Automatic Music Mood Classification Based on Timbre and Modulation Features

Jia-Min Ren, Ming-Ju Wu, and Jyh-Shing Roger Jang, *Member, IEEE*

Abstract—In recent years, many short-term timbre and long-term modulation features have been developed for content-based music classification. However, two operations in modulation analysis are likely to smooth out useful modulation information, which may degrade classification performance. To deal with this problem, this paper proposes the use of a two-dimensional representation of acoustic frequency and modulation frequency to extract joint acoustic frequency and modulation frequency features. Long-term joint frequency features, such as acoustic-modulation spectral contrast/valley (AMSC/AMSV), acoustic-modulation spectral flatness measure (AMSFM), and acoustic-modulation spectral crest measure (AMSCM), are then computed from the spectra of each joint frequency subband. By combining the proposed features, together with the modulation spectral analysis of MFCC and statistical descriptors of short-term timbre features, this new feature set outperforms previous approaches with statistical significance.

Index Terms—Music mood classification, modulation spectrogram, octave-based spectral contrast/valley, spectral flatness/crest measure

1 Introduction

With the rapid growth of digital music available on the Web (*e.g.,* 7digital[1]) and on personal devices, managing large music collections has become an important and challenging issue [1]. Improve the organization and management of music collections usually requires the attachment of various metadata for each music file. Traditional metadata labels, such as artist, album, and title are insufficient for certain applications [2], such as music therapy. Other labels, such as music mood, which describes the inherent emotional expression of a music clip, are more useful in these scenarios [3].

1.1 Emotion Models and Recognition

A considerable amount of work has been dedicated to the modeling of relationships between music and emotions[2] from different disciplines, including psychology, musicology and music information retrieval [4], [5]. Most of the proposed emotion models belong to either the categorical approach or the dimensional approach [4], [6]. Categorical approaches represent emotions as a set of categories that are clearly distinct from each other. For example, Ekman [7] proposed six basic emotion categories based on human facial expressions of anger, fear, happiness, sadness, disgust, and surprise. Another famous categorical approach is Hevner's affective checklist [8], where eight clusters of affective adjectives were discovered and laid out in a circle, as shown in Fig. 1. Each cluster includes similar adjectives, and "the meaning of neighboring clusters varies in a cumulative way until reaching a contrast in the opposite position" [5]. Hu and Downie [9] also derived five emotion categories by clustering affective tags of music from the All Music Guide, as shown in TABLE 1. This emotion taxonomy has been used in the annual MIREX[3] audio music mood classification task since 2007. (More details about this contest are given in Section 4.1.)

While the categorical approach focuses mainly on distinguishing different emotions from music, the dimensional approach characterizes emotions on a small number of emotion dimensions (usually 2 or 3) intended to represent the internal emotions of humans. A famous emotion model is Russell's circumplex model [10] which consists of a circular structure within two dimensions of valence and arousal, as shown in Fig. 2, where inversely correlated emotions are placed across the circle from one another. Supportive evidence of this valence-arousal, circular-structure arrangement was obtained by scaling 28 affective terms [10]–[12].

Since a music piece may evoke more than one emotion, the categorical approach can be cast into

- *J.-M. Ren is with the Data Analytic Technology Department, Intelligent Analytic Technology Division, Computational Intelligence Technology Center, Industrial Technology Research Institute, Hsinchu 30040, Taiwan. (e-mail: jmren@mirlab.org)*
- *M.-J. Wu is with the Department of Computer Science, National Tsing Hua University, Hsinchu 30013, Taiwan. (e-mail: brian.wu@mirlab.org)*
- *J.-S. R. Jang is with the Department of Computer Science and Information Engineering, National Taiwan University, Taipei 10617, Taiwan. (e-mail: jang@mirlab.org)*

1. http://www.7digital.fm
2. Moods and emotions are used interchangably in this work.
3. Music Information Retrieval Evaluation eXchange, see the website http://www.music-ir.org/mirex/wiki/MIREX_HOME for more information.

TABLE 1
Emotion categories used in MIREX audio mood classification contest.

Cluster	Mood
I	rousing, passionate, confident, boisterous, rowdy
II	cheerful, rollicking, fun, sweet, amiable/good natured
III	poignant, literate, wistful, bittersweet, autumnal, brooding
IV	silly, humorous, campy, quirky, whimsical, witty, wry
V	fiery, aggressive, tense/anxious, intense, volatile, visceral

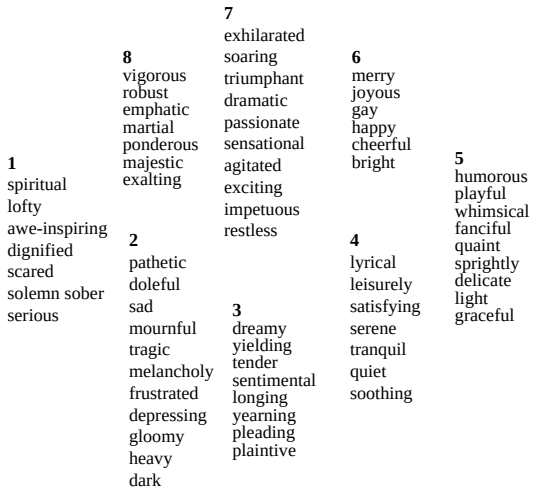


Fig. 1. Hevner's eight clusters of affective terms [8].

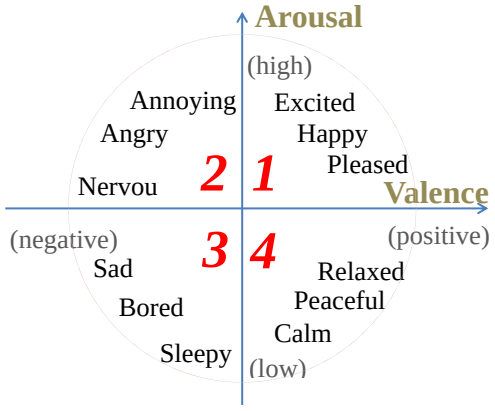


Fig. 2. The 2D valence-arousal emotion space [10].

either a single-label classification problem [13], [14] or a multi-label classification problem [15]. On the other hand, the dimensional approach can be formulated as a regression problem [16] since the output is a point that can move continuously within the emotion space. Regardless of whether a categorical or dimensional approach is used, an effective feature set is always required. Therefore, in this paper, we focus on developing a better feature set to improve the performance of the single-label music mood classification problem. The identified features can be used for the dimensional approach too.

1.2 Audio Features

Many audio features have been proposed for content-based music classification. In general, we can roughly categorize audio features as short-term or long-term [17]. Short-term features, which *e.g.*, capture the timbral characteristics of audio signals, are usually extracted from short time windows (also called frames). Widely used timbre features include the zero crossing rate, spectral centroid, spectral flux, spectral rolloff, spectral skewness, spectral kurtosis, Mel-frequency cepstral coefficients (MFCC), octave-based spectral contrast (OSC) [18], [19], spectral flatness measure (SFM) [20], spectral crest measure (SCM) [20], MPEG-7 normalized audio spectrum envelop (NASE) [21], etc.

On the other hand, long-term audio features, which generally describe temporal evolutions of a music clip or capture the inherent properties of music that humans perceive, are usually generated by aggregating short-term features. Several methods have been proposed to aggregate temporal features: statistical moment [1], [22], entropy or correlation [9], modulation spectral analysis [10], etc. Long-term features used to reveal the human perception of audio properties include tempo [1], [23], melody [1], and rhythm [16].

