MUSIC GENRE CLASSIFICATION USING NOVEL FEATURES AND A WEIGHTED VOTING METHOD

Dalwon Jang, Minho Jin, and Chang D. Yoo

Div. of EE, School of EECS, KAIST, 373-1 Guseong Dong, Yuseong Gu, Daejeon 305-701, Korea
{dal1, jinmho}@kaist.ac.kr and cdyoo@ee.kaist.ac.kr

ABSTRACT

This paper proposes a novel music genre classification system based on two novel features and a weighted voting method. The proposed features, modulation spectral flatness measure (MSFM) and modulation spectral crest measure (MSCM) represent the time-varying behavior of music signals and indicate the strength of beat. The proposed weighted voting method determines a music genre by summarizing the classification results of consecutive time windows. Experimental results show that the proposed features and the weighted voting method improve the music genre classification system.

1. INTRODUCTION

With the recent proliferation of digital music, there is increasing demand for efficient management of a large digital music database, and a music genre is considered as one possible way of management. A music genre, a high-level descriptor of music, is extensively used in music stores, radio stations, and the Internet. The manual classification of the music genre is a laborious and time-consuming work. As an alternative, the automatic content-based music genre classification system is receiving increased attention, and as a consequence, a number of automatic content-based music genre classification systems have been proposed [1]-[6].

For music genre classification system, various content-based features were proposed. The features are categorized into short-term feature and long-term feature. The short-term features, that mainly characterize the spectrum, include spectral centroid[1], spectral rolloff[1], Mel-frequency cepstral coefficient (MFCC)[1], octave-based spectral contrast (OSC) [3], etc. The long-term features, that mainly characterize either the variation of spectral shape or beat information, include low-energy[1], Daubechies wavelet coefficients histogram (DWCH) [5], octave-based modulation spectral contrast (OMSC) [2], beat histogram [1], etc. A feature vector, which is used as the input of multi-class classifier, consists of both the long-term features and the statistics of the short-term features. The short-term features are computed in a short window in which music signal is assumed to be statistically stationary, and their statistics are computed in a long window. In this paper, the short window is denoted as "*analysis window*", and the long window is denoted as "*texture window*" as in [1].

This paper proposes two long-term features: modulation spectral flatness measure (MSFM) and modulation spectral crest measure (MSCM). They are modified versions of spectral flatness measure (SFM) and spectral crest measure (SCM) where the SFM and the SCM are the features that represent the short-term spectrum. Instead of representing the spectrum, the proposed features represent the long-term subband energy variation using the modulation spectrum analysis [7].

This paper also proposes a novel method to combine the information of the *texture windows* of a music clip. A *texture window* is regarded as a basic unit of classification because a feature vector is obtained in a *texture window* by computing statistics of short-term features. In order to improve the classification performance, it is necessary to combine the information of the *texture windows*. For this, the method to use the statistics of feature vectors of *texture windows* and the majority voting method have been used [1, 4, 6]. In this paper, the weighted voting method using the probability, that each *texture window* is classified into each genre class, is proposed.

The remainder of this paper is organized as follows. Section 2 explains the features that have been proposed in the past and proposes novel features. Section 3 explains the traditional methods to combine the information of the *texture windows* and proposes a weighted voting method. Section 4 presents experimental results, and finally Section 5 concludes the paper.

2. FEATURE EXTRACTION

2.1. Timbral texture feature

In [1], timbral texture features were proposed for audio classification. The features include the MFCC, spectral centroid, spectral rolloff, spectral flux, zero crossings, and low-energy. The first 5 features are short-term features, thus statistics of

This work was partly supported by the IT R&D program of MIC/IITA [2007-S-017-01, Development of user-centric contents protection and distribution technology], and grant No. R01-2007-000-20949-0 from the Basic Research Program of the Korea Science and Engineering Foundation.

features are computed in a *texture window*. The last feature is a long-term feature, thus it is computed in a *texture window*. The MFCC, which have been widely used in speech recognition, represents the spectral characteristics based on Mel-frequency scaling[8], and it is also effective to classify music genre and used in various literatures such as [1, 2, 5]. The spectral centroid is the centroid of amplitude spectrum. The spectral rolloff is the frequency bin below which 85% of the spectral distribution is concentrated. The spectral flux is the squared difference of successive amplitude spectrum. The zero crossings is the number of time domain zero crossing of the music signal. The low-energy is the percentage of *analysis windows* that have energy less than the average energy across the *texture window*.

