Happiness Detection in Music Using Hierarchical SVMs with Dual Types of Kernels

Yu-Hao Chin, Chang-Hong Lin, Ernestasia Siahaan, and Jia-Ching Wang

Department of Computer Science and Information Engineering
National Central University, Taiwan, R.O.C.

***Abstract*- In this paper, we proposed a novel system for detecting happiness emotion in music. Two emotion profiles are constructed using decision value in support vector machine (SVM), and based on short term and long term feature respectively. When using short term feature to train models, the kernel used in SVM is probability product kernel. If the input feature is long term, the kernel used in SVM is RBF kernel. SVM model is trained from a raw feature set comprising the following types of features: rhythm, timbre, and tonality. Each SVM is applied to targeted emotion class with calm emotion as the background class to train hyperplanes respectively. With the eight hyperplanes trained from angry, happy, sad, relaxed, pleased, bored, nervous, and peaceful, each test clip can output four decision values, which are then regarded as the emotion profile. Two profiles are fusioned to train SVMs. The final decision value is then extracted to draw DET curve. The experiment result shows that the proposed system has a good performance on music emotion recognition.**

Keywords- Music emotion, support vector machine, happiness verification

I. INTRODUCTION

Listening to music plays an important role in people's daily life. People can gain much benefit from listening to music. Music can make people have various feelings, such as angry, sad, etc. The property of music can be applied to many areas, such as education, inspiration production, therapy, and marketing. Recently, music information retrieval technologies have made great progress. Most researchers focus on detecting emotion in music. Music emotion can be retrieved by extracting and analyzing the emotion content of music. However, people's emotion is subjective. An objective music emotion evaluation is needed. Therefore, machine learning technology is applied in this topic. In this paper, we try to make a system to detect happiness content in music.

Many researches have been proposed in music emotion detection [1]. Existing research method could be divided into two main kinds, including dimension approach and categorical approach. Dimension approach maps features to a point on the emotion model plane [4], then regard the coordinate index as features to train regressors. Categorical approach works by directly feeding features into a classifier to recognize the corresponding emotion categories [5]. This paper uses the method belonging to the second type.

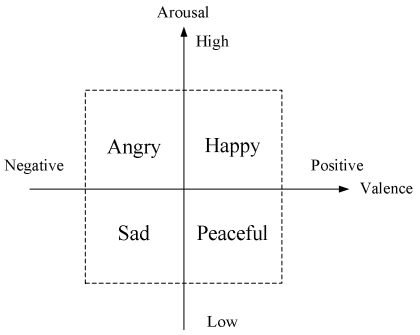

Fig. 1. Thayer's 2-D arousal-valence plane

In previous studies, a number of classifiers have been used, such as support vector regression (SVR) [5], Gaussian mixture models (GMM) [6], KNN [24], and support vector machines (SVMs) [7].

We consider Thayer's arousal-valence plane (See Fig. 1) in this paper [8]. According to the plane, human's emotion can be classified into four main classes. In Thayer's model, the *x*-axis refers to valence, and the *y*-axis refers to arousal. Valence refers to the grade of positive or negative perceived emotion. Arousal refers to the intensity of perceived emotion [7] [9].

It's noted that emotion perception is not based on a single feature but a combination of features [4] [18]. In our feature set, Legendre-Based Trend Coefficients (LBTCs) is given in the form of long-term statistics, and others are short-term feature. If all features' temporal are similar, they can be combined by simply merging feature vectors. However, it's harder to solve the problem when features' temporal are different. In previous work, most solution is to calculate long-term statistics of short-term features. Then combine short-term feature with long-term feature [25]. Instead of calculating statistics of short-term features, we train SVM models with dual types kernel based on two terms of feature respectively. Because of the characteristic of short-term feature, it is trained by using probability product kernel. Long-term feature is modeled with SVM using RBF kernel.

II. System Overview

Our system's process is presented in Fig. 2. Firstly, raw feature is extracted from music signal by using MIR toolbox. Because each feature's scale is different, we use central limit theorem [16] to normalize them. After normalization, features are divided into short-term feature and long-term classes. We use SVM to transform both two terms of features into emotion profiles with different kernel. Short term feature is trained by using probability product kernel, and long term feature is modeled by using RBF kernel. Finally, SVM is adopted to fusion profiles and recognize testing data.

