A Novel Method based on OMPGW Method for Feature Extraction in Automatic Music Mood Classification

Shasha Mo, and Jianwei Niu, *Senior Member, IEEE,*

Abstract—Music mood is useful for music-related applications such as music retrieval or recommendation, which represents the inherent emotional expression of music signals. In this paper, a novel technique is proposed for music signal analysis in the view of emotions, which is based on the orthogonal matching pursuit, Gabor functions, and the Wigner distribution function. The technique, called the OMPGW method, consists of three-level schemes: the low-level, the middle-level and the high-level schemes. For the low-level schemes, the orthogonal matching pursuit combined with Gabor functions is proposed to provide an adaptive time-frequency decomposition of music signals. Compared with other algorithms for signal analysis, the proposed algorithm can achieve a higher spatial and temporal resolution and give a better interpret of the music signal structures. In the middle-level schemes, the Wigner distribution function is applied to obtain the time-frequency energy distribution of the results from the low-level schemes. High-level schemes are used to describe the modeling of audio features, the procedure of music mood classification. A classifier based on support vector machines is utilized to model the features that is extracted with the proposed technique regarding the emotion models. Several experiments are conducted with four datasets, and better results are achieved with the proposed method. In music mood classification experiments, music clips are classified into different kinds of mood clusters, and mean accuracy of 69.53% on our dataset can be achieved using the proposed method.

Index Terms—Affective computing, feature extraction, music mood classification, orthogonal matching pursuit, time-frequency analysis, Wigner distribution function

1 Introduction

As the dramatic increase of digital music accessible to the general public, retrieving and managing a vast amount of music collections has become an important and challenging issue. To cope with this necessity, Music Information Retrieval (MIR) as an emerging research area has gained increasing attention during the past few years [1], [2], [3]. Automatic emotion recognition in musical audio is an interesting topic in the MIR field, and it could provide many potential applications to music retrieval. In general, there are two steps for automatic music mood classification: feature extraction and classifier learning [4], [5]. The aim of classifier learning is to map the obtained features to emotions under the condition of minimizing prediction error. For the feature extraction step, a music signal is represented by a small set of parameters through various algorithms. A considerable amount of work has been dedicated to the feature extraction [6], [7], [8]. Audio feature extraction is directly or indirectly based on various signal processing techniques, and it will influence the efficiency of a music mood classification system.

In this paper, a novel technique called the OMPGW method is proposed for automatic music mood classification. The technique is based on the Orthogonal Matching Pursuit (OMP) [9], Gabor functions [10], and the Wigner distribution function [11]. Due to the time-varying behavior of music signals, the OMP algorithm combined with Gabor functions is used to handle this characteristic, which leads to a representation of both rhythmic and transient music signal components. The Wigner distribution function is then applied to the OMP results for obtaining the time-frequency energy distribution of music signals. Compared with other techniques, such as Fourier transform, Short Time Fourier Transform (STFT) [12], Wavelet transform [13], constant-Q transform (CQT) [14], [15], Gammatone filter bank (GTFB) [16], [17] and multiscale spectro-temporal modulations (MSTM) [18] et al., the proposed technique is a multiscale decomposition technique and can provide an adaptive time-frequency analysis with a higher spatial and temporal resolution. Based on these processing schemes, the modeling of audio features are presented and followed with the music mood classification procedure. For the reason that the proposed technique is attributed to a better feature extraction, an improved performance will be obtained in music mood classification.

For the proposed automatic music mood classification system in this paper, the music signal is firstly decomposed into a linear expansion of atoms selected from a time-frequency dictionary using the OMP algorithm. By adding Wigner distributions of the selected atoms, the time-frequency energy distribution of music signals will be obtained. The decomposition coefficients and the time-frequency energy distribution of music are then applied to

- *Shasha Mo is with the State Key Laboratory for Virtual Reality Technology and Systems, School of Computer Science and Engineering, Beihang University, Beijing, 100191, China.*
- *Jianwei Niu and Shihao Wang are also with the State Key Laboratory for Virtual Reality Technology and Systems, School of Computer Science and Engineering, Beihang University, Beijing, 100191, China. E-mail: niujianwei@buaa.edu.cn*

Manuscript received April 19, 2005; revised August 26, 2015.

extract audio features, and a novel feature set based on these obtained features is formed and called ATF features. Finally, the obtained ATF features contribute to a classifier learning, and the classification system is obtained and tested. Experiments on four mood annotated datasets have shown that the proposed approach achieved an excellent performance. A mean accuracy of 69.53% on our dataset is achieved in music mood classification experiments.

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

1) A novel technique called OMPGW is proposed for music mood classification, which provides an adaptive time-varying description of music signals with a higher spatial and temporal resolution. The proposed approach is a combination of three signal processing algorithms and the basis for audio feature extraction.
2) Based on the OMPGW method, ten audio features called ATF features are extracted, including spectral centroid, spectral rolloff, spectral flux, spectral bandwith, spectral contrast, spectral flatness measure, spectral contrast, subband power, frequency cepstrum coefficient, and coefficient histogram. In fact, the ATF features outperform all the other algorithms, indicating the superiority of the proposed approach.
3) In order to evaluate the proposed method, five datasets are employed for music mood classification, which are the Soundtracks dataset [19], [20], [21], the MIREX-T dataset, the MTV dataset [22] and the MediaEval 2015 (emotion in Music task) database [23]. It is worth pointing out that the MIREX-T dataset is adapted from the MIREX-like mood dataset [24] and annotated into four basic mood classes by our research group.

The rest of this paper is organized as follows. Section 2 reviews the related work in emotion models and audio features for music classification. An overview of the proposed music mood classification system is described in Section 3. The OMPGW method is introduced in Section 4. Section 5 deals with the feature extraction based on the proposed approach. Section 6 shows a method for single-label music mood classification. Experiments and performance evaluation of the proposed approach are discussed in Section 7. Finally, conclusions are given in Section 8.

2 Related Work

2.1 Emotion Models

Music mood describes the inherent emotional expression of music, and many researchers have been dedicated to the music emotion modeling from different disciplines. In the music psychology, there are two theoretical frameworks for perceived emotions: the categorical models and the dimensional models. The earliest and still best-known categorical model for creating music mood taxonomy is Hevners adjective circle [25] [26]. For the discrete emotions, Erola and Vuoskoski choose happy, sad, fearful, angry, surprising and tender, which have been widely used in previous music studies [19]. Another well-known model for musical expression is the Thayers model of mood as illustrated in Fig. 1, which is adapted from the Russells circumplex model. Thayers model is a two-dimensional approach and indicates two underlying stimuli, which could influence mood responses: stress and energy corresponding to valence and arousal in Russells model, respectively [27] [28] [29]. According to the level of stress and energy, music mood can be divided into four clusters: Contentment, Depression, Exuberance, and Anxious. With the Thayers model of mood, it is possible to obtain important cues for computational modeling. A visual summary of the two-dimensional models of Russell and Thayer, combined with the categorical models of six basic emotions by Erola and Vuoskoski, is shown in Fig. 2. Due to the infinite adjectives of emotions, Thayers model and the categorical model of Erola and Vuoskoski are adopted to the music classification system in this paper.

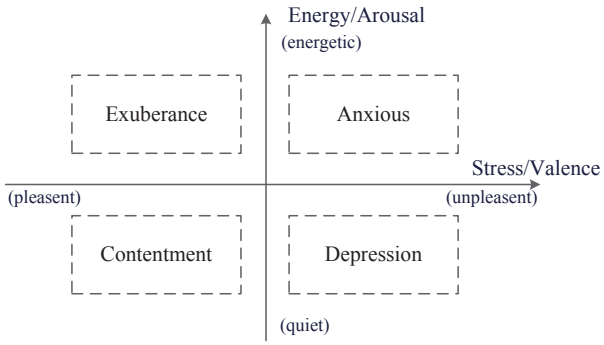


Fig. 1. Thayers model of mood, where four quadrants can be identified: calm-energy (exuberance), clam-tiredness (contentment), stress-energy (anxious), and stress-tiredness (depression).