2.2. Octave-based spectral contrast(OSC)

The OSC represents the strength of spectral peaks and spectral valleys in each sub-band separately, so that it could represent the relative spectral characteristics [3]. Spectral peaks and spectral valleys are obtained from the amplitude spectrum, thus discrete Fourier transform (DFT) is first applied on the music signal, and the spectrum is divided into octave-based subband. Spectral peaks and spectral valleys are estimated by averaging the values in the small neighborhood around maximum and minimum values of the amplitude spectrum respectively. The spectral contrast is defined as difference between spectral peak and spectral valley. The statistics of spectral peaks and spectral contrasts are used for the feature vector of *texture window*.

2.3. Octave-based modulation spectral contrast (OMSC)

The OMSC[2] is extracted using long-term modulation spectrum analysis[7]. The amplitude spectrum of music signal is divided into octave-based subband. For each subband, the modulation spectrum is obtained by applying DFT on the sequence of the sum of the amplitude spectrum. From spectral peaks and spectral contrasts of the modulation spectrum, the OMSC is obtained.

2.4. Proposed features

The MSFM and the MSCM, obtained in a *texture window*, are proposed for music genre classification. They are modified version of the SFM and the SCM. The SFM and the SCM, which are obtained in an *analysis window*, represent the short-term spectrum, and the detailed explanation about the SFM and the SCM is found in [9]. The MSFM and the MSCM are obtained from long-term modulation spectrum[7], and they describe the time-varying behavior of subband energy, thus they are comparable with the OMSC. To define the MSFM and the MSCM, the energy in the ith octave-based subband and the qth *analysis window* ($q = 1, 2, \cdots, Q$), $E_q[i]$ is first defined as

$$E_q[i] = \sum_{f=L_i}^{H_i} |M[f]|^2 \tag{1}$$

where $M[f]$ is the amplitude spectrum of music signal at the fth bin, and L_i and H_i are the lowest and the highest frequency bins of the ith subband. The MSFM for the ith subband is defined as

$$MSFM[i] = \frac{\sqrt[Q/2]{(\prod_{q=1}^{Q/2} A_q[i])}}{\frac{\sum_{q=1}^{Q/2} A_q[i]}{Q/2}}, \tag{2}$$

and the MSCM for the ith subband is defined as

$$MSCM[i] = \frac{\max(A_q[i])_{q=1}^{Q/2}}{\frac{\sum_{q=1}^{Q/2} A_q[i]}{Q/2}} \tag{3}$$

where $A_q[i]$ is the amplitude of modulation spectrum, which is obtained by applying DFT on the sequence of $E_q[i]$ along q axis, at the qth bin. When the beat of a input music is strong, the value of $MSFM[i]$ is close to 0, and the value of $MSCM[i]$ is high. Thus, the beat strength in a *texture window* can be represented by the proposed features, and the features help to divide the music clips which have the strong beat and the music clips which does not.

3. WEIGHTED VOTING

Various methods have been used to efficiently combine the information of *texture windows*, and as shown in Fig. 1, the methods can be classified into two categories: statistics and voting methods. In statistics method shown in Fig. 1 (a), the feature vector of a music clip is obtained by computing the statistics of feature vectors of *texture windows*, and the music clip is classified using the feature vector. The method to compute the mean of feature vectors of *texture windows* [1] and the method to compute both the mean and the standard deviation of feature vectors of *texture windows* [6] fall under the category of statistics method. In voting method shown in Fig. 1 (b), the classification results of *texture windows* are combined via voting. The majority voting [4] and the proposed weighted voting fall under the category of voting method.

3.1. Mean of feature vectors (MF) [1]

In this method, feature vectors of *texture windows* are averaged, and the mean of feature vectors is used as the feature vector of a music clip.

3.2. Mean and standard deviation of feature vectors (MSF) [6]

In this method, the feature vectors of *texture windows* are collected, and the mean and standard deviation of feature vectors are used as the feature vector of a music clip.

