Automatic Mood Detection and Tracking of Music Audio Signals

Lie Lu, *Member, IEEE*, Dan Liu, and Hong-Jiang Zhang, *Fellow, IEEE*

***Abstract*—Music mood describes the inherent emotional expression of a music clip. It is helpful in music understanding, music retrieval, and some other music-related applications. In this paper, a hierarchical framework is presented to automate the task of mood detection from acoustic music data, by following some music psychological theories in western cultures. The hierarchical framework has the advantage of emphasizing the most suitable features in different detection tasks. Three feature sets, including intensity, timbre, and rhythm are extracted to represent the characteristics of a music clip. The intensity feature set is represented by the energy in each subband, the timbre feature set is composed of the spectral shape features and spectral contrast features, and the rhythm feature set indicates three aspects that are closely related with an individual's mood response, including rhythm strength, rhythm regularity, and tempo. Furthermore, since mood is usually changeable in an entire piece of classical music, the approach to mood detection is extended to mood tracking for a music piece, by dividing the music into several independent segments, each of which contains a homogeneous emotional expression. Preliminary evaluations indicate that the proposed algorithms produce satisfactory results. On our testing database composed of 800 representative music clips, the average accuracy of mood detection achieves up to 86.3%. We can also on average recall 84.1% of the mood boundaries from nine testing music pieces.**

***Index Terms*—Affective computing, hierarchical framework, mood detection, mood tracking, music emotion, music information retrieval, music mood.**

I. Introduction

MOST PEOPLE enjoy music in their leisure time. At present there is more and more music on personal computers, in music libraries, and on the Internet. In order to facilitate music organization, music management, and other related music applications, such as music search and music play-list generation, various metadata need to be created for each music piece. Although traditional information such as the artist, the album, or the title of a musical work remains important, these tags have limited applicability in many music-related applications. For example, when an individual comes back home from work, he may want to listen to some relaxing light music; while when he is at gymnasium, he may want to choose some exciting music with a strong beat and fast tempo. Traditional metadata is not helpful in these scenarios. Therefore, more semantic information is expected to be extracted to archive and search music, such as beat, tempo, genre, and mood [1].

Beat and tempo detection and genre classification have been developed in a few research works, using different features and different models. For example, Goto *et al.* [2] implemented a real-time beat tracking system based on a multiple-agent architecture, and Scheirer [3] proposed a beat tracking system using a bank of band-pass filters and parallel comb filters. In the genre identification field, Tzanetakis *et al.* [4] presented a music genre classification system using the features of timbre texture, rhythmic content, and pitch content, while Jiang *et al.* [5] implemented a similar algorithm to classify music into classical, pop, jazz, and rock by introducing some new features, i.e., octave-based spectral contrast. A number of related works on content-based music analysis are reported in the proceedings of ISMIR [6], a conference dedicated to music information retrieval.

However, compared with many works on the above research topics, few works have focused on automatic music mood detection. In this paper, we will try to build a computational framework to automatically estimate the mood inferred in each music clip. It is noted that, in most psychology textbooks, "emotion" usually means a short but strong experience while "mood" is a longer but less strong experiences. Therefore, we mainly choose to use the word "mood" in this paper. However, the words "affect," "emotion" and "emotional expression" are still used in order to keep the same usage as those used in the references.

A. Related Works on Music Mood and Emotion

Emotions usually play a critical role in rational decision-making, perception, human interaction, and human intelligence [7]. Traditional research on mood and emotion has a long history toward building intelligent machines which can perceive and express emotions, such as affective computing [7]. These research works often focus on discovering the physiological or psychological signals which are related to emotional expression, and then detecting or synthesizing emotions based on these signals. For example, Picard *et al.* [8] proposed an approach to affective detection using various affective psychological signals, including electromyogram, blood-volume pressure, heart rate, skin conductivity, and respiration.

Most of the research works on music mood and emotion are related to music composition and music expressivity [15]. These works are deployed to discover different aspects of music composition and music performance which can communicate emotions and influence listeners' emotional responses. For example,

Manuscript received January 24, 2005; revised September 20, 2005. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Dan Ellis.

L. Lu and H.-J. Zhang are with Microsoft Research Asia, Beijing 100080, China (e-mail: llu@microsoft.com; hjzhang@microsoft.com).

D. Liu was with Microsoft Research Asia, Beijing 100080, China. She is now with the Department of Cognitive Science, University of California at San Diego, La Jolla, CA 92093-0515 USA (e-mail: dliu@cogsci.ucsd.edu).

Digital Object Identifier 10.1109/TSA.2005.860344

in [9] and [10], Repp presented a quantitative analysis of timing and dynamics in the musical expression of Chopin's Etude in E major. Gabrelsson *et al.* [11] presented an overview of the relations between musical structures and different emotional expressions. Juslin [12] studied the utilization of the acoustic cues in communication of music emotions by performers and listeners and measured the correlation between various emotional expressions (including *anger*, *sadness*, *happiness*, and *fear*) and various acoustic cues (including *tempo*, *sound level*, *spectrum*, and *articulation*). With the findings, Julsin *et al.* further built a computational model of emotional expression in music performance [13], and synthesized five emotional expressions, including *happiness*, *sadness*, *anger*, *fear*, and *tenderness* [14]. Bresin *et al.* [16] also presented a work on emotion synthesis by using Director Musices program.

However, to our knowledge, only a few papers addressed automatic music mood detection from a piece of music, based on the cues used by performers and listeners to communicate emotion. Liu *et al.* [17] has presented a mood recognition system, where a fuzzy classifier was adopted to classify the mood of Johann Strauss's waltz centos into five clusters. In this system, tempo, loudness, pitch change, note density, and timbre were extracted from musical instrument digital interface (MIDI) files and used as the primitives to recognize the mood of music. Katayose *et al.* [18] also proposed a sentiment extraction system for pop music, where monophonic acoustic data was first transcribed into music codes, and then primitives of music such as melody, rhythm, harmony, and form were extracted from these music codes. Other relevant works include Friberg *et al.* [19], in which a linear regression model was used for musical expression extraction based on several low-level cues such as sound level, tempo, articulation, attack velocity, and spectral content; and Mion [20], in which a Bayesian network was employed to automatically recognize the expressive content of piano improvizations. These works have led to some preliminary results. However, most of them concentrated on MIDI or symbolic representations due to the difficulty of extracting useful features from acoustic musical data. A recent work of Lemon *et al.* [21] presented some low-level acoustic features such as loudness, centroid, and interonset interval, and studied their correlation with three factors of emotional expressions (valence, activity, and interest). However, the paper did not present an automatic mood detection algorithm.

B. Issues in Mood Detection From Music Audio Signals

In order to build a good computational model for automatic mood detection from acoustic music data, there are several issues that need to be considered beforehand.

1) One common objection to music mood detection is that the emotional expression and perception of music is subjective and it depends on many factors including culture, education, and personal experience. Thus, for the same music piece, different musicians may have different performances, while different individuals might have different perceptions. Therefore, it is usually argued that the mood is too subjective to be detected. However, as much research [11], [22] has shown, musical sounds, with certain patterns or structures, usually have inherent emotional expression. Moreover, Juslin [12] indicated that these emotions are able to be communicated. It was also found that, within a given cultural context, there are major agreements among individuals regarding the mood elicited by music [23]. Therefore, it is possible to build a mood detection system in a certain context, for example, the classical music in western culture. In this paper, our research will also focus on classical music.
2) Another issue relates to mood taxonomy. There is debate over whether emotions are categories or continua. It is also not clear what are the basic emotions that music can express and humans can perceive. Researchers usually use adjectives to describe moods. However, the adjectives are used quite freely and the adjective list is usually immense. Fortunately, several fundamental research works on the basic emotion dimensions [24], [25] provide the basis for music mood taxonomy, as well as giving some important cues for computational modeling of music moods. In our approach, Thayer's model of mood [25] is adopted as the basis of mood taxonomy and mood detection.
3) The third issue is over the acoustic features. Except for some features extracted from MIDI or symbolic representations, there are few acoustic features available to represent various moods. However, most music clips in the real world are in the form of recorded acoustic waveforms and there is no available transcription system that can translate them well into symbolic representations. Therefore, it is necessary to deal with the acoustic data directly. Although there have been many works in development of music and audio features, most of them are not suitable to exactly or directly represent the emotional content of a music signal. Some features developed for music analysis in current literatures, such as mel-frequency cepstral coefficients (MFCC), short-time energy (STE), and zero-crossing rate (ZCR) [26], are originally proposed for speech and audio analysis. Some features used for speech emotion detection, such as pitch or F_0 [27], are not feasible for music mood representation. Some music-specific features, such as the timbre texture and rhythm content proposed by Tzanetakis *et al.* [4], are used for a different goal of music genre classification. Moreover, the used timbre features therein are all based on standard features proposed for music-speech discrimination and speech recognition, and the rhythm features are not designed to specifically represent the primitives of different moods. Therefore, to build a good computational model of music mood detection, we should extract acoustic features to exactly represent the primitives of various moods. In this paper, three acoustic feature sets including intensity, timbre, and rhythm are extracted, based on the basic emotional dimensions and mood taxonomy.
4) The fourth consideration is related to mood detection. In order to get better performance for mood detection, in this paper, we present a hierarchical framework, utilizing the most suitable features in different steps of mood classification. Furthermore, since the mood is usually changeable in an entire piece of classical music [28], we extend

the algorithm to mood tracking by dividing the music into several independent segments, each of which contains a constant mood.

With the above considerations and solutions, we form our approach to mood detection and mood tracking. An earlier version of this work was published in [29]. The current paper has a full discussion of the approach from that paper. The rest of this paper is structured as follows. Section II presents the music mood taxonomy, and Section III describes three feature sets we used to represent emotion primitives, including intensity, timbre, and rhythm. The detailed mood detection process, with a hierarchical computation model, is presented in Section IV. In Section V, an approach to automatic music mood boundary detection is presented for mood tracking in an entire piece of music. Section VI deals with the empirical experiments and performance evaluations of the proposed algorithms, and Section VII is with the conclusions and future directions.

