Automatic Music Genre Classification Based on Modulation Spectral Analysis of Spectral and Cepstral Features

Chang-Hsing Lee, Jau-Ling Shih, Kun-Ming Yu, and Hwai-San Lin

***Abstract*—In this paper, we will propose an automatic music genre classification approach based on long-term modulation spectral analysis of spectral (OSC and MPEG-7 NASE) as well as cepstral (MFCC) features. Modulation spectral analysis of every feature value will generate a corresponding modulation spectrum and all the modulation spectra can be collected to form a modulation spectrogram which exhibits the time-varying or rhythmic information of music signals. Each modulation spectrum is then decomposed into several logarithmically-spaced modulation subbands. The modulation spectral contrast (MSC) and modulation spectral valley (MSV) are then computed from each modulation subband. Effective and compact features are generated from statistical aggregations of the MSCs and MSVs of all modulation subbands. An information fusion approach which integrates both feature level fusion method and decision level combination method is employed to improve the classification accuracy. Experiments conducted on two different music datasets have shown that our proposed approach can achieve higher classification accuracy than other approaches with the same experimental setup.**

***Index Terms*—Mel-frequency cepstral coefficients, modulation spectral analysis, music genre classification, normalized audio spectrum envelope, octave-based spectral contrast.**

I. Introduction

With the development of computer networks, it becomes more and more popular to purchase and download digital music from the Internet. Since a typical music database often contains millions of music tracks, it is very difficult to manage such a large music database. It will be helpful in managing a vast amount of music tracks when they are properly categorized. The retail or online music stores often organize their collections of music tracks by categories such as genre, artist, and album. Music genre labels are often attached to an artist or an album and do not reflect the characteristic of a particular music track. The category information is often manually labeled by experienced managers or by the consumers (for example, CDDB,[1] MusicBrainz,[2] etc.). Perrot and Gjerdingen reported a human subject study in which college students were tested to make music genre classification among a total of ten music genres [1]. Their research found that humans with little to moderate musical training can achieve about 70% classification accuracy, based on only 300 ms of audio. Another study conducted the experiments on music genre classification by 27 human listeners. Each one will listen to the central 30 s of each music track and be asked to choose one out of six music genres. These listeners achieved an inter-participant genre agreement rate of only 76% [2]. These results indicate that there is significant subjectivity in genre annotation by humans. That is, different people classify music genres differently, leading to many inconsistencies. In addition, manual annotation is a time-consuming and laborious task. Thus, a number of supervised classification techniques have been developed for automatic classification of unlabeled music tracks [3]–[16]. In this study, we focus on the music genre classification problem which is defined as genre labeling of music tracks. In general, automatic music genre classification plays an important and preliminary role in a music organization or music retrieval system. A new album or music track can be assigned to a proper genre in order to place it in the appropriate section of an online music store or music database.

To determine the genre of a music track, some discriminating audio features have to be extracted through content-based analysis of the music signal. In general, the audio features developed for audio classification can be roughly categorized into three classes: short-term features, long-term features, and semantic features. Short-term features, typically describing the timbral characteristics of audio signals, are usually extracted from every short time window (or frame) during which the audio signal is assumed to be stationary. The timbral characteristics generally exhibit the properties related to instrumentations or sound sources such as music, speech, or environment sounds. The most widely used timbral features include zero crossing rate (ZCR), spectral centroid, spectral bandwidth, spectral flux, spectral rolloff, Mel-frequency cepstral coefficients (MFCC), discrete wavelet transform coefficients [4], [13], [17], octave-based spectral contrast (OSC) [5], [6], MPEG-7 normalized audio spectrum envelope (NASE) [18], [19], etc.

Generally, music genres not only correspond to the timbre of the music but also to the temporal structure of the music. That is, the time evolution of music signals will provide some useful information for music genre discrimination. To characterize the temporal evolution of a music track, long-term features can be generated by aggregating the short-term features extracted from several consecutive frames within a time window. The methods developed for aggregating temporal features include statistical

Manuscript received June 30, 2008; revised January 23, 2009. First published April 28, 2009; current version published May 15, 2009. This work was supported in part by the National Science Council of R.O.C. under contract NSC-96-2221-E-216-043. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Gerald Schuller.

The authors are with the Department of Computer Science and Information Engineering, Chung Hua University, Hsinchu 300, Taiwan (e-mail: chlee@chu.edu.tw; sjl@chu.edu.tw; yu@chu.edu.tw; m09502029@chu.edu.tw).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TMM.2009.2017635

[1]http://www.cddb.com

[2]http://musicbrainz.org

moments [3], [8], [12], [20], entropy or correlation [20], [21], nonlinear time series analysis [20], autoregressive (AR) models or multivariate autoregressive (MAR) models [11], modulation spectral analysis [8], [12], [16], [20], [22], etc.

Semantic features give meanings to audio signals in human-recognizable terms which generally reveal the human interpretation or perception of certain audio properties such as mood, emotion, tempo, genre, etc. The most widely known semantic descriptors include tempo (measured in beats-per-minute, BPM), rhythmic information, melody, pitch distribution, and so on. Rhythmic features can provide the main beat(s) and the corresponding strength of a music track. Several beat-tracking algorithms have been proposed to estimate the main beat and the corresponding strength [3], [23]. Pitch features, mainly derived from the pitch histogram [3], [24], can describe the harmonic contents or the pitch range of the music.

Once the features are extracted from a music track, a classifier will be employed to determine the genre of the given music track. Several supervised learning approaches, such as K-nearest neighbor (KNN) [3], [4], linear discriminant analysis (LDA) [4], Gaussian mixture models (GMM) [3], [4], [6], hidden Markov models (HMM) [18], Adaboost [14], regularized least-squares framework [15], and support vector machines (SVM) [4], [16], [25], have been employed for audio classification. Li *et al.* [4] evaluated the performance of different classifiers, including SVM, KNN, GMM, and LDA. SVM is always the best classifier for music classification in their comparative experiments. Grimaldi *et al.* [13] used a set of features based on discrete wavelet packet transform (DWPT) to represent a music track. The classification performance was evaluated by four alternative classifiers: KNN, 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 tried to use a specific classifier to improve the classification performance. However, the improvement is limited. In fact, it has been shown that employing effective feature sets will have much more effect on the classification accuracy [8].

In this paper, long-term modulation spectral analysis [26], [27] of MFCC [28], OSC [5], [6], and MPEG-7 NASE [18], [19] features will be used to characterize the time-varying behavior of music signals. A modulation spectrogram corresponding to the collection of modulation spectra of all MFCC (OSC or NASE) features will be constructed. Discriminative features are then extracted from each modulation spectrogram for music genre classification. Our contributions are as follows:

- proposal of some novel features based on long-term modulation spectral analysis of OSC and MPEG-7 NASE;
- extraction of effective and compact modulation spectral features from statistical aggregations of the modulation spectral contrasts (MSCs) and modulation spectral valleys (MSVs) computed from the modulation spectrograms of MFCC, OSC, and NASE;
- investigation of the importance of various modulation frequency ranges for music genre classification;
- development of an information fusion approach which integrates both feature level fusion and decision level fusion in order to improve the classification accuracy.

II. Proposed Music Genre Classification System

The proposed music genre classification system consists of two phases: the training phase and the classification phase. The training phase is composed of three main modules: feature extraction, linear discriminant analysis (LDA) [29], and information fusion. The classification phase consists of three modules: feature extraction, LDA transformation, and classification. A detailed description of each module will be described below.

A. Feature Extraction

A novel feature set derived from modulation spectral analysis of the spectral (OSC and NASE) as well as cepstral (MFCC) feature trajectories of music signals is proposed for music genre classification.

1) Mel-Frequency Cepstral Coefficients (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 [9], [28].

Before computing the MFCC of a music track, an input music signal is first pre-emphasized in order to amplify the high-frequency components, using a first-order FIR high-pass filter as follows:

$$s'(n) = s(n) - \alpha s(n-1) \tag{1}$$

where $s(n)$ is the input signal, $s'(n)$ is the emphasized signal, a typical value for the constant α is 0.95. The emphasized signal is then divided into a number of overlapped frames. To minimize the ringing effect, we multiply each frame by a Hamming window $w(n)$ given as follows:

$$w(n) = 0.54 - 0.46\cos\left(\frac{2\pi n}{N-1}\right),\ 0 \le n \le N-1 \tag{2}$$

where N is the length of the Hamming window. FFT is then applied on each pre-emphasized, Hamming windowed frame to obtain the corresponding spectrum. In this paper, the frame size is 23 ms (1024 samples for sampling frequency 44.1 kHz) with 50% overlap. The spectrum is then decomposed into a number of subbands by using a set of triangular Mel-scale band-pass filters. Let $E(b)$, $0 \le b < B$, denote the sum of power spectrum coefficients within the bth subband, where B is the total number of filters (B is 25 in this study). MFCC can be obtained by applying DCT on the logarithm of $E(b)$ as follows:

$$MFCC(l) = \sum_{b=0}^{B-1} \log_{10}(1+E(b))\cos\left(l\frac{\pi}{B}(b+0.5)\right),\ \ 0 \le l < L \tag{3}$$

where L is the length of MFCC feature vector (L is 20 in the study). Note that the offset of "1" in the calculation of MFCC is provided to get a positive logarithmic energy for any positive value $E(b)$. As a result, the MFCC feature vector can be represented as follows:

$$\mathbf{x}_{\mathrm{MFCC}} = [MFCC(0), MFCC(1), \ldots, MFCC(L-1)]^{\mathrm{T}}. \tag{4}$$

TABLE I
FREQUENCY RANGE OF EACH OCTAVE-SCALE BAND-PASS FILTER (Sampling rate = 44.1 kHz)

Filter number	Frequency range (Hz)
0	[0, 0]
1	(0, 100]
2	(100, 200]
3	(200, 400]
4	(400, 800]
5	(800, 1600]
6	(1600, 3200]
7	(3200, 6400]
8	(6400, 12800]
9	(12800, 22050]

2) Octave-Based Spectral Contrast (OSC): OSC was developed to represent the spectral characteristics of a music signal [5], [6]. It considers the spectral peak and valley in each subband independently. In general, spectral peaks correspond to harmonic components and spectral valleys the non-harmonic components or noise in music signals. Therefore, the difference between spectral peaks and spectral valleys will reflect the spectral contrast distribution.