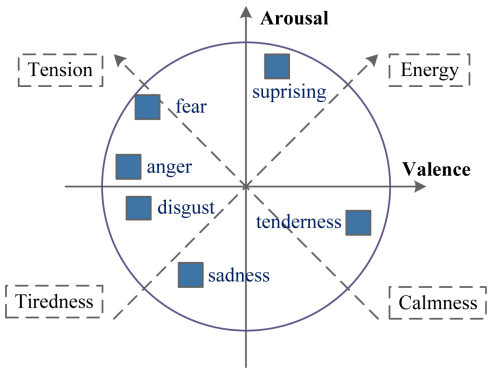


Fig. 2. The dimensional models of Russell and Thayer with common basic emotion categories.

2.2 Audio Features

Many literatures have been focused on the audio feature extraction for content-based music analysis, and several taxonomies are used for audio feature categorization. In the view of music understanding, audio features can be roughly categorized into two levels: low-level and mid-level features [1], [30], [31]. Low-level features are extracted directly from signal processing techniques, while mid-level

features are obtained on the basis of low-level features. Besides, semantic labels, providing information about how humans understand and interpret music, are regarded as top-level labels such as mood, genre, etc. The top-level labels cannot be easily extracted from lower level features for the reason of the semantic gap that needs to be bridged by inferring the labels from lower level features.

A summary of audio features is shown in Table 1 [15], [17], [18], [32]–[38]. For most of existing studies, audio features are extracted using spectral or cepstral analysis. The Fourier transform is an important building block for conventional spectral-based audio feature extraction. In addition, cepstral analysis is valuable for analyzing and developing spectral features, which is computed with the logarithm of the signal spectrum by talking the Inverse Fourier transform of it. The modulation spectral analysis is another audio feature extraction algorithm that is based on Fourier transform. Both STFT and Wavelet transform are time-frequency transforms that have the ability to provide the frequency along with the associated time of music signals. They have slight benefits over the Fourier transform in the case of examining specific frequencies, thus a considerable improvement in MIR will be obtained. CQT, GTFB and MSTM are auditory filters with a logarithmic uniform distribution of the central frequency in the auditory system, and they can also provide a time-frequency spectrogram. The time-frequency transforms, such as STFT and Wavelet transform, have slight benefits over the Fourier transform in the case of examining specific frequencies. The three auditory filters can characterize speech more like the human ear-brain combination than STFT and Wavelet transform, and an improvement has been made in describing audio signals. All of these signal processing techniques can be seen as a technique that decomposes signals over a family of functions called time-frequency atoms. Nevertheless, these time-frequency atoms have to be decided a prior choice before the analysis, and an improper choice may severely bias the analysis of music signals. In this case, using matched time-frequency analysis technique, i.e. adaptive time-frequency analysis, can achieve a better performance.

TABLE 1
Summary of features in music mood classification.

Feature Class	Feature Name
Time domain	Zero Crossing Rate
Fourier transform (Frequency domain & Cepstral domain)	Spectral Bandwidth Spectral Centroid Spectral Flux Spectral Rolloff Spectral Crest Factor Spectral Flatness Measure Octave-based Spectrum Envelop Amplitude Spectrum Envelop Mel-frequency Cepstrum Coefficient Fourier Cepstrum Coefficient Linear Predictive Cepstrum Coefficient
STFT	Short-time Fourier Transform features
Wavelet transform	Daubechies Wavelet Coef Histogram
CQT	Auditory cortical representations
GTFB	Gammatone frequency cepstral coefficients
MSTM	Time-varying cortical representations

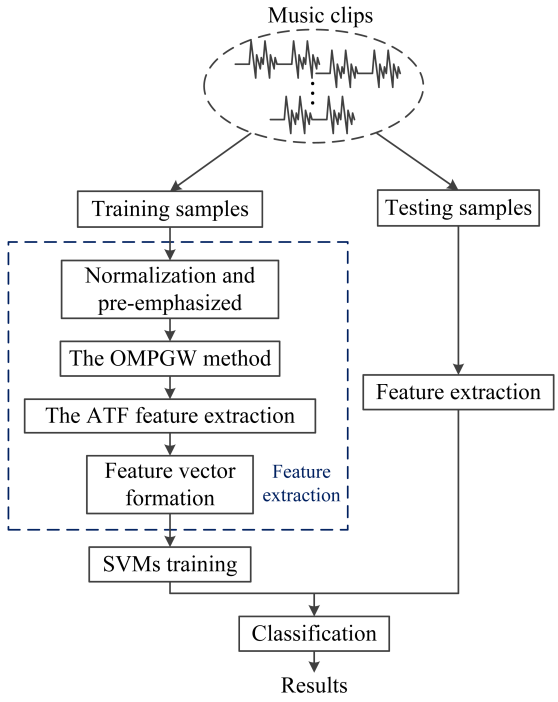


Fig. 3. Framework of the proposed music mood classification system.

3 System Overview

The framework of the proposed music mood classification system using the OMPGW method is shown in Fig. 3. This system contains three main parts: (1) After a preprocessing performed on music clips, the proposed method is used for signal analysis. Firstly, the OMP algorithm combined with Gabor functions is used to provide an adaptive time-frequency representation of the preprocessed music clips, and some decomposition coefficients along with the corresponding functions are obtained. The Wigner distribution function is then applied to obtain the adaptive time-frequency energy distribution of music clips, which is based on the decomposition coefficients and the selected atoms. (2) ATF features based on the OMPGW results are computed from each input music clip, and a normalization is applied to each feature value. (3) The ATF features are concatenated with the Thayers model of mood for SVMs. During testing, the same features are extracted, and the mood of the music clip is classified using the pre-trained SVMs.

4 The OMPGW Method

Before the content-based audio feature extraction, music clips will be represented by a signal processing method called the OMPGW method, which could capture the inherent structure of signals. In this section, the proposed method is presented for generating an adaptive time-frequency estimate of music signals. The proposed method is based on the OMP algorithm, Gabor functions and the Wigner distribution function. Firstly, the matching pursuit algorithm with Gabor functions is applied to music signals, and the Wigner distribution function is used to obtain the time-frequency energy distribution of music signals. The

OMP algorithm is then introduced to acquire an optimal result when compared with the matching pursuit algorithm. Finally, the processing step of the OMPGW method used for music signals is presented. The proposed method has outstanding performance compared with the conventional signal processing technique used in the representation of a non-stationary signal.

4.1 Matching Pursuit Combined with Gabor Functions

To accurately capture the time-frequency characteristics of music signals, the OMP algorithm [39]–[42] is applied to decompose music signals into a linear expansion of atoms. These atoms are selected from a redundant dictionary of functions in order to best match signal structures. The matching pursuit algorithm as the basis of OMP will be introduced in this subsection.

Let $\mathbf{L}^2(\mathbf{R})$ be a Hilbert space, and a complete dictionary is defined as a redundant family $\mathbf{D} = (g_{\gamma n})_{\gamma n \in \Gamma}$ of vectors in $\mathbf{L}^2(\mathbf{R})$, such that $\overline{span}\,(\mathbf{D}) = \mathbf{L}^2(\mathbf{R})$ and $||g_{\gamma n}|| = 1$. For a music signal $f(t) \in \mathbf{L}^2(\mathbf{R})$, a linear expansion of $f(t)$ over a family of vectors selected from $\mathbf{D}$ can be written in the following form:

$$f(t) = \sum_{n=0}^{k-1} \langle R_n f(t), g_{\gamma n}(t) \rangle g_{\gamma n}(t) + R_n f(t) \tag{1}$$

where k is the number of selected atoms. $\langle R_n f(t), g_{\gamma n}(t) \rangle$ is the inner product of $(R_n f(t), g_{\gamma n}(t))$. $R_k f(t)$ is the residual term after approximating $R_{k-1} f(t)$ in the direction of $g_{\gamma(k-1)}(t)$. The matching pursuit algorithm is an iterative procedure by projecting the residual term on a vector of $\mathbf{D}$ that matches the residual term almost at best. At each iteration, $g_{\gamma n}(t)$ is selected to best match the inner structures of $R_n f(t)$ and orthogonal to the residual term $R_{n+1} f(t)$, hence

$$||R_n f(t)||^2 = |\langle R_n f(t), g_{\gamma n}(t) \rangle|^2 + ||R_{n+1} f(t)||^2. \tag{2}$$

