AUTOMATIC MUSIC GENRE CLASSIFICATION USING MODULATION SPECTRAL CONTRAST FEATURE

Chang-Hsing Lee, Jau-Ling Shih, Kun-Ming Yu, and Jung-Mau Su

Department of Computer Science and Information Engineering,
Chung Hua University, HsinChu, Taiwan, R.O.C.
{chlee, sjl, yu, m09302040}@chu.edu.tw

ABSTRACT

In this paper, we proposed a novel feature, called octave-based modulation spectral contrast (OMSC), for music genre classification. OMSC is extracted from long-term modulation spectrum analysis to represent the time-varying behavior of music signals. Experimental results have shown that OMSC outperforms MFCC and OSC. If OMSC is integrated with MFCC and OSC, the classification accuracy is 84.03% for seven music genre classification.

1. INTRODUCTION

With the development of computer network, it becomes more and more popular to purchase and download digital music from the Internet. Therefore, effective management of a large digital music database is a substantial matter. It is helpful to manage a vast amount of music tracks if they are properly classified into a set of music genres in advance. To determine the music genre of a music track by experienced managers is a laborious and time-consuming work. Therefore, a number of automatic content-based music classification systems have been developed to deal with this problem [1-5]. In general, automatic music genre classification plays an important and preliminary role in a music information retrieval system.

To determine the music genre of a music piece, a number of content-based features are first extracted. In general, the features employed for music classification can be roughly categorized into three classes: timbral texture, rhythmic features, and pitch content features [1]. Timbral features are generally characterized by the properties related to instrumentations or sound sources such as music, speech, or environment signals. The features used to represent the timbral texture of a music track include zero crossing, spectral centroid, spectral flux, spectral rolloff, Mel-frequency cepstral coefficients (MFCC) [1, 6, 7], Daubechies wavelet coefficients histograms (DWCH) [2], and octave-based spectral contrast (OSC) [3, 4], etc. Rhythmic features provide the main beat and its strength of a music track. Several beat-tracking algorithms have been proposed to estimate the main beat and the corresponding strength [1, 8, 9]. Pitch features, mainly derived from the pitch histogram [1, 10], describe the melody of the music.

After the features are extracted from each music track, a classifier is employed to determine the music genre of the music track. Li et al. [2] have evaluated the performance of different classifiers, including support vector machines (SVM), *k*-nearest neighbor (k-NN), Gaussian mixture models (GMM), and linear discriminant analysis (LDA). SVM is always the best classifier for music classification in their comparative experiments. Pedro et al. [11] use self-organising neural maps (SOM) for music genre classification. Grimaldi et al. [12] use a set of features based on discrete wavelet packet transform (DWPT) to represent a music track. The classification performance is evaluated by four alternative classifiers: simple *k*-nearest neighbor, one-against-all, Round-Robin, and feature-subspace based ensembles of nearest neighbor classifiers. The best result is achieved by using the feature-subspace based ensembles of nearest neighbor classifiers. A number of studies try to use a specific classifier to improve the classification performance. However, the improvement is limited. In fact, employing effective feature sets will have much more effect on the classification accuracy [13]. In this paper, a feature based on modulation spectrum will be proposed for automatic music genre classification.

2. FEATURE EXTRACTION

A novel feature, called Octave-based modulation spectral contrast (OMSC), is proposed for music genre classification. In addition, two existing features, including MFCC and OSC, will be combined with OMSC to improve the classification accuracy.

2.1. Mel-Frequency Cepstral Coefficient (MFCC)

MFCC have been widely used for speech recognition due to their ability to represent the speech spectrum in a compact form. In fact, MFCC have been proven to be very effective in automatic speech recognition and in modeling the subjective frequency content of audio signals.

To derive the MFCC of a music piece, an input music signal is divided into a number of frames. MFCC are then computed for each frame. To characterize the whole music piece, the mean and standard deviation of MFCC of all frames in a music piece are used as features.

2.2. Octave-Based Spectral Contrast (OSC)

OSC was developed to represent the spectral characteristics of a music piece. It considers the spectral peak and valley in each sub-band separately. In general, spectral peaks correspond to harmonic components and spectral valleys correspond to non-harmonic components or noise in a music piece. Therefore, the difference between spectral peaks and spectral valleys will reflect the spectral contrast distribution.