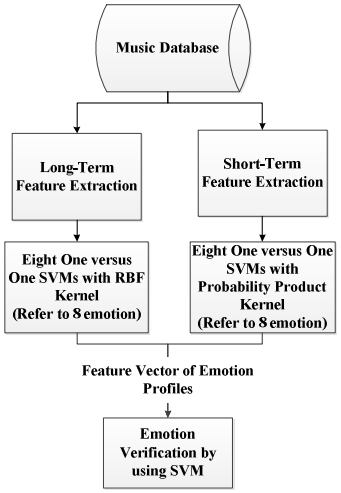

Fig. 2. Processes of proposed method

III. Feature Extraction and Emotion Profile

A. Feature Extraction

In previous work, Cooper and Foote extract Mel-Frequency Cepstrum Coefficients (MFCCs) from music signal, and they found that MFCCs is similar to music timbre expression [26]. There are 8 kinds of feature extracted from music clips in this paper, which are divided into short term feature and long term feature. Short-term feature includes of Eventdensity, Zerocross, Chromagram, MFCC, Spectrum Centroid, Spectrum Spread, and Spectrum Flatness. Long term-feature includes LBTCs. Features are extracted from the data by using MIR toolbox [2].

1) Short-Term Feature

The feature set comprise of three main types, i.e. Rhythm, Timbre, and Tonality. Rhythm includes Eventdensity. Timbre includes Zerocross, Chromagram, MFCC, Spectrum Spread, Spectrum Centroid, and Spectrum Flatness. Tonality consists of LBTCs. Eventdensity means peaks picked per second on onset curve [2]. Zerocross refers to the number of times a signal changes signs. Chromagram is also called Harmonic Pitch Class Profile. Chromagram presents the distribution of energy along the pitches or pitch classes [2]. Mel-frequency cepstrum coefficients (MFCC) also have good performance on the system. This feature models the human auditory perception system. The origin of MFCCs is the power of Mel windows. Spectrum centroid is an economical description of the shape of the power spectrum [20-22]. Additionally, it is correlated with a major perceptual dimension of timbre; i.e. sharpness. Spectrum spread is an economical descriptor of the shape of the power spectrum that indicates whether it is concentrated in the vicinity of its centroid, or else spread out over the spectrum [20-22]. It allows differentiating between tone-like and noise-like sounds. Spectrum Flatness describes the flatness properties of the spectrum of an audio signal within a given number of frequency bands. The flatness of a band is defined as the ratio of the geometric mean and the arithmetic mean of the spectral power coefficients within the band [20-22]. A high deviation from a flat shape may indicate the presence of tonal components.

2) Long-Term Feature

Long-term feature includes LBTCs. With the MFCCs and the subband powers, this study further proposes new audio features based on Legendre polynomial [23]. These new features which are called Legendre-based trend coefficients (LBTCs) are derived by computing the trend of each MFCC coefficient and 4^{th} subband power using Legendre polynomial. To trend of each one dimensional feature, we use (4) to derive its LBTC.

$$LBTC(\mathbf{x}(n)) = \sum_{n=0}^{\Gamma} \mathbf{x}(n) P_i(n), 0 \le i \le 3 \tag{4}$$

$$p_0(n) = (\frac{n}{\Gamma})(\Gamma+1)^{-0.5} \tag{5}$$

$$p_1(n) = \left[\frac{12M}{(\Gamma+1)(\Gamma+2)}\right](\frac{n}{\Gamma} - 0.5) \tag{6}$$

$$p_2(n) = \left[\frac{180\Gamma^3}{(\Gamma-1)(\Gamma+1)(\Gamma+2)(\Gamma+3)}\right]^{0.5} \times \left(\frac{n^2}{\Gamma^2} - \frac{n}{\Gamma} - \frac{\Gamma-1}{6\Gamma}\right) \tag{7}$$

$$p_3(n) = \left[\frac{2800\Gamma^5}{(\Gamma-2)(\Gamma-1)(\Gamma+1)(\Gamma+2)(\Gamma+3)}\right]^{0.5} \times \left[\frac{n^3}{\Gamma^3} - \frac{3n^2}{2\Gamma^2} + \frac{(6\Gamma^3 - 3\Gamma + 2)n}{10\Gamma^3} - \frac{(\Gamma-1)(\Gamma-2)}{20\Gamma^2}\right] \tag{8}$$

where n is the sample index, Γ is the total sample number, $P_i(n)$ and $\mathbf{x}(n)$ are the i^{th} order of Legendre polynomial and the signal to extract its trend.