In order to minimize $||R_{n+1} f(t)||$, the atom $g_{\gamma n}(t)$ is chosen to maximize $|\langle R_n f(t), g_{\gamma n}(t) \rangle|$. In some cases, it is only possible to find a sub-optimal choice of $g_{\gamma n}(t)$, which is close to the maximum in the sense that

$$|\langle f(t), g_{\gamma n}(t) \rangle| \geq \alpha \sup_{\gamma \in \Gamma} |\langle f(t), g_{\gamma}(t) \rangle| \tag{3}$$

where α is an optimality factor and $0 < \alpha \leq 1$. An energy equation can be yielded from equation (2)

$$||f(t)||^2 = \sum_{n=0}^{k-1} |\langle R_n f(t), g_{\gamma n}(t) \rangle|^2 + ||R_k f(t)||^2. \tag{4}$$

It can be seen from equation (4) that a strong convergence $\lim_{k \to \infty} ||R_k f(t)||^2 = 0$ along with a good approximation of the music signal $f(t)$ will be obtained. For a complete dictionary, the matching pursuit algorithm converges to $f(t)$, which is presented as follows

$$f(t) = \sum_{n=0}^{\infty} \langle R_n f(t), g_{\gamma n}(t) \rangle g_{\gamma n}(t). \tag{5}$$

A large redundant and complete dictionary of Gabor time-frequency atoms, i.e. sine-modulated Gaussian functions, is proposed for the matching pursuit algorithm. The Gaussian functions are compact and have a low time-frequency bandwidth product, ensuring a higher spatial and temporal resolution. For any scale $s > 0$, frequency modulation ξ and translation μ, the index is denoted as $\gamma = (\mu, s, \xi)$, and phase is $\phi \in [0, 2\pi]$. The Gabor atom can be expressed as

$$g_{\gamma}(t) = K(\gamma) \exp \left\{ -\pi \left(\frac{t-\mu}{s} \right)^2 \right\} \cos \left\{ \xi (t-\mu) + \phi \right\} \tag{6}$$

where $K(\gamma)$ is the normalization factor that ensures $||g_{\gamma}(t)|| = 1$. It is indicated that the above equation describes a general family of time-frequency atoms that can be generated by scaling, translating and modulating a single Gabor function. Suppose the signal length is N, a dyadic dictionary $\mathbf{D}_{\alpha} = (g_{\gamma n})_{\gamma n \in \Gamma_{\alpha}}$ is created with the following values: $\gamma = \{a^j, pa^j \Delta\mu, ka^{-j} \Delta\xi\}$ with $a = 2$, $\Delta\mu = 0.5$, $\Delta\xi = 2$, $0 < j < log_2 N$, $0 \leq p < N 2^{-j+1}$ and $0 \leq k < 2^{j+1}$. The dyadic dictionary $\mathbf{D}_{\alpha}$ is a subset of the complex Gabor dictionary, and the number of atoms in $\mathbf{D}_{\alpha}$ is $O(N log_2 N)$. In addition, the phase ϕ is estimated from the coefficients, which is computed by taking the inner product of the residual signal with the Gabor atoms.

With the varying sclaes $\gamma = (\mu, s, \xi)$, the matching pursuit algorithm provides an adaptive adjustment of the spatial and temporal resolution to best match the input signals, in contrast to the other algorithms that use fixed resolution cells. In general, the matching pursuit algorithm is a procedure for computing adaptive signal representations through searching over the time-frequency dictionary $\mathbf{D}_{\alpha}$ to find an approximate parameter scale, time and frequency localization of the main signal structures. The complexity for each matching pursuit iteration is $O(N log_2 N)$, and each iteration requires approximately as much CPU time as a Fast Fourier transform.

4.2 Wigner Distribution Function

The music signal $f(t)$ has been decomposed into a series of atoms as shown in equation (5). From the decomposition of the music signal $f(t)$ within a time-frequency dictionary, the time-frequency energy distribution of $f(t)$ will be derived by adding Wigner distributions of selected atoms

$$Ef(t, \omega) = \sum_{n=0}^{\infty} |\langle R_n f(t), g_{\gamma n}(t) \rangle|^2 W g_{\gamma n}(t, \omega) \tag{7}$$

where $W g_{\gamma n}(t, \omega)$ is the Wigner distributions of selected atoms $g_{\gamma n}(t)$. The Wigner distribution function as a powerful tool for time-frequency analysis of signals was proposed by Wigner in 1932 [43].It can be expressed as the signal energy distribution both in time and frequency.For a given selected atom $g_{\gamma n}(t)$, the corresponding Wigner distribution can be evaluated from the time domain

$$W g_{\gamma n}(t, \omega) = \int_{-\infty}^{\infty} e^{-j\omega\tau} g_{\gamma n}(t + \tau/2) g^*_{\gamma n}(t - \tau/2) d\tau. \tag{8}$$

4.3 Orthogonal Matching Pursuit

The convergence of the matching pursuit algorithm in its originally proposed form relies essentially on the fact that

$|\langle R_n f(t), g_{\gamma n}(t)\rangle| = 0$, i.e. the residual error is only orthogonal to the last selected atom. In this sense, the result of the matching pursuit algorithm will in general be suboptimal, and the residual error may be quite large. To cope with this problem, a modification of the matching pursuit algorithm called the OMP algorithm is proposed. In OMP, the criterion for atom selection is the same as the matching pursuit approach, although such a selection is not optimal. The selected atoms can be orthogonalized using the Gram-Schmidt process at each iteration, and then the residual error will be always orthonormal to the span of the selected atoms

$$\begin{aligned} f(t) &= \sum_{n=0}^{k-1} \langle R_n f(t), g_{\gamma n}(t)\rangle g_{\gamma n}(t) + R_k f(t), \\ with &|\langle R_n f(t), g_{\gamma n}(t)\rangle| = 0, n = 0, 1, ..., k-1. \end{aligned} \tag{9}$$

With an orthonormal set of atoms, the OMP approach has an improved convergence rate, and an optimal approximation of signals will be obtained.

4.4 Adaptive Time-frequency Analysis of music signal

Musical sounds with a wider band of frequency are more complicated than human vocal sounds. Most of music signals consist in a short transient part followed by a stationary part that eventually slowly fades out, therefore time-frequency methods are the optimal choice for music signal analysis. An adaptive time-frequency analysis of music signal with the OMPGW method is proposed for a better feature extraction in the music mood classification system. Compared with other time-frequency analysis algorithms, such as STFT, Wavelet transform, CQT, GTFB and MSTM, the OMPGW method gives a multiscale adaptive time-frequency decomposition of signals with a redundant dictionary of Gabor functions. Moreover, the time and frequency components of a signal are determined with arbitrary precision and selected according to the input signal in the proposed method, thus the proposed method will best match the signal structures. The Gabor functions have a low time-frequency bandwidth product and the spectral and temporal resolution will be improved. After applying the OMPGW method, the outputs, including the atoms along with the corresponding decomposition coefficients and the time-frequency energy distribution of music signals, will be saved for the feature extraction. The OMPGW method is comprised of the following steps, and the pseudo-code of it is shown in Algorithm. 1:

1) Initialization. For a music signal $f(t)$, the residual error is initialized as $R_0 f(t) = f(t)$ and the set of selected atoms is $\mathbf{D}_c = \emptyset$. Let the iteration counter $k = 1$.
2) Atom selection. The inner product of $\{\langle R_k f(t), g_{\gamma k}(t)\rangle;\ g_{\gamma k}(t) \in \mathbf{D}_a\}$ is computed, and the atom $g_{ck}(t)$ is found for which the inner product is sufficiently large as shown in equation (3). The set of selected atoms is then reordered $\mathbf{D}_{ck} = \mathbf{D}_{c(k-1)} \cup \{g_{ck}(t)\}$, and the selected atom $g_{ck}(t)$ is taken as the column and in the k position, where the selected atom is the columns.
3) Orthogonalization. The orthogonal projection operator with the span of the columns of $\mathbf{D}_{ck}$ is defined as $\mathbf{P}_k = \mathbf{D}_{ck}(\mathbf{D}_{ck}^* \mathbf{D}_{ck})^{-1} \mathbf{D}_{ck}^*$. The orthogonal projection operator $\mathbf{P}_k$ is then applied to the residual error $R_k f(t) = (\mathbf{I} - \mathbf{P}_k) R_k f(t)$, where $\mathbf{I}$ is the identity matrix.
4) Update. The model is updated:
$f(t) = \sum_{n=0}^{k-1} \langle R_n f(t), g_{cn}(t)\rangle g_{cn}(t)$
$R f_k(t) = f(t) - f_k(t)$.
5) Convergence. The algorithm is considered to be converged, if the stop condition is achieved, for example, the residual error is less than a small constant. Otherwise, the iteration counter k is incremented by 1, and steps 1 to 5 are repeated.
6) Outputs. By adding the Wigner distribution of each selected atom, the time-frequency energy distribution of $f(t)$ will be obtained as in equation (6). The corresponding coefficient $a_k = \langle R_n f(t), g_{\gamma n}(t)\rangle$ of each iteration is saved.