II. Music Mood Taxonomy

As mentioned in the previous section, one issue with mood detection is the mood taxonomy. In music psychology, traditional approaches usually use adjectives to describe mood responses, such as *Pathetic*, *Hopeful*, and *Gloomy*. However, the set of adjectives relating to emotions is immense [7]; and the adjectives vary quite freely in different theories, research, and applications. Currently, there is not a standard mood taxonomy system accepted by all. For example, Katayose *et al.* [18] uses some adjectives including *Gloomy*, *Urbane*, *Pathetic*, and *Serious*; in All Music[1] , mood is clustered into more classes, such as *Angry*, *Bitter*, *Cheerful*, *Dreamy*, *Gentle*, *Hungry*, *Messy*, *Plaintive*, *Unsetting*, and so on. Hevner's adjective checklist [30] has served as a basis for some subsequent research on mood response to music. This checklist is composed of 67 adjectives from eight clusters, including *Sober*, *Gloomy*, *Longing*, *Lyrical*, *Sprightly*, *Joyous*, *Restless*, and *Robust*, each of which is selected as a representative adjective of each cluster.

Instead of describing emotion as a set of dimensions as above, there is other evidence that these emotion dimensions are interrelated in a highly systematic fashion [24]. Accordingly, Russell [24] proposed a circumplex model of affect based on two bipolar dimensions instead of several mono-polar states. The two dimensions are called pleasant-unpleasant and arousal-sleep. Thus, each affect word could be defined as some combination of the pleasure and arousal components. Later, Thayer [25] adapted Russell's model to music using a two-dimensional energy-stress mood model, as shown in Fig. 1. Here, the energy dimension corresponds to the arousal, while stress corresponds to pleasure in Russell's model. Actually, the precise name of each dimension varies much in different literatures. For example, in Juslin's work [12], the dimensions are valence and activity level, while in affective computing [7], they are valence and arousal. However, the basic meaning of each dimension is very similar. The dimension of arousal, energy, or activity level represents from quiet to energetic (or in a reverse direction), while the dimension of stress, pleasure or valence represents from unpleasant to pleasant, or from negative to positive, or in reverse.

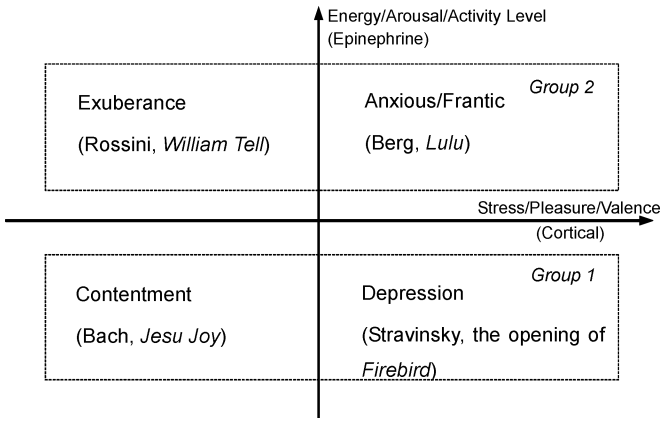


Fig. 1. Illustration of Thayer's two-dimensional model of mood, indicating the two underlying stimuli that influence mood responses: stress and energy; where left to right represents pleasant to unpleasant, and from bottom to top shows quiet to energetic. The four quadrants represent happy–energetic (*Exuberance*), happy–quiet (*Contentment*), tense–energetic (*Anxious/Frantic*), and tense–quiet (*Depression*). It is noted that the precise name of each dimension and each cluster varies greatly in different literatures. For example, the dimension of energy is also called as arousal or activity level, while the dimension of stress is also called pleasure or valence.

Unlike Hevner's checklist that uses individual adjectives which collectively form a mood pattern, this dimensional approach indicates two underlying stimuli that influence mood responses: stress/pleasure/valence and energy/arousal/activity level. It is of great importance for computational modeling of each mood.

Based on the level of stress and energy, Thayer's model divides music mood into four clusters (each in one quadrant): *Contentment*, *Depression*, *Exuberance* and *Anxious/Frantic*, as illustrated in Fig. 1, where the horizontal dimension (from left to right) represents from pleasant to unpleasant, while the vertical dimension (from bottom to top) shows from quiet to energetic. In these four clusters, *Contentment* refers to happy and quiet music, such as Bach's "Jesus, Joy of Man's Desiring," *Depression* refers to tired and tense music, such as the opening of Stravinsky's "Firebird," *Exuberance* refers to happy and energetic music such as Rossini's "William Tell Overture," and *Anxious/Frantic* refers to tense and energetic music, such as Berg's "Lulu." These four clusters almost cover the basic mood response to music; and they are usually in the most highly rated emotions as discovered in [31], [32], with equivalent adjectives (or nouns) such as *Happiness*, *Sadness*, *Desire*, and *Unrest* in Kreutz [31] and *Joy*, *Sadness*, *Anxiety*, *Calm*, and *Tension* in Lindstrom [32]. Moreover, these four clusters are explicit and discriminable, and the two-dimension structure gives important cues for computational modeling. Therefore, this emotion model (with corresponding mood taxonomy) is applied in our mood detection system. It is noted that there are many equivalent adjectives (or nouns) to describe these four clusters, such as *Tenderness*, *Sadness*, *Happiness*, and *Fear/Anger* in Juslin *et al.* [13], and the aforementioned adjectives in [31] and [32]. These adjectives may seem more reasonable to describe music mood. However, in this paper, we still follow the adjectives in Thayer's two-dimensional model in order to remain consistent.

[1]All Music. http://www.allmuisc.com.

A further model of emotion has three dimensions, including arousal, valence, and control/dominate. The third "control" dimension addresses the internal or external sources of the emotion [7]. However, the effect of control is quite small, while the effects of valence and arousal account for most of the independent variance in emotional responses [33], [34]. Therefore, the control dimension is neglected in this paper, as it was in [35].

III. Feature Extraction

It was indicated that mode, intensity, timbre, and rhythm are of great significance in arousing different music moods in many previous works [11], [22], [23], [30]. Juslin [12] also found that tempo, sound level, spectrum, and articulation are highly related to various emotional expressions. These findings are very similar although the exact words are different, such as rhythm versus tempo, and intensity versus sound level. Different emotional expressions are usually associated with different patterns of acoustic cues. For example, *Contentment* usually associates with slow tempo, low sound level, and soft timbre, while *Exuberance* is with fast tempo, fairly high sound level, and bright timbre. However, of the factors given above that influence emotional expressions, mode is very difficult to obtain from acoustic data (although some preliminary research is in [36]), and articulation is also extremely difficult to measure as pointed out in [9]. Therefore, in our approach, only the features of intensity (sound level), timbre (spectrum), and rhythm (tempo) are extracted and used in the proposed mood detection system. Compared with the two dimensions in Thayer's model of mood, intensity is correlated to "energy" or "arousal," while both timbre and rhythm are corresponding to "stress" or "valence" [12], [21].

In feature extraction, each music clip is first down-sampled into a uniform format: 16 kHz, 16 bits, mono channel, and divided into nonoverlapping frames of 32-ms length. In each frame, an octave-scale filter-bank is used to divide the spectral domain into several subbands in order to get more details of spectrum, as

$$\left[0, \frac{\omega_0}{2^n}\right), \left[\frac{\omega_0}{2^n}, \frac{\omega_0}{2^{n-1}}\right), \ldots \left[\frac{\omega_0}{2^2}, \frac{\omega_0}{2^1}\right] \tag{1}$$

where ω_0 refers to the sampling rate and n is the number of subband filters. In the experiments, seven subbands are segmented from the fast Fourier transform (FFT) frequency domain directly for a simple implementation. Then, the intensity features and timbre features are extracted from each frame. Their means and standard deviations are calculated across the music clip and then constitute the feature sets of timbre and intensity, respectively. Meanwhile, the rhythm features are extracted at the level of the whole music clip, instead of each frame, since rhythm is a long-term pattern. In order to remove the correlation among these raw features, the Karhunen–Loeve (K–L) transform [37] is performed on each feature set over the entire training corpus. The K–L transform maps each of three feature vectors into an orthogonal space, and the corresponding covariance matrix becomes diagonal in the new feature space. Details of the feature extraction process are addressed in the following subsections.

A. Intensity Features

Intensity is an essential feature in mood detection. For example, the intensity of *Contentment* and *Depression* is usually little, while that of *Exuberance* and *Anxious/Frantic* is usually large, based on the Thayer's mood model. It is also consistent with the acoustic cues of various emotional expressions in [12]. Huron [38] pointed out that, in the two factors in Thayer's model, energy (or intensity) is more computationally tractable and can be estimated using simple amplitude-based measures. Following this argument, in this system, the intensity feature of each frame is composed of the spectrum sum of the signal and the spectrum distribution in each subband, which are defined by

$$I(n) = \sum_{k=0}^{\frac{\omega_0}{2}} A(n,k) \tag{2}$$

$$D_i(n) = \frac{1}{I(n)} \sum_{k=L_i}^{H_i} A(n,k) \tag{3}$$

where $I(n)$ is the intensity of the nth frame and $D_i(n)$ is the intensity ratio of the ith subband, L_i and H_i are the lower and upper bounds of the ith subband, respectively, and $A(n,k)$ is the absolute value of the kth FFT coefficients of the nth frame.

B. Timbre Features

A number of previous research works [4], [26] utilized MFCC and so-called spectral shape features to represent the timbre of audio and music signals. The spectral shape features, including brightness, bandwidth, rolloff, and spectral flux, are important in discriminating different moods. For example, the brightness of *Exuberance* music is usually higher than that of *Depression*, since *Exuberance* music generally has a larger spectral energy in the high subbands than *Depression* music, based on our observations. Spectral flux represents the spectrum variations between adjacent frames. From this view, it might be roughly correlated to the articulation, which is an important factor of emotional expression as addressed in [12]. Therefore, the spectral shape features are first used in our approach.