B. Emotion Profile

The emotional profiles express the confidence of each of the eight emotion classes. It is an approach to interpret the emotional content of natural human expression by providing multiple probabilistic class labels, rather than a single hard label. For example, happy emotion not only contains

happiness content, but also consists of feature properties that are similar to the content of peaceful. The similarity to peaceful may cause data to be recognized as an incorrect class. Therefore, the evidence for happiness and peaceful are conveyed through representation in an emotion profile. The profile can help determine which emotion is most easily be perceived by human. In this paper, emotion content of signal is presented in terms of a set of simple emotion classes: happy, angry, sad, bored, nervous, relaxed, pleased, and peaceful [15].

IV. Hierarchical SVMs with Dual Types of Kernels

The proposed system includes two sets of SVM. One is based on short term feature with probability product kernel. The other is based on long term feature with RBF kernel. Each set of classifiers consists of eight SVMs, and classifier outputs are used to construct emotion profiles.

A. Support Vector Machine

The SVM theory is an effective statistical technique and has drawn much attention on audio classification tasks [7]. A SVM is a binary classifier that creates an optimal hyperplane to classify input samples. This optimal hyperplane linearly divides the two classes with the largest margin [10]. Denote $T=\{(\mathbf{x}_i,y_i),\ i=1,\cdots,N\}$ as a training set for SVM, each pair $(\mathbf{x}_i,y_i)$ means training sample $\mathbf{x}_i$ belongs to a class $\mathbf{y}_i$ where $\mathbf{y}_i$ $\{+1,-1\}$. The fundamental concept is to choose a hyperplane which can classify T accurately while maximizing the distance between the two classes. Finally, the decision function of classifying a new data point x can be written as:

$$\begin{aligned} f(\mathbf{x}) &= \text{sign}((\mathbf{w},\mathbf{x})+b) \\ &= \text{sign}(\sum_{i=1}^{m}\alpha_i\mathbf{x}_i\cdot\mathbf{x}+b) \\ &= \text{sign}(\sum_{i=1}^{m}\alpha_i k(\mathbf{x}_i\cdot\mathbf{x})+b) \end{aligned} \tag{9}$$

B. Probability Product Support Vector Machine

Functions that satisfy Mercer's theorem can be used as kernels. and is a kernel function. Using Mercer's theory, we can introduce a mapping function $\varphi(\mathbf{x})$, such that $k(\mathbf{x}_j,\mathbf{x}_i)=\phi(\mathbf{x}_j)\phi(\mathbf{x}_i)$. This provides the ability of handling nonlinear data, by mapping the original input space $\mathbf{R}^d$ into some other space.

We utilize the probability product kernel [27] defined as

$$k(p_i,p_j)=\int_{\mathbf{x}} p_i^{\rho}(\mathbf{x})p_j^{\rho}(\mathbf{x})d\mathbf{x}=\left\langle p_i^{\rho},p_j^{\rho}\right\rangle_{L_2} \tag{10}$$

where L_2 is a Hilbert space, and ρ is a positive constant. Probability product kernel allows us to introduce prior knowledge of data.

In this paper, we assume a d-dimensional Gaussian distribution of our data in the form

$$\begin{aligned} p(\mathbf{x}) &= N(\boldsymbol{\mu},\boldsymbol{\Sigma}) \\ &= (2\pi)^{-d/2}\,|\boldsymbol{\Sigma}|^{-1/2}\exp(-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})) \end{aligned} \tag{11}$$

The probability product kernel between two Gaussian distributions $p_i(\mathbf{x})=N(\boldsymbol{\mu},\boldsymbol{\Sigma}_i)$ and $p_j(\mathbf{x})=N(\boldsymbol{\mu},\boldsymbol{\Sigma}_j)$ thus becomes

$$\begin{aligned} k(p_i,p_j) &= \int_{\mathrm{R}^d} p_i^{\rho}(\mathbf{x})p_j^{\rho}(\mathbf{x})dx \\ &= (2\pi)^{(1-2\rho)d/2}\rho^{-d/2}\,|\boldsymbol{\Sigma}_{\varsigma}|^{1/2}|\boldsymbol{\Sigma}_i|^{-\rho/2}|\boldsymbol{\Sigma}_j|^{1-\rho/2}\times \\ &\quad \exp(-\frac{\rho}{2}(\boldsymbol{\mu}_i^T\boldsymbol{\Sigma}_i^{-1}\boldsymbol{\mu}_i+\boldsymbol{\mu}_j^T\boldsymbol{\Sigma}_j^{-1}\boldsymbol{\mu}_j+\boldsymbol{\mu}_{\varsigma}^T\boldsymbol{\Sigma}_{\varsigma}^{-1}\boldsymbol{\mu}_{\varsigma})) \end{aligned} \tag{12}$$

where $\boldsymbol{\Sigma}_{\varsigma}=(\boldsymbol{\Sigma}_i^{-1}+\boldsymbol{\Sigma}_j^{-1})^{-1}$, and $\boldsymbol{\mu}_{\varsigma}=(\boldsymbol{\Sigma}_i^{-1}\boldsymbol{\mu}_i+\boldsymbol{\Sigma}_j^{-1}\boldsymbol{\mu}_j)^{-1}$. This paper further assumes that $\boldsymbol{\Sigma}=\sigma^2\mathbf{I}$ in (4).