To compute the OSC features, FFT is first employed to obtain the spectrum of each audio frame. This spectrum is then divided into a number of subbands by the set of octave scale filters shown in Table I. Let $(P_{b,1}, P_{b,2}, \ldots, P_{b,N_b})$ denote the power spectrum within the bth subband, N_b is the number of FFT frequency bins in the bth subband. Without loss of generality, let the power spectrum be sorted in a decreasing order, that is, $P_{b,1} \geq P_{b,2} \geq \cdots \geq P_{b,N_b}$. The spectral peak and spectral valley in the bth subband are then estimated as follows:

$$Peak(b) = \log\left(\frac{1}{\alpha N_b}\sum_{i=1}^{\alpha N_b} P_{b,i}\right) \tag{5}$$

$$Valley(b) = \log\left(\frac{1}{\alpha N_b}\sum_{i=1}^{\alpha N_b} P_{b,N_b-i+1}\right) \tag{6}$$

where α is a neighborhood factor (α is 0.2 in this study). The spectral contrast is given by the difference between the spectral peak and the spectral valley as follows:

$$SC(b) = Peak(b) - Valley(b). \tag{7}$$

The feature vector of an audio frame consists of the spectral contrasts and the spectral valleys of all subbands. Thus, the OSC feature vector of an audio frame can be represented as follows:

$$\mathbf{x}_{\mathrm{OSC}} = [Valley(0), \ldots, Valley(B-1), SC(0), \ldots, SC(B-1)]^{\mathrm{T}} \tag{8}$$

where B is the number of octave scale filters.

3) Normalized Audio Spectral Envelope (NASE): NASE was defined in MPEG-7 for sound classification [18], [19]. The NASE descriptor provides a representation of the power spectrum of each audio frame. Each component of the NASE feature vector represents the normalized squared magnitude of a particular frequency subband.

To extract the NASE features, each audio frame is pre-emphasized and multiplied by a Hamming window function and analyzed using FFT to derive its spectrum, notated $X(k)$, $1 \leq k \leq N$, where N is the size of FFT. The power spectrum is defined as the normalized squared magnitude of the DFT spectrum $X(k)$ as follows:

$$P(k) = \begin{cases} \frac{1}{N \cdot E_w}|X(k)|^2, & k = 0, \frac{N}{2} \\ \frac{2}{N \cdot E_w}|X(k)|^2, & 0 < k < \frac{N}{2} \end{cases} \tag{9}$$

where E_w is the energy of the Hamming window function $w(n)$ of size N_w:

$$E_w = \sum_{n=0}^{N_w-1} |w(n)|^2. \tag{10}$$

To reduce the number of spectral features, the power spectrum is divided into logarithmically spaced subbands spanning between 62.5 Hz ("*loEdge*") and 16 kHz ("*hiEdge*") over a spectrum of 8 octave interval centered at a frequency of 1 kHz (see Fig. 1). The number of logarithmic subbands within the frequency range [*loEdge*, *hiEdge*] is given by $B = 8/r$, where r is the spectral resolution of the frequency subbands ranging from 1/16 of an octave to 8 octaves as follows:

$$r = 2^j \text{octaves}, \ -4 \leq j \leq 3. \tag{11}$$

In this study, $r = 1/2$ is used for subband decomposition (i.e., $B = 16$), which provides adequate frequency decomposition for music genre classification in our experiments. When the resolution r is high, in the narrower low-frequency subbands a small frequency change will make it move from one subband to another. A reasonable solution is to assume that a power spectrum coefficient whose distance to a subband edge is less than half the FFT resolution will contribute to the audio spectral envelope (ASE) coefficients of both neighboring subbands. In this paper, a linear weighting method is used to compute the contribution of such a coefficient shared by its two neighboring subbands. In MPEG-7, the ASE coefficients consist of one coefficient representing power between 0 Hz and *loEdge*, a series of B coefficients representing power in logarithmically spaced subbands between *loEdge* and *hiEdge*, and a coefficient representing power above *hiEdge*. Therefore, a total number of $B+2$ ASE coefficients will be generated. The ASE coefficient for the bth subband is defined as the sum of power spectrum coefficients within this subband as follows:

$$ASE(b) = \sum_{k=loK_b}^{hiK_b} P(k), \ 0 \leq b \leq B+1 \tag{12}$$

where loK_b and hiK_b are, respectively, the integer frequency bins corresponding to the lower edge and higher edge of the bth subband. Each ASE coefficient is then converted to the decibel scale as follows:

$$ASE_{\mathrm{dB}}(b) = 10\log_{10}(1 + ASE(b)), \ 0 \leq b \leq B+1. \tag{13}$$

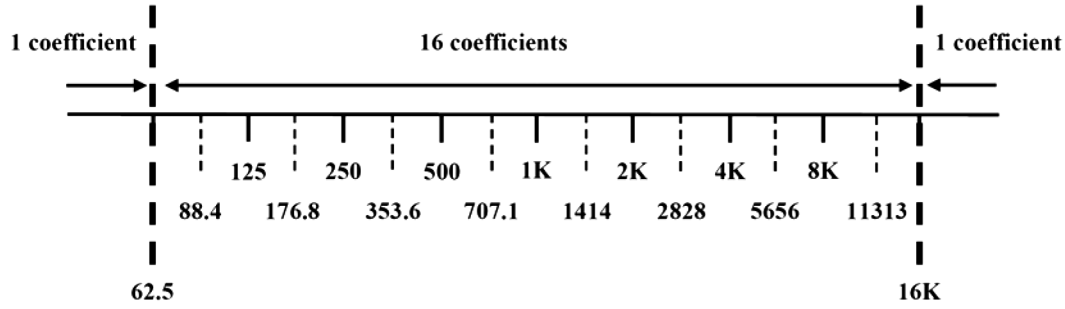

Fig. 1. MPEG-7 NASE subband decomposition with spectral resolution $r = 1/2$.

The NASE coefficient is derived by normalizing each decibel-scale ASE coefficient with the root-mean-square (RMS) norm gain value, R, as follows:

$$NASE(b) = \frac{ASE_{\mathrm{dB}}(b)}{R}, \ 0 \le b \le B+1 \tag{14}$$

where the RMS-norm gain value R is defined as follows:

$$R = \sqrt{\sum_{b=0}^{B+1} (ASE_{\mathrm{dB}}(b))^2}. \tag{15}$$

In MPEG-7, the features vector of an audio frame consists of the RMS-norm gain value R and the NASE coefficients $NASE(b)$, $0 \le b \le B+1$. Thus, the NASE feature vector of an audio frame will be represented as follows:

$$\mathbf{x}_{\mathrm{NASE}} = [R, NASE(0), NASE(1), \ldots, NASE(B+1)]^{\mathrm{T}}. \tag{16}$$

In MPEG-7, the NASE feature vectors are projected onto a lower-dimensional representation using de-correlated basis functions such as PCA or ICA. In this study, modulation spectral analysis will be used to characterize the temporal variations of NASE along time axis.

4) Modulation Spectral Analysis of MFCC, OSC, and NASE: MFCC, OSC, and NASE capture only short-term frame-based features of audio signals. To characterize the variations of a sound within a longer audio segment or the whole music track, long-term features must be generated from a time series of short-term features. In this study, long-term modulation spectral analysis of MFCC, OSC, and NASE is employed to capture the time-varying behavior of the music signals.

Modulation spectral analysis has been used for speech recognition [26], [30], speaker recognition [31], speaker separation [32], audio identification [27], and sound classification [8], [12], [20], [22]. Modulation spectral analysis tries to capture long-term spectral dynamics within an acoustic signal. Typically, a modulation spectral model is a two-dimensional joint "acoustic frequency" and "modulation frequency" representation [31], [33]. Acoustic frequency means the frequency variable of conventional spectrogram whereas modulation frequency captures time-varying information through temporal modulation of the signal. It has been suggested that in speech signals the modulation frequencies range from 2 to 8 Hz reflect syllabic and phonetic temporal structure [34]. The periodicity in music signals will cause some nonzero terms in the joint frequency representation. Typically, modulation spectrum in the range of 1–2 Hz is on the order of musical beat rates, 3–15 Hz is on the order of speech syllabic rates, and higher ones contribute to perceptual roughness revealing musical dissonance [8].

The computation of the joint acoustic-modulation frequency spectrum is generally carried out in two phases. First, the spectrogram is computed using FFT on each pre-emphasized, Hamming windowed overlapping frame. The pre-emphasis function and the Hamming window function are shown in (1) and (2), respectively. Let $S(n,k)$ denote this time-frequency representation, where n is the time variable (frame number) and k is the acoustic frequency variable (FFT bins). The second FFT (or DCT) is then applied on the FFT amplitude envelope of each acoustic frequency (or frequency subband) along time axis to produce the amplitude modulation spectrogram $M(n,k,q)$, where q is the modulation frequency index. Lower q's correspond to slower spectral changes while higher q's correspond to faster spectral changes.

In addition to capture the spectral dynamics through modulation spectral analysis of each acoustic frequency (or frequency subband), modulation spectral analysis of cepstral quefrencies such as MFCC was also used for speech recognition [30] or audio classification [8], [12], [16], [20], [22]. In this study, we will apply modulation spectral analysis to MFCC, OSC, and NASE in order to capture the time-varying behavior of these different music features. To the best of our knowledge, this is a new proposal that extracts discriminative features from the modulation spectrogram of OSC or NASE for sound classification.