For a music piece, it is first decomposed into a number of overlapped frames and the spectrum of each frame is obtained by FFT. The spectrum of each frame is then divided into a number of subbands by the octave scale filters shown in Table 1. Let $(x_{b,1}, x_{b,2}, \ldots, x_{b,N_b})$ denote the spectrum of the b-th sub-band, N_b is the total number of FFT frequency bins in the b-th sub-band. The spectrum is then sorted in a non-ascending order and is represented as $(x_{b,d(1)}, x_{b,d(2)}, \ldots, x_{b,d(N_j)})$ where $x_{b,d(1)} \geq x_{b,d(2)} \geq \ldots \geq x_{b,d(N_j)}$. The spectral peak and the spectral valley in the b-th sub-band is estimated as follows:

$$Peak_b = \log(\frac{1}{\alpha N_b}\sum_{i=1}^{\alpha N_b} x_{b,d(i)}), \quad (1)$$

$$Valley_b = \log(\frac{1}{\alpha N_b}\sum_{i=1}^{\alpha N_b} x_{b,d(N_b-i+1)}), \quad (2)$$

where α is a neighborhood factor (α is 0.2 in this paper). The spectral contrast between the spectral peak and the spectral valley of the b-th subband is given by:

$$SC_b = Peak_b - Valley_b. \quad (3)$$

To represent the whole music piece, mean and standard deviation of the spectral contrast and spectral peak of all frames are used as the spectral contrast features. In this paper, the number of subbands B is 8. Therefore, the feature dimension of OSC is $4B = 32$.

2.3. Octave-Based Modulation Spectral Contrast (OMSC)

MFCC and OSC capture only the short-term features of each music signal. To characterize the whole music piece, the most widely used approach is to compute the mean and standard deviation of all short-term features. In this paper, OMSC is extracted from long-term modulation spectrum analysis [14, 15] to describe the time-varying behavior of the music signals.

First, each music piece is decomposed into a number of overlapped frames and the spectrum of each frame is obtained by FFT. In this paper, the frame size is 1024 samples and the overlapping size between every two neighboring frames is 256 samples. The spectrum is divided into a number of subbands by a set of octave scale band-pass filters. Let $E_q[b]$ denote the sum of amplitude spectrum of all FFT bins within the b-th subband:

$$E_q[b] = \sum_{k=1}^{N_b} \left| X_{q,b}[k] \right|, 0 \leq b < B, \quad (4)$$

where B is the number of band-pass filters, N_b is the number of FFT bins in the b-th subband, $X_{q,b}[k]$ is the FFT bin value in the b-th subband of the q-th frame.

For each subband, the modulation spectrum is obtained by applying DFT on $E_q[b]$ along the time-trajectory of each subband within a texture window of P frames:

$$M_b[n] = \sum_{q=0}^{P-1} E_q[b] e^{-j2\pi\frac{q}{N}n}, 0 \leq n < P, 0 \leq b < B \quad (5)$$

where $M_b[n]$ is the modulation spectrum of the b-th subband. In this paper, the length of the texture window, P, is 512, which is about 6 seconds. For each subband, the average modulation spectrum of all texture windows of a music piece is computed:

$$\overline{M}_b[n] = \frac{1}{T}\sum_{t=1}^{T} \left| M_{t,b}[n] \right|, 0 \leq n < P \quad (6)$$

where $M_{t,b}[n]$ is the modulation spectrum of the b-th subband in the t-th texture window, T is the number of texture windows. Then, the average modulation spectrum of each subband is divided into J modulation frequency subbands (in this paper J is 32). Thus, the number of modulation frequency bins in each modulation frequency subband, B_M, is P/J. The modulation spectral contrast (MSC) and the modulation spectral valley (MSV) are then evaluated for the first $J/2$ subbands:

$$MSC_b[j] = \max_{j \times B_M \leq n < (j+1) \times B_M} (\overline{M}_b[n]) - \min_{j \times B_M \leq n < (j+1) \times B_M} (\overline{M}_b[n]), \quad (7)$$

$$MSV_b[j] = \min_{j \times B_M \leq n < (j+1) \times B_M} (\overline{M}_b[n]), \quad 0 \leq j < J/2. \quad (8)$$