Once audio features are extracted from music clips of different moods, our next task is to construct classifiers for mood classification. Several supervised learning approaches have been proposed for music classification of various kinds, including Gaussian mixture models (GMM) [1], [24], hidden Markov models (HMM) [21], Adaboost [25], linear discriminant analysis (LDA) [24], k-nearest neighbor classifier (KNNC) [1], [24], and support vector machine (SVM) [24], [26].

1.3 Long-Term Modulation-Based Features

Although the use of different classifiers will affect the accuracy of music classification, feature sets have been shown to have a more significant effect on accuracy [27]. Therefore, some of the recent work on music classification has focused on discovering long-term discriminative features [28]–[30]. A representative approach is Lee *et al.*'s method [28] for analyzing the modulation spectra of timbre features extracted from short time frames. Fig. 3 shows how Lee *et al.* derived modulation features from a music clip. Short-time

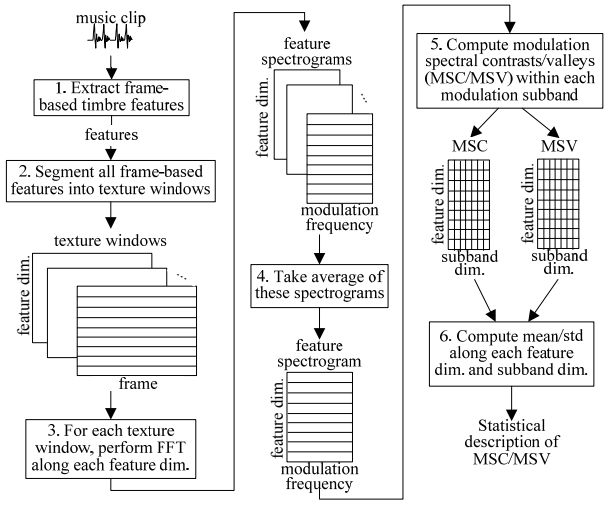


Fig. 3. Flowchart for extracting modulation features from a music clip.

timbre features, such as MFCC and OSC, are extracted from audio frames. These frame-based timbre features are segmented into texture windows. To capture temporal variations of these music features, fast Fourier transform (FFT) is applied along each feature dimension of a texture window to obtain a feature spectrogram. In this way music features with slow and fast spectral changes are respectively represented as non-zero terms at low and high modulation frequencies. A representative feature spectrogram is further created by averaging feature spectrograms obtained from all texture windows. Modulation spectral contrasts/valleys (MSC/MSV) are then computed within each modulation subband, reflecting the strength of rhythm in the music. Finally, the mean and standard deviation along each row and each column of the MSC and MSV matrices are computed. We can then concatenate these statistical features to form a compact feature vector for each music clip.

However, in Lee *et al.*'s approach, the averaging process (to compute the representative feature spectrogram) and the summarization operation (to compute the mean and standard deviation of MSC/MSV matrices) are likely to smooth out useful modulation information, which may degrade classification performance. To deal with this problem, this paper proposes the use of joint frequency features computed from a joint frequency representation, which is defined as a two-dimensional representation of acoustic frequency and modulation frequency [31]. These joint frequency features, including acoustic-modulation spectral contrast/valley (AMSC/AMSV) and acoustic-modulation spectral flatness/crest measure (AMSFM/AMSCM), are computed from spectra of each joint frequency subband. Without taking the average of feature spectrograms and computing statistical descriptors of MSC/MSV matrices, the proposed features retain more modulation information for better classification.

1.4 Contribution

The main contributions of this paper can be summarized as follows.

1) We have proposed a feature set for music mood classification, which combine modulation spectral analysis of MFCC, OSC, and SFM/SCM, and statistical descriptors of short-term timbre features. By employing these features for SVMs, our submission to the MIREX 2011 audio mood classification task was ranked #1. In fact, the submission outperformed all the other submissions of the task from 2008 to 2014, indicating the superiority of the proposed feature sets.

2) Moreover, based on a part of the aforementioned feature sets, we have also proposed another new feature set that combines the newly proposed joint frequency features (including AMSC/AMSV and AMSFM/AMSCM), together with the modulation spectral analysis of MFCC, and statistical descriptors of short-term timbre features. Experiments conducted on three mood datasets demonstrate that the proposed feature set even outperforms our MIREX 2011 submission with statistical significance.

The remainder of this paper is organized as follows. Section 2 gives an overview of the proposed music mood classification system. The used audio features are described in Section 3. Our submission to MIREX 2011 audio mood classification contest is presented in Section 4, where the submission was based on Lee *et al.*'s modulation spectral analysis, together with statistical descriptors of short-term timbre features for SVMs. The proposed joint frequency features are introduced in Section 5. A visiualization of a music clip for the proposed joint frequency features and Lee *et al.*'s modulation features is shown in Section 5.2. Experimental results for the proposed features are discussed in Section 6. Finally, we conclude this work and discuss future work in Section 7.

2 System Overview

Fig. 4 shows a flowchart of the proposed audio mood classification system. For the extraction of short-term timbre features, statistical spectrum descriptors (SSD), MFCC, OSC, and SFM/SCM are computed from audio frames. We then compute the mean and standard deviation along each feature dimension over all frames of a music clip to obtain a compact feature vector for each music clip. For the extraction

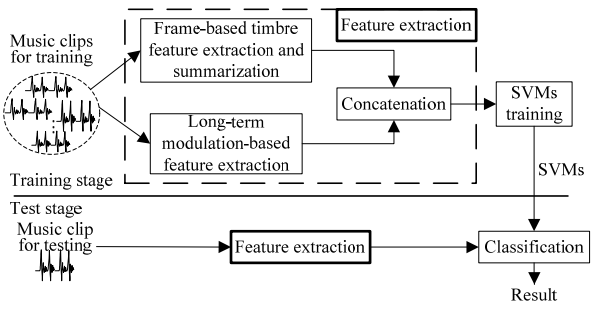


Fig. 4. Flowchart of the proposed audio mood classification system.

of long-term modulation-based features, we either perform the modulation spectral analysis on MFCC, OSC, and SFM/SCM (which were used in our MIREX 2011 submission), or compute joint-frequency features from a joint frequency representation (which were used in the extended experiments, as described in Section 6). Finally, we concatenate these statistical descriptors of short-term timbre features with long-term modulation-based features for SVMs. In the test stage, the same set of features is extracted from a test music clip; we then use the pre-trained SVMs to classify the test music clip.

3 Audio Feature Extraction

This section first describes short-term timbre features, followed by long-term modulation spectral analysis.

3.1 Short-Term Timbre Features

To reliably capture spectral characteristics of audio signals, music clips are usually divided into short-time frames within which the signals can be assumed to be stationary. In this study, we segmented music clips into 46 ms frames (1,024 samples with a sample rate of 22,050 Hz) with 50% overlap. Each frame is pre-emphasized and then multiplied by a Hamming window to deal with the ringing effect. Spectral analysis using FFT was then applied to the Hamming-windowed frame. To measure the spectral distribution of audio signals, statistical spectrum descriptors are computed from the magnitude spectrum. This study also uses three types of timbre features (including MFCC, OSC, and SFM/SCM) that have been proven to be effective in music classification.

- *SSD (Statistical Spectrum Descriptors)*
  SSD consists of spectral centroid (SC), spectral flux (SF), spectral rolloff (SR), spectral skewness (SS), and spectral kurtosis (SK). These features are generally used to measure the spectral shape, spectral change, and spectral distribution of audio signals. More details about these features can be found in [1] and [20].