Without loss of generality, let

$$\mathbf{x}_n = [x_n(1), x_n(2), \ldots, x_n(D)]^{\mathrm{T}}$$

denote the feature vector extracted from the nth audio frame of a music signal, where D is the length of the feature vector. The feature vector $\mathbf{x}_n$ can be the MFCC, OSC, NASE feature vector, or a combination of these feature vectors by concatenating them together. The modulation spectrogram is obtained by applying FFT independently on each feature value along the time trajectory within a texture window of length W as follows:

$$M_t(m,d) = \sum_{n=0}^{W-1} x_{(t\times W/2)+n}(d) e^{-j2\pi(n/W)k} \qquad 0 \le m < W, \ 0 \le d < D \tag{17}$$

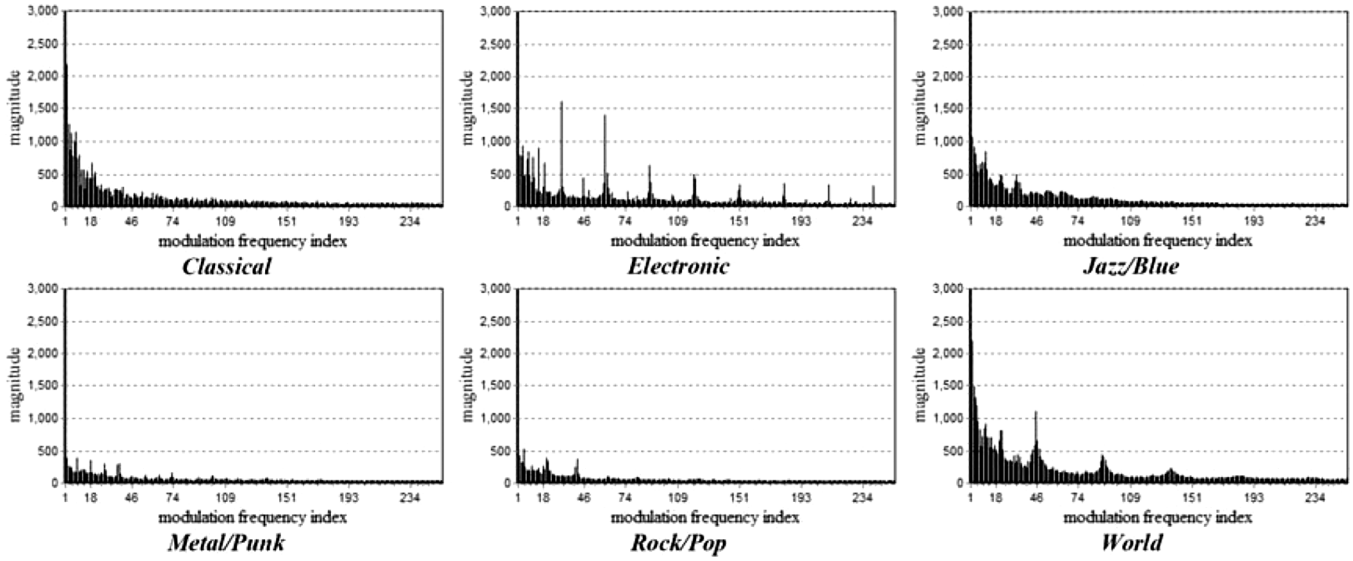

Fig. 2. Modulation spectra of the first MFCC feature value, MFCC(1), of different music genres.

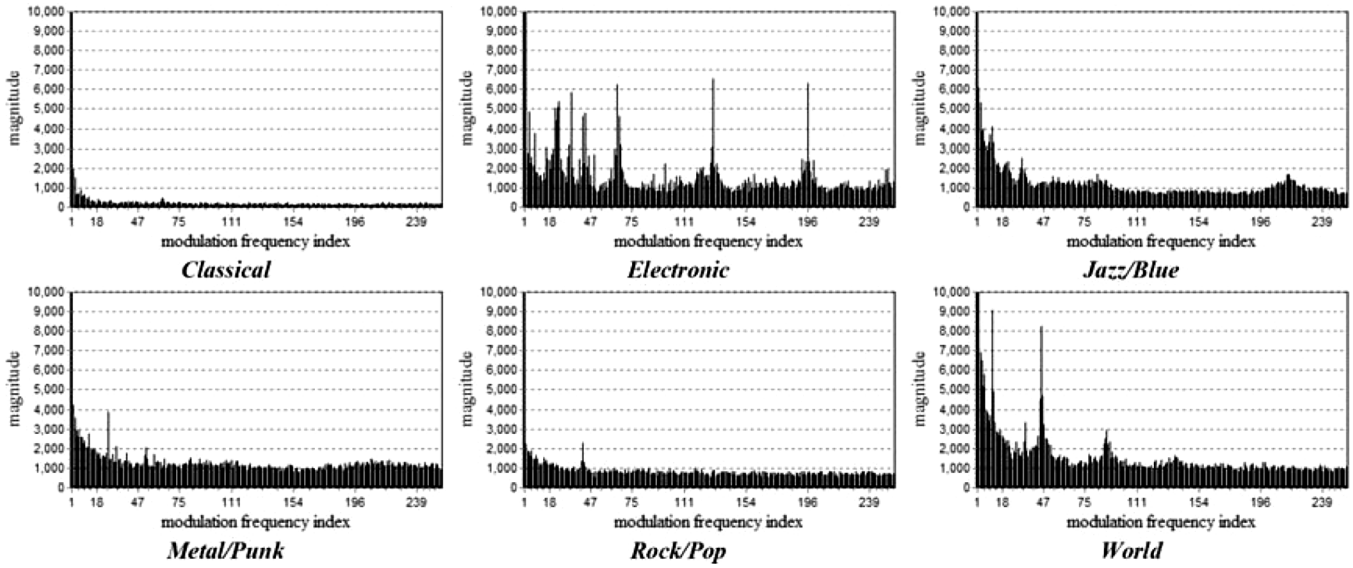

Fig. 3. Modulation spectra of the third valley coefficient, *Valley*(3), of OSC feature vector of different music genres.

where $M_t(m,d)$ is the modulation spectrogram for the tth texture window, m is the modulation frequency index. In this study, the window length W is 512 (about 6 s) with 50% overlap between two successive texture windows. The representative modulation spectrogram of a music track is derived by time averaging the magnitude modulation spectrograms of all texture windows as follows:

$$\overline{M}(m,d) = \frac{1}{T}\sum_{t=1}^{\mathrm{T}} |M_t(m,d)|, \quad 0 \le m < W, \quad 0 \le d < D \tag{18}$$

where T is the total number of texture windows in the music track.

Figs. 2–4 show some modulation spectra of MFCC, OSC, and NASE feature values. From these figures, it is clear that there exist regular large peaks for *Electronic* music followed by the *World* music. For the modulation spectrum of the first MFCC value (see Fig. 2), the lower modulation frequency components of *Classical* music and *Jazz/Blue* music are larger than those of *Metal/Punk* music and *Rock/Pop* music whose modulation spectra are smaller and smoother than others. For the modulation spectrum of the third valley coefficient of OSC feature vector (see Fig. 3), the modulation spectra of *Jazz/Blue* music, *Metal/Punk* music and *Rock/Pop* music are similar, whereas the modulation spectrum of *Classical* music has smaller magnitude values than others. For the modulation spectrum of the second NASE coefficient (see Fig. 4), the lower modulation frequency components of *Jazz/Blue* music is larger than those of *Metal/Punk* music and *Rock/Pop* music. Similarly, the modulation spectrum of *Classical* music has smaller magnitude values than others. From these figures, we can see that the modulation spectra of different music genres bear some different distributions. That is, each of these three types of modulation features can encompass different information for

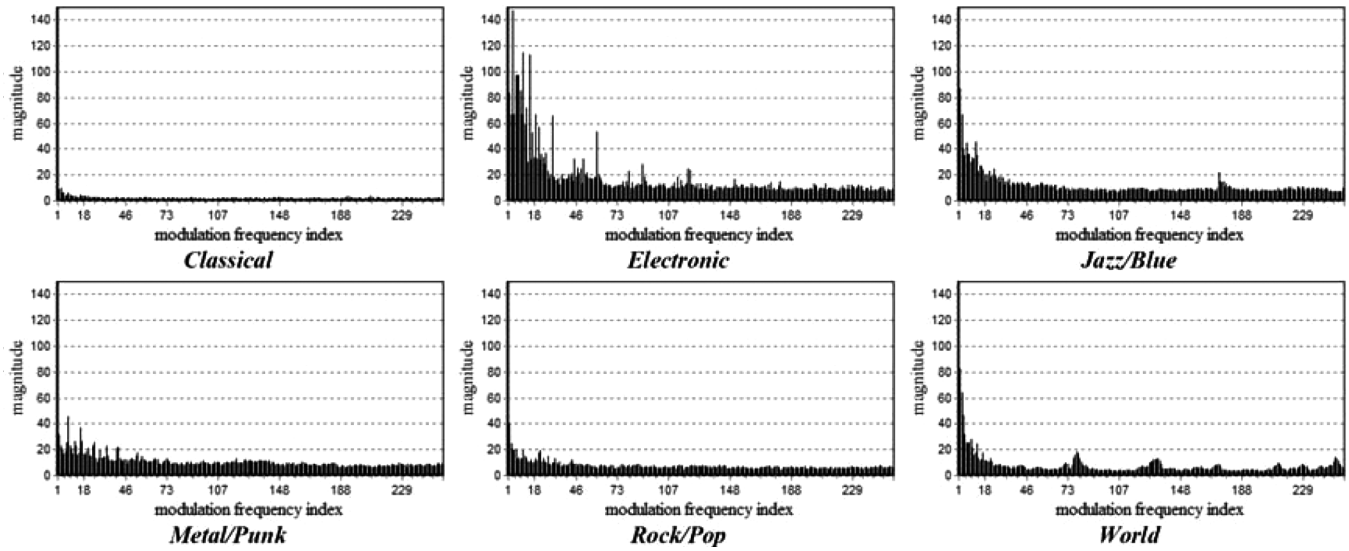

Fig. 4. Modulation spectra of the second NASE coefficient, NASE(2), of different music genres.

music genre classification. Thus, in this paper an information fusion approach which provides both feature level fusion (by concatenating these three different types of modulation spectral features) and decision level combination (by combining the classification results with each type of modulation spectral feature as the input of a classifier) is employed in an attempt to improve the classification accuracy.