As for MFCC, it is used with great success in general speech and audio processing, and is also used appropriately as one of timbre texture features in a music genre classification system [4]. However, it averages the spectral distribution in each subband, and, thus, loses the relative spectral information. To complement for this disadvantage, *octave-based spectral contrast*, the feature that was proposed in our previous work [5], is utilized instead. The feature considers the spectral peak, spectral valley, and their dynamics in each subband and roughly reflects the relative distribution of the harmonic and nonharmonic components in the spectrum. The evaluations on a music genre recognition system [5] indicated its better performance than MFCC.

Octave-based spectral contrast is also extracted from the FFT. In order to ensure the steadiness of the feature, the strength of spectral peaks and spectral valleys are estimated by the average of a percent of the largest values and the lowest values in the spectrum, respectively, instead of the exact maximum and minimum values.

TABLE I
DEFINITION OF TIMBRE FEATURES

Feature Name		Definition
Spectral Shape Features	Brightness	Centroid of the short-time Fourier amplitude spectrum.
	Bandwidth	Amplitude weighted average of the differences between the spectral components and the centroid.
	Roll off	95^{th} percentile of the spectral distribution.
	Spectral Flux	2-Norm distance of the frame-to-frame spectral amplitude difference.
Spectral Contrast Features	Sub-band Peak	Average of a percent of the largest amplitude values in the spectrum of each sub-band.
	Sub-band Valley	Average of a percent of the lowest amplitude values in the spectrum of each sub-band.
	Sub-band Contrast	The difference between the Peak and Valley in each sub-band.

Suppose the FFT vector of kth subband is $\{x_{k,1}, x_{k,2}, \cdots, x_{k,N}\}$. After sorting it in a descending order, the new vector can be represented as $\{x'_{k,1}, x'_{k,2}, \cdots, x'_{k,N}\}$, where $x'_{k,1} > x'_{k,2} > \cdots > x'_{k,N}$. Thus, the strength of the spectral peaks and spectral valleys are estimated as

$$\text{Peak}_k = \log\left\{\frac{1}{\alpha N}\sum_{i=1}^{\alpha N} x'_{k,i}\right\} \tag{4}$$

$$\text{Valley}_k = \log\left\{\frac{1}{\alpha N}\sum_{i=1}^{\alpha N} x'_{k,N-i+1}\right\} \tag{5}$$

where α is used as a small neighborhood factor and set to 0.2 following [5]. The corresponding spectral contrast is defined as

$$SC_k = \text{Peak}_k - \text{Valley}_k. \tag{6}$$

In this paper, both the spectral shape features and spectral contrast feature are used and compose a 25-dimension timbre feature. A summary of the used features and the corresponding definitions is listed in Table I. More details on the computation of all the listed features can be found in references [4], [5], and [26].

C. Rhythm Features

In general, three aspects of rhythm are closely related with people's mood response: rhythm strength, rhythm regularity, and tempo [12], [29]. For example, it is usually observed that, in the *Exuberance* cluster, the rhythm is usually strong and steady, and the tempo is fast, while the *Depression* music is usually slow and does not have distinct rhythm pattern. In our mood detection system, five novel features are proposed to represent the above mentioned three aspects of rhythm.

The process of rhythm feature analysis is illustrated in Fig. 2. After the FFT, each frame is divided into seven octave-based subbands as (1), and the amplitude envelope of each subband is calculated by convolving with a half Hanning (raised cosine) window, as

$$A'_i(n) = A_i(n) \otimes h_w(n) \tag{7}$$

where $A_i(n)$ is the amplitude or intensity of the ith subband, $A'_i(n)$ is the corresponding amplitude envelope, and $h_w(n)$ is

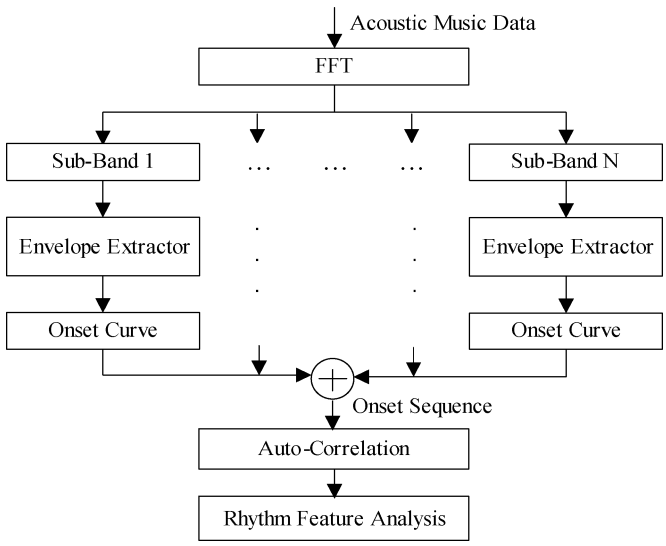


Fig. 2. Basic process of the rhythm features extraction.

the coefficients of a half-Hanning window. Half Hanning has a low-pass characteristic, and is usually used for envelope extraction while still keeping the sharp attacks in the amplitude curve [39]. It is defined as

$$h_w(n) = 0.5 + 0.5\cos\left(2\pi \cdot \frac{n}{(2L-1)}\right), \quad n \in [0, L-1] \tag{8}$$

where L is the length of half-Hanning window, and it is empirically chosen as 12 (in units of our 32-ms frames) in this implementation, to have a tradeoff between the resolution and variation of the obtained amplitude envelope.

After obtaining the amplitude envelope, a Canny operator [40], which is usually used for edge detection in image processing, is used for onset sequence detection by calculating the variance of the amplitude envelope

$$O_i(n) = A'_i(n) \otimes C(n) \tag{9}$$

where $O_i(n)$ is the onset sequence of the ith subband, and $C(n)$ is the Canny operator with a Gaussian kernel

$$C(n) = \frac{n}{\sigma^2} e^{-\frac{n^2}{2\sigma^2}} \quad n \in [-L_c, L_c] \tag{10}$$

and L_c is the length of Canny operator and σ is used to control the operator's shape, which are experimentally set to 12 and 4 in our implementation, respectively. Unlike previous approaches to onset detection that use first-order difference [39], in our approach, the Canny operator considers a larger range of points; thus, it could detect more potential onsets and smooth out of the noise which may be got from simple first-order differences.

Finally, the onset curve of each subband is summed and used to represent the rhythm information of a music clip. From the curve, onsets can be chosen from the peaks which are larger than some threshold which is set as 0.8 in our experiments. Fig. 3 illustrates two examples of the onset sequence from music clips with different moods, where (a) is from an *Exuberance* music clip and (b) is from a *Depression* music clip. From Fig. 3, it can be seen that the *Exuberance* music clip has

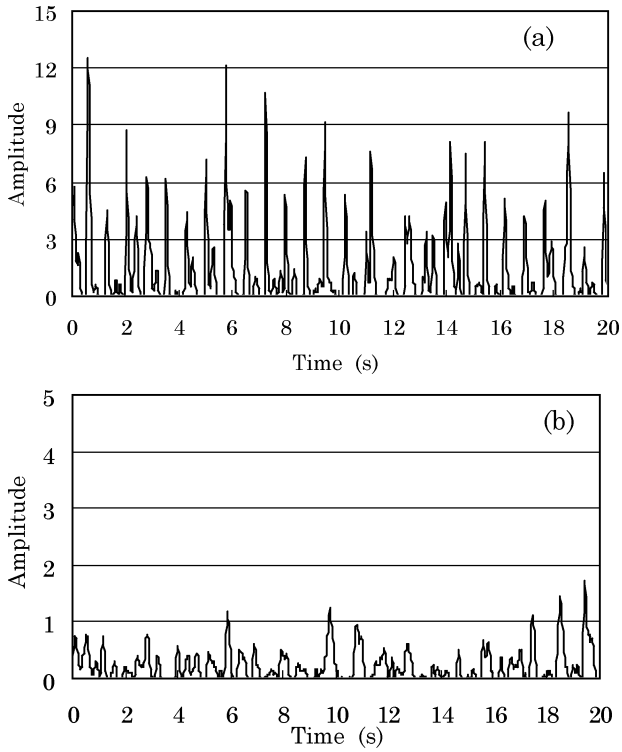


Fig. 3. Two examples of onset sequence from music clips with different mood. (a) is from an *Exuberance* music clip, which has strong rhythm and regular rhythm pattern, while (b) is from a *Depression* music clip, which has weak rhythm and no obvious rhythm pattern.

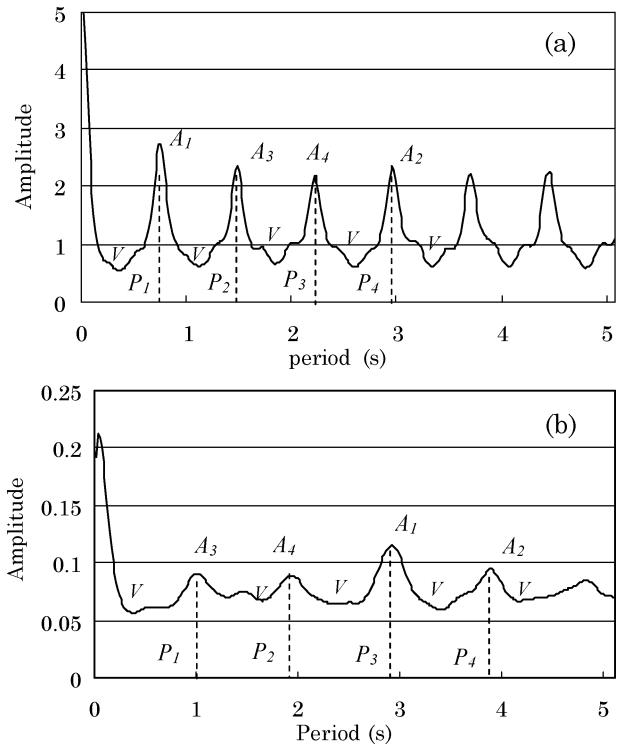


Fig. 4. Examples of autocorrelation curve corresponding to the onset sequence in Fig. 3. (a) An *Exuberance* music clip. (b) A *Depression* music clip.

a much stronger rhythm than the *Depression* music clip. Most of the onset strengths of Fig. 3(a) are larger than 3.0, while those of Fig. 3(b) are usually less than 1.5. Accordingly, the first component of rhythm feature set, *Rhythm Strength*, can be intuitively correlated to the average strength of onsets, where the onset strength can be assumed as the values of peaks in the onset sequence.