Algorithm 1 Orthogonal Matching Pursuit Algorithm

Input: Music signal $f(t)$.
Output: The set of selected atoms $\mathbf{D}_c$, Atoms index set Λ, Decomposition coefficients set a_k.

Set iteration counter $k \leftarrow 1$, the residual error $R_0 f(t) \leftarrow f(t)$, the set of selected atoms $\mathbf{D}_c \leftarrow \emptyset$, index set $\Lambda \leftarrow \emptyset$, $A \leftarrow \emptyset$, where $\emptyset$ denotes empty set, and τ is a small constant.
while $||R_k f(t)|| > \tau$ **do**
    Find the best matching atom, $i.e.$ the maximum inner product between $R_k f(t)$ and $g_{\gamma(k)}(t)$ by exploiting $ck \leftarrow \arg\max_{\gamma k \notin \Lambda_{k-1}} |\langle R_k f(t), g_{\gamma k}(t)\rangle|$.
    Update the set of selected atoms $\mathbf{D}_{ck} \leftarrow \mathbf{D}_{c(k-1)} \cup \{g_{ck}(t)\}$ and the index set $\Lambda_k \leftarrow \Lambda_{k-1} \cup \{ck\}$.
    Compute the orthogonal projection operator by using the Gram-Schmidt process $\mathbf{P}_k \leftarrow \mathbf{D}_{ck}(\mathbf{D}_{ck}^* \mathbf{D}_{ck})^{-1} \mathbf{D}_{ck}^*$ and orthogonalize the residual error $R_k f(t) \leftarrow (\mathbf{I} - \mathbf{P}_k) R_k f(t)$.
    Update the representation model of signals $f(t) \leftarrow \sum_{n=0}^{k-1} \langle R_n f(t), g_{cn}(t)\rangle g_{cn}(t)$ and the residual error $R f_k(t) \leftarrow (f(t) - f_k)$, where $a_k \leftarrow \langle R_k f(t), g_{\gamma k}(t)\rangle$.
    $k \leftarrow k + 1$.
end while
return $\mathbf{D}_c,\ \Lambda,\ a_k$

5 Feature Extraction

The extraction of ATF features based on the OMPGW method will be introduced in this section. The ATF features are first extracted from each input music signal. A novel feature vector is then constructed from the extracted ATF features for further mood classification with the conventional processing techniques, such as cepstral analysis, modulation spectral analysis. Detailed feature extraction processes will be addressed in the following subsections.

5.1 ATF Feature Extraction

In the proposed music mood classification system, the original music signals are first down-sampled into a uniform format, with a sample rate of 16kHz and a resolution of 16-bit. Different from most of existing techniques, the

OMPGW method does not require any signal segmentation techniques before the feature extraction due to its adaptive time-frequency nature. Thus, each music signal does not need to be divided into frames, and the ringing effect can be avoided.

An input music signal is to be pre-emphasized [44], [45], whose aim is to increase the amplitude of the high-frequency components with respect to the magnitude of other frequencies. Pre-emphasis could improve the overall signal-to-noise ratio by minimizing the radiation effects of sounds. A first-order FIR high-pass filter can be used to realize the pre-emphasis with the transfer function as follows

$$H(z) = 1 - \alpha z^{-1}. \tag{10}$$

where α is the emphasized coefficient with a typical value of 0.95 [38]. After the pre-emphasis, the result of the input signal at the mth sample time is

$$s'(m) = s(m) - \alpha s(m-1). \tag{11}$$

where $s(m)$ is the input signal and $s'_0 = s_0$.

After the pre-emphasis, the OMPGW method is applied to analyze the preprocessed music signal. The decomposition coefficients $a_k (k = 1, ..., n)$ from the OMPGW analyzing results are sorted according to the iteration order, which are observed and contain significant information of music signals. Several features are then extracted from the decomposition coefficients a_k and the time-frequency energy distribution $Ef(t, \omega)$. Basic process of the ATF feature extraction is presented in Fig. 4, and details are shown as follows:

Spectral Centroid: The spectral centroid is the center of mass of the decomposition coefficients and has a robust connection with the impression of brightness, similarly to the Fourier spectrum. The definition of ATF spectral centroid is defined as

$$Scentroid = \frac{\sum_{k=1}^{n} a_k \cdot k}{\sum_{k=1}^{n} a_k}. \tag{12}$$

Spectral Rolloff: The spectral rolloff is defined as the decomposition coefficients below which of the magnitude distribution is concentrated. It is also a measure of spectral shape that defined by the decomposition coefficients, and the ATF spectral rolloff is defined as

$$Srolloff = 0.85 \times \sum_{k=1}^{n} a_k. \tag{13}$$

Spectral Bandwidth: The bandwidth of a signal is the difference between the upper and lower frequencies in a continuous set of frequencies. In this paper, the ATF bandwidth is defined as the number n of the decomposition coefficients a_k.

Spectral Contrast: The spectral contrast is the different between the spectral peak and the spectral valley, and the ATF spectral contrast is computed with the decomposition coefficients as:

$$Scontrast = \log\left(\frac{1}{\beta n}\sum_{k=1}^{\beta n} a_k\right) - \log\left(\frac{1}{\beta n}\sum_{k=1}^{\beta n} a_{n-k+1}\right). \tag{14}$$

where the largest β percentage spectra and the smallest β percentage spectra are used to computed the spectral peak and the spectral valley, respectively.

Spectral Flatness Measure: The spectral flatness is also called tonality coefficient and is used to characterize an audio spectrum. The ATF spectral flatness is calculated in the following form

$$Sflatness = \frac{\sqrt[n]{\prod_{k=1}^{n} a_k}}{\frac{\sum_{k=1}^{n} a_k}{n}}. \tag{15}$$

Spectral Crest Factor: The spectral crest factor is used to distinguish noise-like and tone-like sounds to their characteristic spectral shapes, and it is inversely proportional to the flatness. The spectral crest factor computed with decomposition coefficients of the proposed method is as follows

$$Scf = \frac{\max\{|a_k^2|\}}{\frac{1}{n}\sum_{k=1}^{n} |a_k|^2}. \tag{16}$$

Subband Power: Three subband power sections are calculated for the time-frequency energy distribution $Ef(t, \omega)$ in this paper. Suppose the half sampling frequency is ω_0, then the frequency band is divided into three intervals as $[0, \omega_0/8]$, $[\omega_0/8, \omega_0/4]$ and $[\omega_0/4, \omega_0/2]$, corresponding to the time-frequency energy distribution $Ef(t, \omega)$ of a given music signal. The subband power is defined as

$$Subpower_j = \sum_{k_j} z_j^2(k_j). \tag{17}$$

where $z_j(k_j)$ is the kth coefficient of the time-frequency energy distribution $Ef(t, \omega)$ for subband j.

Coefficient Histogram:The histogram technique is an efficient means of estimating the probability distribution of a variable. For the time-frequency energy distribution $Ef(t, \omega)$, the histogram of the coefficient $z_j(k_j)$ is constructed with each subband j. The histogram characteristic function $H_j(n_j)\,(n_j = 1, 2, \cdots, N_j)$ of each subband j is obtained through a discrete Fourier transform, and M-dimensional statistical characteristics of histogram characteristic function is extracted [46]

$$Chf_m = \frac{\sum_{n_j=1}^{N_j/2} n_j^m \left|H_j(n_j)\right|}{\sum_{n_j=1}^{N_j/2} \left|H_j(n_j)\right|}, m = 1, 2, \cdots, M \tag{18}$$

where N_j is the sampling number of $H_j(n_j)$.