Previous study on modulation frequency selectivity has suggested that the human perception for modulation frequency follows a logarithmic frequency scale with resolution consistent with a constant-Q effect [35]. Based on this finding, the averaged modulation spectrum of each spectral/cepstral feature value will be decomposed into J logarithmically spaced modulation subbands. In this study, the number of modulation subbands is 8 ($J = 8$). Table II shows the frequency interval of each modulation subband. For each spectral/cepstral feature value, the modulation spectral peak (MSP) and modulation spectral valley (MSV) within each modulation subband are then evaluated as follows:

$$MSP(j,d) = \max_{\Phi_{j,l} \le m < \Phi_{j,h}} \left(\overline{M}(m,d)\right) \tag{19}$$

$$MSV(j,d) = \min_{\Phi_{j,l} \le m < \Phi_{j,h}} \left(\overline{M}(m,d)\right) \tag{20}$$

where $\Phi_{j,l}$ and $\Phi_{j,h}$ are, respectively, the low modulation frequency index and high modulation frequency index of the jth modulation subband, $0 \le j < J$. The MSPs correspond to the dominant rhythmic components and MSVs the non-rhythmic components in the modulation subbands. Therefore, the difference between MSP and MSV will reflect the modulation spectral contrast distribution as follows:

$$MSC(j,d) = MSP(j,d) - MSV(j,d). \tag{21}$$

A music signal that returns a high modulation spectral contrast value will have large MSP value and a small MSV value and is likely to represent a signal with strong rhythmic content. A music signal that returns a low modulation spectral contrast value will likely be a signal without any significant rhythmic pattern. As a result, all MSCs (or MSVs) will form a $D \times J$ matrix which contains the modulation spectral contrast (or modulation spectral valley) information. Each row of the MSC (or MSV) matrix corresponds to the variant modulation frequency components of identical spectral/cepstral feature value, which reflects the beat interval of a music signal. Each column of the MSC (or MSV) matrix corresponds to the same modulation subband of different spectral/cepstral feature values.

TABLE II
FREQUENCY RANGE OF EACH MODULATION SUBBAND

Subband number	Modulation frequency index range	Modulation frequency range (Hz)
0	[0, 2)	[0, 0.33)
1	[2, 4)	[0.33, 0.66)
2	[4, 8)	[0.66, 1.32)
3	[8, 16)	[1.32, 2.64)
4	[16, 32)	[2.64, 5.28)
5	[32, 64)	[5.28, 10.56)
6	[64, 128)	[10.56, 21.12)
7	[128, 256)	[21.12, 42.24)

To derive a compact feature vector, the mean and standard deviation along each row (and each column) of the MSC and MSV matrices will be computed as the modulation feature values of a music track. Along each row of the MSC or MSV matrix, the mean and standard deviation correspond to the average and variation of the rhythmic strength over variant modulation subbands for a specific spectral/cepstral feature value. Along each column of the MSC or MSV matrix, the mean and standard deviation correspond to the average and variation of the rhythmic strength over different spectral/cepstral feature values on a specific modulation subband. In summary, the modulation spectral feature values derived from the dth ($0 \le d < D$) row of the MSC and MSV matrices can be computed as follows:

$$u_{MSC}^{row}(d) = \frac{1}{J}\sum_{j=0}^{J-1} MSC(j,d) \tag{22}$$

$$\sigma_{MSC}^{row}(d) = \left(\frac{1}{J}\sum_{j=0}^{J-1}\left(MSC(j,d) - u_{MSC}^{row}(d)\right)^2\right)^{1/2} \tag{23}$$

$$u_{MSV}^{row}(d) = \frac{1}{J}\sum_{j=0}^{J-1} MSV(j,d) \tag{24}$$

$$\sigma_{MSV}^{row}(d) = \left(\frac{1}{J}\sum_{j=0}^{J-1}\left(MSV(j,d) - u_{MSV}^{row}(d)\right)^2\right)^{1/2}. \tag{25}$$

Thus, for a music track the modulation spectral feature vector derived from the D rows of the MSC and MSV matrices is of size $4D$ and can be represented as follows:

$$\begin{aligned}\mathbf{f}^{row} =& [u_{MSC}^{row}(0), \sigma_{MSC}^{row}(0), u_{MSV}^{row}(0), \sigma_{MSV}^{row}(0)\ldots,\\ & u_{MSC}^{row}(D-1), \sigma_{MSC}^{row}(D-1)\, u_{MSV}^{row}\\ & (D-1), \sigma_{MSV}^{row}(D-1)]^{\mathrm{T}}.\end{aligned} \tag{26}$$

Similarly, the modulation spectral feature values derived from the jth ($0 \le j < J$) column of the MSC and MSV matrices can be computed as follows:

$$u_{MSC}^{col}(j) = \frac{1}{D}\sum_{d=0}^{D-1} MSC(j,d) \tag{27}$$

$$\sigma_{MSC}^{col}(j) = \left(\frac{1}{D}\sum_{d=0}^{D-1}\left(MSC(j,d) - u_{MSC}^{col}(j)\right)^2\right)^{1/2} \tag{28}$$

$$u_{MSV}^{col}(j) = \frac{1}{D}\sum_{d=0}^{D-1} MSV(j,d) \tag{29}$$

$$\sigma_{MSV}^{col}(j) = \left(\frac{1}{D}\sum_{d=0}^{D-1}\left(MSV(j,d) - u_{MSV}^{col}(j)\right)^2\right)^{1/2}. \tag{30}$$

Thus, the modulation spectral feature vector derived from the J columns of the MSC and MSV matrices is of size $4J$ and can be represented as follows:

$$\begin{aligned}\mathbf{f}^{col} =& [u_{MSC}^{col}(0), \sigma_{MSC}^{col}(0), u_{MSV}^{col}(0), \sigma_{MSV}^{col}(0)\\ & \ldots, u_{MSC}^{col}(J-1), \sigma_{MSC}^{col}(J-1), u_{MSV}^{col}(J-1), \sigma_{MSV}^{col}(J-1)]^{\mathrm{T}}.\end{aligned} \tag{31}$$

In this paper, these two modulation spectral feature vectors ($\mathbf{f}^{row}$ and $\mathbf{f}^{col}$) are concatenated together to yield the modulation spectral feature vector of a music track, which is of size $(4D + 4J)$ as follows:

$$\mathbf{f} = [(\mathbf{f}^{row})^{\mathrm{T}}, (\mathbf{f}^{col})^{\mathrm{T}}]^{\mathrm{T}}. \tag{32}$$

5) Feature Vector Normalization: Since the dispersion is not identical for each modulation spectral feature value, a linear normalization will be independently applied to each modulation spectral feature value to make its range between 0 and 1 as follows:

$$F(m) = \frac{f(m) - f_{\min}(m)}{f_{\max}(m) - f_{\min}(m)} \tag{33}$$

where $F(m)$ denotes the normalized mth modulation spectral feature value, $f_{\max}(m)$ and $f_{\min}(m)$ denote, respectively, the maximum and minimum of the mth modulation spectral feature values of all training music tracks. These reference values, $f_{\max}(m)$ and $f_{\min}(m)$, are computed during the training phase and are stored for later reference. In the classification phase, for actual normalization, each modulation spectral feature value extracted from the current music signal is modified using the reference maximum and minimum values to obtain its corresponding normalized values according to (33).

B. Linear Discriminant Analysis (LDA)

Linear discriminant analysis (LDA) [29] aims at improving the classification accuracy at a lower dimensional feature vector space. LDA deals with the discrimination between various classes rather than the representation of all classes. The objective of LDA is to minimize the within-class distance while maximize the between-class distance. In LDA, an optimal transformation matrix that maps an H-dimensional feature space to an h-dimensional space ($h \le H$) has to be found in order to provide higher discriminability among various music classes.

Let $\mathbf{S_W}$ and $\mathbf{S_B}$ denote the within-class scatter matrix and between-class scatter matrix, respectively. The within-class scatter matrix is defined as follows:

$$\mathbf{S_W} = \sum_{c=1}^{C}\sum_{n=1}^{N_c}(\mathbf{x}_{c,n} - \overline{\mathbf{x}}_c)(\mathbf{x}_{c,n} - \overline{\mathbf{x}}_c)^{\mathrm{T}} \tag{34}$$

where $\mathbf{x}_{c,n}$ is the nth feature vector labeled as class c, $\overline{\mathbf{x}}_c$ is the mean vector of class c, C is the total number of music classes, and N_c is the number of training vectors labeled as class c. The between-class scatter matrix is given by the following:

$$\mathbf{S_B} = \sum_{c=1}^{C} N_c(\overline{\mathbf{x}}_c - \overline{\mathbf{x}})(\overline{\mathbf{x}}_c - \overline{\mathbf{x}})^{\mathrm{T}} \tag{35}$$

where $\overline{\mathbf{x}}$ is the mean vector of all training vectors. The most widely used transformation matrix is a linear mapping that maximizes the so-called Fisher criterion J_F defined as the ratio of between-class scatter to within-class scatter as follows:

$$J_F(\mathbf{A}) = tr((\mathbf{A}^{\mathrm{T}}\mathbf{S_W}\mathbf{A})^{-1}(\mathbf{A}^{\mathrm{T}}\mathbf{S_B}\mathbf{A})) \tag{36}$$

where $\boldsymbol{A}$ is a transformation matrix. From the above equation, we can see that LDA tries to find a transformation matrix that maximizes the ratio of between-class scatter to within-class scatter in a lower-dimensional space.