- *Rhythm Strength:* The average onset strength in the onset sequence. The stronger the rhythm is, the higher the value is.

From Fig. 3, we also find there is a big difference of the rhythm regularity between (a) and (b). Fig. 3(a) has a more regular rhythm than Fig. 3(b). In order to describe rhythm regularity more accurately, an autocorrelation is performed on the onset sequence curve. Fig. 4 illustrates the corresponding autocorrelation curve, with the maximum lag of 5 s.

It is clear that if a music clip has an obvious and regular rhythm, the peaks of the corresponding autocorrelation curve will be obvious and strong as well, and vice versa. According to this fact, the following two feature components are extracted to represent the distinctness and regularity of the rhythm.

- *Average Correlation Peak:* The average strength (amplitude) of the local peaks in the auto-correlation curve. The more regular the rhythm is, the higher the value is.
- $avr(\mathbf{A})/avr(\mathbf{V})$: The ratio between the average peak strength and average valley strength. The more obvious the rhythm is, the higher the value is. Here a valley is determined as the local minimum between two adjacent peaks.

In our implementation, only the four largest peaks and the neighboring valleys are selected from the autocorrelation curve for *Rhythm Regularity* calculation, in order to avoid selecting the small peaks. Fig. 4 also illustrates the positions of the top four maximum peaks P_1, P_2, P_3, and P_4, from near to far from the origin, and the corresponding valleys V around these peaks.

Furthermore, in order to represent the speed of the music performance, average tempo is also estimated from the autocorrelation curve.

- *Average Tempo:* Represents the average speed of the music performance. Similar to many approaches to fundamental frequency detection [41], average tempo is estimated as the maximum common divisor of the detected peaks, assuming that the average tempo does not vary much in the music clip, as follows:

$$T = \arg\min_{P_k} \sum_{i=1}^{N} \left| \frac{P_i}{P_k} - \mathrm{round}\left(\frac{P_i}{P_k}\right) \right| \qquad (11)$$

where P_i is the ith detected peak, and the operator $\mathrm{round}(x)$ is to get the nearest integer to x.

In our implementation, (11) is minimized over a predefined range of plausible tempi, and the average tempo is further normalized by 120 beat per minute (BPM). However, such an average tempo only represents the occurrence frequency of beats. It could not represent the frequency of the underlying onsets, which is also an important cue of music performance. Therefore, another feature component is also extracted as follows.

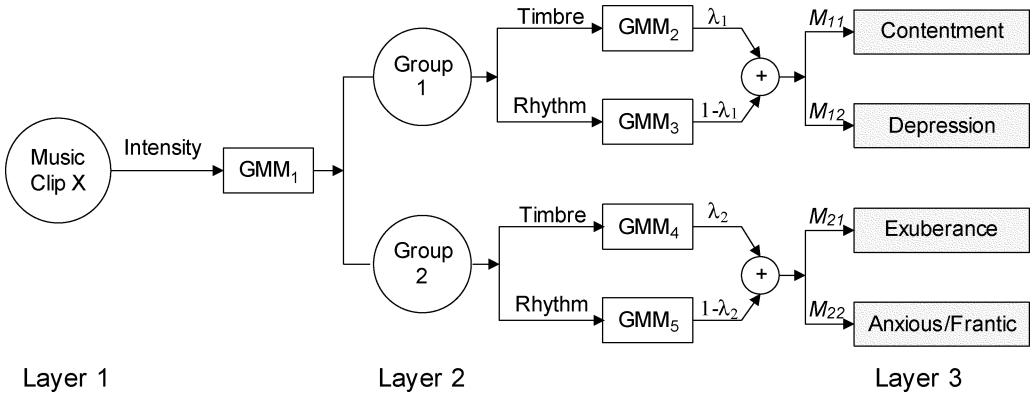


Fig. 5. Hierarchical mood detection framework, which can place emphasis on different features in different classification tasks, and make better use of sparse training data.

- *Average Onset Frequency:* Estimated from the onset sequence shown in Fig. 3. It can be easily calculated as the ratio between the number of onsets and the corresponding time duration. The larger value it is, the faster the performance is.

The aforementioned five feature components compose a five-dimension rhythm feature set, which can perform quite well on music moods discrimination. For example, *Exuberance* music usually has large rhythm strength, large onset frequency, and large average autocorrelation peak, while *Depression* music has a weak strength, slow onset frequency or tempo, and low autocorrelation peak.

D. Feature Representation of a Music Clip

We must now consider the problem of combining above three feature sets to represent a music clip, since the intensity and timbre features are obtained from each frame while the rhythm features are from an entire clip. Moreover, it is not appropriate to simply concatenate the components of each feature set into a feature vector, since the characteristics and dynamics of these feature components are so different. Therefore, a normalization process is first performed on each feature component to make their scale similar. The normalization (also called standardization or z-scores) is processed as $x_i' = (x_i - \mu_i)/\sigma_i$, where x_i is the ith feature component, the corresponding mean μ_i and standard deviation σ_i can be calculated from the ensemble of the training data. Then, the K–L transform [37] is performed on each feature set across the entire training data, in order to remove the correlation among these normalized features. As a result of the K–L transform, each feature vector is mapped into an orthogonal space, and its corresponding covariance matrix becomes diagonal in the new feature space. This procedure helps to achieve a better performance in the later mood classification. It is noted that, in our approach, the K–L transform is performed on each feature set independently, since in our hierarchical mood detection framework (Section IV), each feature set is considered individually instead of concatenating them into one feature vector.

To this end, the basic statistics (mean and standard deviation) of the frame-based intensity features and timbre features, and the final rhythm features are used to represent the mood primitives of a music clip.

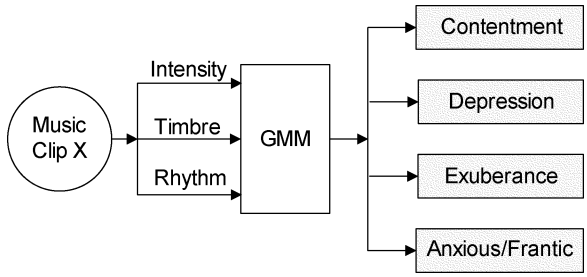


Fig. 6. Corresponding nonhierarchical mood detection framework (see Table III).

IV. Mood Detection

Based on Thayer's model of mood, a hierarchical framework is proposed for mood detection, as illustrated in Fig. 5. As Huron [38] pointed out, energy (or intensity) is more computationally tractable in the two factors in Thayer's model. Therefore, the intensity features are first used to classify a music clip into one of two mood groups. The basic rule could be, if its energy is low, the music clip will be classified into Group 1 (*Contentment* and *Depression*); otherwise, it is classified into Group 2 (*Exuberance* and *Anxious/Frantic*). Subsequently, the remaining features, including timbre and rhythm, are used to determine which exact mood the music clip is. Meanwhile, as Juslin [12] pointed out, the relative importance of each acoustic cue is different for different emotional expressions. This means that the performance of different features is not consistent in discriminating different pairs of mood clusters. Accordingly, the proposed framework considers this fact, and makes it possible to use the most suitable features for different tasks by setting appropriate weightings for different features. Moreover, compared with its nonhierarchical counterpart as shown in Fig. 6, the hierarchical framework can make better use of sparse training data [42], which is very important especially when the training data is limited.

In the framework, a Gaussian mixture model (GMM) with 16 mixtures is utilized to model each feature set regarding each mood cluster (group). In constructing each GMM, the *Expectation Maximization* (EM) algorithm is used to estimate the parameters of Gaussian components and mixture weights, and K-means are employed for initialization. For more details on these techniques see, e.g., [43].

With the obtained GMM models, the detailed mood classification can be performed in the following two steps. In the first step, a music clip is classified into different mood groups, i.e., Group 1 (*Contentment* and *Depression*) and Group 2 (*Exuberance* and *Anxious/Frantic*), by employing a simple hypothesis test with the intensity features, as

$$\lambda = \frac{P(G_1|I)}{P(G_2|I)} \begin{cases} \geq 1, & \text{Select} \quad G_1 \\ < 1, & \text{Select} \quad G_2 \end{cases} \tag{12}$$

where λ is the likelihood ratio, G_i represents different mood group, I is the intensity feature set, and $P(G_i|I)$ is the probability that the testing music clip belongs to mood group G_i given its intensity features, which can be calculated from the GMM model.

In the second step, the music clip in Group 1 is further classified into *Contentment* and *Depression*, while that in Group 2 is further classified into *Exuberancek* and *Anxious/Frantic*, based on the timbre and rhythm features. In each group, the probability of the testing clip belonging to an exact mood $M_{i,j}$ can be calculated as

$$P(M_{i,j}|G_i, T, R) = \lambda_i \times P(M_{i,j}|T) + (1 - \lambda_i) \times P(M_{i,j}|R), \quad i, j = 1, 2 \tag{13}$$

where $M_{i,j}$ is the jth mood cluster in ith mood group, T and R represent timbre and rhythm features, respectively, and λ_1 and λ_2 are two weighting factors to represent different importances of timbre and rhythm features in the mood detection among different mood groups, as Juslin pointed out in [12]. After obtaining these probabilities (likelihood), a simple hypothesis test, similar to (12) is employed again to classify the music clip into exact mood cluster.

