Automatic Classification of Musical Genres Using Inter-Genre Similarity

Ulaş Bağcı, Student Member, IEEE, and Engin Erzin, Senior Member, IEEE

Abstract—Musical genre classification is an essential tool for music information retrieval systems and it has potential to become a highly demanded application in various media platforms. Two important problems of the automatic musical genre classification are feature extraction and classifier design. In this letter, we propose two novel classifiers using inter-genre similarity (IGS) modeling and investigate the use of dynamic timbral texture features in order to improve automatic musical genre classification performance. Inter-genre similarity is modeled over hard-to-classify samples of the musical genre feature space. In the classification, samples within inter-genre similarity class are eliminated to reduce inter-genre confusion and to improve genre classification performance. Experimental results show that the proposed classifiers provide better classification rates than the existing methods.

Index Terms—Inter-genre similarity (IGS) modeling, Mel-frequency cepstral coefficients (MFCC), musical genre classification.

I. INTRODUCTION

GENRE classification is crucial for the categorization of musical pieces. Automatic musical genre classification has important applications in professional media production, radio stations, audio-visual archive management, entertainment and recently on the Internet. Although musical genre classification is done manually and it is hard to precisely define the specific content of a musical genre, it is generally agreed that audio signals of music belonging to the same genre contain certain common characteristics since they have similar harmonic and rhythmical language. These common characteristics have motivated recent research activities to improve automatic musical genre classification [1]–[8]. The problem is inherently challenging as the human identification rates after listening 3 s samples are reported to be around 70% [9].

Feature extraction and classifier design are two important problems of the automatic musical genre classification. Timbral texture features, which represent short-time spectral information, rhythmic content features including beat and tempo, and pitch content features are thoroughly investigated in [1]. High-level musical features including instrumentation, texture, rhythm, dynamics, pitch statistics, melody and chords are investigated in [5]. Another novel feature extraction method is proposed in [3], in which local and global information of music signals are captured by computation of histograms on their Daubechies wavelets coefficients for the characterization of genre, emotion, style, and similarity informations. A comparison of human and automatic musical genre classification is presented in a recent work [4]. Mel-frequency cepstral coefficients (MFCC) are used for modeling and discrimination of music signals [1]. Linear prediction cepstrum coefficients, zero-crossing rates, Mel-frequency cepstral coefficients, spectral power, amplitude envelope, spectrum flux, and cepstrum flux are investigated as features to characterize music content for automatic classification of pure and vocal music [7]. Various classifiers such as K-nearest neighbor (KNN) and Gaussian mixture model (GMM) classifiers [1], [3], construction of decision trees [10], multi-class AdaBoost [8] and support vector machines (SVM) [3], [7] are employed for automatic musical genre classification. In another study, radial basis function (RBF) networks with a combination of unsupervised and supervised initialization methods are used for fast training and classification of musical genre [6].

In this study, we extend the boosting idea, which was proposed in [11], to use inter-genre similarity information for the design of discriminative classifiers. In order to improve discrimination between musical genre classes, hard-to-classify samples are used to model an inter-genre similarity class. During the classification, the samples within the inter-genre similarity class are discarded to reduce inter-genre confusion and to improve the classification rates. Experimental results over the musical genre dataset compiled by Tzanetakis [1] show that the proposed method provides better classification rates than the existing methods.

The organization of the letter includes a brief description of the feature extraction in Section II. The discriminative musical genre classification which uses the inter-genre similarity is discussed in Section III. Experimental results are provided in Section IV followed by discussions and conclusions.

II. FEATURE EXTRACTION

In the literature, features for musical genre classification are examined under two groups: timbral texture and rhythmic content features. Timbral features represent short-time properties, such as spectral information and zero-crossings. Rhythmic content features represent long-term properties including beat, tempo, pitch content, etc. The instrumentation of a music performance has a significant influence in genre recognition. The timbre of constituent sound sources is reflected in the spectral distribution of a music signal. Hence, spectral features
are widely used in the literature. An extensive investigation of both timbral texture and rhythmic content features can be found in [1].