In this study, a whitening procedure is integrated with LDA transformation such that the multivariate normal distribution of the set of training vectors becomes a spherical one [29]. First, the eigenvalues and corresponding eigenvectors of $\mathbf{S_W}$ are calculated. Let $\mathbf{\Phi}$ denote the matrix whose columns are the orthonormal eigenvectors of $\mathbf{S_W}$, and $\mathbf{\Lambda}$ the diagonal matrix formed by the corresponding eigenvalues. Thus, $\mathbf{S_W\Phi} = \mathbf{\Phi\Lambda}$.

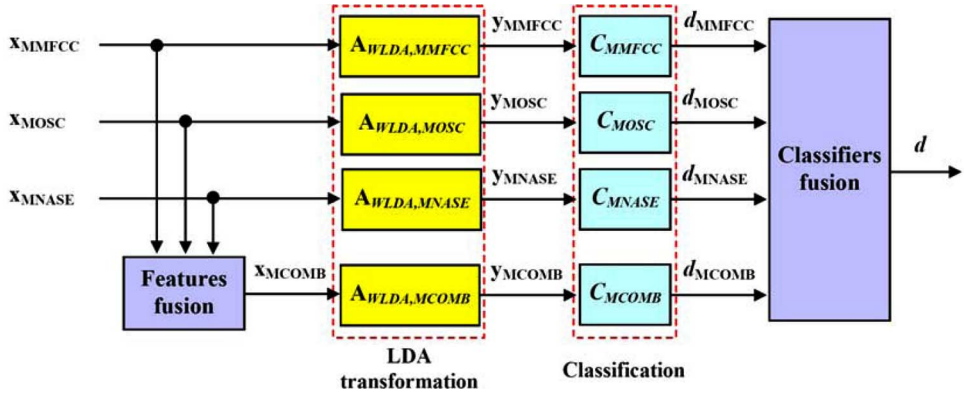

Fig. 5. Whole information fusion structure.

Each training vector $\mathbf{x}$ is then whitening transformed by $\mathbf{\Phi\Lambda}^{-1/2}$ as follows:

$$\mathbf{x}^w = (\mathbf{\Phi\Lambda}^{-1/2})^{\mathrm{T}}\mathbf{x}. \tag{37}$$

It can be shown that the whitened within-class scatter matrix $\mathbf{S}_{\mathbf{W}}^w = (\mathbf{\Phi\Lambda}^{-1/2})^{\mathrm{T}}\mathbf{S}_{\mathbf{W}}(\mathbf{\Phi\Lambda}^{-1/2})$ derived from all the whitened training vectors will become an identity matrix $\mathbf{I}$. Thus, the whitened between-class scatter matrix $\mathbf{S}_{\mathbf{B}}^w = (\mathbf{\Phi\Lambda}^{-1/2})^{\mathrm{T}}\mathbf{S}_{\mathbf{B}}(\mathbf{\Phi\Lambda}^{-1/2})$ contains all the discriminative information. A transformation matrix $\mathbf{\Psi}$ can be determined by finding the eigenvectors of $\mathbf{S}_{\mathbf{B}}^w$. Assuming that the eigenvalues are sorted in a decreasing order, the eigenvectors corresponding to the $(C-1)$ largest eigenvalues will form the column vectors of the transformation matrix $\mathbf{\Psi}$. Finally, the optimal whitened LDA transformation matrix $\mathbf{A}_{WLDA}$ is defined as follows:

$$\mathbf{A}_{WLDA} = \mathbf{\Phi\Lambda}^{-1/2}\mathbf{\Psi}. \tag{38}$$

$\mathbf{A}_{WLDA}$ will be employed to transform each H-dimensional feature vector to be a lower h-dimensional vector. Let $\mathbf{x}$ denote the H-dimensional feature vector, the reduced h-dimensional feature vector can be computed by the following:

$$\mathbf{y} = \mathbf{A}_{WLDA}^{\mathrm{T}}\mathbf{x}. \tag{39}$$

C. Information Fusion Phase

In this paper, an information fusion approach which integrates both feature level fusion and decision level combination [36] is employed to improve the performance of music genre classification. At the stage of feature level fusion, a combined modulation spectral feature vector is obtained by concatenating the three different types of modulation spectral feature vectors. At the stage of decision level combination, each modulation spectral feature vector (including the concatenated modulation spectral feature vector) will serve as the input of a specific classifier and the classification results will be combined to determine the classified music class. The whole information fusion structure is depicted in Fig. 5.

Let $\mathbf{x}_{\mathrm{MMFCC}}$, $\mathbf{x}_{\mathrm{MOSC}}$, and $\mathbf{x}_{\mathrm{MNASE}}$ denote the three modulation spectral feature vectors extracted from an input music track. At the stage of feature level fusion, a new combined feature vector $\mathbf{x}_{\mathrm{MCOMB}}$ is obtained by concatenating $\mathbf{x}_{\mathrm{MMFCC}}$, $\mathbf{x}_{\mathrm{MOSC}}$, and $\mathbf{x}_{\mathrm{MNASE}}$ together as follows:

$$\mathbf{x}_{\mathrm{MCOMB}} = [\mathbf{x}_{\mathrm{MMFCC}}^{\mathrm{T}},\ \mathbf{x}_{\mathrm{MOSC}}^{\mathrm{T}},\ \mathbf{x}_{\mathrm{MNASE}}^{\mathrm{T}}]^{\mathrm{T}}. \tag{40}$$

In this paper, each individual modulation spectral feature vector ($\mathbf{x}_{\mathrm{MMFCC}}$, $\mathbf{x}_{\mathrm{MOSC}}$, and $\mathbf{x}_{\mathrm{MNASE}}$) as well as the concatenated modulation spectral feature vector ($\mathbf{x}_{\mathrm{MCOMB}}$) will be transformed using its corresponding LDA transformation matrix in order to derive the transformed feature vectors ($\mathbf{y}_{\mathrm{MMFCC}}$, $\mathbf{y}_{\mathrm{MOSC}}$, $\mathbf{y}_{\mathrm{MNASE}}$, and $\mathbf{y}_{\mathrm{MCOMB}}$) as follows:

$$\mathbf{y}_{\mathrm{MMFCC}} = \mathbf{A}_{WLDA-\mathrm{MMFCC}}^{\mathrm{T}}\,\mathbf{x}_{\mathrm{MMFCC}} \tag{41}$$

$$\mathbf{y}_{\mathrm{MOSC}} = \mathbf{A}_{WLDA-\mathrm{MOSC}}^{\mathrm{T}}\,\mathbf{x}_{\mathrm{MOSC}} \tag{42}$$

$$\mathbf{y}_{\mathrm{MNASE}} = \mathbf{A}_{WLDA-\mathrm{MNASE}}^{\mathrm{T}}\,\mathbf{x}_{\mathrm{MNASE}} \tag{43}$$

$$\mathbf{y}_{\mathrm{MCOMB}} = \mathbf{A}_{WLDA-\mathrm{MCOMB}}^{\mathrm{T}}\,\mathbf{x}_{\mathrm{MCOMB}} \tag{44}$$

where $\mathbf{A}_{WLDA-\mathrm{MMFCC}}$, $\mathbf{A}_{WLDA-\mathrm{MOSC}}$, $\mathbf{A}_{WLDA-\mathrm{MNASE}}$, and $\mathbf{A}_{WLDA-\mathrm{MCOMB}}$ are the whitened LDA transformation matrices corresponding to the modulation spectral feature vectors $\mathbf{x}_{\mathrm{MMFCC}}$, $\mathbf{x}_{\mathrm{MOSC}}$, $\mathbf{x}_{\mathrm{MNASE}}$, and $\mathbf{x}_{\mathrm{MCOMB}}$, respectively. Note that these LDA transformation matrices are distinct for these four different types of modulation spectral feature vectors and are separately computed using the corresponding type of modulation spectral feature vector alone. To evaluate the similarity between two music tracks, a specific classifier is designed for each type of transformed feature vector. To combine the results from the four different classifiers, the sum rule will be employed to compute the overall distance between two music tracks. For the cth ($1 \le c \le C$) music genre class, let $\mathbf{y}_{\mathrm{MMFCC}}^c$, $\mathbf{y}_{\mathrm{MOSC}}^c$, $\mathbf{y}_{\mathrm{MNASE}}^c$, and $\mathbf{y}_{\mathrm{MCOMB}}^c$ denote its four different types of representative feature vectors and let $d_{\mathrm{MMFCC}}(c)$, $d_{\mathrm{MOSC}}(c)$, $d_{\mathrm{MNASE}}(c)$, and $d_{\mathrm{MCOMB}}(c)$ denote, respectively, the distances between each type of transformed vector of the input music track and the corresponding representative feature vector of the cth music genre class as follows:

$$d_{\mathrm{MMFCC}}(c) = d^2(\mathbf{y}_{\mathrm{MMFCC}},\ \mathbf{y}_{\mathrm{MMFCC}}^c) \tag{45}$$

$$d_{\mathrm{MOSC}}(c) = d^2(\mathbf{y}_{\mathrm{MOSC}},\ \mathbf{y}_{\mathrm{MOSC}}^c) \tag{46}$$

$$d_{\mathrm{MNASE}}(c) = d^2(\mathbf{y}_{\mathrm{MNASE}},\ \mathbf{y}^c_{\mathrm{MNASE}}) \tag{47}$$

$$d_{\mathrm{MCOMB}}(c) = d^2(\mathbf{y}_{\mathrm{MCOMB}},\ \mathbf{y}^c_{\mathrm{MCOMB}}) \tag{48}$$

where $d^2(\bullet,\bullet)$ denotes the distance between two vectors, which is measured by the squared Euclidean distance. The overall distance evaluated for the cth ($1 \le c \le C$) music genre is defined as the sum of each individual distance as follows:

$$d(c) = d_{\mathrm{MMFCC}}(c) + d_{\mathrm{MOSC}}(c) + d_{\mathrm{MNASE}}(c) + d_{\mathrm{MCOMB}}(c). \tag{49}$$