Actually, in the Group 1, the tempo of both mood clusters is usually slow and the rhythm pattern is generally not steady, while the timbre of *Contentment* is usually much brighter (with higher brightness) and more harmonic than that of *Depression*. Therefore, the timbre features are more important than the rhythm features in discriminating *Contentment* from *Depression* in Group 1. On the contrary, in Group 2, rhythm features are more important. *Exuberance* music usually has a more distinguishable and steady rhythm than *Anxious/Frantic* music, while their timbre features are similar, since both mood clusters usually have similar dominant instruments, such as brass. Based on these facts, λ_1 is usually set as larger than 0.5, while λ_2 is less than 0.5. The optimal values of λ_1 and λ_2 will be given in the experiments (Section VI).

V. Mood Tracking

In the previous section, we present a hierarchical framework on mood detection for a given music clip, which contains a consistent mood type. However, for a piece of classical music, the mood may well change one or more times within a single piece. Therefore, it is not appropriate to detect an exclusive mood for an entire piece of music. A usual method to track the mood changes is using a sliding window of a certain length and overlap. However, such a window may contain mixed moods that cannot be recognized correctly. It would be better if we could first find potential mood change boundaries and then divide the music into several independent segments, each of which contains a constant mood. Thus, we can detect a unique mood from each segment correspondingly. With this two-step mood tracking scheme, in this section, we mainly propose an unsupervised approach to potential mood boundary detection in a music piece. Subsequently, the mood tracking in an entire piece of music can be easily performed by identifying the mood in each independent segment, using the approach presented in Section IV.

Since the intensity, timbre, and rhythm are main primitives in mood detection, their changes also provide the main cues for a new mood event. Thus, all of these three feature sets are used to complement each other in mood boundary detection (mood segmentation). In our approach, an intensity outline is first implemented to coarsely detect potential boundaries, and then the timbre and rhythm features are used to detect possible mood changes in each contour of the intensity outline. These two steps are similar to the hierarchical process in mood detection discussed in the above section.

According to some music theory [28], one musical paragraph is often composed of 16 bars. Although the limitation is adjustable, in our implementation, the minimum length of a segment containing a constant mood is set to 16 s, assuming that each musical paragraph always contains a constant mood and a very fast tempo is about 1 bar/second in the classical music.

A. Intensity Outline

Classical music usually has an obvious pattern in the energy development, with alternation between strong intensity and weak intensity. These alternations are good indicators of potential mood change. In this section, an intensity outline is first detected to temporally divide a music piece into several segments, where each segment contains an almost constant intensity or is corresponding to a single energy contour that grows from low to high and then back to low energy. These segments can roughly represent the mood development along the timeline. The basic process of the intensity outline detection is illustrated in Fig. 7, with an example classical music, Tchaikovsky's 1812 overture.

In order to divide the energy envelope into such segments, a threshold is usually needed to discriminate the contour peaks (convexes) from the contour valleys (concaves). However, one constant threshold does not always work well, as Fig. 7(a) shows, where the threshold is set to the average of the energy envelope, and the height of the outline is the average intensity of the corresponding segment. It can be seen that some contours are confused and, thus, merged, especially in the middle part of the music. To solve such a problem, in our approach, two thresholds are defined to preliminarily quantize the energy envelope into three levels instead of binaries, thereby generating a larger number of potential segment boundaries whenever either threshold is crossed with the energy envelope. The thresholds are adaptively set based on the statistics of the energy envelope of the entire piece of music, as $Th1 = \mu - 0.5\sigma$, $Th2 = \mu + 0.5\sigma$, where μ and σ are the corresponding mean and standard deviation, respectively. Our experiments show that slight changes of these thresholds do not have much impact on the intensity outline, indicating that this scheme is relatively

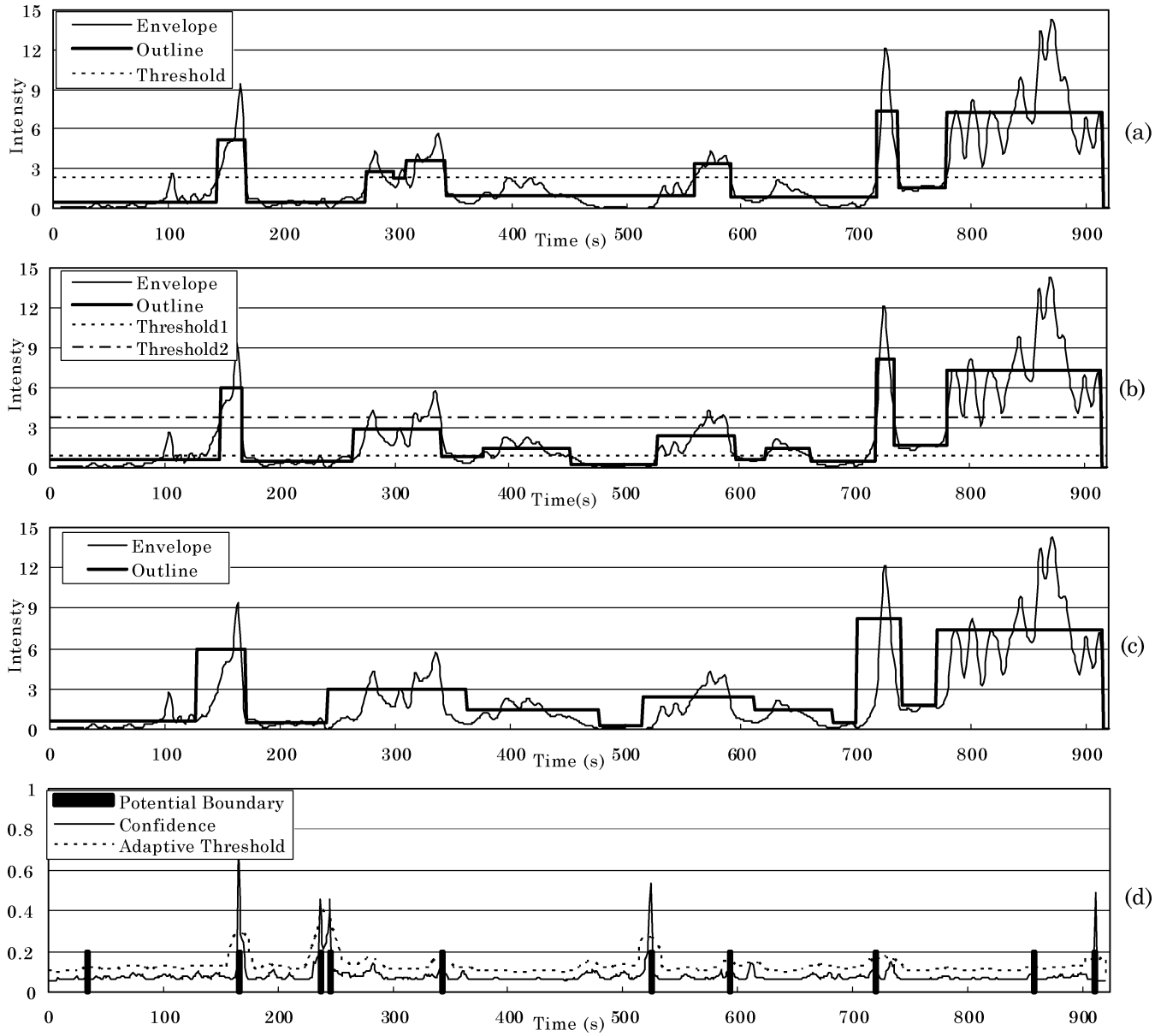


Fig. 7. Illustration of some temporary results in each step of potential mood boundary detection, with an example classical music, Tchaikovsky's 1812 overture. (a) Intensity outline with one threshold. (b) Intensity outline with two thresholds. (c) Intensity outline after boundary adjustment. (d) Potential boundaries and the corresponding confidence and adaptive threshold sequences based on timbre and rhythm features.

robust. However, such segmentation may cause some trivial segments shorter than the assumed minimum length. These short segments are merged to their neighboring segments based on the similarity of average intensity. Fig. 7(b) shows the intensity outline after the merging process, with the corresponding energy envelope and two thresholds. From the figure, it can be seen that the resulting segmentation appears more appropriate for this example. However, the boundaries of some segments might be biased. In order to get more reliable results, the boundary of each segment is aligned to the nearest valley of the energy envelope, where a valley is defined as a point of local minimum. The final intensity outline is illustrated in Fig. 7(c), where each change of the outline is a potential mood change boundary. It is noted that after boundary adjustment, some transition segments, such as the one at about the 600th second (shown in Fig. 7(b)), are further removed.

B. Potential Boundary Detection Based on Timbre and Rhythm

Furthermore, we detect the potential mood boundaries in each intensity segment, based on the timbre and rhythm features. A basic rule is, if there is a large dissimilarity of timbre and rhythm between two adjacent segments, there is a potential mood boundary. In our implementation, the dissimilarity of timbre and rhythm at each second is measured by comparing the two windows before and after this time slot, and the length of each window is empirically set as 8 s in order to keep the comparison range not larger than the assumed minimum length of a musical paragraph. Thus, a dissimilarity sequence is obtained with the resolution of 1 s. It is easy to increase the resolution at the cost of more computation load.

To measure the timbre dissimilarity between every pair of adjacent windows, in our approach, *divergence shape* [44] is

utilized, supposing the feature vectors are Gaussian distributed. Divergence shape is an efficient distance measure between two feature distributions, as indicated in many previous works. It is defined as

$$D = \frac{1}{2} tr\left[(C_i - C_j)\left(C_j^{-1} - C_i^{-1}\right)\right] \tag{14}$$

where $\mathrm{tr}[\cdot]$ is to calculate the trace of a matrix, C_i and C_j is the estimated covariance matrices of the frame-based timbre features in the ith and jth windows, respectively, and, thus, represent the feature distributions of the corresponding windows. On the other hand, to measure the rhythm dissimilarity between two windows, Euclidean distance is simply utilized in our approach, since rhythm features are segment-based instead of frame-based.

