Music Genre Classification Algorithm Based on Dynamic Frame Analysis and Support Vector Machine

Shih-Hao Chen
Dept. of Information Engineering
I-Shou University
Kaohsiung County 840, Taiwan
e-mail: s94631115@stu.edu.tw

Shi-Huang Chen
Dept. of Computer Science & Information Engineering
Shu-Te University
Kaohsiung County 824, Taiwan
e-mail: shchen@stu.edu.tw

Rodrigo Capobianco Guido
Institute of Physics of São Carlos - University of São Paulo,
São Carlos,
São Paulo 13566-590, Brazil.
e-mail: guido@ifsc.usp.br

***Abstract*—This paper proposes a new music genre classification algorithm based on dynamic music frame analysis and support vector machine (SVM). The dynamic music frame analysis could cover the long-term and the short-term music genre features which can represent the time-varying behavior of music signals. The music genre features used in this paper are mel-frequency cepstral coefficient (MFCC) and log energy with dynamic frame length. The dynamic music frame analysis will be applied to train an optimized non-linear decision rule for music genre classifier via SVM. Experimental results show that the proposed new music genre classification algorithm could achieve the average classification accuracy rate of 98% for the six different music genres, including classic, dance, lullaby, Bossa, piano, and blue.**

Keywords- music genre classification; mel-frequency cepstral coefficient (MFCC); support vector machine (SVM)

I. INTRODUCTION

With the rapidly expansion of digital music contents, the music information retrieval (MIR) system has been one of popular research topics in the past years. The automatic analysis of music signals is one of the most important parts of MIR [4]. In general, the first step of automatic music analysis is to make use of several characteristics that can capture the information about music content. Among these characteristics, music genre information is regarded as a principal one. The music genres are the top-level descriptors or labels used by music dealers and librarians for categorizing and describing the vast universe of music [2]. It can be used to describe music as well as to structure music database [3]. Usually, music genres have the properties that are related to the instrumentation and rhythmic structure of music. However, due to music genres do not have strict definitions, most of current musical genre annotations are performed manually [1, 2]. It goes without saying that the manual classification of music genres is quite time-consuming and demanding. Therefore, many different approaches of automatic music genre classification have been proposed in recent years [6].

Most of these proposed systems are based on pattern recognition techniques. Each music signal is represented by numerical features that are then used to train a music classifier via traditional statistical models such as Gaussian mixture models (GMM) and *K*-nearest neighbors (KNN) [2]. On the other hand, various content-based timbral features were proposed for music genre classification. These timbral features could be categorized into short-term and long-term features [2]. The short-term features, which can represent the spectrum of music, include spectral centroid, spectral rolloff, mel-frequency cepstral coefficient (MFCC), and etc. The long-term features, which can characterize either the variation of spectral shape or beat information, include low-energy [1], and beat histogram, and etc [2][5].

Based on the above discussions, this paper proposed a new music genre classification method using dynamic frame analysis of timbral features and support vector machine (SVM). The proposed dynamic frame analysis consists of both the long-term and short-term timbral features. They are log energy and MFCC with dynamic frame length and overlap. In addition, the proposed method uses more recent powerful classification method, namely SVM, as a classifier instead of traditional statistical model. Various experimental results show that the use of SVM can perform significant improvements in music genre classification accuracy.

The remainder of this paper is organized as follows. The introductions to the timbral features and dynamic frame analysis are described in Sections II and III, respectively. Section IV will briefly review SVM and the training process used in this paper. Section V illustrates various experimental results on the six kinds of different music genres. Finally, conclusions are given in Section VI.

II. TIMBRAL FEATURES

Music is a harmonic composition of different sounds from various musical instruments with different intensities. Generally, the bandwidth of music signal is between 40 Hz and 15K Hz. Figure 1 shows the sample spectrums of the six kinds of music used in this paper. They are classic, dance, lullaby, Bossa, piano, and blue notes. It can be summarized that music is composed of vocal, percussion, brass, woodwinds, strings, and pipe organ. Each type of musical instruments or sounds has their own frequency range.