D. Music Genre Classification Phase

In the classification phase, the four types of modulation spectral feature vectors are first extracted from each input music track. The same linear normalization using (33) is applied to each feature value. Each type of normalized feature vector is then transformed to be a lower-dimensional feature vector by using its corresponding whitened LDA transformation matrix. Each type of classifier is employed to compute the distances between the transformed feature vector and the representative feature vectors of all music class. In this study, the representative feature vector of the cth music genre is defined as the mean of all whitened LDA transformed feature vectors computed from all training music tracks labeled as the cth music genre as follows:

$$\overline{\mathbf{y}}_c = \frac{1}{N_c}\sum_{n=1}^{N_c}\mathbf{y}_{c,n} \tag{50}$$

where $\mathbf{y}_{c,n}$ denotes the whitened LDA transformed feature vector of the nth music track labeled as the cth music genre, $\overline{\mathbf{y}}_c$ is the representative feature vector of the cth music genre, and N_c is the number of training music tracks labeled as the cth music genre. The sum of each individual distance computed from each type of classifier is regarded as the overall distance for the cth music genre, $d(c)$. Thus, the subject code s that denotes the identified music genre is determined by finding the music class that has the minimum overall distance as follows:

$$s = \arg\min_{1\le c\le C} d(c). \tag{51}$$

III. Experimental Results

In this section, we first describe the two datasets used for performance comparison. Second, evaluation methodologies for these two datasets are presented. Third, investigation of the importance of various modulation frequency ranges for music genre classification is demonstrated. Fourth, comparison of classification accuracy with different classifiers (KNN, GMM, and LDA) will be presented. Finally, we compare the proposed method with other approaches in terms of classification accuracy.

A. Datasets

Two different datasets widely used for music genre classification are employed for performance comparison. The first dataset (GTZAN) consists of ten genre classes: *Blues*, *Classical*, *Country*, *Disco*, *HipHop*, *Jazz*, *Metal*, *Pop*, *Reggae*, and *Rock*. Each class consists of 100 recordings of music pieces of 30-s duration. This dataset was collected by Tzanetakis [3] and was used for performance evaluation by many researchers [3], [4], [12], [16]. These excerpts were taken from radio, compact disks, and MP3 compressed audio files. Each item was stored as a 22 050 Hz, 16-bit, mono audio file. The second dataset (ISMIR2004 GENRE) was used in the *ISMIR2004 Music Genre Classification Contest* [37]. This dataset consists of 1458 music tracks in which 729 music tracks are used for training and the other 729 tracks for testing. The audio file format is 44.1-kHz, 128-kbps, 16-bit, stereo MP3 files. In this study, each stereo MP3 audio file was first converted into a 44.1-kHz, 16-bit, mono audio file before classification. These music tracks are classified into six classes: *Classical*, *Electronic*, *Jazz/Blue*, *Metal/Punk*, *Rock/Pop*, and *World*. In summary, the music tracks used for training/testing include 320/320 tracks of *Classical*, 115/114 tracks of *Electronic*, 26/26 tracks of *Jazz/Blue*, 45/45 tracks of *Metal/Punk*, 101/102 tracks of *Rock/Pop*, and 122/122 tracks of *World* music genre.

B. Evaluation Methodology

The classification performance on the GTZAN dataset is evaluated based on a randomized ten-fold cross-validation repeated ten times. The dataset was randomly divided into ten folds, of which nine are used for training and the remaining one is used to test the classification accuracy. The classification accuracy is evaluated ten times on the ten different combinations of training/testing sets. The overall classification accuracy is calculated as the average of ten independent ten-fold cross-validations.

In order to be able to compare our proposed method with the results from the *ISMIR2004 Music Genre Classification Contest*, our experiment on the ISMIR2004 GENRE dataset used the same training and testing set partition as in the contest. In the contest, the classification performance is evaluated based on a 50:50 training and testing set partition instead of a ten-fold cross validation. Since the music tracks per class in the ISMIR2004 GENRE dataset are not equally distributed, the overall accuracy of correctly classified genres is evaluated as follows:

$$CA = \sum_{1\le c\le C} P_c \times CA_c \tag{52}$$

where P_c is the probability of appearance of the cth music genre, CA_c is the classification accuracy for the cth music genre.

C. Investigation of the Importance of Various Modulation Frequency Ranges

To investigate the importance of various modulation frequency ranges for music genre classification, experimental results in terms of classification accuracy for different band-pass filters applied in the modulation frequency domain will be presented. In this experiment, the band-pass filter is defined in terms of subband number. That is, a pair of subband numbers (MS_L, MS_H), $1 \le MS_L \le MS_H \le 8$, is used to define the

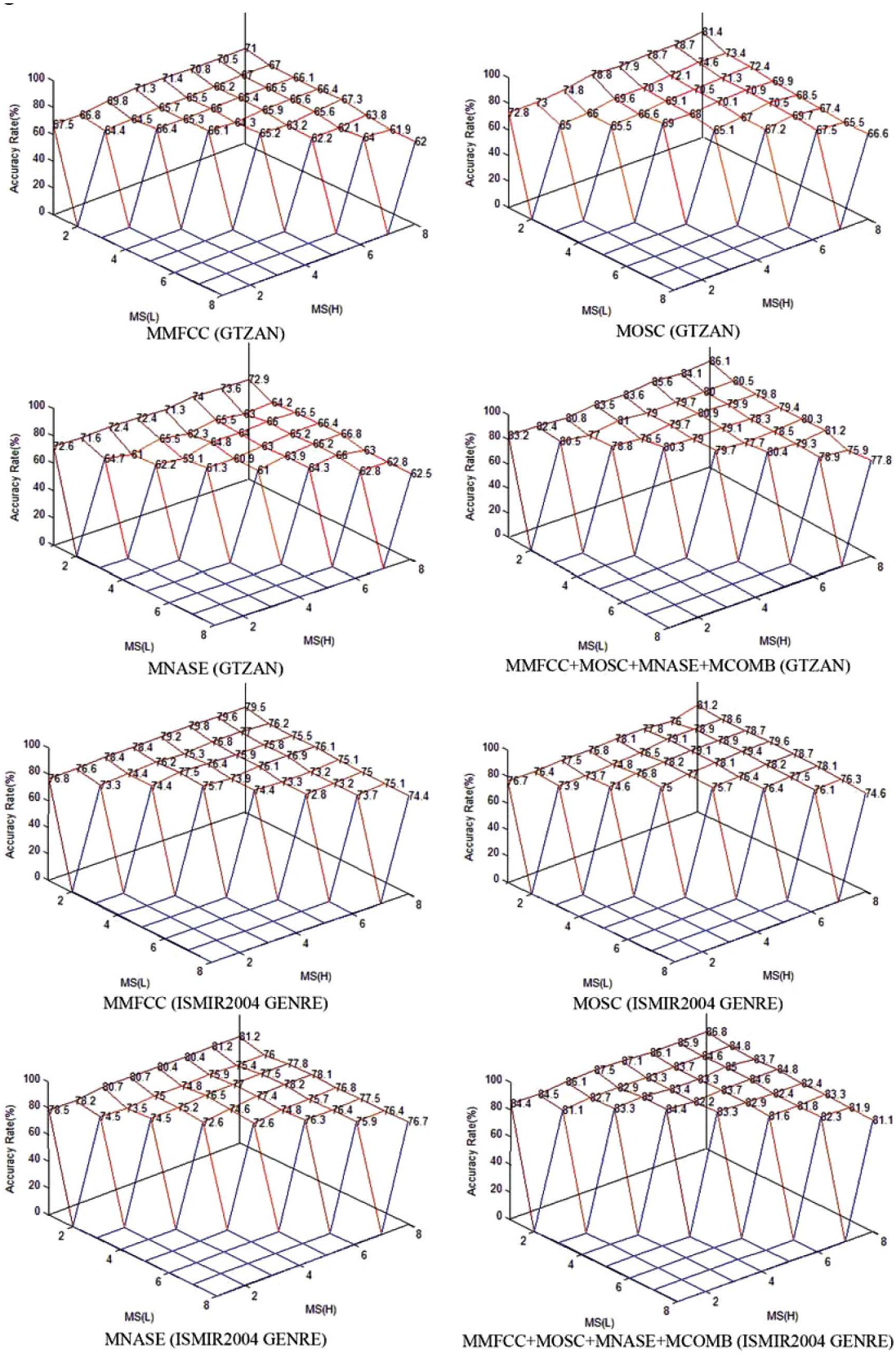

Fig. 6. Comparison of the importance of various modulation frequencies for music genre classification.

lower cutoff modulation frequency (the low edge of subband number MS_L) and higher cutoff modulation frequency (the high edge of subband number MS_H). Fig. 6 shows the classification results of different modulation band-pass filters using different modulation spectral feature vectors on the two datasets. The vertical axis shows the classification accu-