In this study, timbral texture features, which characterize short-term spectral properties, are considered to represent musical genres. Short-time analysis are performed over 25 ms overlapping audio windows for the extraction of timbral texture features for each 10 ms frame. Hamming window of size 25 ms is applied to the audio window to remove edge effects. The timbral texture features, including 13-dimensional Mel-frequency cepstral coefficient (MFCC) vector $F_M$ and four-dimensional spectral shape vector $F_S$ with spectral centroid, spectral roll-off, spectral flux and zero-crossing rate, are extracted from the analysis window. The resulting combined timbral texture feature vector is represented by $F_T = [F_M | F_S]^T$.

Temporal changes in the spectra play an important role in human hearing perception. One way to capture this information is to extract derivative features, which measure the change in short-term spectra over time. Although the derivative features are widely used in speech recognition, they are not considered for musical genre classification. In this study, we also investigate the use of derivative features for musical genre classification. The derivative feature of the $i$th analysis window is calculated using the following regression formula:

$$
\Delta F(i) = \sum_{k=1}^{K} k [F(i+k) - F(i-k)]
$$

where the number of analysis windows in the regression computation is set to $2K + 1 = 5$. The timbral texture feature vector is extended to include the first- and second-order derivative features, and the resulting dynamic feature vector is represented as $F_D = [F | \Delta F | \Delta \Delta F]^T$.

III. MUSIC CLASSIFICATION USING INTER-GENRE SIMILARITY

Music signals that belong to the same genre contain certain common characteristics, as they are composed of similar types of instruments with similar rhythmic patterns. These common characteristics are captured with statistical pattern recognition methods to build automatic musical genre classifiers [1]–[4]. Since the boundaries between genre types are not clearly defined, musical genre classification is a challenging problem. One can expect to observe similar spectral content and rhythmic patterns across different musical genre types, which increase confusion and mis-classification rates. Inter-genre similarity (IGS) modeling is proposed to decrease the level of confusion across similar musical genre types. In IGS modeling, first an inter-genre similarity class is extracted over hard-to-classify samples of the musical genre feature space. Then, the samples within inter-genre similarity class are discarded in the classifier to reduce the inter-genre confusion and to improve the genre classification performance. Inter-genre similarity modeling is further improved with iterative IGS modeling (IIGS). The IGS and IIGS models are described in the following subsections.

A. Inter-Genre Similarity (IGS) Modeling

The timbral texture features represent the short-term spectral content of music signals. Since music signals may include similar instruments and similar rhythmic patterns, no sharp boundaries exist between certain genre types. The inter-genre similarity modeling is proposed to capture the similar spectral contents among different genre types. Once the IGS clusters are statistically modeled, the IGS frames can be captured and removed from the decision process to reduce the inter-genre confusion.

Let $\lambda_1, \lambda_2, \ldots, \lambda_N$ be the $N$ different genre models in the database. Also let $p(f|\lambda)$ be the class-conditional probability density function, where $f$ and $\lambda$ are respectively the feature vector and the genre class model. The class-conditional probability density function is extracted by the GMMs with diagonal covariance matrices. The IGS clusters are constructed and the class-conditional statistical models are extracted using the following steps.

1) Perform the statistical modeling of each genre in the database using the available training data of the corresponding musical genre class.
2) Perform frame based genre identification task over the training data, and label each frame as a correct-classification or mis-classification.
3) Construct the statistical model, $\lambda_{IGS}$, for the IGS class over all the mis-classified frames among all the musical genre types.
4) Update all the $N$-class musical genre models, $\lambda_n$, over the correctly-classified frames.