TABLE 2
Frequency Ranges of Octave-Scale Band-Pass Filters (Sample Rate 22 050 Hz)

Filter Number	Frequency Range (Hz)
1	[0, 100]
2	(100, 200]
3	(200, 400]
4	(400, 800]
5	(800, 1600]
6	(1600, 3200]
7	(3200, 6400]
8	(6400, 11025]

- *MFCC (Mel-Frequency Cepstral Coefficients)*
  MFCC was originally proposed for speech processing, and now has been successfully used in both speech recognition and music classification due to its ability to model the subjective frequency contents of audio signals [32]. The steps for computing MFCC can be found in [28]. Note that, although typically 13 dimensions are utilized for MFCC in speech representation, here we used 20-dimension MFCC to follow the work of [28]. We also used the energy from each frame since it is found to be useful for classifying music contents [18].
- *OSC (Octave-Based Spectral Contrast)*
  OSC was proposed to represent the spectral characteristics of music signals [18]. Compared to MFCC, which was computed by averaging the spectral distribution in each Mel-scale filter, OSC considers the spectral peak (SP), spectral valley (SV) and their difference in each octave-scale filter. In general, the SP represents the harmonic component and the SV corresponds to the non-harmonic component or noise in the spectrum. The difference between the SP and SV roughly reflects the relative spectral distribution in music signals.
  To compute OSC, we use the octave-based band-pass filters (as listed in TABLE 2) to divide the spectrum into several subbands. Suppose the magnitude spectra within the a-th subband are $(P_{a,1}, P_{a,2}, ..., P_{a,N_a})$, where N_a denotes the number of FFT frequency bins within the a-th subband, and $1 \leq a \leq A$ (A is 8 in this study). Here, without loss of generality, we can assume that these spectra are sorted in a descending order. Afterwards, to ensure these extracted features are steady, we estimate the strength of the spectral peak and the spectral valley by averaging values in the largest α percentage spectra and that in the smallest α percentage spectra as follows [18],

$$Peak\,(a) = \log\left(\frac{1}{\alpha N_a}\sum_{i=1}^{\alpha N_a} P_{a,i}\right), \qquad (1)$$

$$Valley\,(a) = \log\left(\frac{1}{\alpha N_a}\sum_{i=1}^{\alpha N_a} P_{a,N_a-i+1}\right), \quad (2)$$

where α is a neighborhood factor (0.2 in this study, identical to that used in [28]). The spectral contrast is then computed as the difference between the spectral peak and the spectral valley:

$$SC(a) = Peak(a) - Valley(a). \quad (3)$$

Following [18] and [28], a feature vector consisting of the spectral valleys and spectral contrasts of all subbands is used to represent the OSC features extracted from an audio frame.

- *SFM/SCM (Spectral Flatness Measure/Spectral Crest Measure)*
  SFM/SCM are proposed to measure the degree of noisiness (or flatness) and sinusoidality of the spectra [20]. Similar to OSC, in this study SFM/SCM are also computed within each octave-scale subband. SFM is defined as the ratio of the geometric mean to the arithmetic mean of the magnitude spectra,

$$SFM\,(a) = \frac{\sqrt[N_a]{\prod_{i=1}^{N_a} B_{a,i}}}{\frac{1}{N_a}\sum_{i=1}^{N_a} B_{a,i}}, \quad (4)$$

  where $B_{a,i}$ is the *i*-th magnitude spectrum in the *a*-th subband. Audio signals with SFM close to 1 indicate a similar amount of power in all spectral bands. An example of this case is white noise. For tonal signals, *e.g.*, a mixture of sine waves, SFM will be close to 0. Similarly, SCM is defined as the ratio of the maximum value within the *a*-th subband to the arithmetic mean of the magnitude spectra within the *a*-th subband,

$$SCM\,(a) = \frac{\max\limits_{i=1,\ldots,N_a}(B_{a,i})}{\frac{1}{N_a}\sum_{i=1}^{N_a} B_{a,i}}. \quad (5)$$

- *Statistical Descriptors of Short-Term Timbre Features*

  To summarize short-term timbre features (including SSD, MFCC, OSC, and SFM/SCM), we compute the mean and standard deviation along each feature dimension over all frames [1]. These statistical descriptors are then concatenated to form a compact feature vector. Thus, we can construct a feature vector of length $116 (= 2 \times 5 + 2 \times 21 + 2 \times 16 + 2 \times 16)$ for each music clip of arbitrary length.

3.2 Long-Term Modulation Spectral Analysis of MFCC, OSC, and SFM/SCM (MMFCC, MOSC, and MSFM/MSCM)

Frame-based timbre features can only capture short time spectral properties of audio signals. To characterize long time spectral variations within a longer audio segment, Lee *el al*. [28] first performed a long-term modulation spectral analysis on short-term timbre features (*e.g.*, MFCC, OSC, and NASE). It is worth noting that the same analysis has been successfully applied to speech recognition [33], speaker recognition [34], and sound classification [22], [27]. Typically, the modulation spectral analysis is based on a two-dimensional representation of acoustic frequency and modulation frequency [31]. Here the acoustic frequency is the standard frequency in FFT, while the modulation frequency can be used to capture the time-varying information through the temporal modulation of audio signals. Previous work [31] has shown that the periodicity of music signals will cause some nonzero terms in the joint frequency representation. In general, modulation spectra in the range of 1-2 Hz and 3-15 Hz respectively represent musical beat rates and the order of speech syllabic rates [27].

TABLE 3
Frequency Ranges of Each Modulaton Subband

Subband Number	Modulation Frequency Range (Hz)
1	[0, 0.33)
2	[0.33, 0.66)
3	[0.66, 1.32)
4	[1.32, 2.64)
5	[2.64, 5.28)
6	[5.28, 10.56)
7	[10.56, 21.03]

* Given a frame duration of 46 ms and overlap of 23 ms, the frame rate is 42.06 frames/sec. Thus the maximal modulation frequency is 21.03 Hz.

Traditional modulation spectral analysis is usually carried out in three steps [34]. First, a spectrogram is computed using FFT on each pre-emphasized, Hamming-windowed frame. Then, a modulation spectrogram is obtained by performing FFT again on the magnitude spectrum of each acoustic frequency over a texture window of W frames. In this way low and high modulation frequencies respectively correspond to slow and fast spectral changes. Finally, modulation spectrograms computed from all texture windows of a music clip are averaged to obtain a representative modulation spectrogram.

To perform modulation spectral analysis on MFCC, OSC, and SFM/SCM to obtain new features (respectively denoted as MMFCC, MOSC, and MSFM/MSCM after modulation computation), we adopted Lee *et al*.'s approach [28] as follows.

- *Step 1:*
  Apply FFT on each feature dimension independently within a texture window of W frames to obtain the feature spectrogram. Here a texture window is a two-dimensional matrix consisting of either MFCC, OSC, or SFM/SCM features extracted from W audio frames, and $W/2$ is the number of modulation frequency bins. In our

MIREX 2011 submission, W is 256 (around 6 seconds) with 50% overlap.