Fig.1 gives an example of the modulation spectrum of the first subband. We can see that there exist regular peaks for Dance, Hip-Hop, and Rock music, whereas the modulation spectrum is smoother for Chamber and Orchestra music. From the modulation spectrum, the mean and standard deviation of MSC and MSV of all modulation subbands is computed as the features of the whole music piece:

$$u_{MSC}[b] = \frac{2}{J}\sum_{j=0}^{J/2-1} MSC_b[j] \quad (9)$$

$$\sigma_{MSC}[b] = \frac{2}{J}\sum_{j=0}^{J/2-1} (MSC_b[j] - u_{MSC}[b])^2 \quad (10)$$

$$u_{MSV}[b] = \frac{2}{J}\sum_{j=0}^{J/2-1} MSV_b[j] \quad (11)$$

$$\sigma_{MSV}[b] = \frac{2}{J}\sum_{j=0}^{J/2-1} (MSV_b[j] - u_{MSV}[b])^2 \quad (12)$$

As a result, the length of the feature vector is $4B$.

After the features are extracted from a music piece, a linear normalization process is applied to each feature value. Then, the normalized feature vector is transformed by LDA such that the classification accuracy can be further improved at a lower-dimensional feature space [16].

3. EXPERIMENTAL RESULTS

In the experiments, there are 1783 music tracks derived from compact disks. All music tracks in our database are

44.1 kHz, 16 bits, stereo wave files. Half of the music tracks are used for training and the others for testing. All the music tracks are classified into seven classes including 342 tracks of chamber (Ch), 405 tracks of dance (D), 183 tracks of hip-hop (H), 203 tracks of jazz (J), 178 tracks of orchestra (O), 201 tracks of popular (Po), and 271 tracks of rock (R) music. The nearest neighbor classifier is used for music genre classification. The performance is measured in terms of the classification accuracy:

$$CA = \frac{N_C}{N}, \tag{15}$$

where N_C is the number of music tracks that is correctly classified, and N is the total number of test music tracks.

The comparison of average classification accuracy and reduced feature dimension for various features with/without LDA is shown in Table 2. From this table, we can see that the performance of the proposed OMSC feature is better than MFCC and OSC. Furthermore, the classification accuracy can be improved up to 84.03% when OMSC is combined with MFCC and OSC. Table 3 shows the confusion matrices of the classification results of different features. Each row gives the classification results for a specific music genre. It can be shown that the classification accuracy of the dance music and the hip-hop music are better than other music types whereas the classification accuracy of the chamber music, jazz music, and pop music are worse than others.

4. CONCLUSION

A novel features, called octave-based modulation spectral contrast (OMSC), is proposed for music genre classification. OMSC is extracted from long-term modulation spectrum analysis to capture the time-varying behavior of music signals. Experimental results have shown that OMSC outperforms MFCC and OSC. If OMSC is integrated with MFCC and OSC, the classification accuracy is 84.03% for seven music genre classification.

5. REFERENCES

[1] G. Tzanetakis and P. Cook, "Musical genre classification of audio signals", *IEEE Trans. on Speech and Audio Processing*, Vol. 10, No. 3, pp. 293-302, July 2002.

[2] T Li, M. Ogihara, and Q. Li, "A Comparative study on content-based music genre classification", *ACM Conf. on Research and Development in Information Retrieval,* 2003, pp. 282-289.

[3] K. West and S. Cox, "Features and classifiers for the automatic classification of musical audio signals", *ISMIR'2004*.

[4] D. N. Jiang, L. Lu, H. J. Zhang, J. H. Tao, and L. H. Cai, "Music type classification by spectral contrast feature", *ICME'2002*, Vol. 1, pp. 113-116.

[5] K. Umapathy, S. Krishnan, and S. Jimaa, "Multigroup classification of audio signals using time-frequency parameters", *IEEE Trans. on Multimedia*, Vol. 7, No. 2, pp. 308-315, April 2005.

[6] E. Scheirer and M. Slaney, "Construction and evaluation of a robust multifeature speech/music discriminator", *ICASSP'1997*, Vol. 2, pp. 1331-1334.

[7] R. Vergin, D. O'Shaughnessy, and V. Gupta, "Compensated mel frequency cepstrum coefficients", *ICASSP'1996*, Vol. 1, pp. 323-326.

[8] M. E. P. Davies and M. D. Plumbley, "Beat tracking with a two state model", *ICASSP'2005*, Vol. 3, pp. 241-244.

[9] W. A. Sethares, R. D. Robin, and J. C. Sethares, "Beat tracking of musical performance using low-level audio feature", *IEEE Trans. on Speech and Audio Processing,* Vol. 13, No. 12, pp. 275-285, Mar. 2005.

