A TEMPO FEATURE VIA MODULATION SPECTRUM ANALYSIS AND ITS APPLICATION TO MUSIC EMOTION CLASSIFICATION

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

Samsung Advanced Institute of Technology

{yy.shi, xuan.zhu, hyounggook.kim, kiwan.eom}@samsung.com

ABSTRACT

This paper proposes a tempo feature extraction method based on the long-term modulation spectrum analysis. To transform the modulation spectrum to a condensed feature vector, the log-scale modulation frequency coefficients are introduced. This idea aims at averaging the modulation frequency energy via the constant-Q filter-banks. Further it is pointed out that the feature can be extracted directly from the perceptually compressed data of digital music archives. To verify the effectiveness of the feature and its utility to music applications, the feature vector is used in a music emotion classification system. The system consisting two layers of Adaboost classifiers. In the first layer the conventional timbre features are employed. Then by adding the tempo feature in the second layer, the classification precision is improved dramatically. By this way the discriminability of the classifier based on the given features can be exploited extremely. The system obtains high classification precision on a small corpus. It proves that the proposed feature is very effective and computationally efficient to characterize the tempo information of music.

1. INTRODUCTION

In musical terminology, tempo is the beat rate of music corresponding to human perceived music speed and it is measured by the number of beats per minute. It is one of the basic elements of music. Musicians compose tempo and its movements according to their tastes and moods.

Much endeavour has been put to estimate the tempo from music signals, but not very successfully. MIREX'05 reports that the highest tempo extraction precision is 96% (at least one tempo correct), 55.71% (two tempos correct), 25% (at least one phase correct) and 5% (two phases correct), respectively, at the 140 wave files [1]. Although the accurate estimation is very difficult, it has become a major topic in the music information retrieval community and an indispensable part in the MIREX evaluations [2].

Unlike the tempo estimation method several other approaches are proposed, which bypass the quantity estimation and present the tempo information in a feature set perceptually. Among them the two most typical methods are the beat spectrum [3] and the beat histogram [4]. Specifically, the beat spectrum is estimated from the diagonal sum of the similarity matrix computed on an audio feature vector. The strong periodicity indicated by the strong peaks at the beat spectrum corresponds to the length of a note phase. While in the beat histogram method, the autocorrelation function of the energy envelop is calculated and its peaks are detected. Consequently, the periodicities corresponding to the peaks are accumulated in a histogram. Both methods try to estimate a "spectrum" whose coefficients present strength at the periodicities.

In this paper we propose a computationally efficient tempo feature extraction method based on the long-term modulation spectrum analysis via fast Fourier transform (FFT). To verify the feature's effectiveness and its utility to music signal processing applications, it is applied to automatic 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 piece. The topic here is trying to tell the emotion expressed in the music solely relying on the audio data.

In the milestone work of music and emotion research [5], the structural features are suggested as the most important factor. Structural features can be subdivided into two types: segmental and suprasegmental features. Segmental features consist of the acoustic characteristics, which are described by the duration, energy, pitch and timbre or harmonic structure of the tones. Their effects on emotion are relatively stable and universal. Suprasegmental features in music are melody, tempo, rhythm, harmony and other aspects of musical structure and form. They carry emotional information primarily through symbolic coding, as based on a process of historically evolved, sociocultural conventionalization. So basically the structural features should be extracted from the audio signals and employed for the automatic emotion classification.

Previous work can be found in [6] and [7]. [6] 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

1-4244-0367-7/06/$20.00 ©2006 IEEE

strength, stability and ratio of onsets. The system obtains from 76.6% to 94.5% precision for different emotion category. [7] 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 from 25% to 86%. So the tempo or rhythm feature is a decisive factor.

Unlike the previous work, the proposed tempo feature is employed rather than the specific estimate of beat rate, or information of onsets. Other novelties exist in that all the features are extracted directly from the perceptually compressed data and Adaboost is adopted for high precision classifier training and feature selection.

In Section 2 the tempo feature and the compression domain extraction method are introduced. Section 3 explains the emotion classification system in detail. Section 4 describes the experiment settings and results. The final Section draws the conclusion.