Frequency Cepstrum Coefficient: The cepstrum is defined as the inverse Discrete Fourier Transform (DFT) of the log magnitude of the DFT of a signal. The cepstrum analysis for the time-frequency energy distribution $Ef(t, \omega)$ is

$$Fcep_1 = \sqrt{\frac{2}{n}}\sum_{k=1}^{n}\{\log_{10} z(k)\} \cdot \cos l(k-0.5)\pi/n \quad l = 1, ..., L \tag{19}$$

where $Fcep_1$ is the L-order coefficients of cepstrum.

5.2 Feature Vector Formation

To represent a music clip, the extracted ATF features are used to form a feature vector. There are in total $9+L+3\times M$ features derived from the results of the OMPGW method, as

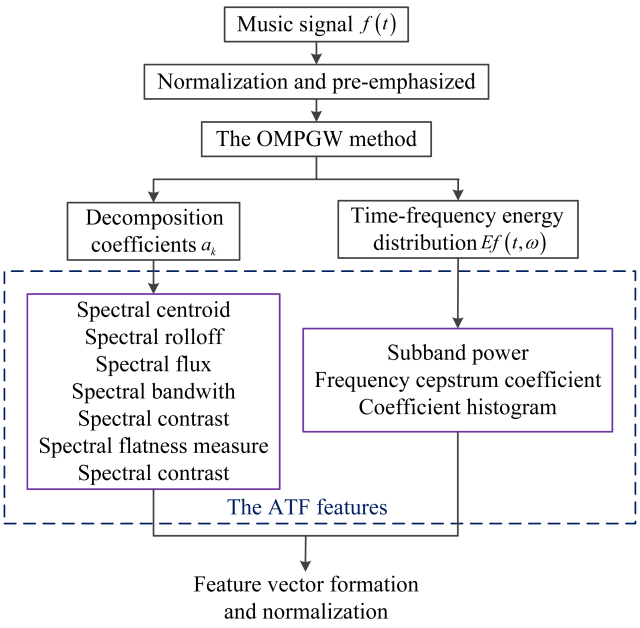


Fig. 4. Basic process of the ATF features extraction.

shown in Fig. 4. The ATF features contain 7 features extracted from the decomposition coefficients a_k and $2+L+3\times M$ features extracted from the time-frequency energy distribution $Ef(t,\omega)$. Furthermore, since the dispersion of each music signal is not identical, a normalization process will be performed on each feature before obtaining the feature vector.The normalization is processed with the Z-score standardization [47] as follows

$$f_i' = \frac{f_i - \mu_i}{\sigma_i} \tag{20}$$

where f_i denotes the ith feature component. μ_i and σ_i, computed from the ensemble of the training data, are the mean and standard deviation of f_i, respectively. The Karhunen-Loeve (K-L) transform will then be performed on the normalized features to remove the relativity among them. The K-L transform is to map the feature vector into an orthogonal space, and the corresponding covariance matrix is diagonal in the new space. This procedure makes the classifying procedure easier and leads to a better performance in music mood classification. For the classification phase, each ATF feature value is extracted from the current music signal and normalized with the reference mean and standard deviation values according to equation (19).

6 Music Mood Classification

After extracting the OMPGW-based features, a classifier will be employed in music mood classification systems. The purpose of music mood classification is to classify music clips into different emotional categories like contentment, depression, exuberance and anxious. In this paper, Support Vector Machines (SVMs) is utilized to model the proposed features regarding mood clusters. SVMs has been successfully used in audio classification with a high accuracy. Different from Linear Discriminant Analysis (LDA) that is suitable for a lower dimensional feature vector space, SVMs aims at improving the classification accuracy with a high-dimensional data and is flexibility in modeling diverse sources of data. In order to cope with the dimensional models of Russell, the bi-directional Long Short-Term Memory recurrent neural networks (BLSTM-RNNs) [48] is applied to predict the music emotion. Two BSLTM-RNNs with 6hidden layers are trained for arousal and valence regression separately. More details about SVMs and BLSTM-RNNs see [45], [49]–[52].

7 Experiments

In order to evaluate the performance of the proposed approach, experiments are conducted with five mood annotated datasets. In this section, the five datasets used for music mood classification experiments are first described, and then methodologies for constructing these experiments are presented. Finally, performance investigation of the proposed approach is demonstrated in terms of mood classification accuracy.

7.1 Datasets

To evaluate the performance of the proposed approach, five different datasets are employed. The first dataset consists of six discrete mood classes: happy, sad, fearful, angry, surprising, and tender, which is named Soundtracks dataset. These classes include 30 recordings of music clips lasting about 18 to 30 seconds, and emotions of each music clip are rated on a scale of 1 to 7. All music clips are played in a random order, which is the same for the whole group. The Soundtracks dataset is a public music database and has been used by many researchers.

The second dataset is from the MIREX-like mood dataset and re-annotated into the Thayers mood of model by our research group. The Thayers mood of model includes four basic mood classes as illustrated in Fig. 2, which are contentment, depression, exuberance, and anxious. This new dataset, called MIREX-T dataset, has a total of 903 music clips with a length of 30 seconds, and each music clip belongs to one mood classes. In order to remove the influence induced by recording conditions, all music clips are normalized with zero mean amplitude and unit variance.

The third is the MTV database that consists of 195 instances or14.2 h of music. The music in the MTV database is selected from top ten of 20 years (1981C2000) of MTV Europe Most Wanted and covers a wide variety of popular music genres. Five subjects, including four male and one female, aged from 22 to 33 with an average age of 25.5 have classified all songs into Thayers mood of model with four classes. The ground truth is obtained by averaging all the subjects votes and reduced to two classes: arousal and valence, respectively.

The fourth is the MediaEval 2015 database, which has a task on time-continuous estimation of emotion in music. This database consists of 431 music clips, and each clip has a length of 45 seconds. The dimensional model as illustrated in Fig. 2 will be employed to describe emotions in music on two orthogonal axes: valence and arousal. The annotators are asked to submit the valence and arousal scores for each music clip in a time-continuous fashion, and best quality annotations are selected. The ground truth is created by annotators.

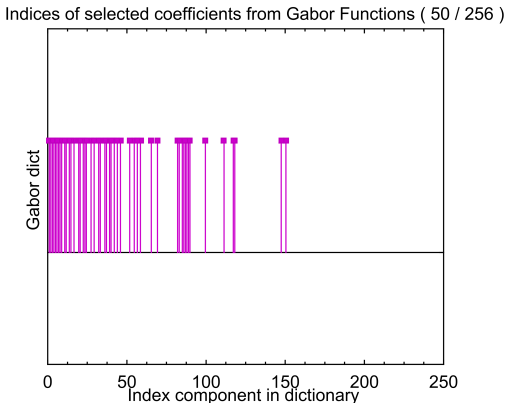
This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TAFFC.2017.2724515, IEEE Transactions on Affective Computing

7.2 Feature Vector Formation

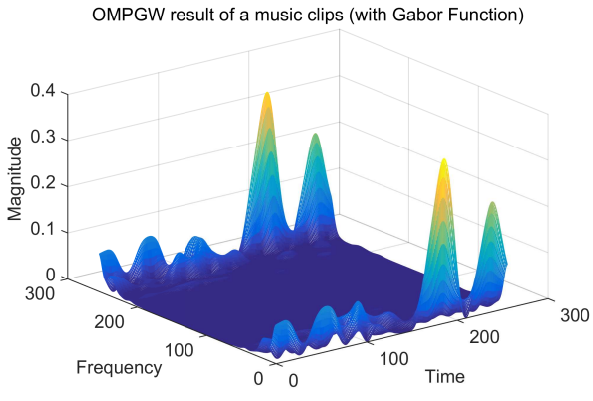
The classification performance of the proposed approach on the five datasets is evaluated, which is based on a 10-fold cross-validation and repeated ten times. Music clips of the five datasets are sorted in a random order in each mood class, respectively. Ninety percentage of music clips in each dataset are used for training, and the remaining ten percentage music clips are applied to test the classification accuracy. The overall classification accuracy is calculated by averaging all the 10-fold cross-validations.

