

# COMPARISON OF THE STATISTICAL AND INFORMATION THEORY MEASURES: APPLICATION TO AUTOMATIC MUSICAL GENRE CLASSIFICATION

‡Hassan Ezzaidi and †Jean Rouat

‡†Ermetis, Université du Québec à Chicoutimi, Chicoutimi, Québec, Canada, G7H 2B1

†NECOTIS, IMSI, Université de Sherbrooke, Sherbrooke, Québec, Canada, J1K 2R1

hezaidi@uqac.quebec.ca, Jean.Rouat@ieee.org

## ABSTRACT

Recently considerable research has been conducted to retrieve pertinent parameters and adequate models for automatic music genre classification using different databases. Many of previous works are derived from speech and speaker recognition techniques. In this paper, four measures are investigated for mapping the features space to decision space. The first two measures are derived from second-order statistical models and last measures are based upon information theory concepts. A Gaussian Mixture Model (GMM) is used as a baseline and reference system. For all experiments, the file sections used for testing have never been used during training. With matched conditions all examined measures yield the best and similar scores (almost 100%). With mismatched conditions, the proposed measures yield better scores than the GMM baseline system, especially for the short testing case. It is also observed that the average discrimination information measure is most appropriate for music category classifications and on the other hand the divergence measure is more suitable for music subcategory classifications.

## 1. INTRODUCTION

The considerable advances in audio technologies, the accessibility to information via the Internet and the increasing production of digital music create many new needs to exploit this large musical universe. One of these very popular needs is automatic genre classification.

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 different taxonomies are used to classify the same musical genres. This is related to the fact that many different descriptors and semantic ambiguities exist to determine genre classification [6]. For example, names of categories associated to each genre are not always similar or coherent including the hierarchical structure (sub-categories) itself. The manufacturing of new instruments

and the realization of new albums continue to accentuate this tendency. In the future, genre taxonomy will remain in an elastic and dynamical structure. As argued by Aucouturier et Pachet [6], genre may be used in intentional or extensional concept. For each concept, genre taxonomy is related to different interpretations at the descriptor level or at the semantical level. The authors describe three approaches to establish musical genre classification: manual classification (projection of human or expert knowledge), *prescriptive* approach that relies on supervised learning using signal processing techniques (classify genre as they are found) and finally emergent classification approaches that are based on similarity measures to automatically produce the hierarchical genre structure. We propose (for musical genre classification) to investigate new classification techniques based on second order and information measures. The *prescriptive* approach (classify genre as they are found) is adopted in this work. The measures were derived from statistical models and information theoretical concept. They have been 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 and on 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. The parameters can be recursively estimated when the time duration of musical piece is long enough. Experiments are carried out according to different strategies as matched and mismatched conditions, long or short testing with either long or short training. All proposed measures are evaluated in their asymmetrical form with two prescriptive genre taxonomies. Results are compared to a Gaussian Mixture Model (GMM) recognizer system.

## 2. RELATED WORK

Several works are proposed to extract genre information features from the musical signal. The majority of them are

inspired from speech/speaker recognition and music/speech discrimination systems. Several features and models proposed and experimented for genre classification can be found in [10] [7] [8].

Generally, parameters can be divided into three feature families. The *first family* represents the timbral texture of audio signal and usually comprises the following features [8] [9] [7]: Spectral Centroid, Spectral Centroid, Spectral Rollof, Zero crossing Rate, Fast Fourier transform, Spectral Flux, Mel Frequency Cepstral and Linear Prediction Coefficients.

The *second family* represents the rhythmic content features as proposed in [9]. These features are based on detecting the periodicities of the signal. For the extraction of these features, a discrete wavelet transform, envelope extraction, autocorrelation function and finally the peak detection are elaborated to built a beat histogram.

The *third family* is based on pitch content features [9] [10]. The pitch features are based on a pitch histogram obtained from multiple pitch detections. The pitch histogram describe the harmonic content of the musical signal. Five features are extracted from the pitch histogram and used for musical genre classification.

*Other features* have been proposed and experimented for automatic genre classification such as the audio low-level descriptor in the context of MPEG-7 standard, Root mean square, periodicity rate, various order central moments [3].

The majority of proposed works in automatic musical genre classification, use Gaussian Mixture models (GMM) which we use here as a reference system. The reader is referred to other alternatives in [10], and [9].

### 3. PROPOSED CLASSIFICATION MEASURES

#### 3.1. Second-order statistical measures

Recognition based on second-order statistical measures was initially proposed and tested in the context of text independent speaker identification by Bimbot et al. [1].

Let  $\{m_R(i)\}_{1 \leq i \leq M}$  be a sequence of  $M$  independent parameter vectors related to the source information noted as  $R$ , extracted from an acoustical signal. All vectors are  $p$ -dimensional, assumed to be independent and distributed like a Gaussian function. Therefore, they are characterized in the parametric form only by 2 parameters: a mean vector noted  $\overline{m_R}$  and a covariance matrix noted  $\Sigma_R$  as:

$$\overline{m_R} = \frac{1}{M} \sum_{i=1}^{i=M} m_R(i)$$

$$\Sigma_R = \frac{1}{M} \sum_{i=1}^{i=M} (m_R(i) - \overline{m_R})^T (m_R(i) - \overline{m_R})$$

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

Similarly, a sequence of  $N$  vectors  $(\{m_T(i)\}_{1 \leq i \leq N})$  corresponds to a target information source to be classified and that obeys to the same properties as the reference source.

Hence, the target source can be represented by 2 parameters: the mean vector  $\overline{m_T}$  and the covariance matrix  $\Sigma_T$ .