From the viewpoint of speech recognition, these instruments and sounds can be regarded as a spectrum expanded version of speech sounds. Therefore, the dynamic frame analysis used to represent music genre can be based on

978-0-7695-4217-1/10 $26.00 © 2010 IEEE
DOI 10.1109/ISM.2010.61

the features proposed for speech recognition or music-speech discrimination. This research selects two common usage features, namely mel-frequency cepstral coefficient (MFCC) and log energy, with dynamic frame length as the timbral features for music genre classification.

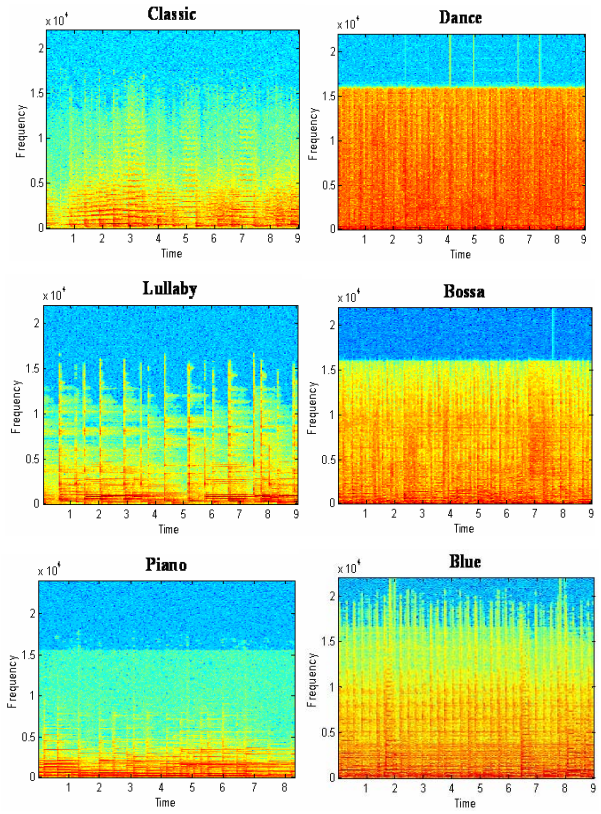

Figure 1. The sample spectrums of the six kinds of music used in this paper.

A. MFCC

It is shown that MFCC can capture the acoustic characteristics for speech recognition, music classification, and other audio/speech related applications [7-8]. According to the previous psychophysical studies, human perception of the frequency content of sounds follow a subjectively defined nonlinear scale called the "mel" scale [9] defined as,

$$f_{\text{mel}} = 1125 \ln(1 + \frac{f}{700}) \tag{1}$$

where f is the actual frequency in Hz. This leads to the definition of MFCC and its calculation process is given as follows.

Let $s(n)$, n = 1~N, be a music signal frame that is pre-emphasized and Hamming-windowed [8, 9]. First, the time domain signal, $s(n)$, is transferred into frequency domain by an M point discrete Fourier transform (DFT). The resulting energy spectrum can be represented as

$$|S(k)|^2 = \left| \sum_{n=1}^{M} s(n) \cdot e^{\left(\frac{-j2\pi nk}{M}\right)} \right|^2 \tag{2}$$

where $1 \le k \le M$. Then, the triangular filter banks, whose frequency bands are linearly spaced in the mel scale, are imposed on the spectrum obtained in (2). The outputs $e(l)$, l = 1~Q, of the mel-scaled band-pass filters can be calculated by a weighted summation between respective filter response $H_i(k)$, i = 1~M, and the energy spectrum $|S(k)|^2$ as

$$e(i) = \sum_{k=1}^{M} |S(k)|^2 \cdot H_i(k) \tag{3}$$

where $H_i(k)$ is defined as

$$H_i(k) = \begin{cases} 0, & \text{for} \quad k < f_{b(i-1)} \\ \dfrac{(k - f_{b(i-1)})}{(f_{b(i)} - f_{b(i-1)})}, & \text{for} \quad f_{b(i-1)} \le k < f_{b(i)} \\ \dfrac{(f_{b(i+1)} - k)}{(f_{b(i+1)} - f_{b(i)})}, & \text{for} \quad f_{b(i)} \le k < f_{b(i+1)} \\ 0, & \text{for} \quad k > f_{b(i+1)} \end{cases} \tag{4}$$