Based on the obtained dissimilarity measure at each second, the corresponding confidence, which means the probability of there being a potential mood boundary at the time slot, is simply defined as

$$\mathrm{Conf}_F(i) = \frac{1}{A_d} \exp\left(\frac{D_i - \mu_d}{3\sigma_d}\right) \tag{15}$$

where the subscript F represents the feature set (timbre or rhythm), D_i is the dissimilarity (distance) of the corresponding feature at the ith second, μ_d and σ_d are the mean and standard deviation of the entire distance sequence, and A_d is a normalization constant to make the maximum confidence 1. Thus, the total confidence based on both timbre and rhythm can be set as

$$\mathrm{Conf}(i) = \alpha_C \times \mathrm{Conf}_T(i) + (1 - \alpha_C) \times Conf_R(i) \tag{16}$$

where $\mathrm{Conf}_T(i)$ and $\mathrm{Conf}_R(i)$ are the confidence of there being a potential boundary at the ith second, based on timbre features and rhythm features respectively; α_C is a confidence weighting factor and set to 0.5 in our implementation, assuming the same importance of these two feature sets in the mood boundary detection.

To this end, a mood boundary can be found at the ith second, if the following conditions are satisfied

$$\begin{aligned} &\mathrm{Conf}(i) > \mathrm{Conf}(i+1), \\ &\mathrm{Conf}(i) > \mathrm{Conf}(i-1), \\ &\mathrm{Conf}(i) > Th_i \end{aligned} \tag{17}$$

where Th_i is a threshold. The first two conditions guarantee that a local peak exists, and the last condition can prevent very low peaks from being detected. However, the threshold is not appropriate to set as a constant value which may not work well for all the music pieces. In our implementation, the threshold is adaptively set according to its context, as

$$Th_i = \alpha_T \times \frac{1}{2N} \sum_{n=-N}^{N} \mathrm{Conf}(i+n) \tag{18}$$

where N is the number of the previous and succeeding confidences used to predict the threshold, and α_T is a threshold amplifier. In our approach, N is set to 8 so that the threshold is automatically set according to its neighborhood of 16 s, and α_T is experimentally chosen as 1.5 in order to obtain optimal result.

An example of the detected potential boundary, with the corresponding confidence sequence and adaptive threshold, is shown in Fig. 7(d). Comparing Fig. 7(d) and (c), it is noted that most of the boundaries detected by timbre and rhythm are matched with those from intensity outline. It indicates the threshold works well in detecting mood boundaries, and different energy contours may also contain different timbre or rhythm.

Integrating boundaries of timbre/rhythm and those of intensity outline may again cause some short segments. Therefore, we still need to merge the short segments to their neighboring segments based on the intensity similarity. It should also be noted that a music piece is prone to be over-segmented with our mood boundary detection scheme. This is intentional since we more prefer to recall all possible candidate mood boundaries. False alarms can be further removed with mood detection on each segment, using the hierarchical mood detection algorithm discussed in Section IV.

TABLE II
MUSIC MOODS AND CORRESPONDING MUSIC STYLES

Moods	Corresponding Music Styles
Contentment	Church Music and Serenade
Depression	Cello Concerto, Piano Concerto, Funeral Part
Exuberance	March, Overture, Dance Music
Anxious/Frantic	Overture, Epilogue

VI. EXPERIMENTS

In this section, the proposed approach is evaluated with our testing database. We first present the performance of mood classification on a selected music set, and then the approach to mood tracking is evaluated with some famous musical works.

A. Mood Detection on Music Clips

Our database contains about 250 pieces of music, composed mainly in the classical period and romantic period. Choir, orchestra, piano, and string quartet are all included to ensure the diversity of music style in the database. Table II shows the main music styles we used to select music clips of different mood clusters. For instance, the music clips in the *Contentment* cluster are mainly selected from Church music and Serenades, while the *Exuberance* clips are mainly from Overtures, Marches, and Dance music.

Three experts participated in selecting and annotating representative 20-s music clips from the database into the predefined four mood clusters: *Contentment*, *Depression*, *Exuberance*, and *Anxious/Frantic*. In the annotation process, the experts usually agreed on most of the music clips. If the experts had different opinions and cannot get an agreement on one music clip, the clip

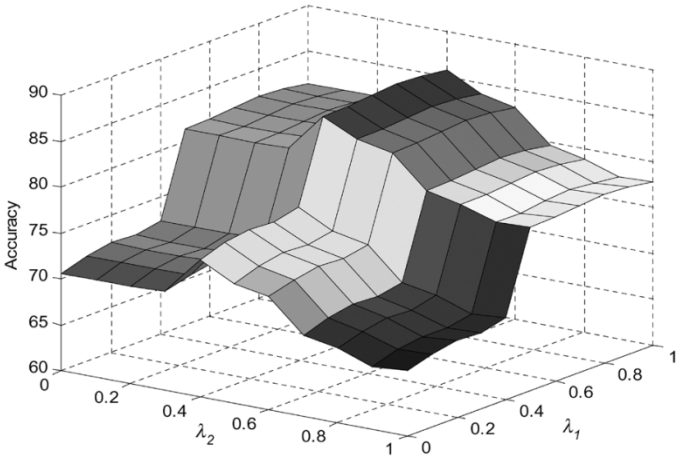


Fig. 8. Average accuracy of mood detection, corresponding to different settings of the weighting factor set, where x-y represents the weighting set (λ_1, λ_2), and the z-axis represents the mood detection accuracy (unit: 100%).

would not be selected in our training or testing data set. That means we only selected the music clips over which the experts agreed on their mood type in order to avoid false annotations and ensure that the perceived mood of each selected music clip is consistent and representative. The music clips which are not representative enough, i.e., the ambiguous cases, were removed. (It is noted that this scheme may have a potential problem since there are many ambiguous cases in the real world). An example is Suppe's "Light Cavalry," from which only two representative clips of *Depression* and three clips of *Exuberance* were selected. It might be argued that the same piece of music may be performed with different tempi or instruments to communicate different emotions, and, thus, we should carefully choose one performance of a music clip. However, this fact is of little relevance to our data selection and annotation, since our mood annotation is based on human perception instead of the music expression which is related to music composition. Finally, 800 representative music clips of 20 s are selected from the 250 music pieces for mood classification evaluation, and each mood cluster has 200 music clips, respectively.

In our experiments, the detection results are evaluated using a cross-validation where the dataset is randomly partitioned so that 75% of the data is used for training and the remaining 25% is used for testing. Ten random partitions are performed and the results are averaged.

As mentioned previously, the timbre and rhythm features have different importance in mood detection in different mood groups. Accordingly, in our experiments, we first evaluate the detection performance corresponding to each weighting factor set (λ_1, λ_2) as defined in (13), by exhaustively searching in the (λ_1, λ_2) space $([0,1] \times [0,1])$ with 10 points in each dimension. The detailed results are illustrated in Fig. 8. It can be seen that the optimal average accuracy of 86.3% is achieved when $\lambda_1 = 0.8$ and $\lambda_2 = 0.4$. It indicates that the timbre features are much more important than the rhythm features in classifying *Contentment* and *Depression* in Group 1, while the rhythm features are slightly more important to discriminate *Exuberance* from *Anxious/Frantic* in Group 2. It is also noted

TABLE III
COMPARISON OF MOOD DETECTION RESULTS WITH DIFFERENT WEIGHTING FACTOR SET AND DIFFERENT FRAMEWORK (UNIT: 100%)

(λ_1, λ_2)	(1,1)	(0,0)	(0.8, 0.4)	None
Contentment	74.3±8.1	53.4±10.0	**76.6±7.6**	75.0±11.8
Depression	94.5±3.4	72.2±9.5	**94.5±3.4**	94.2±2.6
Exuberance	71.2±22.7	67.4±19.6	**85.5±3.2**	64.7±20.5
Anxious/Frantic	71.3±20.1	89.4±4.0	**88.5±6.7**	88.3±7.9
Average	77.8	70.6	**86.3**	80.6

TABLE IV
MOOD DETECTION CONFUSION MATRIX BASED ON HIERARCHICAL FRAMEWORK (UNIT: 100%)

	Contentment	Depression	Exuberance	Anxious/Frantic
Contentment	**76.6±7.6**	21.8±7.2	0.5±0.8	1.2±1.2
Depression	4.0±3.5	**94.5±3.4**	0±0	1.5±2.5
Exuberance	0±0	0.8±1.3	**85.5±3.2**	13.7±4.8
Anxious/Frantic	0±0	0±0	11.5±6.7	**88.5±6.7**

that there is a big accuracy jump when λ_1 is equal to 0.5; while the accuracy varies little when λ_1 is larger than 0.5.

To clearly show the discrimination power of the individual timbre or rhythm feature set in mood detection, the corresponding performance on each mood cluster is listed in Table III, where the part before $\pm$ is the average accuracy, and the part after it is the standard deviation across 10 testing sets. From Table III, it can be seen that when timbre features are used only [with weighting factor (1, 1)], the average accuracy is 77.8%, while 70.6% is achieved when the rhythm features are used solely [with weighting factor (0, 0)]. When both of these two feature sets are used and the weightings are optimally set to (0.8, 0.4), the average accuracy improves to 86.3%, a relative improvement of 22.2% over using rhythm features only and 10.9% over using timbre features only.

In our subsequent experiment, we compared the performance between our hierarchical framework and its nonhierarchical counterpart which integrates three feature sets and carries on classification directly, as illustrated in Fig. 6. The performance of this system is also shown in Table III, in the final column (marked "None"). From the numbers, it is noted that the hierarchical framework improves the average accuracy by 7.1% relative, compared with the nonhierarchical framework. Meanwhile, it also can be seen that the hierarchical framework is more robust since the obtained deviations are much smaller than those of nonhierarchical framework. The average deviation decreases from 10.7% to 5.2%.

To show more detailed classification results of these two frameworks, Tables IV and V show the confusion matrices among the different mood clusters, where each row corresponds to the actual mood cluster and each column to the predicted cluster.