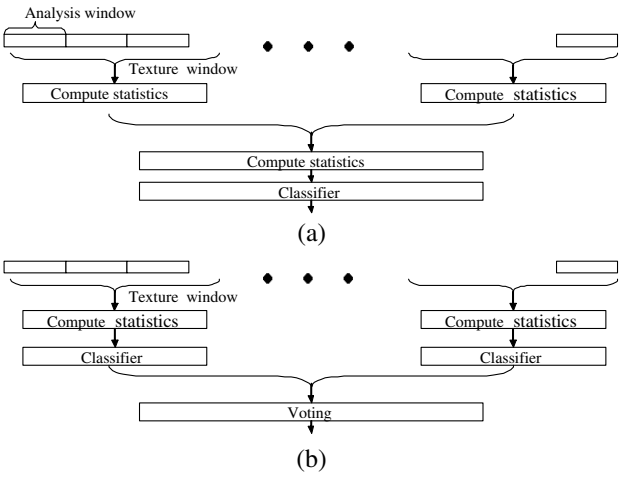

Fig. 1. (a) Statistics method: The feature vector of a *texture window* is obtained by computing the statistics of features of *analysis windows*. After that, the feature vector of a music clip is obtained by computing the statistics of feature vectors of *texture windows*. Using the feature vector, the music clip is classified. (b) Voting method: After obtaining the feature vectors of *texture windows*, *texture windows* are classified. By voting the classification results of *texture windows*, the genre class of a music clip is determined.

3.3. Majority voting (MV) [4]

In this method, after the feature vector of a *texture window* are extracted, classification is performed for each *texture window*. Thus, a number of classification results are obtained for a music clip where the classification result of a *texture window* is one-best genre class. A genre of the input music clip is determined by majority voting of the classification results.

3.4. Weighted voting (WV)

In the proposed weighted voting method, each *texture window* is classified, and the probability that the *texture window* is classified into each genre class is obtained and used as the weight. Given K genre classes of data, the probability that the nth *texture window* ($n = 1, \cdots, N$) is classified as the kth genre is defined as

$$p_k^{(n)} = P(y^{(n)} = k|\mathbf{x}^{(n)}),\ k = 1, 2, \cdots, K \qquad (4)$$

where $\mathbf{x}^{(n)}$ and $y^{(n)}$ are the feature vector and the classification result of the nth *texture window*. The genre class which has the maximum of $\sum_{n=1}^{N} p_k^{(n)}$ is selected as a classification result of a music clip. The probabilities are estimated by combining all pairwise comparisons [10], and they are derived from the following optimization problem:

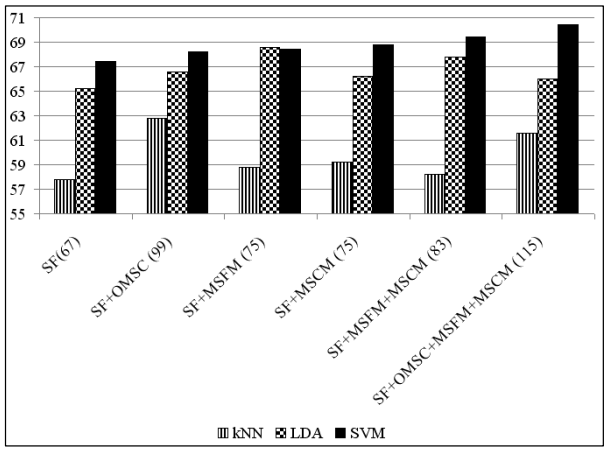

Fig. 2. Classification accuracy of various feature sets using 3 classifiers. The number within parentheses indicates dimension of feature vector. "SF" denotes two features of the OSC and timbral texture features. Most of SF features are short-term features.

$$\min \frac{1}{2} \sum_{k=1}^{K} \sum_{j:j \neq k} (r_{jk}^{(n)} p_k^{(n)} - r_{kj}^{(n)} p_j^{(n)})^2$$
$$\text{subject to } \sum_{k=1}^{K} p_k^{(n)} = 1, p_k^{(n)} \geq 0, \forall i \qquad (5)$$

where $r_{kj}^{(n)}$ is the estimated pairwise class probabilities which is mathematically expressed by $r_{kj}^{(n)} \approx P(y^{(n)} = k|y^{(n)} = k \text{ or } j, \mathbf{x}^{(n)})$. The solution of the optimization problem is detailed in [10].

4. EXPERIMENTAL RESULTS

4.1. Experimental setup

Our data set contains 1000 songs over 10 genres: 100 songs per a genre. The genres are Classic, Jazz, R&B, Country, Rock, Hiphop, Metal, Dance, Newage, and Electronica. For classification, 30 second of music after initial 30 second is used. The audio files in data set are converted to 44100Hz, 16-bit, and mono signal.

The length of *analysis window* is about $93ms$, and it has 50% overlap. The length of *texture window* is about $3s$, thus it contains 63 *analysis windows*. Eight octave bands referring to [2] are used. The classification results are obtained using LIBSVM[11], which is a library for support vector machine (SVM), and MATLABArsenal toolkit[12].