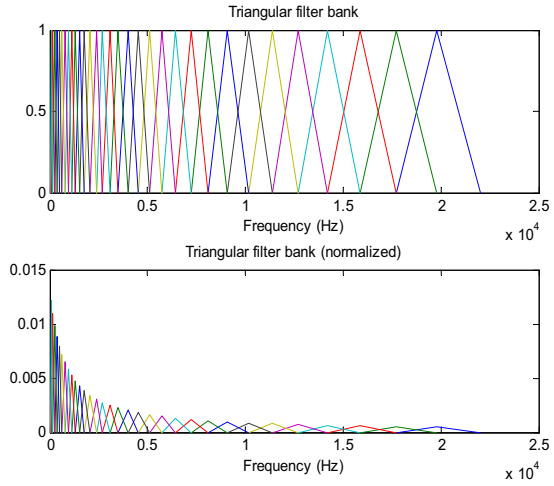

Figure 2. Original (upper one) and normalized (lower one) mel-space triangular filter bank (M=32)

In (4), $f_{b(i)}$ are the boundary points of the filters and are depended on the sampling points Fs and the number of points N in DFT:

$$f_{b(i)} = \left(\frac{N}{F_s}\right) \cdot f_{\text{mel}}^{-1}\left(f_{\text{mel(low)}} + i\frac{f_{\text{(mel)high}} - f_{\text{(mel)low}}}{M+1}\right). \tag{5}$$

Here, $f_{\text{mel(low)}}$ and $f_{\text{mel(high)}}$ are respectively the low and high boundary frequencies for the entire filter bank. f_{mel}^{-1} is the inverse to (1) transformation, formulated as

$$f_{\text{mel}}^{-1} = 700\left[e^{\left(\frac{f_{\text{mel}}}{1125}\right)} - 1\right] \tag{6}$$

Figure 2 shows the original as well as normalized mel space triangular filter bank with M = 32. Finally, discrete cosine transform (DCT) is taken on the log filter bank energies, log[$e(l)$], and the MFCC coefficients C_m can be written as,

$$C_m = \sqrt{\frac{2}{Q}}\sum_{p=0}^{Q-1}\log[e(p+1)]\cdot\cos\left[m\cdot\left(\frac{2p-1}{2}\right)\cdot\frac{\pi}{Q}\right] \tag{7}$$

where $0 \le m \le M-1$. The summary of MFCC calculation process is illustrates given in Figure 3.

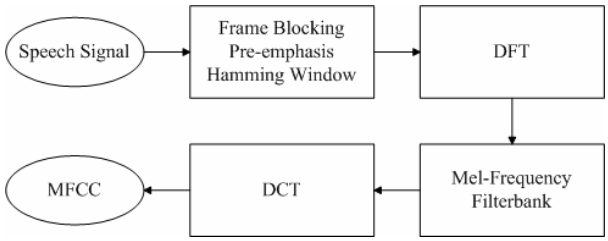

Figure 3. The block diagram of MFCC calculation process.

B. Log Energy

The log energy is usually cooperated with MFCC for applications of speaker recognition and audio segmentations [10]. The definition of log energy used in this paper is defined in (8).

$$E = \log\left(\sum_{n=0}^{N-1} s[n]^2\right) \tag{8}$$

where N is the number of music samples in a frame.

III. Dynamic Frame Analysis

Because the traditional music genre classification is depended on the features which are extracted from a fix frame length or signal time-frequency analysis, it can not well represent the various kinds of music genre. Therefore, this paper developed a new technology, called dynamic frame analysis, to solve this problem. The dynamic parameters used in this technology include frame size (FS), frame overlap (FO), triangular filter banks (TFB), and cepstral vector dimension (CVD). The detail descriptions of these four parameters are given as follows.

1. FS: Depending on the sampling rate of music signal, the FS used in this paper is changed from 11ms to 372ms.
2. FO: Similar to FS, the FO used in this paper could be changed from 1ms to 186ms.
3. TFB: The TFB is the M parameter shown in equations (2) and (3). In this paper, the TFB could be varied from 16 to 64.
4. CVD: It follows from equation (7) that the results of DCT, C_m, and delta cepstrum will be affected by the parameter m. Hence, this paper varies m from 12 to 50.