C. Emotion Profile Construction and Happiness Detection

Emotion profile presents perception probability of each of the eight emotion-specific decisions. In this paper, the emotion profiles are constructed using decision value. The decision value of SVM represents the degree of similarity between model and testing data. The goodness of similarity measure can be used to find out which model fits the data most accurately. Using the emotion profile, the most probably perceived emotion in music can be detected [15]. SVM is adopted to fusion profiles and detect happiness emotion in testing data.

V. Experimental Results

Our database is constructed by collecting music clips from two websites: All Music Guide [11] and Last.fm [13]. The database consists of nine classes of emotion, including happy, angry, sad, bored, nervous, relaxed, pleased, calm, and peaceful. Calm is taken as the models' opposite site when training models. Each emotion class contains 120 music clips. Half of the songs are used as training data, and the others are used as testing data. In this paper, 480 music clips are tested. All of songs the are western music, and are encoded in 16 KHz WAV format.

The research's goal is to provide a system which can verify particular emotion in music. Happiness is the class to be recognized, while peaceful, angry, and sad are considered as the other classes. It is noted that support vector machines (SVMs) perform well on this topic [7]. We also use SVM, which is based on LIBSVM library [14]. For facing the challenge of having multiple emotions in a song [15], two emotion profile are constructed by decision value in support vector machine (SVM). Then both of two profile are

fusioned by SVM. Afterwards, we then decide if happy emotion is most probably perceived in a song.

We evaluate the performance of our proposed system in terms of equal error rate (EER). As mentioned before, music may contain multiple emotions. If we know which emotion class a song most likely belongs to, we may know the main emotion of the song.

The whole feature set dimension is 82. Fig. 4 shows the DET curve of our experiment. Clearly, the proposed method can achieve 15% EER. From our results, we see that the system performs well on music happy emotion verification.

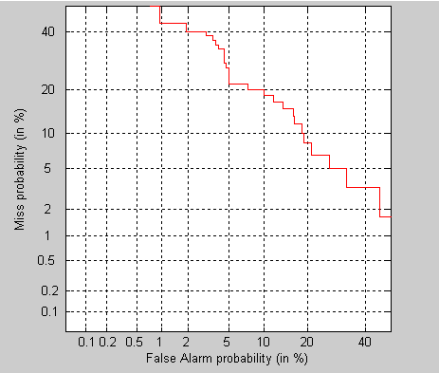

Fig. 4. DET curve of experimental result

VI. Conclusion

In this paper, we have proposed a system for detecting happy emotion in music. This work developed a novel feature extraction approach for emotional music recognition. To create a more distinguishable feature set, we extracted two emotion profiles by using the decision value of SVMs with different kernels, which are then regarded as new features to train new SVMs.

Detecting emotion in music has become the concern of many researchers in recent years. The technology can be applied in various areas, such as education, inspiration production, therapy, and marketing. In marketing, it is noted that a store's background music can be helpful in improving the store's image and make its employees happier [19].

References

[1] C. H. Yeh, H. H. Lin, and H. T. Chang, "An efficient emotion detection scheme for popular music," *IEEE Trans. Circuits and Systems, 2009. ISCAS 2009. Symposium*, pp. 1799-1802, 24-27 May 2009.

[2] O. Lartillot and P. Toiviainen, "MIR in Matlab (II): A toolbox for musical feature extraction from audio," in *Proc. Int. Conf. Music Information Retrieval*, pp. 127–130, 2007 http://users.jyu.fi/_lartillo/mirtoolbox/

[3] P. Saari, T. Eerola, and O. Lartillot, "Generalizability and simplicity as criteria in feature selection: application to mood classification in music," *IEEE Trans. Audio, Speech, and Language Processing,* vol. 19, no. 6, pp. 1802-1812, Aug. 2011.