Following these steps, $N$-class musical genre models and a single-class IGS model are constructed. In the musical genre classification process, one can find the most likely musical genre class, $\lambda^*$, given a sequence of features, $\{f_1, f_2, \ldots, f_k\}$, which are extracted from a decision window of music signal, by maximizing the weighted joint class-conditional probability

$$
\lambda^* = \arg \max_{\lambda_n} \frac{1}{\sum \omega_k} \sum_{k=1}^{K} \omega_k \log p(f_k|\lambda_n).
$$

The weights $\omega_k$ are defined based on the class-conditional IGS model

$$
\omega_k^{IGS} = \left\{ \begin{array}{ll}
1, & \text{if } p(f_k|\lambda_n) > p(f_k|\lambda_{IGS}) \\
0, & \text{otherwise}
\end{array} \right.
$$

where index $n$ runs over all $N$ genre classes and index $k$ runs over all features in the decision window.

The proposed weighted joint class-conditional probability maximization discards the IGS frames for each musical genre from the decision process. Therefore, the inter-genre confusion decreases and the genre classification rate increases with the resulting discriminative decision process. Supportive experimental results that show the improvements in the classification are presented in Section IV.

B. Iterative Inter-Genre Similarity (IIGS) Modeling

Inter-genre similarity modeling can be repeatedly used to fine tune the detection of hard-to-classify samples in the training data. In each iteration, a new IGS model is formed over the new set of mis-classified samples. In the decision process a frame is discarded if it matches any one of the IGS classes. The $T$-step iterative inter-genre similarity (IIGS) model is constructed using the following steps:

Authorized licensed use limited to: UNIVERSITY OF NOTTINGHAM. Downloaded on February 4, 2010 at 09:45 from IEEE Xplore.  Restrictions apply.
TABLE I  
AVERAGE CORRECT CLASSIFICATION RATES OF THE FLAT CLASSIFIER FOR VARYING NUMBER OF GMM MIXTURES AND DECISION WINDOW SIZES

<table>
<thead>
<tr>
<th>Feature</th>
<th>Flat Classification Rates for Varying Number of GMM Mixtures (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>3sec Decision Window</td>
</tr>
<tr>
<td>Sets</td>
<td>8 16 24 32 48</td>
</tr>
<tr>
<td>$F_M$</td>
<td>53.89 55.51 56.12 56.38 57.48</td>
</tr>
<tr>
<td>$F_D$</td>
<td>59.34 62.76 63.25 63.42 64.25</td>
</tr>
<tr>
<td>$F_T$</td>
<td>55.52 56.35 57.28 58.45 59.41</td>
</tr>
<tr>
<td>$F_P$</td>
<td>58.97 61.48 62.69 63.44 64.56</td>
</tr>
</tbody>
</table>

1) Perform IGS modeling, and get $\lambda_{IGS_i}$ and updated musical genre models $\lambda_n$ for all $N$ classes. Set $t = 2$.  
2) Perform frame based genre identification task with IGS modeling over the training data, and label each frame as a correct-classification, mis-classification or correct-IGS classification.  
3) Construct the statistical model, $\lambda_{IGS}$, over all the mis-classified frames among all the musical genre types.  
4) Update all the $N$-class musical genre models, $\lambda_n$, over the correctly-classified frames.  
5) Increment iteration counter $t$, and if $t \leq T$ go to step 2.  

The above creates $N$-class musical genre models and $T$-class IGS models. In the musical genre classification process, the decision is taken by maximizing the weighted joint class-condition probability in (2). In the IGS modeling weights of the joint class-condition probability, $\omega_n$, are defined as

$$\omega_{n}^{IGS} = \begin{cases} 1, & \text{if } p(f_k|\lambda_n^t) > p(f_k|\lambda_{IGS}) \forall t \\ 0, & \text{otherwise.} \end{cases} \hspace{1cm} (4)$$