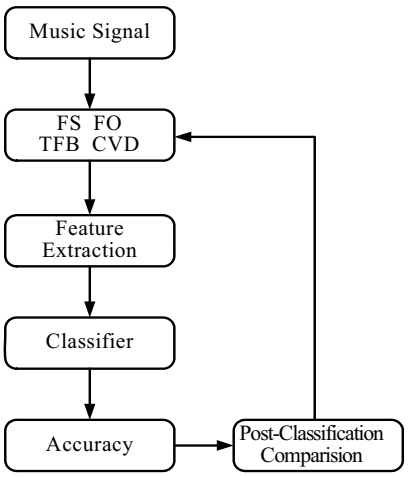

Figure 4. The flowchart of the dynamic frame analysis.

Figure 4 shows the flowchart of the proposed dynamic frame analysis. First, it will set up the ranges of the dynamic parameters, namely frame size (FS), frame overlap (FO), Triangular filter banks (TFB), and cepstral vector dimension (CVD). Then it will extract timbral features from the music frame with the dynamic parameters. The classifier will generate a training model and output a pre-classification result. This pre-classification result will compare to the post-classification comparison and adjust these four dynamic parameters. This processing will be terminated when the accuracy could not be further improved.

IV. Support Vector Machine

It has been shown that SVM has superb performance at binary classification tasks and handle large dimensional feature vectors better than other classification methods [7]. Basically, SVM is designed to search for a hyper-plane that can separate two classes (+1 and −1 classes usually labeled) with maximum margin. SVM can be constructed from sums of a known kernel function $K(\cdot, \cdot)$ to define such a hyper-plane [11].

$$f(\mathbf{x}) = \sum_{i=1}^{N}\alpha_i y_i K(\mathbf{x}, \mathbf{x}_i) + b \tag{9}$$

where $y_i \in \{1,-1\}$ are the target values, $\sum_{i=1}^{N}\alpha_i y_i = 0$, and $\alpha_i > 0$. The vectors $\mathbf{x}_i \subseteq R^n$ are the support vectors and

could be obtained from the training.

Many hyper-planes can achieve the above separation purpose. However, the SVM used in this paper is aimed to find the one that could maximize the margin (the minimal distance from the hyper-plane to each point). To achieve this purpose, the soft-margin SVM, which includes slack variables $\xi_i \geq 0$, is used in this paper. Figure 5 shows the slack variables, where ξ_i is defined as

$$\xi_i = \max(0, \gamma - y_i(< w, x_i > + b)). \quad (10)$$

where the parameter ξ_i can measure the amount by which the training set fails to have margin γ, and take into account any misclassification of the training data. Consequently, the training process tolerates some points misclassified and is suitable in most classification cases.

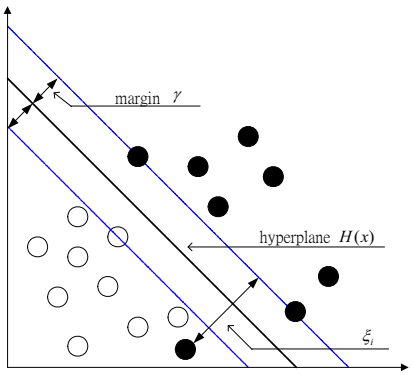

Figure 5. The margin and the slack variable for a classification problem.

Based on the above definitions, SVM is a two-class classifier and therefore is not suitable for the task of music genre classification. To modify the original two-class SVM for multi-class music genre classification, this paper makes use of pair-wise comparison. In pair-wise comparison, a classifier is trained for each possible pair of classes. For example, if there has K music genre classes, this results in $(K-1)K/2$ binary classifiers. Then one can execute the multi-class classification by evaluating all of these $(K-1)K/2$ individual classifiers and assigning the instance to the class which gets the highest number of votes.

TABLE I. THE FOUR COMMON KERNEL FUNCTIONS USED IN SVM.