2. TEMPO FEATURE EXTRACTION

2.1. Motivation

The basic idea is to decompose the tempo periodicities to different frequency components. It can be implemented directly by doing the FFT in a long-term analysis window. Figure 1 visualizes the decomposition effects, where (a) shows the spectrum of music pieces, (b) illustrates the waveforms of band-pass filtered signals, and (c) is the result of doing long-term FFT on (b). The figures in the left column come from a piece of piano solo with 0.6Hz tempo; while the figures in the right column are extracted from the music of African drum beating with 2.4Hz tempo. Clearly the peaks at the amplitude spectrum correspond to the tempo frequency and its harmonics.

There are two key processes:

(1) Band-pass filtering: it can emphasize the energy envelop of the rhythm instrument;

(2) Long-term Fourier analysis: the several-seconds analysis window can guarantee the decomposition effects on the tempo information.

Theoretically, the above processes are the modulation spectrum analysis. The decomposed tempo frequency is equivalent to the modulation frequency in that paradigm. The amplitude spectrum is named as modulation spectrum.

2.2. Tempo feature

According to [8], the human perception of the modulation frequency also abides by the constant-Q effect. An efficient way to simulate this effect is to apply a set of logarithmic-scale triangular filters on the modulation spectrum. The output of the triangular filter-banks can represent the perceptual modulation spectral shape, which is named as log-scale modulation frequency coefficients (LMFC) here.

LMFC can be extracted from the audio signals easily. Also it can be extracted directly from the most popular music compression data, MPEG-1 layer3 (MP3) archive. And the method can be extended to other compression formats, for example, AC-3, Ogg Vorbis and AAC.

The tempo feature estimation requires taking the long-term modulation spectrum analysis on the low-pass filtered signals. The modified discrete cosine transform coefficients (MDCT), which can be accessed via an MP3 partial decoder, are sub-band filtered signals of music audio data. The MDCT coefficients are down-sampled by a factor of 576 with around 13ms time interval for frames. MDCT-LMFC can be extracted by performing the long-term Fourier analysis on the low-frequency MDCT coefficients.

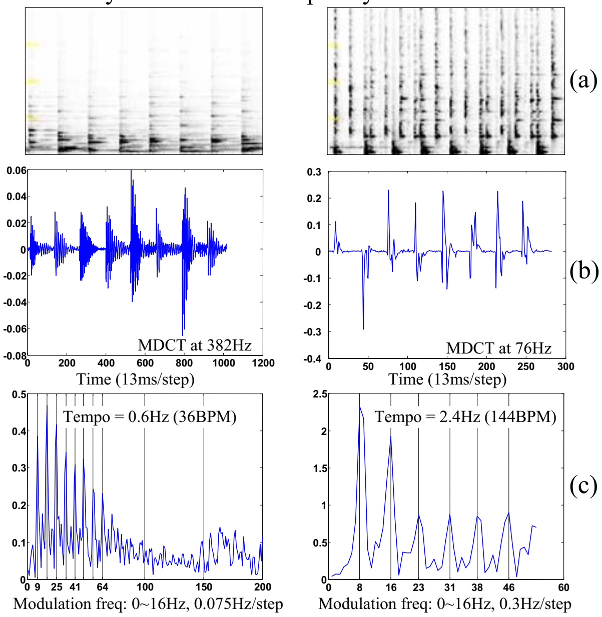

Figure 1: Decomposition of tempo frequency

3. EMOTION CLASSIFICATION

3.1. System structure

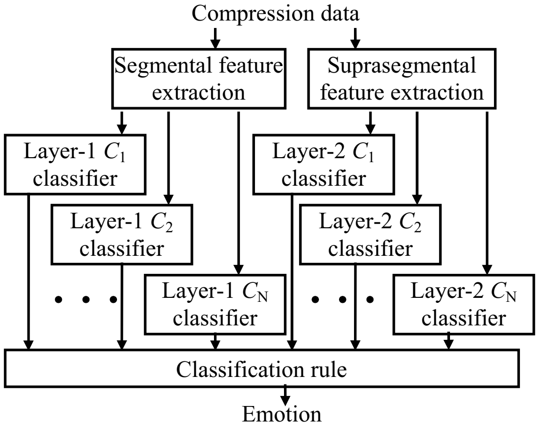

Figure 2: System structure

Figure 2 gives the system structure of emotion classification. It is composed of three parts: feature extraction, hierarchical classifier and classification rule. Segmental features and suprasegmental features are extracted, and then input into the first and the second layers of the classifier respectively. Each single classifier in any layer is trained by Adaboost [9]. Finally the emotion is output according to the classification rule.