7.3 Results

To visualize what kind of information conveyed by the proposed OMPGW method, several results are presented in Figs. 5-6. Fig. 5 shows the atoms selected from the dictionary of Gabor functions, where Fig. 5(a) presents the basis Gabor functions and Fig. 5(b) is the distribution of selected atoms of the proposed method after several iterations. Fig. 6 gives the direct information of the OMPGW method calculated from a music clip. The decomposition coefficients calculated from a music clip is shown in Fig. 6(a), and the corresponding time-frequency energy distribution is depicted in Fig. 6(b).



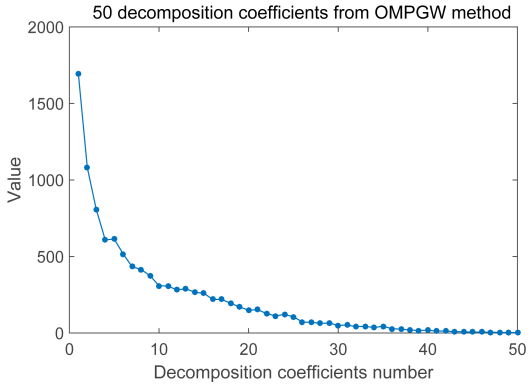
(a) Basis Gabor functions



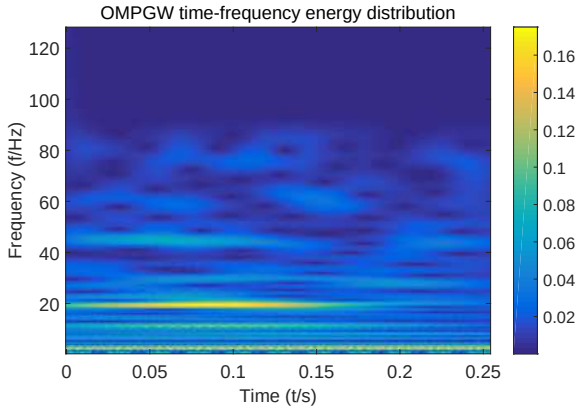
(b) Gabor dictionary

Fig. 5. Information of atoms used in the OMPGW method.

A comparison experiment is conducted to validate the performance of the proposed approach and other algorithms in presenting music signal. Fig. 7 presents a sample piece of music signal and its reconstructed version using Fourier



(a) Decomposition coefficients



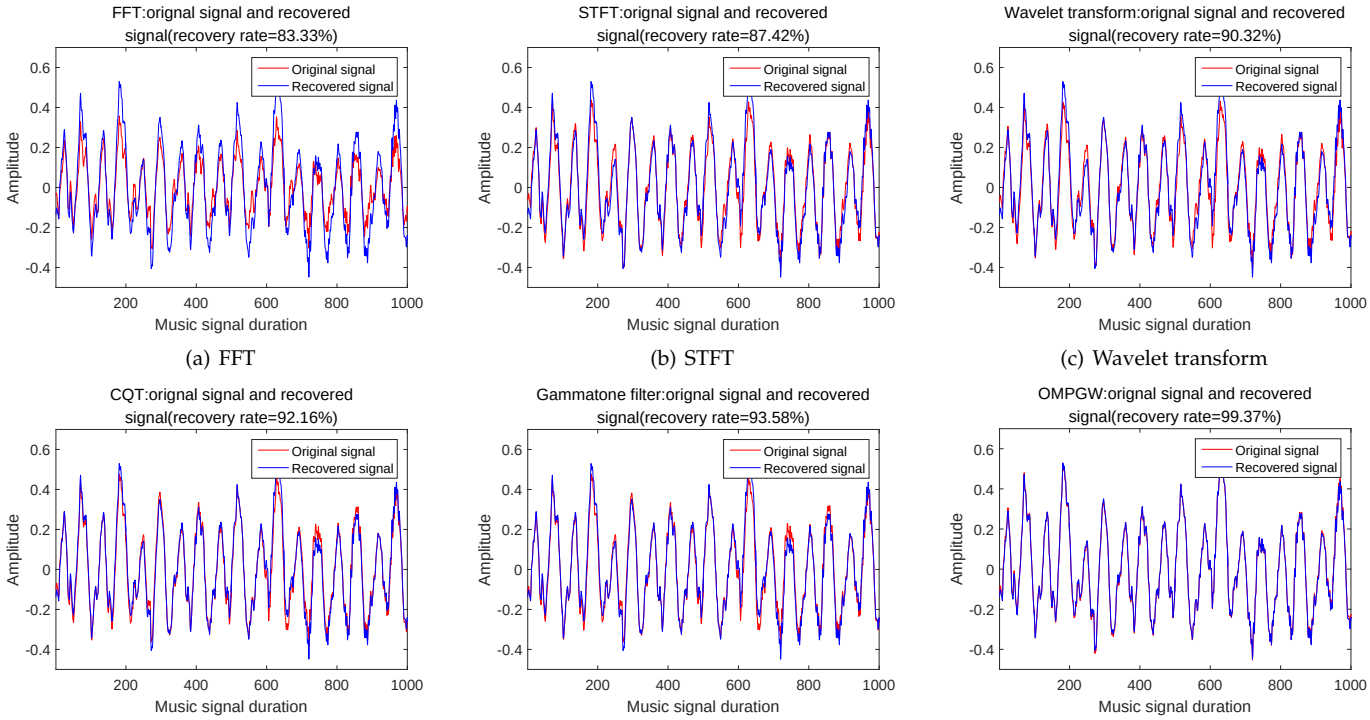
(b) Time-frequency energy distribution

Fig. 6. Direct information of the OMPGW method calculated from a music clip.

transform, STFT, Wavelet transform (here using Daubechies Wavelet [4] as the Wavelet function), CQT, Gammatone filter and the OMP algorithm. The feature extraction based on MSTM is essentially based on the constant-Q filters and the Wavelet transform, thus it will not be presented in the comparison in Fig. 7. For the aim of keeping experimental conditions to be consistent, the number of decomposition coefficients of the six algorithms is set to be uniform. It can be seen from Fig. 7 that the reconstructed signal of the OMP algorithm has the highest similarity with the original signal, while the reconstructed signal of the other comparison algorithms does not exactly suit the original signal. Moreover, the average fitting degree of these six algorithms between the original and the reconstructed music signal (95% confidence interval) is shown in Fig.8. Based on the Gabor functions, the OMPGW method can provide an adaptive time-frequency decomposition of the input signal with high spectral and temporal resolution, and it has the ability to perfectly describe signals. Thus, the average fitting degree of the proposed OMPGW method is the highest of all under the same experimental conditions as shown in Fig. 7 and Fig. 8. In general, the OMPGW method has an outstanding performance in signal decomposition and reconstruction, especially for transient signal processing.

In the following experiments, the proposed ATF features are compared with the ones extracted with other analysis algorithms in term of mood classification accuracy. The anal-

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TAFFC.2017.2724515, IEEE Transactions on Affective Computing



(a) FFT (b) STFT (c) Wavelet transform

(d) CQT (e) Gammatone filter (f) The OMPGW method

Fig. 7. A sample music signal and its reconstructed version with different analysis algorithms.

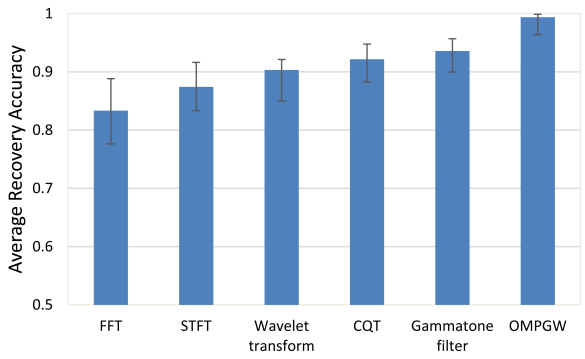


Fig. 8. Average recovery accuracy with different analysis algorithms (95% confidence interval).

ysis methods used for comparison include Fourier transform, STFT, Wavelet transform, CQT, GTFB and MSTM. The features based on these six comparison methods are presented in Table 1. Experimental results with the four datasets presented in subsection 7.1 have been shown in Tables 2-5. The first column in Tables 2-5 gives the comparison methods, and the second column lists the feature dimension. For different datasets, the number of music clips is different, and then the feature dimension of each comparison method is different.