4.2. Performance of the proposed features

In Fig. 2, the classification accuracy of various feature sets are presented using three different classifiers: k-nearest neighbor

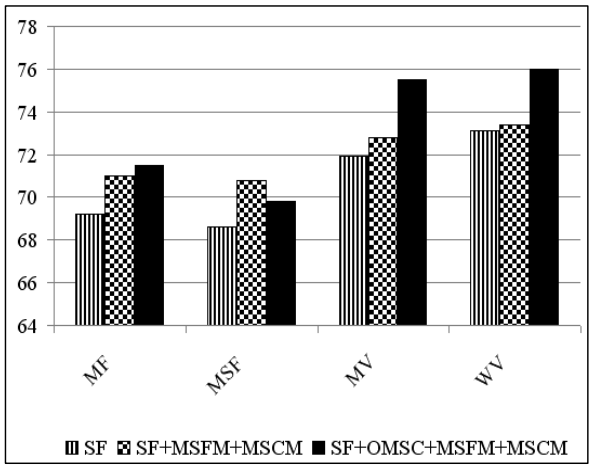

Fig. 3. Classification accuracies of various methods to combine the information of *texture windows*.

(k-NN), linear discriminant analysis (LDA) , and SVM classifiers. For k-NN classifier, we set $k = 5$, and for SVM classifier, we use the pairwise SVM classifier with linear kernel. For training, 50 songs for each genre are chosen, and the other 50 songs are used for test. The MF method is used in this experiment. The figure shows that addition of the proposed features increases the accuracy and that the proposed features are more efficient than the OMSC when combined with the OSC and timbral texture features. When the proposed features are combined with the all features explained in Section 2, the best result 70.4% is obtained using SVM classifier.

4.3. Performance of the weighting voting method

Fig. 3 compares the accuracy of four methods using some feature sets and a pairwise SVM classifier with linear kernel. The accuracy are obtained via 10-fold cross validation. As shown in the figure, the weighted voting method obtains the highest accuracy. The accuracies of the voting methods are always higher than those of the statistics methods. The accuracy of the weighted voting method is higher than that of the majority voting method. Using the all features explained in Section 2 and the WV method, the best result 76.0% is obtained.

5. CONCLUSION

This paper proposes a music genre classification system based on both two novel features and a weighted voting method. Two proposed features represent the time-varying behavior of octave-based subband energy. In the weighted voting method, which intends to combine the information of consecutive time windows, the probability that each time window is classified into each genre class is computed, and the genre of a music clip is determined using the probability. Experimental results show that the proposed features are more efficient than the OMSC when combined with traditional features and that the weighted voting method is better than other methods. Thus, the proposed features and the proposed weighted voting method improve the music genre classification system.

6. REFERENCES

[1] 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.

[2] C-H. Lee, J-L. Shih, K-M. Yu, and J-M Su, "Automatic music genre classfication using modulation spectral contrast feature," *Proc. ICME 07*, 2007

[3] D. N. Jiang, L. Lu, H. J. Zhang, J. H. Tao, and L. H. Cai, "Music type classification by spectral contrast feature," *Proc. ICME 02*, vol. 1, pp. 113-116, 2002.

[4] K. West and S. Cox, "Features and classifiers for the automatic classification of musical audio signals," *ISMIR04*, 2004

[5] T. Li, M. Ogihara, and Q. Li, "A comparative study on content-based music genre classification" *Proc. ACM Conf. on Research and Development in Information Retrieval*, pp. 282-289, 2003

[6] G. Tzanetakis, "MARSYAS Submissions to MIREX 2007," [Online] Available: http://www.music-ir.org/mirex/2007/abs/AI_CC_GC_MC_AS_tzanetakis.pdf

[7] S. Sukittanon, L. E. Atlas, and J. W. Pitton, "Modulation-scale analysis for content identification," *IEEE Trans. on Signal Processing*, vol. 52, no. 10, pp. 3023-3035, Oct., 2004.

[8] X. Huang, A. Acero and H.-W. Hon, "Spoken Language Processing," Prentice Hall PTR, 2001

[9] G. Peeters, "A large set of audio features for sound description (similarity and classification) in the CUIDADO project," *CUIDADO I.S.T Project Report*, 2004

[10] T.-F. Wu, C.-J. Lin, and R.C. Weng, "Probability estimates for multi-class classification by pairwise coupling," *Journal of Machine Learning Research*, 5:975-1005, 2004.

[11] C.-C. Chang and C.-J. Lin, "LIBSVM - A Library for Support Vector Machines," [Online] Available: www.csie.ntu.edu.tw/˜cjlin/libsvm/

[12] R. Yan, "MATLABArsenal: A MATLAB Package for Classification Algorithms," [Online] Available: http://finalfantasyxi.inf.cs.cmu.edu/MATLABArsenal/MATLABArsenal.htm