Function Type	**Definition**
Linear function	$K(x, \bar{x}) = x^T \bar{x}$
Exponential radial basis function (ERBF)	$K(x, \bar{x}) = \exp(-\lvert x - \bar{x} \rvert / 2\sigma^2)$
Polynomial function	$K(x, \bar{x}) = (< x, \bar{x} > + 1)^d$

Table I lists the three common kernel functions for the SVM feature mapping where parameter σ^2 is the variance of the Gaussian function, and parameter d is the degree of the polynomial. The performances of proposed music genre classification using these three kernel functions are given in Table III.

The dynamic frame analysis proposed in this paper consists of MFCC and log energy with dynamic frame length to cover short-term as well as long term features of music genre. In addition, the MFCC used in this paper has variable mel-scale triangular filter banks from 16 to 64.

TABLE II. THE TRAINING SET USED IN THIS PAPER.

Genre	**Track**
Classical	PIANO CONCERTO NO.1 IN B FLAT MINOR THE NUTCRACKER SWAN LAKE TRIO SONATA FOR ORGAN BRANDENBURG CONCERTO NO.2
Dance	KEEN ON DISCO / MISSY EVA GOOD FRIEND / DJ DOVE FEAT. SUNSHINE IN THE RAIN / CHERRY TIME FOR DANCE / DJ DOVE FEAT. MR. V APOLOGIZE / DJ ELVIS JUMPING ALL OVER THE WORLD / SCOOTER DISCOTHEQUE / YAMBOO BIG GIRLS DON'T CRY / NICK SKITZ FEAT. DANNI
Lullaby	ANGELS/ ROBBIE WILLIAMS BETTER MAN/ ROBBIE WILLIAMS FEEL/ ROBBIE WILLIAMS LOVE SOMEBODY/ ROBBIE WILLIAMS SING/ CARPENTERS SUPERSTAR/ CARPENTERS YESTERDAY ONCE MORE/ CARPENTERS PLEASE MR. POSTMAN/ CARPENTERS
Bossa	CATUPIRY LITORAL DE SOL O SAMBA CLEA O CORSO ORIXA O GUARDA-CHUVA AZUL TAIYO NO KODOMOTACHI
Piano	ENDINGS - MICHAEL JONES SUNRISE - KOSTIA DARK EYES - WAYNE GRATZ BELOVED - DAVID LANZ WATER CIRCLES - MIA JANG MINOR TRUTHS - FRED SIMON POETIC JUSTICE - SHEILA LARKIN BETHEL - PAUL CARDALL
Blue	CANTALOOP (FLIP FANTASIA)/US3 STEPPIN' INTO TOMORROW/MADLIB WON'T YOU OPEN UP YOUR SENSES/HORACE SILVER LEMON/TROUBLEMAKERS MIDNIGHT BLUE/KENNY BURRELL MY FUNNY VALENTINE/CHET BAKER BLUE TRAIN/JOHN COLTRANE MOANIN/ART BLAKEY & THE JAZZ MESSENGERS

The training set used in this paper is listed in Table II. There are totally 45 music songs applied to SVM training. Each music track is digitized using 44.1 KHz sampling rate with 16-bit resolution. For each music track in the set for

training, several timbral feature vectors can be obtained. Assume that n_T is the number of the obtained MFCC and log energy vectors. The training set T is then defined to be the $n_T \times (M+1)$ array with row vectors being these M-order MFCC vectors.

Let $T(i, j)$denote the (i, j)-position of T. Use this array T to construct another $n_T \times (M+1)$ array T' whose (i, j) position $T'(i, j)$ is defined to be $T'(i, j) = T(i, j) - \mu_j$, where $\mu_j = \sum_i T(i, j) / n_T$ is the mean of column j. Next, one can normalize T' by computing $T^N(i, j) = T'(i, j) / m_j$, where m_j is the maximum of the absolute value of elements in column j. Thus, each timbral feature vector will have similar weights after the normalization process.

V. Experimental Results

The experimental results of the proposed music genre classification method are achieved by using 300 songs over 6 genres, *i.e.*, classic, dance, lullaby, Bossa, piano, and blue notes. Among these 300 songs, the 45 songs as listed in Table II are used in SVM training process and the remaining 255 songs are applied to evaluate classification performance of the proposed method. All of these songs are noisy-free and are sampled at 44.1 KHz with 16-bit resolution. Each testing song consists of 9 seconds long music. The accuracy rate used in this paper is defined as

$$\frac{\text{Number of corrected frames}}{\text{Totoal number of frames}} \times 100\% \qquad (11)$$

TABLE III. The average accuracy rates of the proposed music genre classification method using 45 training songs and SVM with linear, polynomial, and exponential radial basis kernel functions.

