THE JOURNAL OF THE ACOUSTICAL SOCIETY OF KOREA VOL.25, NO.2E 2006. 6 pp. 74-78

Automatic Emotion Classification of Music Signals Using MDCT-Driven Timbre and Tempo Features

Kim, Hyoung-Gook*, Eom, Ki-Wan*
*Samsung Advanced Institute of Technology Computing Lab
(Received May 2 2006; Revised Jun 9 2006; Accepted Jun 13 2006)

Abstract

This paper proposes an effective method for classifying emotions of the music from its acoustical signals. Two feature sets, timbre and tempo, are directly extracted from the modified discrete cosine transform coefficients (MDCT), which are the output of partial MP3 (MPEG 1 Layer 3) decoder. Our tempo feature extraction method is based on the long-term modulation spectrum analysis. In order to effectively combine these two feature sets with different time resolution in an integrated system, a classifier with two layers based on AdaBoost algorithm is used. In the first layer the MDCT-driven timbre features are employed. By adding the MDCT-driven tempo feature in the second layer, the classification precision is improved dramatically.

Keywords: *Modified discrete cosine transform coefficients, MP3 decoder, Adaboost algorithm*

I. Introduction

Due to the rapid growth of digital music database, how to find out favorite songs in the vast data becomes a big problem for users.

A straightforward method of solving this problem is to categorize and index the songs via the ID3 tag information in the compressed digital music archives, such as song's title, artist, album, and genre. But if the ID3 tag information in our favorite music collections is incomplete or incorrect, this method does not work. In this case, a feasible way is to introduce music similarity measures, which cluster similar acoustic signals of music songs into one category.

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.

The milestone work of music and emotion research is described in [1]. Previous most works on automatic music emotion classification are described in [2] and [3]. [2] uses intensity, timbre and rhythm features and GMM classifiers to recognize the four emotional states of exuberance, anxiousness, contentment and depression. [3] recognizes the four emotions of happiness, sadness, anger and fear by a simple neural network classifier and three kinds of features, which are mean and variance of the silence ratio and the beat rate estimated by a beat tracking algorithm.

Nowadays, most conventional algorithms in the domain of music content-based retrieval extract features from the waveform of audio signals. Unlike the conventional methods, the proposed features in this paper are extracted directly from the perceptually compressed data of digital music archives. Specially, a computationally efficient tempo feature extraction method based on modulation spectrum estimation is employed rather than the specific estimate of beat rate, or information of onsets. The AdaBoost is adopted for high precision classifier training and feature selection. The AdaBoost [8] is an adaptive algorithm to boost a sequence of weak classifiers by dynamically changing the weights associated with the examples based on the errors in the

Corresponding author: Kim, Hyoung-Gook
(hyounggook.kim@samsung.com)
SAIT, Mt, 14-1 Nongseo-dong, Giheung-gu, Younggin-si
Gyunggi-do, 449-712

previous learning so that more attention will be paid to the wrongly classified examples.

Section II explains the proposed automatic emotion classification. Section III describes the experiment settings and results. The final Section draws the conclusion.

II. Automatic Emotion Classification

For the large amount of compressed music archives such as MP3 music files, there are three necessary steps in the stage of the feature extraction: decoding compressed data to waveform, performing Discrete Fourier Transform (DFT), and then extracting timbre or rhythmic features. But in fact, the partial decoding process of compressed music archives has embedded sub-band filtered signals; for example, the Modified Discrete Cosine Transform (MDCT) coefficients, which are available during decoding MP3, AC-3, Ogg Vorbis, AAC, etc. files. Instead of the magnitude spectrum of the PCM-based audio signals, directly getting features from the MDCT can remarkably improve the efficiency of feature extraction. As reported in [4], extracting cepstral features from the compressed data is approximately six times faster than traditional feature extraction scheme. The comparison of two kinds of feature extraction methods is illustrated in Figure 1. The detailed MP3 decoding description can be found in [9].