[10] G. Tzanetakis, A. Ermolinskyi, and P. Cook, "Pitch histogram in audio and symbolic music information retrieval", *IRCAM'2002*.

[11] P. J. P. de Leon and J. M. Inesta, "Feature-driven recognition of music style", *PRIA'2003,* Vol. 2652, pp. 773-781.

[12] M. Grimaldi, P. Cunningham, and A. Kokaram, "An evaluation of alternative feature selection strategies and ensemble techniques for classifying music", *ECML'2003*.

[13] M. F. McKinney and J. Breebaart, "Features for audio and music classification", *ISMIR'2003*.

[14] S. Sukittanon, L. E. Atlas, J. W. Pitton, "Modulation-scale analysis for content identification", *IEEE Trans. on Signal Processing*, Vol. 52, No. 10, pp. 3023-3035, Oct. 2004.

[15] B. Kingsbury, N.Morgan, and S. Greenberg, "Robust speech recognition using the modulation spectrogram", *Speech Commun.*, Vol. 25, No. 1, pp. 117-132, 1998.

[16] R. Duda, P. Hart, and D. Stork, *Pattern Classification*, New York:Wiley, 2000.

Table 1. Frequency range of each octave-scale band-pass filter. (Sampling rate = 44.1 kHz)

Filter	Low Frequency (Hz)	High Frequency (Hz)
ϕ_1	0	100
ϕ_2	100	200
ϕ_3	200	400
ϕ_4	400	800
ϕ_5	800	1600
ϕ_6	1600	3200
ϕ_7	3200	8000
ϕ_8	8000	22050

Table 2. Average classification accuracy (*CA*, %) and reduced feature dimension (*D*).

Feature	without LDA		with LDA	
	CA	*D*	*CA*	*D*
MFCC	40.94	40	55.68	6
OSC	67.72	32	76.94	6
OMSC	72.89	32	78.52	6
MFCC + OSC	69.85	72	79.53	6
MFCC + OMSC	75.93	72	81.10	6
MFCC + OSC + OMSC	77.73	104	84.03	6

Table 3. Confusion matrices of the classification results.
(a) MFCC (b) OSC (c) OMSC (d) MFCC+OSC
(e) MFCC+OMSC (f) MFCC+OSC+OMSC

	Ch	O	D	H	J	Po	R
Ch	**57.31**	28.65	2.34	0.58	8.19	0.58	2.34
O	29.21	**44.94**	0.00	0.00	13.48	5.62	6.74
D	2.97	2.48	**71.78**	14.85	1.49	2.97	3.47
H	1.10	0.00	21.98	**48.35**	3.30	14.29	10.99
J	8.91	8.91	2.97	8.91	**56.44**	12.87	0.99
Po	7.00	6.00	4.00	11.00	11.00	**48.00**	13.00
R	2.96	6.67	14.07	9.63	3.70	16.30	**46.67**

(a)

	Ch	O	D	H	J	Po	R
Ch	**70.18**	25.73	0.00	0.00	4.09	0.00	0.00
O	11.24	**83.15**	0.00	0.00	4.49	1.12	0.00
D	0.00	0.00	**86.14**	9.90	0.00	1.98	1.98
H	0.00	0.00	8.79	**86.81**	0.00	3.30	1.10
J	16.83	1.98	0.00	2.97	**60.40**	14.85	2.97
Po	3.00	1.00	2.00	0.00	15.00	**70.00**	9.00
R	0.00	0.00	1.48	3.70	2.22	14.07	**78.52**

(b)

	Ch	O	D	H	J	Po	R
Ch	**71.93**	21.64	0.00	0.00	5.85	0.00	0.58
O	16.85	**83.15**	0.00	0.00	0.00	0.00	0.00
D	0.00	0.00	**90.10**	4.95	0.00	2.97	1.98
H	0.00	0.00	3.30	**92.31**	1.10	2.20	1.10
J	12.87	2.97	0.00	0.00	**69.31**	11.88	2.97
Po	2.00	1.00	2.00	1.00	12.00	**64.00**	15.00
R	0.74	0.74	0.74	4.44	0.74	17.78	**74.81**

(c)

	Ch	O	D	H	J	Po	R
Ch	**75.44**	21.05	0.00	0.00	3.51	0.00	0.00
O	16.85	**80.90**	0.00	0.00	2.25	0.00	0.00
D	0.00	0.00	**89.60**	6.44	0.50	1.98	1.49
H	0.00	0.00	12.09	**82.42**	0.00	2.20	3.30
J	10.89	3.96	0.00	2.97	**68.32**	9.90	3.96
Po	1.00	1.00	1.00	3.00	14.00	**72.00**	8.00
R	0.00	0.00	1.48	3.70	1.48	12.59	**80.74**

