Automatic Music Genre Classification Using Second-Order Statistical Measures for the Prescriptive Approach

‡Hassan Ezzaidi and ‡Jean Rouat
‡‡Ermetis, Université du Québec à Chicoutimi, Chicoutimi, Québec, Canada, G7H 2B1
†IMSI, Université de Sherbrooke, Sherbrooke, Québec, Canada, J1K 2R1
hezzaidi@uqac.uquebec.ca, Jean.Rouat@ieee.org

Abstract
Several works proposed for the automatic genre musical classification are based on various combinations of parameters, exploiting different models. However, the comparison of all previous works remain impossible since they used different target taxonomy or genre definitions and databases. In this paper, the world largest music database (Real World Computing) is used. Also, different measures related to second-order statistics methods are investigated to achieve the genre classification. Various strategies are proposed for training and testing sessions such as matched conditions, mismatched conditions, long training/testing, long training and short testing. For all experiments, the section of file used in testing has never been presented during the training session. The best classifier achieved 97% and 69% performance when matched and mismatched conditions are used, respectively.

1. INTRODUCTION
The considerable advances in audio technologies (iPod, DVD and MP3 readers...), the accessibility to information via the Internet and the increasing production of the large disk industry (Recording Industry Association of America (RIAA)... create many new needs to exploit this large musical universe. The classification music genre is considered as the most natural and popular need for many people regardless of their origins, languages, cultures, ages or sexes.

Musical Genre is widely used to categorize and label the extremely vast world of music. This task can be achieved by human experts for the music industry or by consumers themselves. As a result, many taxonomies are now available but remain at the same time not closely similar. This reality is related to the fact, that many descriptors and semantic ambiguity can be used to determine this classification [9]. For example, names of categories associated to each classification are not always similar or coherent including the hierarchical structure (subcategories) itself. The manufacturing of new instruments and the realization of new albums continue to complicate this tendency and indicate that taxonomy will remain in the future in an elastic and dynamic structure. As argued by Aucouturier et Pachet [9], genre may be used in intentional or extensional concept. For each concept, genre taxonomies is related to different interpretations on the descriptor level or semantical level. They describe three approaches to establish musical genre classification: manual classification (projection of human or expert knowledge), prescriptive approach (classifying genre as they are found), and finally emergent classification approaches that are based on a similarity measure to automatically produce the hierarchical genre structure.

A comparison between automatic and human genre classification on the same database has been recently investigated by Lippens et al. [7]. They showed that the classification is inherently subjective as reported by previous works [9]. They also experimented a parameter derived from an auditory model but the score remain similar to the one evaluated with the MFCC.

In the case of automatic genre classification, several works are proposed to extract genre information features from the acoustical music signal. The majority of them are inspired from speech/speaker recognition and music/speech discrimination. Several features and models proposed and experimented for genre classification can be found in [11] [12] [13] [6] [7]. Generally, they can be divided into three feature families. The first family represent the timbral texture of audio signal with spectral rolloff, spectral flux, time domain zero crossing, Mel Frequency Cepstral Coefficients (MFCC), Linear Predictive Coding (LPC), texture window and energy. The second family represent the rhythmic content features with wavelet transforms, where the audio signal is analyzed with different resolutions related to several bands. Also, statistical features on time and discrete wavelet time were proposed by Lambrou et al. [5]. The third family is based on pitch content features.

In this work (contribution), a new pattern recognition technique for music genre classification, based on function measures, is investigated. The prescriptive approach (classifying genre as they are found), is adopted in this work. The measures were derived from statistical models in different contexts. They were already used in the context of speaker identification systems but never used for musical genre classification. Moreover, the interest of the proposed technique resides primarily on the simplicity of its mathematical formalism, its potential to be implemented for real or differed time applications. It requires little memory capacity to store the reference prototypes (2 parameters), not much computing time and remains very flexible over the testing/training duration. Experiments are carried out according to different strategies, as matching and mismatching conditions, long training long testing and long training short testing. All the proposed measures are evaluated in their asymmetrical form with two prescriptive genre taxonomies.

2. Proposed function measures
The recognizer is based on second-order statistical methods that were initially proposed and tested in the context of text independent speaker identification by Bimbot et al. [1].

Let \( \{ m_R(i) \}_{i \leq M} \) be a sequence of M independent parameters vectors related to source information noted R, extracted from an acoustical signal. All vectors are p-dimensional, sup-
posed independent and distributed like a Gaussian function. Therefore, they are characterized in the parametric form only by 2 parameters: a mean vector noted \( \mu_R \) and a covariance matrix noted \( \Sigma_R \) as:

\[
\mu_R = \frac{1}{N} \sum_{i=1}^{N} m_R(i),
\]

\[
\Sigma_R = \frac{1}{N} \sum_{i=1}^{N} (m_R(i) - \mu_R) (m_R(i) - \mu_R)^T,
\]

where \( (\cdot)^T \) is the transpose.