Table 2 shows the music mood classification results of various features on the Soundtracks dataset. The results include the classification accuracy and standard deviation of the six emotions that are shown in Fig. 2. Table 3 shows comparison results with the proposed features as well as other six comparison methods on the MIREX-T dataset under the same experimental setup. The music clips are classified into four emotions that are from Thayers model of mood shown in Fig. 1, and classification accuracy of each mood is shown in Table 3. In addition, the mean value of classification accuracy and standard deviation of all emotions is presented in the last column of each Tables 2-3. Table 4 present the classification accuracy of valence and arousal on the MTV database. SVMs is applied to model the features regarding mood clusters for Tables 2-4. It can be seen form Tables 2-4 that the proposed features ensure a higher averaged classification accuracy and a lower standard deviation, which indicates the effectiveness of the proposed OMPGW method.

Table 5 shows the regression results on the MediaEval 2015 database, and the BSLTM-RNNs are used for regressing the valence and arousal values for music clips. In Table 5, the Root-Mean-Square Error (RMSE) is computed to estimate the performance of each comparison method for arousal and valence separately, and the Pearsons correlation (r) of the prediction and the ground truth is also presented. Table 5 shows that the proposed method has the lower RMSE, and the absolute value of r is higher. In general, the OMPGW method has the best overall performance with different emotion models and datasets.

8 Conclusion

In this paper, the OMPGW method, which is based on the orthogonal matching pursuit, Gabor functions, and the Wigner distribution function, was applied to music mood classification systems. In the music classification process,

TABLE 2
Comparison of mood classification Accuracy (%) and Standard Deviation (in parentheses) of various features on the Soundtracks dataset

Feature set	Feature dimension	Emotions (%)						Mean(%)
		Happy	Sad	Fearful	Angry	Surprising	Tender	
Fourier transform	21	36.81	38.20	36.56	37.15	33.78	34.80	36.21(1.63)
STFT	17	38.54	37.31	38.52	37.09	37.21	39.73	38.06(1.04)
Wavelet transform	26	43.93	43.95	42.44	38.65	42.53	44.38	42.62(2.11)
CQT	12	44.12	44.22	41.73	39.87	43.13	42.97	42.67(1.51)
GTFB	12	44.84	42.83	40.92	42.53	41.94	43.69	42.79(1.25)
MSTM	12	45.45	42.61	41.68	42.27	41.43	44.18	43.27(1.17)
ATF features	27	48.41	43.27	48.45	42.90	42.04	43.06	44.68(2.92)

TABLE 3
Comparison of mood classification Accuracy (%) and Standard Deviation (in parentheses) of various features on the MIREX-T dataset

Feature set	Feature dimension	Emotions (%)				Mean(%)
		Contentment	Depression	Exuberance	Anxious	
Fourier transform	43	49.28	62.04	56.30	62.02	57.41(5.24)
STFT	39	50.92	65.98	58.10	75.82	62.70(5.96)
Wavelet transform	48	69.46	74.59	55.50	68.16	66.92(7.02)
CQT	34	57.43	71.34	67.48	70.64	66.73(5.54)
GTFB	34	60.89	72.65	67.29	68.73	67.39(4.24)
MSTM	34	64.12	73.81	65.31	69.52	68.19(3.82)
ATF features	49	66.13	75.09	69.84	67.06	69.53(3.49)

TABLE 4
Comparison of mood classification Accuracy (%) and Standard Deviation (in parentheses) of various features on the MTV dataset

Feature set	Feature dimension	Emotions (%)		Mean(%)
		Arousal	Valence	
Fourier transform	29	65.45	56.42	60.93(4.51)
STFT	25	66.38	58.01	62.19(4.18)
Wavelet transform	34	67.24	59.23	63.23(4.01)
CQT	20	67.95	59.84	63.74(4.05)
GTFB	20	68.16	61.57	64.86(3.29)
MSTM	20	71.82	62.33	67.07(4.74)
ATF features	35	73.94	64.29	69.11(4.82)

TABLE 5
Regression results with different feature sets on the MediaEval 2015 database

Feature set	Feature dimension	Arousal		Valence	
		RMSE	r	RMSE	r
Fourier transform	38	0.289±0.13	0.395±0.31	0.369±0.18	0.013±0.36
STFT	34	0.279±0.12	0.463±0.29	0.352±0.17	0.057±0.52
Wavelet transform	43	0.278±0.12	0.484±0.30	0.346±0.17	0.068±0.46
CQT	29	0.271±0.13	0.527±0.27	0.333±0.16	0.094±0.42
GTFB	29	0.260±0.11	0.539±0.27	0.312±0.17	0.119±0.43
MSTM	29	0.259±0.11	0.581±0.26	0.305±0.15	0.126±0.41
ATF features	44	0.229±0.09	0.676±0.24	0.288±0.13	0.162±0.38

the proposed method was consisted of three-level schemes. Firstly, the OMP algorithm was used to adaptively decompose music signals into a linear expansion of Gabor functions. The time-frequency energy distribution of music signals was then obtained by adding Wigner distributions to the selected Gabor functions. Finally, audio features were extracted from the decomposition coefficients and the time-frequency energy distribution, and the proposed ATF features were applied to music mood classification using a classifier of SVMs. Compared with other analysis algorithms, the proposed approach can match the nonstationary nature of music signals and provide a higher spatial and temporal resolution. Therefore, the ATF features had the ability to improve the mood classification accuracy of music. Experiments conducted on five mood annotated dataset had clarified the efficacy of the proposed approach.

ACKNOWLEDGMENTS

The authors would like to thank...

REFERENCES

[1] 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.

[2] Y. H. Yang and H. H. Chen, "Machine recognition of music emotion: A review," *Acm Transactions on Intelligent Systems & Technology*, vol. 3, no. 3, pp. 338–343, 2012.

[3] G. Tzanetakis and P. Cook, "Musical genre classification of audio signals," *Speech & Audio Processing IEEE Transactions on*, vol. 10, no. 5, pp. 293–302, 2001.

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

[5] W. L. Cheung and G. Lu, "Music emotion annotation by machine learning," in *International Workshop on Multimedia Signal Processing, MMSP 2008, Shangri-la Hotel, Cairns, Queensland, Australia, October*, 2008, pp. 580–585.

[6] T. Li, M. Ogihara, and Q. Li, "A comparative study on content-based music genre classification," *Proc Sigir*, pp. 282–289, 2003.

[7] G. Tzanetakis, R. Jones, and K. Mcnally, "Stereo panning features for classifying recording production style." in *International Conference on Music Information Retrieval, Ismir 2007, Vienna, Austria, September*, 2007, pp. 441–444.

[8] A. Ramalingam and S. Krishnan, "Gaussian mixture modeling using short time fourier transform features for audio fingerprinting," in *IEEE International Conference on Multimedia and Expo, ICME 2005, July 6-9, 2005, Amsterdam, the Netherlands*, 2005, pp. 1146–1149.

[9] Y. C. Pati, R. Rezaiifar, and P. S. Krishnaprasad, "Orthogonal matching pursuit: Recursive function approximation with applications to Wavelet decomposition," in *Conference on Signals*, 1995, pp. 1–3.

[10] L. Rebolloneira and D. Lowe, "Optimized orthogonal matching pursuit approach," *IEEE Signal Processing Letters*, vol. 9, no. 4, pp. 137–140, 2002.

[11] T. T. Cai and L. Wang, "Orthogonal matching pursuit for sparse signal recovery with noise," *IEEE Transactions on Information Theory*, vol. 57, no. 7, pp. 4680–4688, 2011.

[12] Z. Cataltepe and B. Altinel, "Hybrid music recommendation based on different dimensions of audio content and an entropy measure," in *Signal Processing Conference, 2007 European*, 2007.

[13] F. Wang, X. Wang, B. Shao, T. Li, and M. Ogihara, "Tag integrated multi-label music style classification with hypergraph." in *International Society for Music Information Retrieval Conference, Ismir 2009, Kobe International Conference Center, Kobe, Japan, October*, 2009, pp. 363–368.

[14] J. C. Brown, "Calculation of a constant q spectral transform," *Journal of the Acoustical Society of America*, vol. 89, no. 1, pp. 425–434, 1991.

[15] Y. Panagakis and C. Kotropoulos, "Music classification by low-rank semantic mappings," *EURASIP Journal on Audio, Speech, and Music Processing*, vol. 2013, no. 1, p. 13, 2013.