(d)

	Ch	O	D	H	J	Po	R
Ch	**77.78**	18.71	0.00	0.00	3.51	0.00	0.00
O	15.73	**83.15**	0.00	0.00	0.00	0.00	1.12
D	0.00	0.00	**90.59**	4.95	0.00	2.97	1.49
H	0.00	0.00	3.30	**93.41**	1.10	1.10	1.10
J	12.87	1.98	0.00	0.00	**72.28**	8.91	3.96
Po	2.00	0.00	2.00	4.00	11.00	**67.00**	14.00
R	0.74	0.00	0.74	4.44	0.74	14.81	**78.52**

(e)

	Ch	O	D	H	J	Po	R
Ch	**78.95**	18.13	0.00	0.00	2.92	0.00	0.00
O	13.48	**84.27**	0.00	0.00	2.25	0.00	0.00
D	0.00	0.00	**92.57**	4.46	0.00	2.48	0.50
H	0.00	0.00	3.30	**94.51**	0.00	1.10	1.10
J	9.90	1.98	0.00	0.00	**75.25**	8.91	3.96
Po	1.00	0.00	2.00	1.00	11.00	**76.00**	9.00
R	0.00	0.00	0.74	2.96	0.74	12.59	**82.96**

(f)

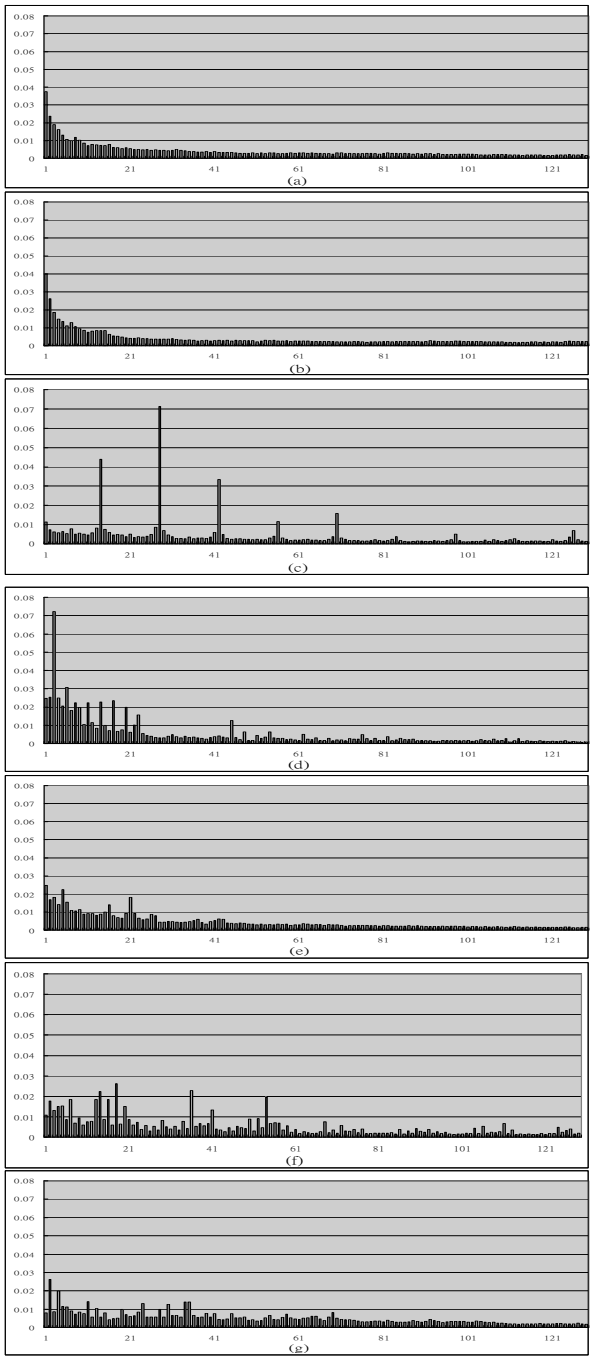

Fig.1 An example of the modulation spectrum of each genre in the first subband. (a) Chamber (b) Orchestra (c) Dance (d) Hip-Hop (e) Jazz (f) Pop (g) Rock