Similarly, a sequence of \( N \) vectors \( \{m_T(i)\}_{1 \leq i \leq N} \) corresponds to a target information source and obeys to the same properties as the reference source. Hence, the target source can be represented by 2 parameters, that are the mean vector \( \mu_T \) and the covariance matrix \( \Sigma_T \).

The asymmetrized similarity measure noted \( \mu_G(T, R) \), derived from the averaged log-likelihood of \( N \) tested observations \[1\] is defined as:

\[
\mu_G(R, T) = \frac{1}{P} [tr(\Sigma_T \Sigma_R^{-1}) - log(\frac{det(\Sigma_T)}{det(\Sigma_R)})] - 1
\]

where:

- \( \mu_T - \mu_R = \Delta_m \), \( tr(\cdot) \) is the matrix trace and \( det(\cdot) \) is the matrix determinant.

A second measure considered as a variant measure of \( \mu_G(R, T) \), and based only on the covariance matrix is also experimented. Originally, this measure was supposed to be robust against channel noise and distortion. In this work, it is considered that the main problem is related to the high intraset variability. This measure is noted \( \mu_{Ge}(R, T) \) and defined as:

\[
\mu_{Ge}(R, T) = \frac{1}{P} [tr(\Sigma_T \Sigma_R^{-1}) - log(\frac{det(\Sigma_T)}{det(\Sigma_R)})] - 1
\]

The arithmetic-geometric sphericity measure is also used and defined as:

\[
\mu_{Se}(R, T) = \log(\frac{tr(\Sigma_T \Sigma_R^{-1})}{P}) - log(\frac{det(\Sigma_T)}{det(\Sigma_R)})^{-1/P}
\]

It was initially proposed and tested in the context of speaker recognition by Grenier \[4\] in the framework of text-dependent experiments and later used by Bimbot et al. \[1\] in the framework of text-independent speaker identification.

An absolute deviation measure proposed by Gish \[3\] computes the total deviation of the eigenvalues \( \lambda_i \) from the unity. The expression of this measure denoted \( \mu_{Dc}(R, T) \) is defined as:

\[
\mu_{Dc}(R, T) = \frac{1}{P} \sum_{i=1}^{P} (\lambda_i - 1)
\]

All measures presented until now, are characterized by the following properties:

- \( \mu_{Ge}(x, y) \geq 0 \); \( \mu_{Ge}(x, x) = 0 \) and \( \mu_{Se}(x, y) \neq \mu_{Se}(y, x) \).

Finally, a Battacharia distance that measure a similarity between two distributions is also used for experiment purposes \[10\]. This distance being defined by:

\[
\mu_{Be} = \frac{1}{8} \sum_{i} (\Sigma_R + \Sigma_T)^{-1} \Delta_m + \frac{1}{2} \log(\frac{det((\Sigma_R + \Sigma_T)/2)}{det(\Sigma_R)\sqrt{det(\Sigma_T)}})
\]

The motivation behind these measures is that musical signal is characterized mainly by rhythmicity and regularity during a long temporal period (without forgetting the rhythmicity periodicity).

### Table 1: Score recognition (in %) for matched conditions; long train/test.

<table>
<thead>
<tr>
<th>Categories</th>
<th>( \mu(G) )</th>
<th>( \mu(Ge) )</th>
<th>( \mu(Se) )</th>
<th>( \mu(Dc) )</th>
<th>( \mu(Be) )</th>
</tr>
</thead>
<tbody>
<tr>
<td>(12)</td>
<td>98</td>
<td>95</td>
<td>95</td>
<td>91</td>
<td>10</td>
</tr>
<tr>
<td>(20)</td>
<td>97</td>
<td>94</td>
<td>95</td>
<td>84</td>
<td>10</td>
</tr>
</tbody>
</table>

### Table 2: Score recognition (in %) for matched conditions; long train/test.

<table>
<thead>
<tr>
<th>Categories</th>
<th>( \mu(G) )</th>
<th>( \mu(Ge) )</th>
<th>( \mu(Se) )</th>
<th>( \mu(Dc) )</th>
<th>( \mu(Be) )</th>
</tr>
</thead>
<tbody>
<tr>
<td>(9)</td>
<td>97</td>
<td>95</td>
<td>90</td>
<td>64</td>
<td>20</td>
</tr>
<tr>
<td>(20)</td>
<td>97</td>
<td>95</td>
<td>90</td>
<td>64</td>
<td>20</td>
</tr>
</tbody>
</table>

### 3. Music Analysis

#### 3.1. Feature vector extraction

Each musical piece is first downsampled from 44.4 kHz to 16 kHz. The musical signal is then divided into frames of 1024 samples. These frames are overlapped by 512 samples. This is done by assuming that the musical signal is more stable and quasi-stationary than the speech signal (coarticulation phenomena). For each frame, a Hamming window is applied without pre-emphasis. The Fourier power spectrum is computed and coefficients are extracted from each 29 triangular Mel-filters. After the application of log operator for each output filters, a discrete cosine transform is applied to extract 12 coefficients in cepstral space.