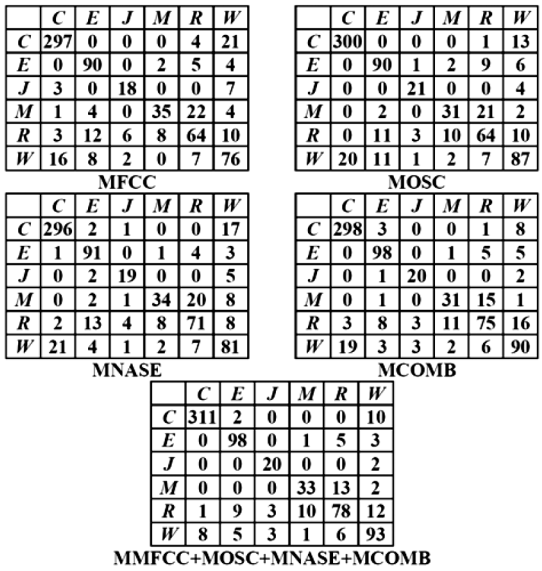

Fig. 7. Confusion matrices for different modulation spectral feature vectors on the ISMIR2004 GENRE dataset.

racy, whereas the other axes indicate the pair of modulation subband numbers, MS_L and MS_H, of the band-pass filter. From this figure, we can see that for most cases the best classification rate is achieved when all modulation subbands are used, that is, set the set of cutoff modulation frequencies $(MS_L, MS_H) = (1, 8)$. Even for some circumstances, different set of cutoff modulation frequencies may yield a better classification result, the improvement is limited. Thus, in the following experiments all modulation subbands will be employed to compute and generate the modulation spectral feature vectors.

D. Classification Results

Fig. 7 shows the confusion matrices of different modulation feature vectors on the ISMIR2004 GENRE dataset. A confusion matrix demonstrates which tracks are correctly classified or not depending on the class. It is read as "$\langle$column$\rangle$ is classified as $\langle$row$\rangle$". For each row or column, the music genres are arranged in the order of *Classical (C)*, *Electronic (E)*, *Jazz/Blue (J)*, *Metal/Punk (M)*, *Rock/Pop (R)*, and *World (W)*. For instance, the number at the first column and last row represents the number of *Classical* music tracks being classified as *World* music. A perfect matrix only contains numbers in the diagonal. In this figure, MMFCC, MOSC, and MNASE denote the approach using modulation spectral feature vector derived from modulation spectral analysis of MFCC, OSC, and NASE, respectively; MCOMB denotes the approach using the combined feature vector by concatenating the feature vectors of MMFCC, MOSC, and MNASE; MMFCC+MOSC+MNASE+MCOMB denotes the information fusion approach which integrates both feature level fusion and decision level fusion. Comparing Fig. 7(a)–(c), we can see that no single feature alone outperforms others for all music genres. For example, MMFCC performs better than MOSC and MNASE for *Metal/Punk* music; MOSC performs best for *Classical*, *Jazz/Blue*, and *World* music; MNASE outperforms the other two for *Electronic* and *Rock/Pop* music. Thus, we expected that a concatenation of these three feature vectors can represent a more generalized feature set with potentially better classification result in a wide range of music genres, as shown in Fig. 7(d). In addition, the classification accuracy can be further improved by using the proposed information fusion approach, as is shown in Fig. 7(e).

TABLE III
AVERAGE CLASSIFICATION ACCURACY (*CA*) FOR VARIOUS FEATURE SETS AS WELL AS DIFFERENT CLASSIFIERS ON THE GTZAN DATASET. GMM(3) AND GMM(4) DENOTE THE GMM CLASSIFIERS WITH THREE AND FOUR GAUSSIAN COMPONENTS, $k = 10$ FOR KNN CLASSIFIER

Feature Set	KNN	GMM(3)	GMM(4)	LDA
AMFCC	22.2%(3.94)	24.3%(1.25)	24.5%(4.01)	28.0%(3.77)
AOSC	61.9%(5.49)	63.4%(5.68)	64.7%(4.50)	62.1%(6.85)
ANASE	56.6%(3.69)	56.9%(5.86)	58.9%(5.28)	57.1%(3.73)
AMFCC+AOSC+ANASE	61.6%(5.36)	59.9%(5.74)	59.5%(4.77)	65.0%(5.72)
MMFCC	61.4%(4.55)	69.6%(4.43)	68.8%(5.09)	71.0%(3.62)
MOSC	67.7%(3.02)	77.5%(2.84)	77.4%(2.68)	81.4%(2.22)
MNASE	67.9%(2.73)	72.8%(4.29)	70.9%(5.45)	72.9%(3.63)
MCOMB	76.4%(1.84)	74.6%(4.25)	74.5%(4.65)	77.2%(3.88)
MMFCC+MOSC	71.9%(3.73)	82.7%(2.50)	81.6%(4.77)	87.4%(2.95)
MMFCC+MOSC+MNASE	76.4%(1.84)	84.6%(2.46)	82.1%(3.73)	**90.6%**(3.06)
MMFCC+MOSC+MNASE+MCOMB	76.4%(1.84)	83.1%(2.23)	81.7%(3.95)	86.1%(2.33)

TABLE IV
CLASSIFICATION ACCURACY (*CA*) OF DIFFERENT FEATURE SETS AS WELL AS DIFFERENT CLASSIFIERS ON THE ISMIR2004 GENRE DATASET

Feature Set	KNN	GMM(3)	GMM(4)	LDA
AMFCC	52.67%	46.36%	47.19%	53.50%
AOSC	71.19%	70.51%	68.72%	67.76%
ANASE	68.31%	67.63%	67.35%	63.10%
AMFCC+AOSC+ANASE	71.33%	69.41%	66.26%	70.23%
MMFCC	78.19%	84.63%	82.58%	79.56%
MOSC	75.17%	82.30%	83.26%	81.34%
MNASE	73.66%	81.07%	82.25%	81.21%
MCOMB	79.42%	85.87%	84.64%	83.95%
MMFCC+MOSC	79.29%	85.60%	84.77%	82.72%
MMFCC+MOSC+MNASE	79.42%	84.22%	84.64%	83.81%
MMFCC+MOSC+MNASE+MCOMB	79.42%	86.42%	85.73%	**86.83%**

The comparisons of classification accuracy on the GTZAN and ISMIR2004 GENRE datasets for various feature sets as well as different classifiers (KNN, GMM, and LDA) are shown in Tables III and IV. In these two tables, AMFCC, AOSC, and ANASE are approaches with their feature vectors derived from simple averaging of the MFCC, OSC, and NASE feature vectors of all frames across the whole music track; AMFCC+AOSC+ANASE, MMFCC+MOSC, and MMFCC+MOSC+MNASE are approaches using only the decision level fusion to determine the classified music class from the results of different classifiers. It is clear that the approaches using modulation spectral features outperform their corresponding approaches using simple averaging features. For the GTZAN dataset, the best classification accuracy is 90.6% using the approach MMFCC+MOSC+MNASE. For the ISMIR2004 GENRE dataset, the best classification accuracy is 86.83% using the approach MMFCC+MOSC+MNASE+MCOMB.

Table V compares our proposed approach with other approaches [3], [4], [12], [16] on the GTZAN dataset with the same experimental setup. Note that the experiment results are evaluated based on a randomized ten-fold cross-validation repeated ten times. It is clear that our proposed approach (MMFCC+MOSC+MNASE) achieves a classification accuracy of 90.6%, which is better than the other approaches.

Table VI shows the comparison with the results from the *ISMIR2004 Music Genre Classification Contest* as well

TABLE V
COMPARISON WITH OTHER APPROACHES ON THE GTZAN DATASET WITH THE SAME EXPERIMENTAL SETUP (TEN-FOLD CROSS-VALIDATIONS)

References	*CA*
Our approach (MMFCC+MOSC+MNASE)	90.6% (3.06)
T. Li *et al.* [4]	78.5% (4.07)
I. Panagakis *et al.* [16]	78.2% (3.82)
T. Lidy & A. Rauber [12]	74.9%
G. Tzanetakis *et al.* [3]	61.0% (4.00)

TABLE VI
COMPARISON WITH THE RESULTS FROM THE *ISMIR2004 MUSIC* GENRE CLASSIFICATION CONTEST AND APPROACHES WITH THE SAME EXPERIMENTAL SETUP (50:50 TRAINING/TESTING SET SPLIT)

References	*CA*
Our approach (MMFCC+MOSC+MNASE+MCOMB)	86.83%
Y. Song *et al.* [15]	84.77%
T. Lidy & A. Rauber [12]	79.70%
E. Pampalk (winner)	84.07%
K. West (2nd rank)	78.33%
G. Tzanetakis (3rd rank)	71.33%
T. Lidy & A. Rauber (4th rank)	70.37%
D. Ellis & B. Whitman (5th rank)	64.00%

as other approaches [12], [15] with the same experimental setup. From this table, we can see that our proposed approach (MMFCC+MOSC+MNASE+MCOMB) performs the best and achieves higher classification accuracy (86.83%) than the winner of the contest with a classification accuracy of 84.07%.

IV. CONCLUSION

A novel feature set, derived from long-term modulation spectral analysis of spectral (OSC and NASE) and cepstral (MFCC) features, is proposed for music genre classification. For each spectral/cepstral feature set, a modulation spectrogram is generated by collecting the modulation spectra of all feature values. Modulation spectral contrast (MSC) and modulation spectral valley (MSV) are then computed from each logarithmically-spaced modulation subband. Statistical aggregations of all MSCs and MSVs are computed to generate effective and compact discriminating features. An information fusion approach which integrates both feature level fusion and decision level combination is employed to improve the classification accuracy. Experiments conducted on the GTZAN dataset have shown that our proposed approach outperforms other approaches with the same experimental setup. In addition, experimental results on the ISMIR2004 GENRE music dataset have also shown that our proposed approach achieves higher classification accuracy (86.83%) than the winner of the *ISMIR2004 Music Genre Classification Contest* with a classification accuracy of 84.07%.