3.2. Feature extraction

Segmental features include the intensity and timbre of music. Suprasegmental feature, LMFC, is the tempo feature proposed in this paper. Besides LMFC, the intensity and timbre features are also extracted in the compression domain.

Intensity is represented by the averaged scale factor (*scf*) of each MP3 frame. Also its delta value (*dscf*) is calculated as in Equation (1).

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

When extracting acoustic features from mp3, *scf* is often used to simulate the frame intensity.

The timbre features presented in [6] is adopted here. They include spectral centroid, spectral bandwidth, roll-off frequency and spectrum flux of the amplitude spectrum. Also the peaks, valleys, arithmetic mean, flatness and crest values on the 7 logarithmic sub-bands are used. The specific calculation can be found in [6]. Here we substitute the MDCT coefficients for the amplitude spectrum. This manipulation is proposed recently to extract the timbre features directly from mp3 frame [10]. Thus, 41 values extracted from the MDCT coefficients and the scale factors of each mp3 frame are used as the segmental features in this paper.

As for the suprasegmental features, we introduce the MDCT-LMFC feature to indicate the tempo information of music. The detailed extraction process is listed as follows in the case of 44100 Hz sampling rate:

(1) The MDCT coefficients in the first 5 MDCT sub-bands (<200Hz) are used as the low-frequency sub-band filtered signals with 38.3Hz bandwidth.

(2) The deviation operation between two adjacent frames is performed on each sub-band of signal to sharpen the fast changing effects of the signals.

(3) The modulation spectrum is obtained by doing a 256-point FFT with a 3-second hamming window on the deviation signal. The time shift of hamming window is 1-second.

(4) The 12-order LMFC is estimated from the 128 amplitude coefficients of the modulation spectrum by applying the log-scale triangular filter-banks on the modulation spectrum.

(5) The final 12×5 coefficients constitute the tempo feature.

3.3. Hierarchical classifier

The classifier includes two layers, each responsible for one type of feature. In every layer there are numbers of pairwise classifiers, each of which is responsible for distinguishing between class C and its anti-class $\tilde{C}$. This structure facilitates to use Adaboost which is a pairwise training method and has the advantages of high classification precision, optimal feature selection and model parameter adjusting.

Adaboost is a boosting procedure in which weak classifiers can be integrated into a strong classifier by adaptively adjusting the “training weight” of each training sample. In this paper each pairwise classifier includes 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.

Each two-class classifier labels a positive (C) or negative ($\tilde{C}$) of the input feature vector. The classification rule determines the emotion of the music piece according to the ratio between the positive frames and the negative frames:

$$I = \arg\max_{j} \left\{ \alpha \frac{N_{1,Cj}}{N_{1,Cj} + \tilde{N}_{1,Cj}} + (1-\alpha) \frac{N_{2,Cj}}{N_{2,Cj} + \tilde{N}_{2,Cj}} \right\} \tag{2}$$

In Equation (2) $N_{i,Cj}$ is the number of positive frames of class *j* in layer *i*, and $\tilde{N}_{i,Cj}$ is the number of negative frames of class *j* in layer *i*. *N* is the number of emotion classes. *I* is the emotion output. α =0.7 in our implementation.

3.4. Music emotion categories

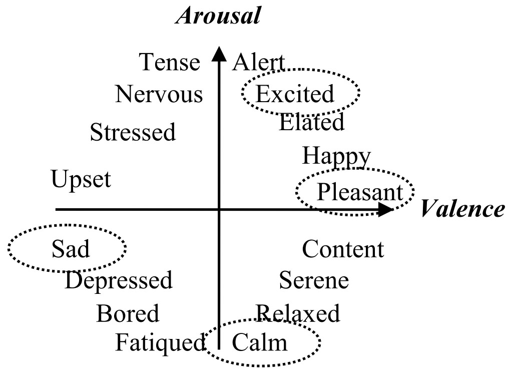

Figure 3: Dimensional map of basic emotions

How to classify the emotion categories and how to label them are always difficult issues for music emotion classification evaluation. There is no consensus on the formulation in musicology and psychology.

Dimensional theory is an accepted approach to define the basic emotions expressed in music. It reduces the various emotion categories to a set of dimensions, mostly in a two-dimensional representation. Typically Russell decomposes emotions along a *valence* dimension from negative to positive and an *arousal* dimension from inactive to active [11]. Figure 3 is an illustration. Among them *calm*, *sad*, *pleasant* and *excited* are selected, because 1) they distribute on the dimensional map separately, and 2) they are able to be labelled consistently by listeners.