Cepstral mean normalization are not performed because as mentioned in \[2\], it removes an important genre attribute that characterizes the piece style. Since one uses a classifier based on measures, this enables us to assume that, the influence of delta and delta-delta MFCC coefficients is of no major importance compared to the static coefficients.

#### 3.2. Music database

The RWC Music Database \[8\] is a copyright-cleared music database and the world’s first large-scale music database compiled specifically for research purposes. It is composed of 100 musical pieces: 73 pieces originally composed and arranged, and 27 pieces coming from the public-domain.

The music database (RWC) is divided into main categories and subcategories of genres as illustrated in figure 1.

### Table 3: Score recognition (in %) for mismatched conditions; long train/test.

<table>
<thead>
<tr>
<th>Categories</th>
<th>( \mu(G) )</th>
<th>( \mu(Ge) )</th>
<th>( \mu(Se) )</th>
<th>( \mu(Dc) )</th>
<th>( \mu(Be) )</th>
</tr>
</thead>
<tbody>
<tr>
<td>(12)</td>
<td>70</td>
<td>65</td>
<td>58</td>
<td>65</td>
<td>18</td>
</tr>
<tr>
<td>(20)</td>
<td>75</td>
<td>55</td>
<td>58</td>
<td>55</td>
<td>18</td>
</tr>
<tr>
<td>(40)</td>
<td>58</td>
<td>50</td>
<td>43</td>
<td>48</td>
<td>8</td>
</tr>
</tbody>
</table>

142
Table 4: Score recognition (in %) for mismatched conditions; long train/test.

<table>
<thead>
<tr>
<th>Categories</th>
<th>( \mu(T, R) )</th>
<th>( \mu(R, T) )</th>
<th>( \mu )</th>
<th>( \mu )</th>
<th>( \mu )</th>
<th>( \mu )</th>
</tr>
</thead>
<tbody>
<tr>
<td>Subcategories</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(9) (29)</td>
<td>95</td>
<td>94</td>
<td>92</td>
<td>92</td>
<td>89</td>
<td>89</td>
</tr>
</tbody>
</table>

Table 5: Score recognition (in %) for matched conditions; long train and Short test.

<table>
<thead>
<tr>
<th>Categories</th>
<th>( \mu(T, R) )</th>
<th>( \mu(R, T) )</th>
<th>( \mu )</th>
<th>( \mu )</th>
<th>( \mu )</th>
<th>( \mu )</th>
</tr>
</thead>
<tbody>
<tr>
<td>Subcategories</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(12) (40)</td>
<td>95</td>
<td>94</td>
<td>92</td>
<td>92</td>
<td>89</td>
<td>89</td>
</tr>
</tbody>
</table>

4. Experiments

4.1. Prescriptive taxonomies

While exploring the RWC database in its hierarchical structural form for genre classification, we found that the number of musical pieces corresponding to each subcategory are not equal. This finding incited us to consider two procedures:

- 1) All subcategories having a number of musical genre different to 3 are rejected. In this case, the categories and subcategories numbers are reduced to 9 and 29, respectively (see fig. 1).
- 2) The number of musical piece is not important, thus, all the existent categories and subcategories are considered. Therefore, the number of categories and subcategories are 12 and 40, respectively (see fig. 1).

4.2. Experiments

Two strategies were investigated as follows:

- Strategy 1 (long training and long testing): In this case, half a musical piece was taken for the training session and the rest (second half piece) for the testing session.
- Strategy 2 (long training and short testing): In this case, the same scenario as strategy 1 is kept for the training session but for the testing session only one minute is extracted from the second half musical piece.

Matched and mismatched conditions are also reported for each strategy. Matched conditions refer to experiments where each musical piece is used in training and testing sessions. In this connection, each piece is represented by one prototype in the reference dictionary. Mismatched conditions refer to experiments where the test musical piece was never used or presented during the training session. As an example, if a subcategory is composed of N musical pieces, one piece is chosen randomly and used for the training session, and then all the remaining pieces (N-1) are presented for the testing sessions. This can be useful to simulate new musical pieces and to evaluate the performance systems for new genre classification.

5. Recognition Criterion

During the training session, mean vectors and covariance matrices are estimated and stored as prototype reference to characterize each musical genre. The indice R was used to design the reference. Similarly, during the testing session, the measures between the test file and all references prototypes of musical genre were evaluated. The prototype style with the minimal distance from the test was assigned to the recognized style.