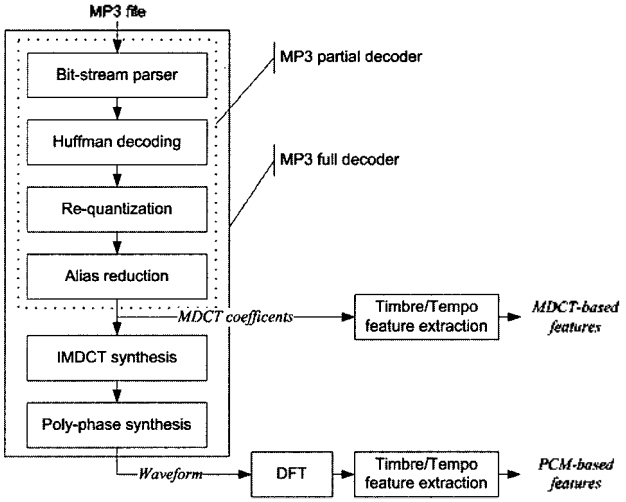

Figure 1. MDCT-based vs. DFT-based Feature Extraction.

The system structure of the emotion classification is shown in Figure 2. It is composed of three parts: feature extraction, two-layer classifier and classification rule. Timbre and Tempo features are extracted directly from the modified discrete cosine transform coefficients (MDCT), which are the output of partial MP3 (MPEG 1 Layer 3) decoder.

And then the features input into the first and the second layers of the classifier respectively. Each single classifier in any layer is trained by AdaBoost. Finally the emotion is the output according to the classification rule.

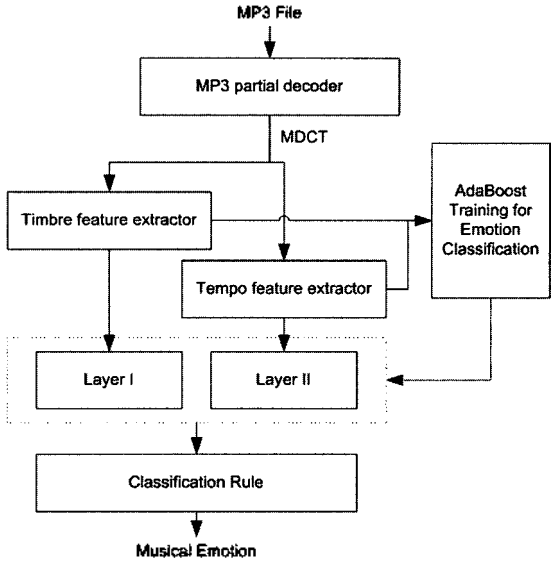

Figure 2. Block diagram of musical emotion classification.

The timbre features are computed from the MDCT coefficients, which are down-sampled by a factor of 576 with around 13ms for frames. The timbre features include the spectral centroid, the spectral bandwidth, the roll-off frequency and the spectrum flux of the MDCT coefficients. All the MDCT-driven timbre features used in this paper are extracted with the frequency range from 65Hz to 8372Hz covering 7 octaves. N_0 and N in the equations (1)-(5) indicate the number of the MDCT coefficients corresponding to 65Hz and 8372Hz, respectively. $S_t(i)$ is used to represent the MDCT coefficient index.

· The MDCT spectral centroid is the first order moment of MDCT coefficients at each frame and indicates an approximation of the perceptual sharpness of the signal.

$$C_t = \frac{\sum_{i=N_0}^{N}\left(|S_t(i)|^2 \times i\right)}{\sum_{i=N_0}^{N}|S_t(i)|^2} \quad (1)$$

· The MDCT spectral bandwidth is the square root of the weighted average difference between MDCT coefficients and the centroid. It indicates the shape of MDCT coefficients on the frequency and it is designed to help differentiating noise-like and

tonal sounds.

$$B_t = \sqrt{\frac{\sum_{i=N_0}^{N}\left(|S_t(i)|^2 \times (i - C_t)^2\right)}{\sum_{i=N_0}^{N}|S_t(i)|^2}} \tag{2}$$

· he MDCT spectral roll-off is defined as the frequency R_t below which 95% of the accumulated power spectrum is concentrated. The value of roll-off is also related to the shape of the magnitude of MDCT coefficients.