In contrast, we find that it is not easy to find the music piece agreed as alert; even the piece has been agreed as excited, it is still possible to be felt as tense by another listener. So our criterion for choosing music emotion categories is only selecting the relatively consistent and widely accepted emotions in music.

4. EXPERIMENT

In order to get the ground truth data of music emotion, here the verbal self-report criterion is adopted. That is, listeners are asked to describe that the music piece is supposed to indicate one of the emotions or none of them in response to different genres of music. 3 females and 3 males (Korean and Chinese) in the ages of 20~35 attend the labelling work in an ordinary office cubic via earphone listening to CD music played by a computer. 194 homogeneous pieces are selected from hundreds of western classical, opera singing, jazz, electronic, popular and rock. They are agreed to indicate one of the 4 emotions by the 6 persons. Among them 43 are calm, consisting of western classical, opera singing and soft pop; 30 are excited, selected from rock and heavy mental; 65 are pleasant, consisting of western and Asian popular, electronic and several pieces of classical; 56 are sad, consisting of classical, jazz and popular. The genre is rough because sometimes it is hard to distinguish among them.

The experiment is carried out on the 194 pieces. The cross-validation is adopted due to the limited data. The experiment has been repeated 5 times. During each time, 80% pieces in every emotion class are used as the training data and the left 20% pieces are used as the testing data. The precision is then averaged from the results of the 5 tests. Table 1 lists the precision.

Table 1: Emotion classification precision

Emotion	Segmental	LMFC	Segmental & LMFC
Calm	86.0%	88.0%	93.0%
Excited	95.0%	95.0%	93.3%
Pleasant	81.5%	85.5%	95.4%
Sad	85.5%	92.5%	85.7%
Average	86.1%	90.2%	92.8%

Clearly the LMFC tempo feature can discriminate the 4 emotions as well as the segmental features can do. The precision is improved when the two types of features are combined into one system.

5. CONCLUSION

This paper proposes a tempo feature extraction method based on the long-term modulation spectrum analysis. To transform the modulation spectrum to a condensed tempo feature vector, the LMFC is introduced to smooth the modulation frequency energy via the constant-Q filter-banks. Also the feature can be extracted directly from the MDCT coefficients.

The MDCT-LMFC is applied in a music emotion classification system to evaluate its effectiveness. The classification precision is improved to a great extent. Although the small corpus and the lack of tempo transcription limits the generalization of the conclusion, it still can be proved that LMFC is a very effective and computationally efficient feature to characterize the long term dynamics of music well, in which the tempo frequency and its harmonics are presented. Moreover, it also can present that a computer can "perceive" the simple emotional state of music by the segmental and suprasegmental features.

6. REFERENCES

[1] M. Alonso, B. David, and G. Richard, "Tempo Extraction for Audio Recordings", *MIREX 2005*.

[2] http://www.music-ir.org/mirex2005/index.php/Main_Page

[3] J. Foote, "The Beat Spectrum: A New Approach to Rhythm Analysis," *ICME 2001*.

[4] G. Tzanetakis, and P. Cook, "Musical Genre Classification of Audio Signals," *IEEE Trans. on Speech and Audio Proc*, 2002.

[5] P.N. Juslin, and J.A. Sloboda, "Music and Emotion: Theory and Research", Oxford Univ. Press, 2001.

[6] D. Liu, L. Lu, and H.J. Zhang, "Automatic Mood Detection from Acoustic Music Data", *ISMIR 2003*.

[7] Y.Z. Feng, Y.T. Zhuang, and Y.H. Pan, "Music Information Retrieval by Detecting Mood via Computational Media Aesthetics," *IEEE/WIC International Conf. on Web Intelligence 2003*.

[8] S. Sukittanon, L.E. Atlas, and J.W. Pitton, "Modulation-Scale Analysis for Content Identification", *IEEE Trans. on Signal Processing*, v.52, n.10, pp.3023~3035, 2003.

[9] R.E. Schapire, "A Brief Introduction To Boosting," *the 16th International Joint Conf. on Artificial Intelligence*, 1999.

[10] Y. Wang, and M.S. Kankanhalli, "Automatic Music Summarization in Compressed Domain," *ICASSP 2004*.

[11] 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.