ACKNOWLEDGMENT

The authors would like to thank the anonymous reviewers for their valuable comments that improved the representation and quality of this paper. The authors would also like to thank Dr. Tzanetakis for kindly sharing his dataset with us.

REFERENCES

[1] D. Perrot and R. Gjerdigen, "Scanning the dial: An exploration of factors in the identification of musical style," in *Proc. Soc. for Music Perception and Cognition*, 1999, Abstract, p. 88.
[2] S. Lippens, J. P. Martens, M. Leman, B. Baets, H. Meyer, and G. Tzanetakis, "A comparison of human and automatic musical genre classification," in *Proc. IEEE Int. Conf. Acoustics, Speech, Signal Processing*, 2004, vol. 4, pp. 233–236.
[3] G. Tzanetakis and P. Cook, "Musical genre classification of audio signals," *IEEE Trans. Speech Audio Process.*, vol. 10, no. 3, pp. 293–302, Jul. 2002.
[4] T. Li and M. Ogihara, "Toward intelligent music information retrieval," *IEEE Trans. Multimedia*, vol. 8, no. 3, pp. 564–573, Jun. 2006.
[5] D. N. Jiang, L. Lu, H. J. Zhang, J. H. Tao, and L. H. Cai, "Music type classification by spectral contrast feature," in *Proc. IEEE Int. Conf. Multimedia and Expo*, 2002, vol. 1, pp. 113–116.
[6] K. West and S. Cox, "Features and classifiers for the automatic classification of musical audio signals," in *Proc. Int. Conf. Music Information Retrieval*, 2004.
[7] K. Umapathy, S. Krishnan, and S. Jimaa, "Multigroup classification of audio signals using time-frequency parameters," *IEEE Trans. Multimedia*, vol. 7, no. 2, pp. 308–315, Apr. 2005.
[8] M. F. McKinney and J. Breebaart, "Features for audio and music classification," in *Proc. Int. Conf. Music Information Retrieval*, 2003, pp. 151–158.
[9] J. J. Aucouturier and F. Pachet, "Representing music genres: A state of the art," *J. New Music Res.*, vol. 32, no. 1, pp. 83–93, 2003.
[10] U. Bağci and E. Erzin, "Automatic classification of musical genres using inter-genre similarity," *IEEE Signal Process. Lett.*, vol. 14, no. 8, pp. 512–524, Aug. 2007.
[11] A. Meng, P. Ahrendt, J. Larsen, and L. K. Hansen, "Temporal feature integration for music genre classification," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 15, no. 5, pp. 1654–1664, Jul. 2007.
[12] T. Lidy and A. Rauber, "Evaluation of feature extractors and psychoacoustic transformations for music genre classification," in *Proc. 6th Int. Conf. Music Information Retrieval*, 2005, pp. 34–41.
[13] M. Grimaldi, P. Cunningham, and A. Kokaram, "A wavelet packet representation of audio signals for music genre classification using different ensemble and feature selection techniques," in *Proc. 5th ACM SIGMM Int. Workshop on Multimedia Information Retrieval*, 2003, pp. 102–108.
[14] J. Bergatra, N. Casagrande, D. Erhan, D. Eck, and B. Kégl, "Aggregate features and Adaboost for music classification," *Mach. Learn.*, vol. 65, no. 2-3, pp. 473–484, Jun. 2006.
[15] Y. Song and C. Zhang, "Content-based information fusion for semi-supervised music genre classification," *IEEE Trans. Multimedia*, vol. 10, no. 1, pp. 145–152, Jan. 2008.
[16] I. Panagakis, E. Benetos, and C. Kotropoulos, "Music genre classification: A multilinear approach," in *Proc. ISMIR*, 2008, pp. 583–588.
[17] C. C. Lin, S. H. Chen, T. K. Truong, and Y. Chang, "Audio classification and categorization based on wavelets and support vector machine," *IEEE Trans. Speech Audio Process.*, vol. 13, no. 5, pp. 644–651, Sep. 2005.
[18] H. G. Kim, N. Moreau, and T. Sikora, "Audio classification based on MPEG-7 spectral basis representation," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 14, no. 5, pp. 716–725, May 2004.
[19] H. G. Kim, N. Moreau, and T. Sikora, *MPEG-7 Audio and Beyond: Audio Content Indexing and Retrieval*. New York: Wiley, 2005.
[20] F. Mörchen, A. Ultsch, M. Thies, and I. Löhken, "Modeling timbre distance with temporal statistics from polyphonic music," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 14, no. 1, pp. 81–90, Jan. 2006.
[21] T. Lambrou, P. Kudumakis, R. Speller, M. Sandler, and A. Linney, "Classification of audio signals using statistical features on time and wavelet transform domains," in *Proc. IEEE Int. Conf. Acoustics, Speech, Signal Processing*, 1998, vol. 6, pp. 3621–3624.
[22] C. H. Lee, J. L. Shih, K. M. Yu, and J. M. Su, "Automatic music genre classification using modulation spectral contrast feature," in *Proc. IEEE Int. Conf. Multimedia and Expo*, 2007, pp. 204–207.
[23] W. A. Sethares, R. D. Robin, and J. C. Sethares, "Beat tracking of musical performance using low-level audio feature," *IEEE Trans. Speech Audio Process.*, vol. 13, no. 2, pp. 275–285, Mar. 2005.

[24] G. Tzanetakis, A. Ermolinskyi, and P. Cook, "Pitch histograms in audio and symbolic music information retrieval," in *Proc. Int. Conf. Music Information Retrieval*, 2002.
[25] C. Xu, N. C. Maddage, and X. Shao, "Automatic music classification and summarization," *IEEE Trans. Speech Audio Process.*, vol. 13, no. 3, pp. 441–450, May 2005.
[26] B. Kingsbury, N. Morgan, and S. Greenberg, "Robust speech recognition using the modulation spectrogram," *Speech Commun.*, vol. 25, no. 1, pp. 117–132, 1998.
[27] S. Sukittanon, L. E. Atlas, and J. W. Pitton, "Modulation-scale analysis for content identification," *IEEE Trans. Signal Process.*, vol. 52, no. 10, pp. 3023–3035, Oct. 2004.
[28] L. Rabiner and B. H. Juang, *Fundamentals of Speech Recognition*. Englewood Cliffs, NJ: Prentice-Hall, 1993.
[29] R. Duda, P. Hart, and D. Stork, *Pattern Classification*. New York: Wiley, 2000.
[30] V. Tyagi, I. McCowan, H. Misra, and H. Bourlard, "Mel-cepstrum modulation spectrum (MCMS) features for robust ASR," in *Proc. Workshop Automatic Speech Recognition and Understanding*, 2003.
[31] T. Kinnunen, "Joint acoustic-modulation frequency for speaker recognition," in *Proc. IEEE Int. Conf. Acoustics, Speech, Signal Processing*, 2006, vol. 1, pp. 14–19.
[32] B. Kollmeier and R. Koch, "Speech enhancement based on physiological and psychoacoustical models of modulation perception and binaural interaction," *J. Acoust. Soc. Amer.*, vol. 95, pp. 1593–1602, 1994.
[33] L. Atlas and S. Shamma, "Joint acoustic and modulation frequency," *EURASIP J. Appl. Signal Process.*, vol. 7, pp. 668–675, 2003.
[34] N. Kanedera, T. Arai, H. Hermansky, and M. Pavel, "On the relative importance of various components of the modulation spectrum for automatic speech recognition," *Speech Commun.*, vol. 28, no. 1, pp. 43–55, May 1999.
[35] S. Ewert and T. Dau, "Characterizing frequency selectivity for envelope fluctuations," *J. Acoust. Soc. Amer.*, vol. 108, pp. 1181–1196, 2000.
[36] J. Kittler, M. Hatef, R. Duin, and J. Matas, "On combining classifiers," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 20, no. 3, pp. 226–239, Mar. 1998.
[37] [Online]. Available: http://ismir2004.ismir.net/ISMIR_Contest.html.

Chang-Hsing Lee was born on July 24, 1968, in Tainan, Taiwan. He received the B.S. and Ph.D. degrees in computer and information science from National Chiao Tung University, Hsinchu, Taiwan, in 1991 and 1995, respectively.

He is currently an Associate Professor in the Department of Computer Science and Information Engineering, Chung Hua University, Hsinchu. His main research interests include audio/sound classification, multimedia information retrieval, and image processing.

Jau-Ling Shih was born on December 13, 1969, in Tainan, Taiwan. She received the B.S. degree in electrical engineering from National Sun Yat-Sen University, Kaohsiung, Taiwan, in 1992, the M.S. degree in electrical engineering from National Cheng Kung University, Tainan, Taiwan, in 1994, and the Ph.D. degree in computer and information science from National Chiao Tung University, Hsinchu, Taiwan in 2002.

She is currently an Associate Professor in the Department of Computer Science and Information Engineering, Chung Hua University, Hsinchu. Her main research interests include image processing, image retrieval, and audio processing.

Kun-Ming Yu received the B.S. degree in chemical engineering from National Taiwan University, Taipei, Taiwan, in 1981, and the M.S. and Ph.D. degrees in computer science from the University of Texas at Dallas in 1988 and 1991, respectively.

He is currently the Dean of the College of Computer Science and Informatics, Chung Hua University, Hsinchu, Taiwan. His research interests include load balancing, distributed and parallel computing, high-performance computing, ad hoc network, computer algorithms, and audio processing.

Hwai-San Lin was born on October 18, 1980, in Taipei, Taiwan. He received the B.S. degree in computer science from Chinese Culture University, Taipei, in 2005 and the M.S. degree in computer science and information engineering from Chung Hua University, Hsinchu, Taiwan, in 2009.

His main research interests include audio/sound classification and image processing.