- *Step 2:*
  Derive an averaged feature spectrogram of a music clip by averaging all feature spectrograms obtained from the previous step:

$$\overline{M}(m,d) = \frac{1}{T}\sum_{t=1}^{T}|M_t(m,d)|, \quad 1 \le m \le \frac{W}{2}, 1 \le d \le D, \tag{6}$$

  where $M_t(m,d)$ represents the feature spectrogram of the t-th texture window, m is the modulation frequency index, and D is the length of the feature vector, which is 21, 16, and 16 for MFCC, OSC, and SFM/SCM, respectively.
- *Step 3:*
  Decompose the averaged feature spectrum of each feature value into logarithmically-spaced modulation subbands (as listed in TABLE 3). This operation is based on the observation that human perception for modulation frequency follows a logarithmic frequency scale with resolution consistent with a constant-Q effect [35]. Then, for each spectral/cepstral feature value, we compute the modulation spectral peak/valley (MSP/MSV) as follows:

$$MSP(b,d) = \max_{\phi_{b,l} \le m < \phi_{b,h}} \left(\overline{M}(m,d)\right), \tag{7}$$

$$MSV(b,d) = \min_{\phi_{b,l} \le m < \phi_{b,h}} \left(\overline{M}(m,d)\right), \tag{8}$$

  where $\phi_{b,l}$ and $\phi_{b,h}$ denote the lowest and highest modulation frequency indices of the b-th modulation subband, $1 \le b \le B$, and B is the number of modulation subbands (7 in this study). Here MSP and MSV respectively correspond to rhythmic and non-rhythmic components within each modulation subband. Thus their difference can be used to measure the modulation spectral contrast (MSC) distribution, which can be defined as

$$MSC(b,d) = MSP(b,d) - MSV(b,d). \tag{9}$$

  Note that MSC, MSP and MSV are matrices of the same size $D \times B$.
- *Step 4:*
  Compute the mean and standard deviation along each row and each column of the MSC and MSV matrices to obtain a summarized modulation feature vector. These statistical descriptors of the modulation features are then concatenated together to form a feature vector of length $(4D+4B)$ of MMFCC, MOSC, or MSFM/MSCM for a music clip.

4 MIREX Audio Mood Classification (AMC) Contest

This section firstly describes the MIREX audio mood classification contest, followed by our submission to this contest and the result of our submission.

4.1 Introduction to the AMC Contest

The audio mood classification (AMC) contest was first conducted within MIREX audio classification (train/test) task in 2007. The goal of AMC is to systematically evaluate algorithms for predicting mood from music. The contest provides a common platform (with common datasets, mood labels, and criteria for performance evaluation) such that different algorithms can be evaluated objectively by the organizer.

The MIREX audio mood dataset involves five clusters, each of which contains 120 clips to form a total of 600 clips. The ground-truth set of this dataset was built based on metadata analysis and human assessments. For more details about how to create the ground-truth set, readers with interest can see [36]. Every clip has a duration of 30 seconds which was encoded as a mono wav file with a sample rate of 22,050 Hz. All submissions were evaluated using three-fold cross-validation and artist filtering was used to produce the training and test sets of both datasets. The evaluation metric is the classification accuracy which is computed as the number of correct classifications divided by the number of test music clips. For each submission, the accuracies of three-fold evaluations are averaged to obtain the final classification accuracy.

4.2 Our MIREX 2011 Winning Method and Result

Our submission was based on Lee *et al.*'s modulation spectral analysis (long-term modulation spectral analysis) of MFCC, OSC, and SFM/SCM, together with statistical descriptors of short-term timbre features for SVMs. By concatenating these two types of features together, we can represent each music clip as a fixed-length feature vector. In the classifier construction stage, SVMs with a radius basis function (RBF) kernel were trained for classification [37]. Here we adopted a grid search to tune the hyper-parameters of SVMs (*e.g.*, cost penalty and gamma) on a three-fold inner cross-validation of the training data. The final tuned parameters are used to train SVMs on the whole training data set. Note that z-normalization was employed for each feature dimension prior to SVM training.

Fig. 5 shows the evaluation results of the MIREX 2011 audio mood classification contest. From this figure, we can observe that our method (JR1) was ranked first out of 16 submissions. TABLE 4 shows the classification accuracy of the contest from 2008 to 2014. From this table, we can observe that, for the audio mood classification contest, our method not only

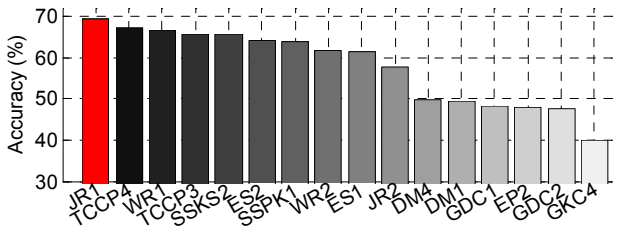


Fig. 5. Evaluation results of the MIREX 2011 audio mood classification contest.

TABLE 4
Comparison of Our MIREX 2011 Submission and Winners of MIREX Audio Mood Classification Contests

Participants	Year	Accuracy (%)	Rank(# of Submissions)
Panda and Paiva	2014	66.33 (%)	1 (12)
Wu and Jang	2013	68.33 (%)	1 (23)
Panda and Paiva	2012	67.83 (%)	1 (20)
Our submission	**2011**	**69.50 (%)**	**1 (17)**
Wang *et al.*	2010	64.17 (%)	1 (36)
Cao and Li	2009	65.67 (%)	1 (33)
Peeters	2008	63.67 (%)	1 (13)

* The results of these submissions are available at http://www.music-ir.org/mirex/wiki/20xx:MIREX20xx_Results, where xx denotes the year of submission (starting from 08 to 14). Note that the results of the same task are comparable since the committees of MIREX used the same training/test splits to evaluate the performance in these years.

won the first place at 2011 but also provides the best result from 2008 to 2014. This indicates the usefulness of long-term modulation frequency analysis for music mood classification.

5 Proposed Joint Frequency Features

This section firstly describes the proposed features, followed by an illustration of the comparison of the proposed and Lee *et al.*'s features.

5.1 Proposed Features

Although applying modulation spectral analysis on music features can achieve good performance for content-based music classification [28], [30], the averaging and summarization operations (see Section 3.2 steps 2 and 4) are likely to smooth out important modulation information, which may degrade the classification performance. To deal with this problem, in this paper we propose the use of joint frequency features computed from a joint frequency representation of an entire music clip as follows. (It should be noted that modulation spectral analysis (proposed by Sukkitanon *et al.* [31]) was used in both Lee *et al.*'s and our approaches. Compared to Lee *et al.*'s approach [28], we propose not to extract the modulation spectrum on local basis for a texture window, but to compute the modulation spectrum on the entire music clip. The advantage of our approach is that there is no need to average or summarize the local modulation features. In contrast, Lee *et al.*'s approach is likely to smooth out important modulation information, leading to a feature set with less discriminative power.)

- *Step 1:*
  Apply FFT on each pre-emphasized, Hamming-windowed frame of a music clip to obtain a conventional spectrogram.
- *Step 2:*
  Perform FFT again for the magnitude spectrum of each acoustic frequency of the entire spectrogram to obtain a joint acoustic and modulation frequency spectrogram (referred to here as a joint frequency representation). Such a representation does not require the use of the texture windows and thus increases modulation frequency resolution so as to extract more discriminative modulation features.
- *Step 3:*
  For the joint acoustic-modulation spectrogram, respectively decompose the modulation spectrum along the acoustic frequency axis and the modulation frequency axis into octave-based and logarithmically spaced modulation subbands. TABLE 2 and TABLE 3 respectively list the frequency ranges of the acoustic and modulation subbands. This allows us to analyze the strength of harmonic (or non-harmonic) components over different musical beat rates in the music signals. To be more specific, for each joint acoustic-modulation frequency subband, we compute the acoustic-modulation spectral peak (AMSP) and the acoustic-modulation spectral valley (AMSV) as follows:

$$AMSP\,(a,b) = \log\left(\frac{1}{\alpha N_{a,b}}\sum_{i=1}^{\alpha N_{a,b}} S_{a,b}[i]\right), \quad (10)$$

$$AMSV\,(a,b) = \log\left(\frac{1}{\alpha N_{a,b}}\sum_{i=1}^{\alpha N_{a,b}} S_{a,b}\left[N_{a,b}-i+1\right]\right). \quad (11)$$