[4] Y. H. Yang and H. H. Chen, "Prediction of the distribution of perceived music emotions using discrete samples," *IEEE Trans. Audio, Speech, and Language Processing,* vol. 19, no. 7, pp. 2184-2196, Sept. 2011.

[5] B. Han, S. Rho, and R. B. Dannenberg, and E. Hwang, "SMERS: Music emotion recognition using support vector regression," in *Proc. Int. Conf. Music Information Retrieval*, Kobe, Japan, 2009.

[6] L. Lie, D. Liu, and H. J. Zhang, "Automatic mood detection and tracking of music audio signals," *IEEE Trans. Audio, Speech, and Language Processing,* vol. 14, pp. 5-18, 2006.

[7] C. Y. Chang, C. Y. Lo, C. J. Wang, and P. C. Chung, "A music recommendation system with consideration of personal emotion," in *Proc. Int. Conf. Computer Symposium* (*ICS*), Dec. 2010, pp. 18-23, 16-18.

[8] R. E. Thayer, *The Biopsychology of Mood and Arousal*. New York: Oxford University Press, 1989.

[9] A. Gabrielsson, "Emotion perceived and emotion felt: Same or different?" *Musicae Scientiae*, pp. 123–147, 2002, special issue.

[10] V. Vapnik, Statistical Learning Theory, New York: Wiley, 1998.

[11] B. Shao, M. Ogihara, D. Wang, and T. Li, "Music recommendation based on acoustic features and user access patterns," *IEEE Transactions on Audio, Speech, and Language Processing,* vol. 17, pp. 1602-1611, 2009.

[12] "The All Music Guide," Available: http://www.allmusic.com.

[13] "Last.fm," Available : http://cn.last.fm/home

[14] C. C. Chang and C. J. Lin, "LIBSVM: a library for support vector machines," 2001. Available: http://www.csie.ntu.edu.tw/ ~cjlin/libsvm

[15] E. Mower, M. J Mataric, and S. Narayanan, "A framework for automatic human emotion classification using emotion profiles," *Audio, Speech, and Language Processing, IEEE Transactions,* vol. 19, no. 5, pp. 1057-1070, July 2011.

[16] M. Chouchane, S. Paris, F. L. Gland, C. Musso, and D. T. Pham, "On the probability distribution of a moving target. Asymptotic and non-asymptotic results," in *Proc. Int. Conf. Information Fusion* (*FUSION*), July 2011, pp. 1-8, 5-8.

[17] M. Zhao, S. Li, "Sparse Representation Classification for Image Text Detection," *Int. Conf. Syst. Computational Intelligence and Design*, 2009, vol. 1, pp.76-79.

[18] K. Hevner, "Expression in music: A discussion of experimental studies and theories," *Psychological Review*, vol. 48, no. 2, pp. 186–204, 1935.

[19] R. E. Milliman, "Using background music to affect the behavior of supermarket shoppers." *Journal of Marketing*, vol. 46, no. 3, pp. 86-91, Summer 1982.

[20] H. G. Kim, N. Moreau, and T. Sikora, MPEG-7 Audio and Beyond: Audio Content Indexing and Retrieval. New York: Wiley, 2005.

[21] M. A. Casey, "MPEG-7 sound-recognition tools," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 11, no. 6, 737-747, 2001.

[22] ISO-IEC/JTCI SC29 WGII Moving Pictures Expeti Group. Information technology - multimedia content description interface - part 4: Audio. Comittee Draft 15938-4, ISO/IEC, 2000.

[23] C. F. Wu, "Bimodal emotion recognition from speech and facial expression," Master Thesis, Department of Computer Science and Information Engineering, National Cheng Kung University, 2002.

[24] F. C. Hwang, J. S. Wang, P. C. Chung, and C. F. Yang, "Detecting emotional expression of music with feature selection approach," in *Proc. Int. Conf. Orange Technologies* (*ICOT*), March. 2013, pp. 282-286, 12-16.

[25] I. Luengo, E. Navas, and I. Hernáez, "Feature analysis and evaluation for automatic emotion identification in speech," *IEEE Trans. Multimedia,* vol. 12, no. 6, pp. 490-501, Oct. 2010.

[26] C. Y. Lin and S. Cheng, "Multi-theme analysis of music emotion similarity for jukebox application," in *Proc. Int. Conf. Audio, Language and Image Processing* (*ICALIP*), July 2012, pp. 241-246, 16-18.

[27] T. Jebara, R. Kondor, and A. Howard, "Probability product kernels," *Journal of Machine Learning Research*, vol. 5, pp. 819–844, July 2004.