$$\sum_{i=N_0}^{Rt}|S_t(i)|^2 = 0.95 \times \sum_{i=N_0}^{N}|S_t(i)|^2 \tag{3}$$

· he MDCT spectral flux is the squared difference between two adjacent frames of the magnitude of MDCT coefficients. It represents the quantity of the MDCT subband components change.

$$F_t = \sum_{i=N_0}^{N}\left(|S_t(i)| - |S_{t-1}(i)|\right)^2 \tag{4}$$

· he MDCT subband flatness is estimated as the ratio between the geometric mean and the arithmetic mean of the magnitude of MDCT coefficients. It expresses the derivation of the signal's power spectrum from the flat shape.

$$L_t = 10 \times \log\left(\frac{\sqrt[(N-N_0+1)]{\prod_{i=N_0}^{N}|S_t(i)|^2}}{\left(\sum_{i=N_0}^{N}|S_t(i)|^2\right)/(N - N_0 + 1)}\right) \tag{5}$$

· The MDCT subband contrast feature set:

The octave-based spectral contrast feature set is designed to provide better discriminability than MFCC in the field of the music classification [5]. It is composed of peak/valley/mean values in octave-scale subbands. In this paper, we extract a similar feature set based on MDCT coefficients.

Here, we choose seven octave-scale subbands: 65-131Hz, 131-262Hz, 262-523Hz, 523-1047Hz, 1047-2093Hz, 2093-4186Hz, and 4186-8372 Hz. For the k-th subband, we can get the corresponding MDCT magnitude vector as the descending order of $\{|S_t(k,1)| > |S_t(k,2)| > \cdots > |S_t(k,N_k)|\}$. Peak/valley/mean values of the MDCT magnitude in the k-th octave-scale subband are calculated as the following equations:

$$P_t(k) = \log\left(\frac{1}{\alpha N_k}\sum_{i=1}^{\alpha N_k}|S_t(k,i)|^2\right) \tag{6}$$

$$V_t(k) = \log\left(\frac{1}{\alpha N_k}\sum_{i=1}^{\alpha N_k}|S_t(k,N_k - i + 1)|^2\right) \tag{7}$$

$$M_t(k) = \log\left(\frac{1}{N_k}\sum_{i=1}^{N_k}|S_t(k,i)|^2\right) \tag{8}$$

where N_k is the number of MDCT components in the k-th subband and α is a constant with the value between 0.02 and 0.2. With the factor α, the peak and valley magnitude can be estimated by averaging a small neighborhood around the maximum and the minimum of the subband.

Each timbre feature vector comprises 26 components described above. It includes the MDCT spectral centroid, the MDCT bandwidth, the MDCT roll-off, flux and the MDCT spectral contrast feature set, which contains 21 feature components. The timbre feature synchronizes with the MP3 frame, so the frame time shift for each feature vector is about 13ms.

The basic idea of the MDCT-driven tempo feature extraction is to decompose the tempo periodicities to different frequency components. As the range of tempo is very narrow from 30 to 300 BPM (0.5~5 Hz), it can be implemented directly by the long-term discrete Fourier analysis on the MDCT subband signal in lower frequency. Since the power of rhythmic instruments is mostly distributed in low frequencies, the energy envelope changing will be sharpened by the low frequency band-pass filtering. In order to emphasize the fast changing in the signal, a full wave rectification and the amplitude deviation between time-neighboring frames are applied on the MDCT subband signal. Hereafter, a 3-order low-pass filter with 10 Hz cut-off frequency is employed to remove the disturbance beyond the range of tempo. Then, a DFT with the long-term analysis window is performed to get the strength of tempo frequency components.

Theoretically, the above process is the modulation frequency analysis. And the decomposed tempo frequency corresponds to the modulation frequency in that paradigm. The amplitude spectrum is named as the modulation spectrum [6].