IV. EXPERIMENTAL RESULTS

Evaluation of the proposed classification algorithms is performed over musical genre dataset of Tzanetakis [1], which includes ten different genre types: blues, classical, country, disco, hiphop, jazz, metal, pop, reggae and rock. In the dataset each genre includes 100 different representative audio segments with duration of 30 s for each of the 10 musical genre, resulting in a total duration of $10 \times 100 \times 30 = 30000$ s. The resulting timbral texture feature vectors are extracted for each 10 ms audio frame with an overlapping window of size 25 ms. The musical genre classification task is performed by maximization of the class-condition probability density function in (2) over two decision window sizes, 3 s and 30 s, where each decision window includes 300 and 3000 analysis windows (feature vectors), respectively. The experiments are performed using two-fold cross validation, where the database is randomly split into two equal subsets and the pair of subsets are used in two alternating train/test experiments. The average correct classification rates are reported. The proposed musical genre classification schemes, IGS and IIGS, are evaluated and compared with a flat classifier structure. The flat classifier is constructed by taking all the weights equal in the joint class-condition probability in (2), as $\omega_{kn} = \frac{1}{N} \forall k, n$.

Table I presents the average correct classification rates of the flat classifier for varying number of GMM mixtures and for different feature vectors. One can draw the following observations from these experiments: 1) The best classification rates are achieved using the dynamic feature vector $F_M^{30}$, which is the MFCC feature vector with its first and second derivatives. 2) There is a significant classification gain in the use of dynamic features. 3) The combined timbral texture feature vector $F_T$ has a better classification performance than the MFCC feature vector $F_M$, but the classification rates for the dynamic features are closer to each other and $F_P^{30}$ does slightly better than $F_M^{30}$. 4) Classification rates of the 30 s decision window are significantly better than the rates of the 3 s decision window and they are close to the rates that are reported in [1].

The average classification rates for the IGS classifier are presented in Table II. Note that, the IGS classification rates for varying number of GMM mixtures and different feature vectors are improved by 10–18 points with respect to the flat classifier. The best classification rates are achieved using the dynamic feature vector $F_M^{30}$ as 79.02% and 88.60% for 3 s and 30 s decision windows, respectively. Table III presents a combined confusion matrix for the flat and IGS classifiers for a better evaluation of the proposed musical genre classifier. Note that, the diagonal of the confusion matrix yields the average correct classification rates for individual genre types. The correct classification rates for the blues, country, disco, pop and reggae improve significantly and reach values higher than 80% with the IGS classifier. The rock genre type has the poorest classification performance at 67% with the IGS classifier, which is improved from 44% of the flat classifier.

The iterative IGS classifier, which improves the IGS classifier at the cost of representing IGS class with multiple GMM classifiers, is evaluated using the $F_M^{30}$ feature vector. Table IV presents the average correct classification rates of the IGS classifier for increasing number of iterations. Note that, in each iteration we observed a classification improvement and the best
The improvements of the IGS and IIGS classifiers are significant when compared to the results of [1] and [3], which report average correct classification rates of 61% and 78%, respectively, for 30 s decision windows using timbral texture features. Note that the proposed IGS and IIGS classifiers achieved 88.60% and 92.40% correct classification rates, respectively.

By considering genre classification decisions as $M$ Bernoulli events, statistical significance of the experimental results can be measured with the expected standard deviation $\sqrt{(p(1-p))/M}$, where $p$ is the average classification accuracy and $M$ is the total number of decisions. With average classification accuracy $p = 0.8$, the expected standard deviation of the classification accuracy is 0.40% for 3 s decision window and 1.26% for 30 s decision window.

### V. Conclusion

In this study, the use of dynamic timbral texture features for musical genre classification is evaluated and two novel classifier structures are proposed to improve automatic genre classification. The best classification rates are achieved by the MFCC feature vector with the first and second derivatives. In the proposed classifier structures, inter-genre similarity class is defined as the population of misclassified features from all musical genre types. Significant classification performance gains are observed by modeling and discarding these hard-to-classify features in the IGS class. The classification improvements strongly indicate that inter-genre spectral similarities across different musical genre types are successfully modeled. Experimental results that show classification improvements are provided.

### ACKNOWLEDGMENT

The authors would like to thank G. Tzanetakis for making available the musical genre database.

### REFERENCES