Here, we assume $S_{a,b}$ is the matrix of the magnitude spectra of the joint a-th acoustic frequency and b-th modulation frequency subbands. For simplicity, we can assume $S_{a,b}$ is a descending sorted vector in which $S_{a,b}[i]$ is the i-th element of $S_{a,b}$, $N_{a,b}$ is the total number of elements in $S_{a,b}$, and α is a neighborhood factor identical to that used in computing OSC. The difference between AMSP and AMSV, denoted as AMSC (acoustic-modulation spectral contrast), can be used to reflect the spectral contrast over a joint frequency subband:

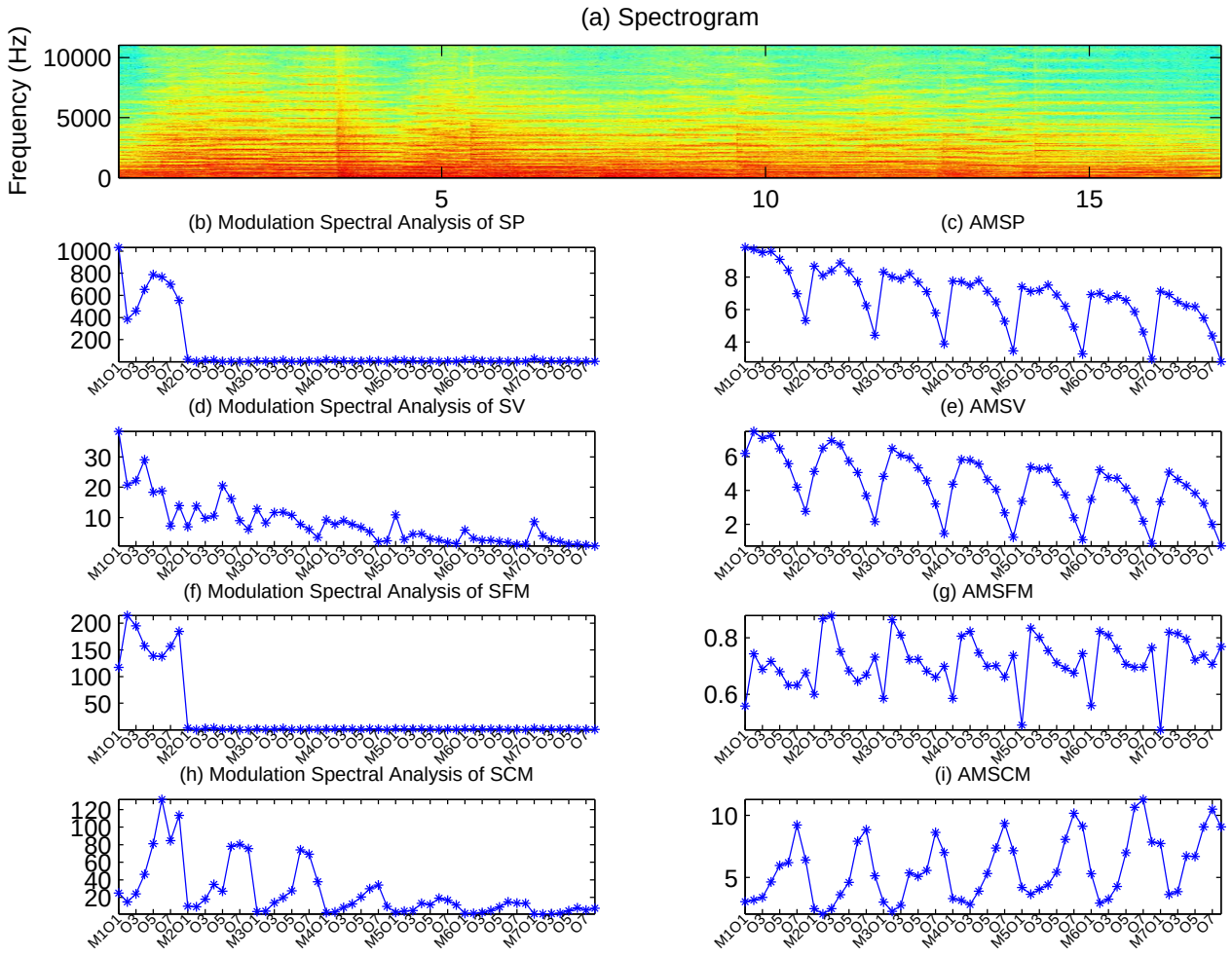


Fig. 6. An illustration of the proposed joint frequency features and Lee *el al.*'s modulation features.

$$AMSC\left(a,b\right)=AMSP\left(a,b\right)-AMSV\left(a,b\right). \quad (12)$$

To measure the noisiness and sinusoidality of the modulation spectra, we further define the acoustic-modulation spectral flatness measure (AMSFM) as the ratio of the geometric mean to the arithmetic mean of the modulation spectra within a joint frequency subband:

$$AMSFM\left(a,b\right)=\frac{\sqrt[N_{a,b}]{\prod_{i=1}^{N_{a,b}}B_{a,b}\left[i\right]}}{\frac{1}{N_{a,b}}\sum_{i=1}^{N_{a,b}}B_{a,b}\left[i\right]}, \quad (13)$$

where $B_{a,b}\left[i\right]$ is the i-th modulation spectrum of the joint a-th acoustic frequency and the b-th modulation frequency subbands. Similarly, the acoustic-modulation spectral crest measure (AMSCM) can be defined as the ratio of the maximum to the arithmetic mean of the modulation spectra within a joint frequency subband,

$$AMSCM\left(a,b\right)=\frac{\max\limits_{i=1,\ldots,N_{a,b}}\left(B_{a,b}\left[i\right]\right)}{\frac{1}{N_{a,b}}\sum_{i=1}^{N_{a,b}}B_{a,b}\left[i\right]}. \quad (14)$$

In summary, for a joint acoustic-modulation spectrogram, we can compute four joint frequency features, namely AMSC, AMSV, AMSFM, and AMSCM, and each of them is a matrix of size $A \times B$.

5.2 Illustration of The Proposed and Lee *et al.*'s Features

Here we used an example of a music clip to show the difference between the proposed joint frequency features and Lee *et al.*'s modulation features. Fig. 6 shows the spectrogram of a 17-second music clip at the top sub-panel, together with the proposed joint frequency features (Figs. 6 (c), (e), (g), and (i)) and Lee *et al.*'s modulation frequency features (Figs. 6 (b), (d), (f), and (h)). The x-axis of these eight sub-panels denotes seven modulation sub-bands (see TABLE 3) and eight octave-based sub-bands (see TABLE 2). For instance, a value computed from *the n-th Modulation*

sub-band and *the first Octave-based sub-band* is denoted as M*n*O*1* in the x-axis label. O*3*, at the right-hand side of M*n*O*1*, represents another value computed from *the same Modulation sub-band*, but *the third Octave-based sub-band*. Similar explanations apply to O*5* and O*7* as well. For the sake of simplicity, we do not display the labels O*2*, O*4*, O*6* and O*8* here.

Compared to Lee *et al.*'s features (modulation spectral analysis of SP and SFM, see Figs. 6 (b) and (f)), the proposed features (AMSP and AMSFM, see Figs. 6 (c) and (g)) have more obvious peaks and valleys for better classification. On the other hand, Lee *et al.*'s features (Figs. 6 (b) and (f)) have relatively larger spectrums in modulation sub-band one (M*1*O*1* to O*8*) than in the other modulation sub-bands (from M*2*O*1* to the last label O*8*), indicating the high-frequency components are smoothed out. The main reason for this difference is that the proposed method performs modulation analysis on the entire music clip, resulting in more obvious peaks and valleys (especially at higher modulation frequencies) than Lee *et al.*'s features where the modulation spectral analysis is only performed on a local texture window. A similar phenomenon can be observed in the comparison of Lee *et al.*'s features (modulation spectral analysis of SV and SCM, see Figs. 6 (d) and (h)) and the proposed features (AMSV and AMSCM, see Figs. 6 (e) and (i)).