Kernel function	Accuracy rate
Linear Function	93.6%
Polynomial Function	96.4%
Exponential Radial Basis Function	99.8%

TABLE IV. Performances of the proposed music genre classification method using SVM with ERBF.

Genre Type	Average Accuracy Rate
CLASSICAL	98.2 %
DANCE	98.5 %
LULLABY	97.4 %
BOSSA	96.1 %
PIANO	98.9 %
BLUE	98.6 %
TOTAL AVERAGE	98.0 %

Table III shows the average accuracy rates of the proposed music genre classification method using 45 training songs and SVM with linear, polynomial, and exponential radial basis kernel functions. In addition, the MFCC order M = 32 and 41 ms frame size are also determined by the dynamic music frame analysis during the SVM training process. It is obvious the use of exponential radial basis kernel function (ERBF) has the best performance and hence was selected for the proposed music genre classification method. Table IV shows the performances of the proposed music genre classification method using SVM with ERBF on the remaining 255 songs. The proposed method can achieve the average accuracy rate of 98% for the classification of six music genres.

VI. Conclusions

In this paper, the dynamic frame analysis and support vector machine (SVM) are applied to the task of music genre classification. The proposed dynamic frame analysis includes two kinds of timbral feature vectors which consist of both the long-term and the short-term features and hence can represent the time-varying behavior of music. Using SVM with the exponential radial basis kernel function, the proposed method can achieve the average accuracy rate of 98% for the classification of six music genres. They are classic, dance, lullaby, Bossa, piano, and blue notes. This paper will try to use other music features and extend the use of other music genres in the future.

References

[1] A. Flexer, “A closer look on artist filters for musical genre classification,” In Proceedings of the 8th International Conference on Music Information Retrieval (ISMIR'07), Vienna, Austria, 2007.

[2] George Tzanetakis and Perry Cook, “Musical genre classification of audio signals,” IEEE Transactions on Speech and Audio Processing, 10(5):pp.293-302, 2002.

[3] Nicolas Scaringella, Giorgio Zoia, and Daniel Mlynek, “Automatic genre classification of music content: a survey, IEEE Signal Processing Magazine, Vol. 23, Issue 2, pp. 133-141, 2006.

[4] Tao Li and G. Tzanetakis, “Factors in automatic musical genre classification of audio signals,” In Proceeding of 2003 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics, pp.143-146, Oct. 2003.

[5] Dalwon Jang; Minho Jin; C.D. Yoo, “Music genre classification using novel features and a weighted voting method,” In Proceeding of the 2008 IEEE International Conference on Multimedia and Expo, pp. 1377-1380, April 2008.

[6] A. Meng, P. Ahrendt, J. Larsen, L.K. Hansen, “Temporal Feature Integration for Music Genre Classification,” IEEE Trans. on Speech and Audio Processing, Vol. 15, No. 5, pp. 1654 - 1664, 2007.

[7] C.C. Lin, S.H. Chen, T. K. Truong, and Yukon Chang, “Audio Classification and Categorization Based on Wavelets and Support Vector Machine,” IEEE Trans. on Speech and Audio Processing, Vol. 13, No. 5, pp. 644-651, Sept. 2005.

[8] L. Rabiner and B. H. Juang, Fundamentals of Speech Recognition, Prentice Hall, 1993

[9] S. B. Davis and P. Mermelstein, ”Comparison of Parametric Representation for Monosyllabic Word Recognition in Continuously Spoken Sentences”, IEEE Trans. On ASSP, vol. ASSP 28, no. 4, pp. 357-365, Aug. 1980.

[10] Chia-Ping Chen and J.A. Bilmes, “MVA Processing of Speech Features,” IEEE Transactions on Audio, Speech, and Language Processing, Vol. 15, No. 1, pp.257-270, Jan. 2007.

[11] V.N. Vapnik, Statistical Learning Theory, John Wiley & Sons, 1998.