[16] D. A. Cooper, "Speech detection using gammatone features and one-class support vector machine," Master's thesis, University of Central Florida, 2009.

[17] J. Qi, D. Wang, Y. Jiang, and R. Liu, "Auditory features based on gammatone filters for robust speech recognition," in *IEEE International Symposium on Circuits and Systems*, 2013, pp. 305–308.

[18] N. Mesgarani, M. Slaney, and S. A. Shamma, "Discrimination of speech from nonspeech based on multiscale spectro-temporal modulations," *IEEE Transactions on Audio Speech & Language Processing*, vol. 14, no. 3, pp. 920–930, 2006.

[19] T. Eerola and J. K. Vuoskoski, "A comparison of the discrete and dimensional models of emotion in music," *Psychology of Music*, vol. 39, no. 1, pp. 18–49, 2011.

[20] J. M. Ren, M. J. Wu, and J. S. R. Jang, "Automatic music mood classification based on timbre and modulation features," *IEEE Transactions on Affective Computing*, vol. 6, no. 3, pp. 1–1, 2015.

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

[22] B. Schuller, C. Hage, D. Schuller, and G. Rigoll, "'mister d.j., cheer me up!': Musical and textual features for automatic mood classification," *Journal of New Music Research*, vol. 39, no. 1, pp. 13–34, 2010, taylor & Francis.

[23] *Emotion in music task at MediaEval 2015*. Wurzen, Germany: In Working Notes Proceedings of the MediaEval 2015 Workshop, September 14-15 2015.

[24] see the website http://www.musicir.org/mirex/wiki/MIREX_home, "Music information retrieval evaluation exchange."

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

[26] ——, "Experimental studies of the elements of expression in music," *American Journal of Psychology*, vol. 48, no. 1, pp. 246–268, 1936.

[27] R. E. Thayer, "The biopsychology of mood and arousal." *Cognitive & Behavioral Neurology*, no. 1, p. 65, 1992.

[28] Z. Xiao, D. Wu, X. Zhang, and Z. Tao, "Music mood tracking based on HCS," in *International Conference on Signal Processing*, 2012, pp. 1171–1175.

[29] L. Lu, D. Liu, and H. J. Zhang, "Automatic mood detection and tracking of music audio signals." *IEEE Transactions on Audio Speech & Language Processing*, vol. 14, no. 1, pp. 5–18, 2006.

[30] F. Morchen, A. Ultsch, M. Thies, and I. Lohken, "Modeling timbre distance with temporal statistics from polyphonic music," *IEEE Transactions on Audio Speech & Language Processing*, vol. 14, no. 1, pp. 81–90, 2006.

[31] J. Bergstra, N. Casagrande, D. Erhan, D. Eck, and B. Kegl, "Aggregate features and AdaBoost for music classification," *Machine Learning*, vol. 65, no. 2, pp. 473–484, 2006.

[32] E. Allamanche and E. Allamanche, "Content-based identification of audio material using MPEG-7 low level description." in *Ismir 2001, International Symposium on Music Information Retrieval, Indiana University, Bloomington, Indiana, Usa, October 15-17, 2001, Proceedings*, 2001.

[33] E. Benetos, M. Kotti, and C. Kotropoulos, "Musical instrument classification using non-negative matrix factorization algorithms and subset feature selection," in *Circuits and Systems, 2006. ISCAS 2006. Proceedings. 2006 IEEE International Symposium on*, 2006, pp. V–V.

[34] C. Hua, "Automatic chord recognition for music classification and retrieval," in *IEEE International Conference on Multimedia and Expo*, 2008, p. 2008.

[35] 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 & Language Processing*, vol. 16, no. 2, pp. 448–457, 2008.

[36] Y. E. Kim and B. Whitman, "Singer identification in popular music recordings using voice coding features," 2002, pp. 164–169.

[37] M. Mandel and E. Dan, "Song-level features and SVMs for music classification," 2007.

[38] 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.

[39] S. G. Mallat and Z. Zhang, "Matching pursuits with time-frequency dictionaries," *IEEE Transactions on Signal Processing*, vol. 41, no. 12, pp. 3397 – 3415, 1994.

[40] R. Gribonval and E. Bacry, "Harmonic decomposition of audio signals with matching pursuit," *IEEE Transactions on Signal Processing*, vol. 51, no. 1, pp. 101–111, 2003.

[41] S. Ray, C. C. Jouny, N. E. Crone, D. Boatman, N. V. Thakor, and P. J. Franaszczuk, "Human ECoG analysis during speech perception using matching pursuit: a comparison between stochastic and dyadic dictionaries." *IEEE Transactions on Biomedical Engineering*, vol. 50, no. 12, pp. 1371–3, 2003.

[42] F. Neves, A. Shmilovici, and J. Aguilar-Martin, "Identification of distillation column operating conditions with orthogonal matching pursuit," in *IEEE International Conference on Fuzzy Systems*, 1997, pp. 1251 – 1257.

[43] E. P. Wigner, *On the Quantum Correction for Themodynamic Equilibrium*. Springer Berlin Heidelberg, 1932.

[44] D. N. Jiang, L. Lu, H. J. Zhang, and J. H. Tao, "Music type classification by spectral contrast feature," in *IEEE International Conference on Multimedia and Expo, 2002. ICME '02. Proceedings*, 2002, pp. 113–116 vol.1.

[45] J. C. Kim and M. A. Clements, "Multimodal affect classification at various temporal lengths," *IEEE Transactions on Affective Computing*, vol. 6, no. 4, pp. 371–384, 2015.

[46] G. Casella and R. L. Berger, *Statistical inference*. Duxbury Press, 2nd edition, 2002.

[47] S. Raschka, "About feature scaling and normalization and the effect of standardization for machine learning algorithms," *Polar Political & Legal Anthropology Review*, vol. 30, no. 1, pp. 67–89, 2014.

[48] A. Graves and J. R. Schmidhuber, "2005 special issue: Framewise phoneme classification with bidirectional lstm and other neural network architectures," *Neural Networks*, vol. 18, no. 5-6, pp. 602–610, 2005.

[49] C. Cortes and V. Vapnik, "Support-vector networks." *Machine Learning*, vol. 20, no. 3, pp. 273–297, 1995.
[50] A. Benhur, "Support vector clustering," *Journal of Machine Learning Research*, vol. 2, no. 2, pp. 125–137, 2002.
[51] L. Sun, S. Kang, K. Li, and H. Meng, "Voice conversion using deep bidirectional long short-term memory based recurrent neural networks," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2015, pp. 4869–4873.
[52] M. Wollmer, Z. Zhang, F. Weninger, and B. Schuller, "Feature enhancement by bidirectional lstm networks for conversational speech recognition in highly non-stationary noise," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2013, pp. 6822–6826.

Jianwei Niu received the PhD degree in computer science from Beijing University of Aeronautics and Astronautics (BUAA, now Beihang University), China, in 2002. He was a visiting scholar at School of Computer Science, Carnegie Mellon University, USA, from Jan. 2010 to Feb. 2011. He is a professor in the School of Computer Science and Engineering, BUAA. He has published more than 100 referred papers, and filed more than 30 patents in mobile and pervasive computing. He served as the Program Chair of IEEE SEC 2008, Executive Co-chair of TPC of CPSCom 2013, TPC members of InfoCom,Percom, ICC, WCNC, Globecom, LCN, and etc. He has served as associate editor of *International Journal of Ad Hoc and Ubiquitous Computing*, associate editor of *Journal of Internet Technology*, editor of *Journal of Network and Computer Applications* (Elsevier). He received the New Century Excellent Researcher Award from Ministry of Education of China 2009, the first prize of technical invention of the Ministry of Education of China 2012, Innovation Award from Nokia Research Center, and won the best paper award in IEEE ICC 2013, WCNC 2013, ICACT 2013, CWSN 2012 and GreenCom 2010. His current research interests include affective computing, mobile and pervasive computing, mobile video analysis. He is a senior member of the IEEE.

Shasha Mo received the B.S. degree in School of Electronic and Information Engineering from Beijing Jiaotong University, Beijing, in 2010 and the Ph.D. degree from the Institute of Electronics, Chinese Academy of Sciences, Beijing, in 2015. She is currently a postdoctoral researcher with the School of Computer Science and Engineering, Beihang University. Her current research interests include affective computing and music mood analysis.