6 Experimental Results

This section presents the experiments conducted on three mood datasets. We compare the proposed joint frequency features with the modulation spectral analysis of OSC and SFM/SCM to show the strength of the proposed features. We also combine the proposed features with MMFCC and statistical descriptors of short-term timbre features to demonstrate how this new set of features outperforms our MIREX 2011 submission.

6.1 Datasets

To evaluate the performance of mood classification, we use the *Soundtracks* [5] dataset, the MIREX-like mood dataset, and our newly collected *MIR-Mood* dataset.[4] *Soundtracks* covers six discrete mood classes, including happiness, sadness, fear, anger, surprise, and tenderness. Each class includes 30 music clips lasting between 18 and 30 seconds. Panda *et. al.* [38] followed the same organization as the one used in MIREX audio mood classification to create the MIREX-like mood dataset. This dataset has a total of 903 30-second clips, each of which belongs to one of the five clusters (as shown in TABLE 1). Each cluster contains different numbers of clips, say, 170 clips in cluster 1,164 clips in cluster 2,215 clips in cluster 3,191 clips in cluster 4, and 163 clips in cluster 5. In addition to these two datasets, we followed ref. [39] to collect social tags from Last.fm[5] and audio files from 7digital to form the *MIR-Mood* dataset. First, four basic mood classes (including angry, happy, relaxed, and sad), which cover the four quadrants of the two-dimensional mood model [39], are used as seeds to retrieve the top 30 tags with the most counts from Last.fm. We then obtained a list of music clips labeled with these retrieved tags. Given the retrieved titles and artists, we used the 7digital API to download preview files. After manually filtering for files overlapping two or more classes, we obtained 553, 587, 619, and 464 music clips, respectively, for the angry, happy, relaxed, and sad classes, totaling 2,223 music clips. The duration of each audio file is around either 30 or 60 seconds. Each music clip of these three datasets is converted into mono with a sample rate of 22,050 Hz. (Note that here we do not use Song *et al.*'s collection [39] to evaluate the performance since they did not provide any trackid, a unique identification of a music clip that allows us to download the audio file).

6.2 Experimental Setup

In our experiments, we used the same strategy as in our MIREX submission to train RBF SVMs. We evaluated the performance of mood classification via ten randomly stratified ten-fold cross-validations of these three datasets. To compare with [40] for evaluating the *Soundtracks* dataset, we did not apply artist filtering here. For the MIREX-like mood dataset, no artist filtering is applied. For the *MIR-Mood* dataset, we applied artist filtering to obtain training and test splits.

6.3 Results

TABLE 5 shows the averaged classification accuracy and standard deviation of various feature sets for these three mood datasets. The first column lists the used feature sets, where the proposed feature sets are in italics. Friedman's test was used to evaluate the significance of the improvements on the same cell consisting of the proposed feature set and Lee *el al.*'s modulation feature (in the column of feature set). The accuracy figures are underlined if the improvement is significant based on Friedman's test.

Three observations for this experiment are as follows.

1) Adding short term timbre features (rows 8 and 9) further improves the classification accuracy. This indicates that MuStd (denoting the concatenation of summarized features of SSD, MFCC, OSC, and SFM/SCM) feature set can effectively

4. MIREX-like mood and MIR-Mood datasets can be downloaded from http://mir.dei.uc.pt/downloads.html and http://mirlab.org/dataSet/public, respectively

5. www.last.fm.com

TABLE 5
Averaged Classification Accuracy (%) and Standard Deviation (in parentheses) of Various Feature Sets on the *Soundtracks*, *MIR-Mood* and *MIREX-like Mood* Datasets

Row Index	Feature Set	Feature Dimension	*Soundtracks*	*MIR-Mood*	*MIREX-like Mood*
1	MuStd[a]	116	39.28 (1.74)	50.38 (0.64)	40.66 (0.93)
2	MOSC	92	37.94 (1.94)	50.68 (0.38)	39.44 (0.57)
3	*AMSC/AMSV*	112	38.56 (2.13)	51.32 (0.63)	41.95 (0.72)
4	MSFM/MSCM	92	32.11 (2.09)	48.84 (0.28)	37.09 (1.15)
5	*AMSFM/AMSCM*	112	32.94 (1.46)	49.11 (0.46)	38.53 (0.72)
6	MOSC+MSFM/MSCM	184	38.33 (2.58)	51.14 (0.36)	39.50 (0.80)
7	*AMSC/AMSV+AMSFM/AMSCM*	224	38.72 (1.78)	51.74 (0.78)	40.71 (0.69)
8	MuStd+MMFCC+MOSC+MSFM/MSCM[b]	412	41.06 (1.58)	52.87 (0.55)	43.22 (0.47)
9	MuStd+MMFCC+*AMSC/AMSV+AMSFM/AMSCM*[c]	452	43.56[d](1.80)	53.01 (0.66)	44.74 (0.79)

[a] MuStd denotes the concatenation of summarized features of SSD, MFCC, OSC, and SFM/SCM.
[b] This feature sed was used in our MIREX 2011 submission.
[c] The best feature set among all.
[d] The underlined numbers indicate that the proposed feature set outperforms the other in the same cell with statistical significance, that is, with $p<0.05$.

TABLE 6
Comparison of the Proposed System with Other Recent Approaches on the Soundtracks Dataset

Approach	Accuracy
The proposed system	**43.56**
Our MIREX 2011 submission	**41.06**
Panagakis and Kotropoulos [40]	39.44

TABLE 7
Comparison of the Feature Set (AMSC/AMSV+AMSFM/AMSCM) With and Without a Pre-emphasis Filter (in Averaged Accuracy (%) and Standard Deviation)

Feature Set	Accuracy (Standard Deviation)
With a Pre-emphasis Filter[a]	50.91[b](0.76)
Without a Pre-emphasis Filter	49.23 (0.38)

[a] The classification system that uses the feature set with a pre-emphasis filter outperforms the other with statistical significance ($p<0.05$).
[b] Since we performed another ten runs of randomly stratified ten-fold cross-validations of MIR-Mood dataset in TABLE 7, a different average accuracy (compared to the row 7 of TABLE 5 on the same dataset) is obtained here due to the randomness in stratified ten folds.

complement the modulation based feature sets for the classification task.

2) The proposed features, AMSC/AMSV (row 3) and AMSFM/AMSCM (row 5), outperform modulation spectral analysis of OSC and SFM/SCM (*e.g.*, MOSC in row 2 and MSFM/MSCM in row 4) in three datasets by small margins. After applying Friedman's test to these results, we found that the proposed features do not have significant improvement in Soundtracks dataset. This phenomenon may be caused by the relatively short duration of the music clips in this dataset (72 out of 180 music clips with durations less than 18 seconds), leading to the difficulty in obtaining effective long-term modulation information. The same observation also applies to MOSC+MSFM/MSCM (row 6) and AMSC/AMSV+AMSFM/AMSCM (row 7).
3) The proposed system that combines MuStd, MMFCC, AMSC/AMSV and AMSFM/AMSCM (row 9) outperforms our MIREX 2011 submission (row 8) on all three datasets with statistical significance ($p<0.05$).

Here we also compared the proposed system with other approaches. As shown in TABLE 6, the proposed system outperforms Panagakis and Kotropoulos' approach [40] which, as far as we know, is the only approach to have evaluated performance on this dataset to classify six moods. This indicates the effectiveness of the proposed joint frequency features.

