004.
6. E. Pampalk, S. Dixon and G. Widmer, "Exploring Music Collections by Browsing Different Views," *Proceeding of International Conference on Music Information Retrieval*, 2003.
7. V. Tyagi, I. McCowan, H. Misra, and H. Bourland, "Mel-Cepstrum Modulation Spectrum (MCMS) Features for Robust ASR," *Proceeding of IEEE Workshop on Automatic Speech Recognition and Understanding*, 2003.
8. G. Peeters, A. L. Burthe, and X. Rodet, "Toward Automatic Music Audio Summary Generation from Signal Analysis," *Proceeding of International Conference on Music Information Retrieval*, 2002.
9. S. Sukittanon, L. E. Atlas, and J. W. Pitton, "Modulation-Scale Analysis for Content Identification," *IEEE Transactions on Signal Processing*, vol. 52, no. 10, pp. 3023~3035, Oct. 2003.
10. Y.Y. Shi, X.Zhuan, H.G.Kim, and K.W.Eom, "A Tempo Feature via Modulation Spectrum Analysis and Its Application to Music Emotion Classification," *Proceeding of the International Conference on Multimedia Expo*, 2006.
11. K. Tanghe, et al, "Collecting Ground Truth Annotations for Drum Detection in Polyphonic Music," *Proceeding of International Conference on Music Information Retrieval*, 2005.
12. P.N. Juslin, and J.A. Sloboda, "Music and Emotion: Theory and Research," Oxford Univ. Press, USA. 2001.
13. D. Liu, L. Lu, and H.J. Zhang, "Automatic Mood Detection from Acoustic Music Data," *Proceeding of International Conference on Music Information Retrieval*, 2003
14. Y.Z. Feng, Y.T. Zhuang, and Y.H. Pan, "Music Information Retrieval by Detecting Mood via Computational Media Aesthetics," *IEEE/WIC International Conference on Web Intelligence*, 2003.
15. N. Cristianini, J. Shawe-taylor, “An Introduction to Support Vector Machines”, Cambridge University Press, 2000.
16. Available at *http://www.csie.ntu.edu.tw/~cjlin/libsvm/*
17. B.L. Feldman, and J.A. Russell, "Independence and Bipolarity in the Structure of Affect," *Journal of Personality and Social Psychology*, v.74, pp.967~984, 1998.