According to [6], the human perception of the modulation frequency also abides by the constant-Q effect. An efficient approach to simulate this effect is to filter the amplitude spectrum by a logarithmic filterbank. Specifically, we use the

filterbank formed by 12 triangular window filters with logarithmically increasing bandwidth and overlapping each other. The outputs of the triangular filters are adopted as our tempo feature, which is named as log-scale modulation frequency coefficients (LMFC).

The tempo feature extraction is achieved by the following steps for a 44.1 kHz MP3 file:

- :ep1. Partially decoding MP3 file to the MDCT coefficients
- :ep2. Keeping MDCT coefficients in the lowest 5 subbands (less than 200 Hz) to form the sub-band filtered signals in the 5 bands with 38 Hz bandwidth;
- :ep3. Applying full wave rectification and amplitude deviation frame-to-frame;
- :ep4. Employing a 3-order low-pass filter with 10 Hz cut-off frequency.
- :ep5. Performing 256-point FFT with 3-second hamming window;
- :ep6. Smoothing with a filterbank composed of 12 triangular window filters.

The final tempo feature vector makes up of 12×5=60 components. On account of the long-time property of tempo feature, the time resolution of LMFC frames is 1 second, which is much longer than timbre features.

Above MDCT-driven two feature sets respectively represent different characteristics of music and have their own specialties. The timbre feature vector consists of 26 components with 13ms time resolution while the tempo feature vector contains 60 components with 1-second time resolution. In order to effectively combine these two feature sets with different time resolution in an integrated system, a classifier with AdaBoost algorithm is used. Comparing with other traditional statistical modeling methods, AdaBoost has the advantages of high classification precision, effective feature selection and optimal model parameter adjusting.

The classifier is composed of two layers: the layer I is responsible for timbre features and the layer II is related to tempo features. In every layer, there are numbers of pairwise classifiers, each of them is responsible for distinguishing between the 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 the 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 either positive (C) or negative ($\tilde{C}$) 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:

$$I = \arg\max_{j} \left\{ \alpha \frac{M_{1,Cj}}{M_{1,Cj} + \tilde{M}_{1,Cj}} + (1-\alpha) \frac{M_{2,Cj}}{M_{2,Cj} + \tilde{M}_{2,Cj}} \right\} \quad (9)$$

In Equation (9), $M_{i,Cj}$ is the number of positive frames of class j in layer i, and $\tilde{M}_{i,Cj}$ is the number of negative frames of class j in layer i. I is the emotion output. $\alpha = 0.7$ in our implementation.

III. Experiment

Typically, Russell decomposes emotions along a valence dimension from negative to positive and an arousal dimension from inactive to active [7]. As the criterion of only selecting relatively consistent and widely accepted music emotions, we choose four music emotion classes: 1) sad, 2) calm, 3) pleasant, and 4) excited in the proposed system.

The dataset is composed by 800 songs, equally 200 songs for each cluster. In order to get the ground truth data of music emotion, 6 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. Independent trained listeners (3 females and 3males in the ages of 20~35) label each song. Only the song, which is consistently agreed as the same emotion category by three listeners, is accepted by the dataset. The length of each song is not less than 30 seconds. These songs are collected from different ways, e.g. compact disks, radio, internet, etc., to cover different recording qualities. Then, all of them are encoded as standard MP3 audio files with 44.1 kHz, stereo, 128kbps properties and stored in the database. All the classification results in this paper are calculated using 20-fold cross-validation evaluation, which means randomly selecting 80% songs for training and the left 20% songs for testing, iterating this process for 5 times, and finally averaging the 5 times of results.

The music emotion classification results of MDCT-driven

features vs. PCM-driven features are shown in Table. 1. Clearly the MDCT-driven LMFC tempo feature can discriminate the 4 emotions as well as the MDCT-driven timbre features can do. The precision is improved when the two types of features are combined into one system. And the combining of MDCT-driven timbre and tempo features yield a slightly low classification precision than the combing of PCM-driven timbre features or tempo features. But the MDCT-driven feature extraction is approximately three times faster than the PCM-driven feature extraction scheme

Table 1. Emotion classification precision.