From Table IV, only 1.6% of the music clips in Group 1 (*Contentment* and *Depression*) are misclassified into Group 2 (*Exuberance* and *Anxious/Frantic*), while only 0.4% of those in

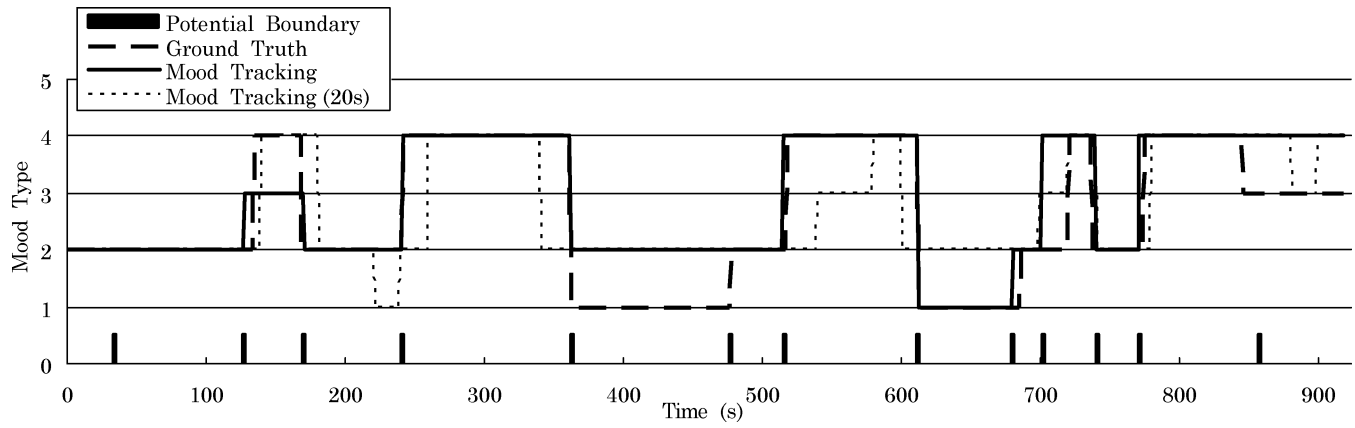


Fig. 9. Mood tracking result on a music piece "1812 overture," where 1–4 in z-axis represents the four mood clusters, that is, 1—contentment, 2—depression, 3—exuberance, and 4—anxious/frantic.

TABLE V
MOOD DETECTION CONFUSION MATRIX BASED ON NONHIERARCHICAL FRAMEWORK (UNIT: 100%)

	Contentment	Depression	Exuberance	Anxious/Frantic
Contentment	**75.0±11.8**	25.0±11.8	0±0	0±0
Depression	5.8±2.6	**94.2±2.6**	0±0	0±0
Exuberance	1.5±2.6	0.7±1.3	**64.7±20.5**	33.0±18.3
Anxious/Frantic	0±0	0±0	11.7±7.9	**88.3±7.9**

Group 2 are misclassified into Group 1. It indicates the good performance of the intensity features in discriminating two mood groups, which serves as the basis for further classification by the timbre and rhythm features.

Comparing Table IV with Table V, it can be observed that, by adopting the proposed hierarchical framework, the overall performance is improved, especially for *Exuberance*, whose accuracy is improved significantly. In the nonhierarchical framework, about 33.0% of *Exuberance* clips are misclassified into *Anxious/Frantic*, which is decreased by more than 50% after using our hierarchical framework. These experimental results show that the proposed hierarchical framework has a better performance than its nonhierarchical counterpart, by using the most efficient features for different mood clusters.

B. Mood Tracking

Several pieces of classical music obtained from CDs and the Internet are used for mood tracking evaluation. The detailed information about the testing data is listed in Table VI. They are more than 1 h in total, with different sampling rate including 48, 44.1, and 32 kHz. In our experiments, each format is converted to 16 kHz and mono-channel before further processing.

Fig. 9 illustrates an example mood tracking result of a testing music piece, "1812 Overture" composed by Tchaikovsky. In the figure, the manually labeled mood development (i.e., ground truth), the detected mood boundaries, and the corresponding mood sequence estimated by the proposed mood tracking algorithm, are all illustrated and compared. The mood tracking results based on simple segmentation every 20 s are also shown in the figure.

TABLE VI
MUSIC DATA FOR MOOD CHANGE DETECTION AND MOOD TRACKING

Index	Title	Composer	Duration
1	1812 Overture	Tchaikovsky	15:23
2	The Sleeping Beauty	Tchaikovsky	03:30
3	Light Cavalry	Suppe	06:16
4	William Tell Overture	Rossini	11:22
5	The Bat Overture	J. Strauss II	08:10
6	Aigrema Overture	Beethoven	08:58
7	Symphony No 5, Movement 1	Beethoven	07:21
8	Carmen Suite	Bizet	01:59
9	Carmina Burana	Orff	02:41

It can be seen from Fig. 9 that almost all of the correct mood boundaries are recalled using the proposed algorithm, although some false alarms also exist. It is also noted that the boundaries obtained by the proposed algorithm are much more accurate than those obtained by simple segmentation with a sliding window of a constant length such as 20 s. Moreover, the mood type of each segment detected based on our approach is also much better. This is because the mood remains consistent in each segment after mood boundary detection, which avoids the confusion caused by mixed moods. Meanwhile, each segment obtained from boundary detection is usually longer and has more data to estimate the inherent mood type.

It should be pointed out that the most important step of mood tracking is to detect the potential mood boundaries. After mood boundaries are detected, the remaining task is to identify the mood type of each segment using the proposed hierarchical algorithm, whose performance is evaluated and presented in the Section VI-A. Therefore, in this section, we mainly evaluate the performance of the mood boundary detection. The detailed results are listed in Table VII.

Table VII lists the number of original mood boundaries (ground truth), and the detected, missed, and false boundaries for each testing music piece, based on which the measures of recall and precision are calculated. In the experiments, a detected mood boundary is assumed correct if the estimated position is less than 3 s away from the ground truth position (the boundary detection resolution is 1 s as mentioned in Section V-B), and the recall is calculated as the ratio between

TABLE VII
MOOD BOUNDARY DETECTION ACCURACY

Index	Original	Detected	Miss	False	Recall	Precision
1	12	13	1	2	91.7%	84.6%
2	3	4	0	1	100%	75.0%
3	5	6	1	2	80.0%	66.7%
4	4	6	0	2	100%	66.7%
5	9	9	1	1	88.9%	88.9%
6	12	11	2	1	83.3%	90.9%
7	12	10	4	2	66.7%	80.0%
8	3	3	1	1	66.7%	66.7%
9	3	3	0	0	100%	100%
All	63	65	10	12	84.1%	81.5%

the number of correctly detected boundaries and that of the original boundaries, while precision is the ratio between the number of correctly detected boundaries and the total number of the detected boundaries. From Table VII, it can be seen that the recall is a little bit larger than precision with our approach, since the false alarms are preferred to the missed boundaries in our system. The overall average recall is 84.1% and the precision is 81.5%. This indicates our approach can achieve satisfying results.

In the mood tracking experiments, we find that sometimes the Thayer's model cannot cover all the mood types inherent in a music piece. For instance, a *Sprightly* mood which usually has small intensity and with a fast tempo, is not included in the current mood taxonomy. We have to classify it as the nearest mood cluster (*Contentment*) in the Thayer's model. This indicates that we need to extend the framework and mood taxonomy to explore other inherent emotional expressions of music clips in future work. We also find that it is still possible that an actual music clip may contain some mixed moods or an ambiguous mood. In such a case, an individual mood may not provide enough information, and it would be better to provide several candidate moods with the corresponding confidences.

VII. CONCLUSION

In this paper, we present an approach to mood detection for acoustic recordings of classical music. Thayer's model of mood is adopted for the mood taxonomy, which is composed of four music moods, *Contentment*, *Depression*, *Exuberance*, and *Anxious/Frantic*. Three efficient feature sets, including intensity, timbre, and rhythm, are proposed and extracted from acoustic data. The intensity feature set is represented by the energy in each subband, the timbre feature set is composed of the spectral shape features and spectral contrast features, and the rhythm feature set indicates three aspects closely related with an individual's mood response, including rhythm strength, rhythm regularity, and tempo. A hierarchical framework is used to detect the mood in a music clip. It first classifies a music clip into mood groups based on intensity features; and then the classification is performed in each group based on timbre and rhythm features. The hierarchal framework can utilize the most suitable features in different tasks and can perform better than its nonhierarchical counterpart.

Furthermore, since the mood is usually changeable in an entire piece of classical music, the approach to mood detection is extended to mood tracking for a music piece by dividing the music into several independent segments, each of which contains a constant mood. In our mood boundary detection algorithm, an intensity outline is first implemented to coarsely detect potential boundaries, and then the timbre and rhythm features are used to detect possible mood changes in an outline segment. Experimental results show that about 84.1% of the boundaries are recalled and the precision is about 81.5%.

There is still much room for future improvements of the proposed algorithm. We will look for ways to extract more powerful acoustic features to better represent music primitives in mood perception, such as mode and articulation which are expected to improve the mood detention accuracy, at least in the valence (i.e., positive–negative) discrimination. Moreover, we will extend the mood taxonomy to cover more mood types which are related to human responses, and incorporate more ambiguous moods in the training set and present the detection results with more information such as confidences. More efficient ways to integrate various features should still be explored due to the unbalanced nature of each feature set. Finally, we will extend the mood detection framework to other music genres, such as pop music.

The proposed approach is a preliminary attempt at mood detection and mood tracking from acoustic music signals. We hope it can inspire more research works on music analysis.