In addition, to verify the validity of the pre-emphasis filter (required by the reviewer), we used the proposed feature set (*i.e.*, AMSC/AMSV+AMSFM/AMSCM) with and without a pre-emphasis filter for the classification task, and briefly discuss the results. Here we only conducted experiments on the MIR-Mood dataset, since this dataset is the largest collection available. The same experimental setup as mentioned in Section 6.2 was used here. TABLE 7 shows the results of averaged accuracy (%) and standard deviation (in parentheses). It is clear that the feature set with a pre-emphasis filter provides improved performance.

The confusion table of this experiment is shown in TABLE 8, where rows denote the ground truth of music moods, and columns denote the computed result of music moods. From this table, we can observe that the use of the pre-emphasis filtering leads to better accuracy for all moods. The pre-emphasis filter is commonly used in speech recognition since it can compensate for the high-frequency part of the speech signal that was suppressed by the human voice production mechanism. For music, most instruments

TABLE 8
Confusion Table of the Feature Set (AMSC/AMSV+AMSFM/AMSCM) With and Without a Pre-emphasis Filter (%)

Preprocess	With a Pre-emphasis Filter				Without a Pre-emphasis Filter			
Mood	Angry	Happy	Relaxed	Sad	Angry	Happy	Relaxed	Sad
Angry	**57.1**	22.2	11.4	9.2	**54.7**	24.1	12.1	9.1
Happy	21.5	**51.4**	18.9	8.2	22.4	**50.6**	18.3	8.7
Relaxed	6.9	17.0	**54.7**	21.4	8.4	16.2	**53.4**	22.1
Sad	13.8	12.6	35.7	**37.9**	13.1	15.0	36.4	**35.4**

can still be modeled by the source-filter model (just like human voice production mechanism), so we suppose that the pre-emphasis filter can be used for both speech and music.

7 Conclusion and Limitation of this Work

In this paper, we found that two operations (which compute the representative feature spectrogram and the mean and standard deviation of the MSC/MSV matrices) in the modulation spectral analysis of short-term timbre features are likely to smooth out useful modulation information, so we propose the use of a joint frequency representation of an entire music clip to extract joint frequency features. These joint frequency features, including acoustic-modulation spectral contrast/valley, acoustic-modulation spectral flatness measure and acoustic-modulation spectral crest measure, outperform the modulation spectral analysis of OSC and SFM/SCM (used in Lee *et al.*'s approach) in three mood datasets by small margins. On the other hand, by combining the proposed features with the modulation spectral analysis of MFCC and statistical descriptors of SSD, MFCC, OSC and SFM/SCM as a new feature set, we found that the set can outperform our MIREX submission on all three datasets with statistical significance (confirmed by Friedman's test).

The advantage of the proposed features is that they can have a better discriminative power due their operation on the entire music, with no averaging over the local modulation features. (In contrast, Lee *et al.*'s approach is likely to smooth out important modulation information, leading to a feature set with less discriminative power.) On the other hand, a drawback of the propose features is that for a music clip with a short duration (*e.g.,* 6 seconds or so), the extracted modulation features are similar to Lee *et al.*'s approach without too much improvement. Future work will explore the possibility of using dimensionality reduction techniques to extract a compact feature set that can achieve equal or better performance. We will also apply these features to multi-label tasks such as auto-tagging and tag-based retrieval.

References

[1] G. Tzanetakis and P. Cook, "Musical genre classification of audio signals," *IEEE Transactions on Speech and Audio Processing*, vol. 10, no. 5, July 2002.

[2] L. Lu, D. Liu, and H. Zhang, "Automatic mood detection and tracking of music audio signals." *IEEE Transactions on Audio, Speech and Language Processing*, vol. 14, no. 1, pp. 5–18, 2006. [Online]. Available: http://dblp.uni-trier.de/db/journals/taslp/taslp14.html#LuLZ06

[3] D. Huron, "Perceptual and cognitive applications in music information retrieval." in *ISMIR*, 2000. [Online]. Available: http://dblp.uni-trier.de/db/conf/ismir/ismir2000.html#Huron00

[4] M. Barthet, G. Fazekas, and M. Sandler, "Multidisciplinary perspectives on music emotion recognition: recommendations for content- and context-based models." *Proc. CMMR*, pp. 492–507, 2012.

[5] T. Eerola and J. K. Vuoskoski, "A comparison of the discrete and dimensional models of emotion in music," *Psychology of Music*, 2010.

[6] Y.-H. Yang and H. H. Chen, "Machine recognition of music emotion: A review," *ACM Trans. Intell. Syst. Technol.*, vol. 3, no. 3, pp. 40:1–40:30, May 2012. [Online]. Available: http://doi.acm.org/10.1145/2168752.2168754

[7] P. Ekman, "Expression and the nature of emotion," *Approaches to emotion*, vol. 3, pp. 19–344, 1984.

[8] K. Hevner, "Expression in music: a discussion of experimental studies and theories." *Psychological Review*, vol. 42, no. 2, pp. 186–204, 1935.

[9] X. Hu and J. S. Downie, "Exploring mood metadata: Relationships with genre, artist and usage metadata." in *ISMIR*. Citeseer, 2007, pp. 67–72.

[10] J. A. Russell, "A circumplex model of affect." *Journal of personality and social psychology*, vol. 39, no. 6, pp. 1161–1178, 1980.

[11] R. T. Ross, "A statistic for circular series," *Journal of Educational Psychology*, no. 5, pp. 384–389, 1938.

[12] R. O. Duda, P. E. Hart, and D. G. Stork, *Pattern Classification (2Nd Edition)*. Wiley-Interscience, 2000.

[13] Z. Fu, G. Lu, K. M. Ting, and D. Zhang, "A survey of audio-based music classification and annotation," *Multimedia, IEEE Transactions on*, vol. 13, no. 2, pp. 303–319, 2011.

[14] Y. E. Kim, E. M. Schmidt, R. Migneco, O. G. Morton, P. Richardson, J. Scott, J. A. Speck, and D. Turnbull, "Emotion recognition: a state of the art review," in *11th International Society for Music Information and Retrieval Conference*, 2010.

[15] Y. Liu, Y. Liu, Y. Zhao, and K. A. Hua, "What strikes the strings of your heart?: Multi-label dimensionality reduction for music emotion analysis," in *Proceedings of the ACM International Conference on Multimedia*, ser. MM '14. New York, NY, USA: ACM, 2014, pp. 1069–1072. [Online]. Available: http://doi.acm.org/10.1145/2647868.2655068

[16] Y.-H. Yang, Y.-C. Lin, Y.-F. Su, and H. H. Chen, "A regression approach to music emotion recognition." *IEEE Transactions on Audio, Speech and Language Processing*, no. 2, pp. 448–457.

[17] Z. Fu, G. Lu, K. M. Ting, and D. Zhang, "A survey of audio-based music classification and annotation." *IEEE Transactions on Multimedia*, vol. 13, no. 2, pp. 303–319, 2011. [Online]. Available: http://dblp.uni-trier.de/db/journals/tmm/tmm13.html#FuLTZ11

[18] D.-N. Jiang, L. Lu, H.-J. Zhang, J.-H. Tao, and L.-H. Cai, "Music type classification by spectral contrast feature." in *ICME (1)*. IEEE, 2002, pp. 113–116. [Online]. Available: http://dblp.uni-trier.de/db/conf/icmcs/icme2002-1.html#JiangLZTC02

[19] K. West and S. Cox, "Features and classifiers for the automatic classification of musical audio signals," in *Proceedings of the 5th International Conference on Music Information Retrieval (ISMIR 2004)*, 2004.