Method	PCM-driven	MDCT-driven		
	timbre & tempo	timbre	tempo	timbre & tempo
Precision	92.3%	86.0%	74.5%	90.5%

The confusion matrix of the result of the proposed method is shown in Table 2. There are two confusion sets in these four music emotion categories: sad vs. calm and pleasant vs. excited, which are also hard to be decided for listeners' perception. Especially, 46 songs of 200 sad songs are false recognized as calm songs.

After checking the errors, we find that some songs express the sad emotion only by the content of lyrics, but our proposed features only can represent the timbre and tempo information of music songs. In this case, it is impossible to distinguish sad songs from calm songs.

Table 2. Music emotion classification confusion matrix.

	Sad	Calm	Pleasant	Excited
Sad	154	46	0	0
Calm	4	192	4	0
Pleasant	2	0	198	0
Excited	0	0	20	180

IV. Conclusions

This paper propose an effective method of classifying emotions of the music signals. A new MDCT-driven tempo feature extraction method (LMFC) is effectively combined with MDCT-driven timbre features by AdaBoost algorithm, which significantly improves the accuracy of the music classification.

Our research in this domain is still in the preliminary stage. Two directions are valuable to be further explored. Due to the fuzzy boundaries between different categories, the current "hard" classification strategy is not suitable for the music classification. With some confidence measures, "soft" classification scheme can be introduced and will be more adequate to these applications.

Our further research will include increasing the size of the music database. We will apply the proposed system to the automatic music genre classification and the similarity search.

References

1. P. N. Juslin, and J. A. Sloboda, "Music and emotion: theory and research," Oxford Univ. Press, 2001.
2. D. Liu, L. Lu, and H.J. Zhang, "Automatic mood detection from acoustic music data," in Proc. of 4th International Conf. on Music Information Retrieval 2003 (ISMIR2003), 26-30, 2003.
3. Y.Z. Feng, Y.T. Zhuang, and Y.H. Pan, "Music information retrieval by detecting mood via computational media aesthetics," in Proc. of IEEE/WIC International Conf. on Web Intelligence 2003, 235-241, 2003.
4. D. Pye, "Content-based methods for the management of digital music," in Proc. of ICASSP2000, 2473-2440, 2000.
5. D. Liu, L. Lu, and H.J. Zhang, "Automatic mood detection from acoustic music data," in Proc. of 4th International Conf. on Music Information Retrieval 2003 (ISMIR2003), 26-30, 2003.
6. S. Sukittanon, L. E. Atlas, and J. W. Pitton, "Modulation-scale analysis for content identification," IEEE Trans. on Signal Processing, **52** (10) 3023-3035, 2003.
7. B.L. Feldman, and J.A. Russell, "Independence and bipolarity in the structure of affect," Journal of Personality and Social Psychology, **74**, 967-984, 1998.
8. Y. Freund, and R.E. Schapire, "A decision-theoretic generalization of on-line learning and an application to boosting," Journal of Computer and system Sciences, **55** (1) 119-139, 1997.
9. ISO/IEC 11172-3, "Coding of moving pictures and associated audio for digital storage media, part3: audio". 1993.

[Profile]

•Kim, Hyoung-Gook

received the diploma degree in electrical engineering and the Ph. D. degree in computer science from the Technical University of Berlin, Berlin, Germany. From 1998 to 1999, he worked on speech recognition at Siemens AG. From 1999 to 2002, he was a Project Leader of the Speech Processing Laboratory at Cortologic AG. From 2002 to 2005, he served as Adjunct Professor of the Communication Systems Department, Technical University of Berlin. Since 2005 he joined Samsung Advanced Institute of Technology as a Project Leader. His research interests include music information retrieval, audiovisual content indexing and retrieval, speech enhancement, and robust speech recognition.

•Eom, Ki-Wan

received the B.S. degree in 1996 from the department of electronic engineering of Gwangju University, Gwangju, Korea, and the M.S. degree in 1998 and the Ph.D. degree in 2006 from the department of electronic engineering of Chonnam National University, Gwangju, Korea. In 2002, he joined Samsung Advanced Institute of Technology as a research staff member. His research interests are audio indexing, music processing, and speech synthesis.