REFERENCES

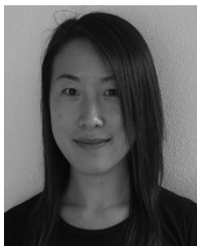
[1] D. Huron, "Perceptual and cognitive applications in music information retrieval," in *Proc. Int. Symp. Music Information Retrieval (ISMIR)*, 2000.
[2] M. Goto and Y. Muraoka, "Beat tracking based on multiple-agent architecture—a real-time beat tracking system for audio signals," in *Proc. Int. Conf. Multi Agent Systems*, 1996, pp. 103–110.
[3] E. Scheirer, "Tempo and beat analysis of acoustic musical signals," *J. Acoust. Soc. Amer.*, vol. 103, no. 1, pp. 588–601, 1998.
[4] G. Tzanetakis and P. Cook, "Musical genre classification of audio signals," *IEEE Trans. Speech Audio Process.*, vol. 10, no. 5, pp. 293–302, Sep. 2002.
[5] D. N. Jiang, L. Lu, H. J. Zhang, J. H. Tao, and L. H. Cai, "Music type classification by spectral contrast features," in *Int. Conf. Multimedia Expo.*, vol. 1, 2002, pp. 113–116.
[6] *Proc. ISMIR: Int. Symp. Music Information Retrieval*, [Online]. http://www.ismir.net/.
[7] R. W. Picard, *Affective Computing*. Cambridge, MA: MIT Press, 1997.
[8] R. W. Picard, E. Vyzas, and J. Healey, "Toward machine emotional intelligence: analysis of affective physiological state," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 23, no. 10, pp. 1175–1191, Oct. 2001.
[9] B. H. Repp, "A microcosm of musical expression. I. Quantitative analysis of pianists' timing in the initial measures of Chopin's Etude in E major," *J. Acous. Soc. Am.*, vol. 104, pp. 1085–1100, 1998.
[10] ——, "A microcosm of musical expression: II. quantitative analysis of pianists' dynamics in the initial measures of Chopin's Etude in E major," *J. Acous. Soc. Am.*, vol. 104, pp. 1972–1988, 1998.
[11] A. Gabrielsson and E. Lindstrom, "The influence of musical structure on emotional expression," in *Music and Emotion: Theory and Research*, P. N. Juslin and J. A. Sloboda, Eds. Oxford, U.K.: Oxford Univ. Press, 2001, pp. 223–248.
[12] P. N. Juslin, "Cue utilization in communication of emotion in music performance: relating performance to perception," *J. Exper. Psychol.: Human Percept. Perf.*, vol. 16, no. 6, pp. 1797–1813, 2000.
[13] P. N. Juslin, A. Friberg, and R. Bresin, "Toward a computational model of expression in music performance: the GERM model," *Musicae Scientiae*, pp. 63–122, 2001–2002. Special issue.

[14] P. N. Juslin, "Perceived emotional expression in synthesized performances of a short melody: capturing the listener's judgment policy," *Music. Sci.*, vol. 1, no. 2, pp. 225–256, 1997.
[15] P. N. Juslin and J. A. Sloboda, *Music and Emotion: Theory and Research.* Oxford, U.K.: Oxford Univ. Press, 2001.
[16] R. Bresin and A. Friberg, "Emotional coloring of computer-controlled music performances," *Comput. Music J.*, vol. 24, no. 4, pp. 44–63, 2002.
[17] D. Liu, N. Y. Zhang, and H. C. Zhu, "Form and mood recognition of Johann Strauss's waltz centos," *Chin. J. Electron.*, vol. 12, no. 4, pp. 587–593 , 2003.
[18] H. Katayose, M. Imai, and S. Inokuchi, "Sentiment extraction in music," in *Int. Conf. Pattern Recognition*, vol. 2, 1988, pp. 1083–1087.
[19] A. Friberg, E. Schoonderwaldt, P. N. Juslin, and R. Bresin, "Automatic real-time extraction of musical expression," in *Proc. Int. Computer Music Conf.*, 2002, pp. 365–367.
[20] L. Mion, "Application of Bayesian networks to automatic recognition of expressive content of piano improvizations," in *Proc. Stockholm Music Acoustics Conf.*, vol. 2, 2003, pp. 557–560.
[21] M. Leman, V. Vermeulen, L. Voogdt, J. Taelman, D. Moelants, and M. Lesaffre, "Correlation of gestural musical audio cues and perceived expressive qualities," in *Gesture-Based Communication in Human-Computer Interaction*, A. Camurri and G. Volpe, Eds. New York: Springer Verlag, 2004. LNAI 2915.
[22] C. L. Krumhansl, "Music: a link between cognition and emotion," *Current Directions Psychological Sci.*, vol. 11, no. 2, pp. 45–50, 2002.
[23] E. Radocy and J. D. Boyle, *Psychological Foundations of Musical Behavior*. Springfield, IL: Charles C. Thomas, 1988.
[24] J. A. Russell, "A circumplex model of affect," *J. Personality Social Psychology*, vol. 39, pp. 1161–1178, 1980.
[25] R. E. Thayer, *The Biopsychology of Mood and Arousal*. Oxford, U.K.: Oxford Univ. Press, 1989.
[26] L. Lu, H.-J. Zhang, and S. Li, "Content-based audio classification and segmentation by using vector machines," *ACM Multimedia Syst. J.*, vol. 8, no. 6, pp. 482–492, 2003.
[27] C. E. Williams and K. N. Stevens, "Emotions and speech: some acoustical correlates," *J. Acoust. Soc. Amer.*, vol. 52, no. 4, pp. 1238–1250, 1972.
[28] R. Kamien, *Music: An Appreciation (5th Edition)*. New York: McGraw-Hill, 1992.
[29] D. Liu, L. Lu, and H.-J. Zhang, "Automatic music mood detection from acoustic music data," in *Proc. Int. Symp. Music Information Retrieval (ISMIR03)*, 2003.
[30] K. Hevner, "Expression in music: a discussion of experimental studies and theories," *Psychol. Rev.*, vol. 42, pp. 186–204, 1935.
[31] G. Kreurz, "Basic emotions in music," in *Proc. 6th Int. Conf. Music Perception Cognition*, 2000.
[32] E. Lindstrom and P. N. Juslin *et al.*, "Expressivity Comes From Within Your Soul: A Questionnaire Study of Student's Perception on Musical Expressivity," *Research Studies in Music Education*, vol. 20, pp. 23–47, 2003.
[33] R. B. Dietz and A. Lang, "Æffective agents: effects of agent affect on arousal, attention, liking & learning," in *Proc. Int. Cognitive Technology Conf.*, 1999, [Online]. http://www.cogtech.org/CT99/dietz.htm.
[34] M. Greenwald, E. Cook, and P. Lang, "Affective judgment and psychophysiological response: dimensional covariation in the evaluation of pictorial stimuli," *J. Psychophysiol.*, vol. 3, pp. 51–64, 1989.
[35] A. Hanjalic and L.-Q. Xu, "Affective video content representation and modeling," *IEEE Trans. Multimedia*, vol. 7, no. 1, pp. 143–154, Feb. 2005.
[36] D. M. Hinn, "The effect of the major and minor mode in music as a mood induction procedure," M.S. thesis, Virginia Polytechnic Inst. State Univ., Blacksburg, 1996.
[37] S. Watanabe, "Karhunen-Loeve expansion and factor analysis, theoretical remarks and applications," in *Trans. 4th Prauge Conf. Information Theory, Statistical Decision Functions, Random Processes*, 1965, pp. 645–660.
[38] D. Huron, "The ramp archetype and the maintenance of auditory attention," *Music Percept.*, vol. 10, no. 1, pp. 83–92, 1992.
[39] E. Scheirer, "Tempo and beat analysis of acoustic musical signals," *J. Acous. Soc. Amer.*, vol. 103, no. 1, pp. 588–601, 1998.
[40] J. Canny, "A computational approach to edge detection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. PAMI-8, no. 6, pp. 679–698, Jun. 1986.
[41] J. Maher and J. Beauchamp, "Fundamental frequency estimation of musical signals using a two-way mismatch procedure," *J. Acoust. Soc. Amer.*, vol. 95, no. 4, pp. 2254–2263, 1994.
[42] A. McCallum *et al.*, "Improving text classification by shrinkage in a hierarchy of classes," in *Proc. Int. Conf. Machine Learning*, 1998, pp. 359–367.
[43] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning: Data Mining, Inference and Prediction*. New York: Springer-Verlag, 2001.
[44] J. P. Campbell, "Speaker recognition: a tutorial," *Proc. IEEE*, vol. 85, no. 9, pp. 437–1462, Sep. 1997.

Lie Lu (M'05) received the B.S. and M.S. degrees in electrical engineering from Shanghai Jiao Tong University, Shanghai, China, in 1997 and 2000, respectively.

Since 2000, he has been with Microsoft Research Asia, Beijing, China, where he is currently an Associate Researcher with the Speech Group. His current research interests include pattern recognition, content-based audio analysis and indexing, and content-based music analysis. He has authored more than 40 publications in these areas and has 10 patents or pending applications.

Mr. Lu served as a member of Technical Program Committee of the IEEE International Conference on Multimedia and Expo in 2004.



Dan Liu received the M.S. degree from the Department of Automation, Tsinghua University, Beijing, China, in 2003. She is currently pursuing the Ph.D. degree in the Department of Cognitive Science, University of California at San Diego, La Jolla.

She was a visiting student at Microsoft Research Asia, Beijing, China, from 2002 to 2003. She is interested in the computational modeling of people's perception by using signal processing, pattern recognition, and statistical learning. She is also interested in applying optimal control theory in modeling people's motor control.

Hong-Jiang Zhang (S'90–M'91–SM'97–F'04) received the B.S. degree from Zhengzhou University, Zhengzhou, China, and the Ph.D. degree from the Technical University of Denmark, Lyngby, Denmark, both in electrical engineering, in 1982 and 1991, respectively.

From 1992 to 1995, he was with the Institute of Systems Science, National University of Singapore, where he led several projects in video and image content analysis and retrieval and computer vision. He also worked at the Massachusetts Institute of Technology Media Laboratory, Cambridge, in 1994 as a Visiting Researcher. From 1995 to 1999, he was a Research Manager at Hewlett-Packard Laboratories, where he was responsible for research and technology transfers in the areas of multimedia management, intelligent image processing, and Internet media. In 1999, he joined Microsoft Research Asia, Beijing, China, where he is currently a Senior Researcher and Assistant Managing Director in charge of media computing and information processing research. He has authored three books, over 200 refereed papers and book chapters, seven special issues of international journals on image and video processing, content-based media retrieval, and computer vision, as well as numerous patents or pending applications.

Dr. Zhang is a member of ACM. He currently serves on the editorial boards of five IEEE/ACM journals and a dozen committees of international conferences.