[20] G. Peeters, "A large set of audio features for sound description (similarity and classification) in the CUIDADO project," Icram, Tech. Rep., 2004.

[21] H.-G. Kim, N. Moreau, and T. Sikora, "Audio classification based on mpeg-7 spectral basis representations," *Circuits and Systems for Video Technology, IEEE Transactions on*, vol. 14, no. 5, pp. 716–725, May 2004.

[22] F. Mrchen, A. Ultsch, M. Thies, and I. Lohken, "Modeling timbre distance with temporal statistics from polyphonic music." *IEEE Transactions on Audio, Speech and Language Processing*, vol. 14, no. 1, pp. 81–90, 2006. [Online]. Available: http://dblp.uni-trier.de/db/journals/taslp/taslp14.html#MorchenUTL06

[23] W. A. Sethares, R. D. Morris, and J. C. Sethares, "Beat tracking of musical performances using low-level audio features." *IEEE Transactions on Speech and Audio Processing*, vol. 13, no. 2, pp. 275–285, 2005. [Online]. Available: http://dblp.uni-trier.de/db/journals/taslp/taslp13.html#SetharesMS05

[24] T. Li and M. Ogihara, "Toward intelligent music information retrieval," *IEEE Transactions on Multimedia*, vol. 8, no. 3, pp. 564–574, 2006.

[25] J. Bergstra, 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, Dec. 2006. [Online]. Available: http://dx.doi.org/10.1007/s10994-006-9019-7

[26] C. Xu, M. Maddage, and X. Shao, "Automatic music classification and summarization," *IEEE Transactions on Speech and Audio Processing*, vol. 13, no. 3, pp. 441–450, May 2005.

[27] M. F. McKinney and J. Breebaart, "Features for audio and music classification." in *ISMIR*, 2003. [Online]. Available: http://dblp.uni-trier.de/db/conf/ismir/ismir2003.html#McKinneyB03

[28] C.-H. Lee, J.-L. Shih, K.-M. Yu, and H.-S. Lin, "Automatic music genre classification based on modulation spectral analysis of spectral and cepstral features." *IEEE Transactions on Multimedia*, vol. 11, no. 4, pp. 670–682, 2009. [Online]. Available: http://dblp.uni-trier.de/db/journals/tmm/tmm11.html#LeeSYL09

[29] Y. Panagakis, C. Kotropoulos, and G. R. Arce, "Non-negative multilinear principal component analysis of auditory temporal modulations for music genre classification." *IEEE Transactions on Audio, Speech and Language Processing*, vol. 18, no. 3, pp. 576–588, 2010. [Online]. Available: http://dblp.uni-trier.de/db/journals/taslp/taslp18.html#PanagakisKA10

[30] S.-C. Lim, S.-J. Jang, S.-P. Lee, and M. YoungKim, "Music genre/mood classification using a feature-based modulation spectrum," in *Mobile IT Convergence (ICMIC), 2011 International Conference on*, Sept 2011, pp. 133–136.

[31] S. Sukittanon, L. E. Atlas, and J. W. Pitton, "Modulation-scale analysis for content identification." *IEEE Transactions on Signal Processing*, vol. 52, no. 10, pp. 3023–3035, 2004. [Online]. Available: http://dblp.uni-trier.de/db/journals/tsp/tsp52.html#SukittanonAP04

[32] J. Aucouturier and F. Pachet, "Representing musical genre: A state of the art," *Journal of New Music Research*, vol. 32, no. 1, pp. 83–93, 2003.

[33] B. E. Kingsbury, N. Morgan, and S. Greenberg, "Robust speech recognition using the modulation spectrogram," *Speech Communication*, vol. 25, no. 13, pp. 117 – 132, 1998. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0167639398000326

[34] T. Kinnunen, "Joint acoustic-modulation frequency for speaker recognition," in *Acoustics, Speech and Signal Processing, 2006. ICASSP 2006 Proceedings. 2006 IEEE International Conference on*, vol. 1, May 2006, pp. 665–668.

[35] S. D. Ewert and T. Dau, "Characterizing frequency selectivity for envelope fluctuations," *The Journal of the Acoustical Society of America*, vol. 108, no. 3, pp. 1181–1196, 2000.

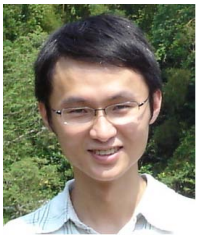
[36] X. H. J. S. Downie, C. Laurier, and M. B. A. F. Ehmann, "The 2007 mirex audio mood classification task: Lessons learned," in *ISMIR 2008: Proceedings of the 9th International Conference of Music Information Retrieval*. Lulu. com, 2008, pp. 462–467.

[37] C. C. Chang and C. J. Lin, "Libsvm: A library for support vector machine," 2010. [Online]. Available: http://www.csie.ntu.edu.tw/~cjlin/libsvm

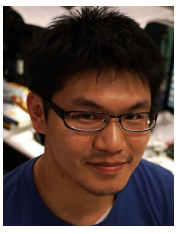
[38] R. Panda, R. Malheiro, B. Rocha, A. Oliveira, and R. Paiva, "Multi-modal music emotion recognition: A new dataset, methodology and comparative analysis," *Proc. CMMR*, 2013.

[39] Y. Song, S. Dixon, and M. Pearce, "Evaluation of musical features for emotion classification," in *Proceedings of the 13th International Society for Music Information Retrieval Conference*, Porto, Portugal, October 8-12 2012, pp. 523–528, http://ismir2012.ismir.net/event/papers/523-ismir-2012.pdf.

[40] Y. Panagakis and C. Kotropoulos, "Automatic music mood classification via low-rank representation," in *Proc*, 2011, pp. 689–693.



Jia-Min Ren received his Ph.D. from the CS Department at National Tsing Hua University, Hsinchu Taiwan. He currently works at the Data Analytic Technology Department at Industrial Technology Research Institute, Hsinchu, Taiwan. His research interests include machine learning, semantic analysis of musical signals, music information retrieval, and analysis of manufacturing data.



Ming-Ju Wu received an M.S. in Computer Science in 2009 from National Chiao Tung University, Hsinchu, Taiwan. In 2015, he received the Ph.D. degree in Computer Science from National Tsing Hua University, Hsinchu, Taiwan. His research interests include information retrieval, image processing, and machine learning..

Jyh-Shing Roger Jang (M'93) received his Ph.D. from the EECS Department at the University of California, Berkeley. He studied fuzzy logic and artificial neural networks with Prof. Lotfi Zadeh, the father of fuzzy logic. As of 2014, Google Scholar shows over 8000 citations for Dr. Jang's seminal paper on adaptive neuro-fuzzy inference systems (ANFIS), published in 1993. After obtaining his Ph.D., he joined MathWorks to coauthor the Fuzzy Logic Toolbox (for MATLAB). He has since cultivated a keen interest in implementing industrial software for pattern recognition and computational intelligence. He was a professor in the CS Dept. of National Tsing Hua Univ., Taiwan, from 1995 to 2012. Since August 2012, he has been a professor in the CSIE Dept. of National Taiwan Univ., Taiwan. He has published one book entitled Neuro-Fuzzy and Soft Computing, two books on MATLAB programming, and one book on JavaScript programming. He has also maintained toolboxes for machine learning and speech/audio signal processing and online tutorials on Data Clustering and Pattern Recognition and Audio Signal Processing and Recognition. His research interests include machine learning and pattern recognition, with applications to speech recognition/assessment/synthesis, music analysis/retrieval, and image identification/retrieval. For further information on Prof. Jang, http://mirlab.org/jang.