6. Results and discussion

All measure families based on second-order statistics and presented in the previous section were used. They are tested in their asymmetrical forms \( U(X, Y) \) and \( U(Y, X) \), except for the Bat-tacharya distance that is originally symmetrical. All the results presented here are supervised learning techniques.

As mentioned above, we are interested in carrying out an automatic classification such that its hierarchical structure is pre-established beforehand. Two structures were defined for the RWC database, each one containing two levels with a different node number on each level. Based on prescriptive approach and statistical measures, we planed to make an automatic classification consisting in learning the genre taxonomy intrinsically as defined previously. The results are reported on tables 1 to 8. Each table illustrates the scores obtained for the five second-order statistical measures for each strategy and specific condition experiments. Each measure \( \mu_G, \mu_{GC}, \mu_{SC} \) and \( \mu_{DC} \) shown in tables, is tested and evaluated for the high hierarchical level where the classification is carried out based on the category label, and on a lower level where the classification is carry out based on the subcategory label.

Without going into deep details, we will rather discuss the results in a more general context. For all experiments the best performance (for all experiments measures and training/testing conditions) is observed for the matched conditions. The Bhattacharyya measure show poor performance in all tests. In light of this observation, it’s no more consider in the further discussion. However, with the \( \mu_G \) measure, we obtain a recognition

Table 6: Score recognition (in %) for matched conditions; long train and Short test.

<table>
<thead>
<tr>
<th>Categories</th>
<th>( \mu(T, R) )</th>
<th>( \mu(R, T) )</th>
<th>( \mu )</th>
<th>( \mu )</th>
<th>( \mu )</th>
<th>( \mu )</th>
</tr>
</thead>
<tbody>
<tr>
<td>Subcategories</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(9) (29)</td>
<td>94</td>
<td>92</td>
<td>91</td>
<td>89</td>
<td>88</td>
<td>88</td>
</tr>
</tbody>
</table>

Table 7: Score recognition (in %) for mismatched conditions; long train and Short test.

<table>
<thead>
<tr>
<th>Categories</th>
<th>( \mu(T, R) )</th>
<th>( \mu(R, T) )</th>
<th>( \mu )</th>
<th>( \mu )</th>
<th>( \mu )</th>
<th>( \mu )</th>
</tr>
</thead>
<tbody>
<tr>
<td>Subcategories</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(12) (40)</td>
<td>65</td>
<td>65</td>
<td>60</td>
<td>60</td>
<td>60</td>
<td>60</td>
</tr>
</tbody>
</table>

Table 8: Score recognition (in %) for mismatched conditions; long train and Short test.

<table>
<thead>
<tr>
<th>Categories</th>
<th>( \mu(T, R) )</th>
<th>( \mu(R, T) )</th>
<th>( \mu )</th>
<th>( \mu )</th>
<th>( \mu )</th>
<th>( \mu )</th>
</tr>
</thead>
<tbody>
<tr>
<td>Subcategories</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(9) (29)</td>
<td>38</td>
<td>36</td>
<td>38</td>
<td>38</td>
<td>38</td>
<td>38</td>
</tr>
</tbody>
</table>
scores between 93% to 98% for genre classification by categories, and recognition rates from 88% to 97% when genre classification was addressed by subcategories. For mismatched conditions, where the test were never seen during training sessions, the score dropped significantly. The recognition score varied from 43% to 75% for the genre classification by categories, and from 24% to 60% by subcategories. The measure $\mu_G$ seems to be more interesting and yields the best score in comparison to the other measures for all experimental conditions. However, the decrease in performance in the case of mismatched conditions can be related to the experts ambiguity in establishing the genre taxonomy.

The good performance obtained with the $\mu_G$ measure confirms that the assumptions are probably verified. Particularly, it was supposed that the parameters are characterized statistically by a mono-Gaussian distribution. Contrary to the speech processing case, this statistical characterization is better adapted to the music and can be checked in a general way by a simple analysis based on histogram.

7. Conclusion

In this work, second order statistical measures were examined to realize the automatic genre taxonomies. They were already used in the context of text independent speaker identification. Five measures were presented in this paper and have been experimented and evaluated. Several strategies as the matching/mismatching conditions, long/short testing and two forms of genre taxonomies, have been considered. The best results were observed in all matching conditions and yielded score up to 98% and 97% for categories and subcategories, respectively. Worst results were observed in all mismatching conditions and decreased to about 65%. In future work, we plan to further the study for the $\mu_G$ measures since it yielded better scores and only two parameters are required, the mean vector ($x_{12}$) and covariance matrix ($12 \times 12$) for each prototype. We project to adapt this technique for unsupervised musical genre classification with the same conditions. The authors will engage in further research to found the thresholds (fixed or adaptive) which will be used to create automatically new category and subcategory and they expect to publish their result in a future work.

8. References