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

$$\mu_G(R, T) = \frac{1}{p} [tr(\Sigma_T \Sigma_R^{-1}) - \log(\frac{det \Sigma_T}{det \Sigma_R}) + \Delta_m^T \Sigma_R^{-1} \Delta_m] - 1 \quad (1)$$

where:  $\overline{m_T} - \overline{m_R} = \Delta_m$ ,  $p = \frac{M(\text{number of Reference features})}{N(\text{number of Test features})}$ ,  $tr()$  is the matrix trace,  $det()$  is the matrix determinant.

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

$$\mu_{Sc}(R, T) = \log(\frac{tr(\Sigma_T \Sigma_R^{-1})}{p}) - \log((\frac{det \Sigma_T}{det \Sigma_R})^{-1/p}) \quad (2)$$

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

#### 3.2. Information theory measures

A discrimination information in the Bayes classifier sense, for class  $\omega_R$  versus class  $\omega_T$ , can be measured by the logarithm of likelihood ratio as defined in [4]:

$$\mu_{R,T} = \ln \left\{ \frac{p_R(x)}{p_T(x)} \right\} \quad (3)$$

where  $p_R(x)$  and  $p_T(x)$  correspond to the probability densities for the reference and target classes, respectively. The averaged information for class  $\omega_R$  versus class  $\omega_T$  is the expectation of  $\mu_{R,T}$  and is defined as:

$$I_{R,T} = \int_x p_R(x) \ln \left\{ \frac{p_R(x)}{p_T(x)} \right\} dx \quad (4)$$

If the distribution of each class is assumed to be Gaussian and multivariate,  $I_{R,T}$  can be expressed in another form according to mean vector and covariance matrix:

$$I_{R,T} = 0.5 \left( \ln \left( \frac{|\Sigma_T|}{|\Sigma_R|} \right) + tr[\Sigma_R(\Delta \Sigma_{TR}^{-1})] + tr[\Sigma_T^{-1} \Delta_m \Delta_m^T] \right) \quad (5)$$

where:  $\Delta_m = \overline{m_R} - \overline{m_T}$ ,  $\Delta \Sigma_{TR}^{-1} = \Sigma_T^{-1} - \Sigma_R^{-1}$ ,  $\Delta \Sigma_{RT}^{-1} = \Sigma_R^{-1} - \Sigma_T^{-1}$  and  $\Delta \Sigma_{RT} = \Sigma_R - \Sigma_T$ .

A divergence measure or *total average information discrimination* is defined as the sum of the average information discrimination  $I_{R,T}$  and  $I_{T,R}$  and can be expressed as:

$$J_{R,T} = \frac{tr[\Delta \Sigma_{RT} \Delta \Sigma_{TR}^{-1}]}{2} + \frac{tr[\Delta \Sigma_{RT}^{-1} \Delta_m \Delta_m^T]}{2} \quad (6)$$

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

		$\mu_G$	$\mu_{Sc}$	$I_{ij}$	$J_{ij}$
Categories (12)	$\mu(T, R)$	98	95	98	99
	$\mu(R, T)$	98	95	98	99
Subcategories (40)	$\mu(T, R)$	97	93	96	97
	$\mu(R, T)$	96	94	97	97

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

		$\mu_G$	$\mu_{Sc}$	$I_{ij}$	$J_{ij}$
Categories (11)	$\mu(T, R)$	98	97	98	99
	$\mu(R, T)$	97	95	98	99
Subcategories (29)	$\mu(T, R)$	97	90	96	98
	$\mu(R, T)$	96	94	97	98

This divergence provides a dissimilarity measure between the normally distributed classes  $\omega_R$  and  $\omega_T$ . The information discrimination  $I_{R,T}$  and divergence  $J_{R,T}$  are tested in the context of automatic musical genre classification. The development and details for equations 5 and 6 can be found in [4]. The motivation for these measures is based on the fact that the musical signal is generally characterized by rhythmicity and regularity which cover a long temporal period<sup>1</sup>. Also, all proposed measures assume that each group or class is characterized by multivariate normal distributions.

## 4. FEATURES EXTRACTION AND THE REFERENCE SYSTEM

### 4.1. Music database

The RWC Music Database [5] is a copyright-cleared music database and it is 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 come from the public-domain. Among the many characteristics of the RWC database, this includes the following: 91.6 hours recording, performance of about 150 instrument bodies, variations for each instrument, variations in instrument manufacturers and musicians, different manufacturer/different musician, wide variety of sounds. The music database (RWC) is divided into main categories and subcategories of genres (see [5]). as illustrated in

### 4.2. Feature vector extraction

Each musical piece is first downsampled from 44.4Khz to 16Khz. Then the musical signal is divided into frames of

<sup>1</sup>without forgetting periodicity of the rhythmicity

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

		$\mu_G$	$\mu_{Sc}$	$I_{ij}$	$J_{ij}$
Categories (12)	$\mu(T, R)$	70	58	75	70
	$\mu(R, T)$	75	58	70	70
Subcategories (40)	$\mu(T, R)$	60	43	58	62
	$\mu(R, T)$	58	48	60	62

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

		$\mu_G$	$\mu_{Sc}$	$I_{ij}$	$J_{ij}$
Categories (11)	$\mu(T, R)$	65	53	70	64
	$\mu(R, T)$	71	50	64	64
Subcategories (29)	$\mu(T, R)$	53	35	49	55
	$\mu(R, T)$	50	38	51	55

1024 samples with 50% overlap. It is assumed that the musical signal is more stable and quasi-stationary than the speech signal where coarticulation is dominant. For each frame, a Hamming window is applied without pre-emphasis. Then, 29 averaged Spectral energies are obtained from a bank of 29 Mel triangular filters followed by a discrete cosine transform, yielding 12 Mel frequency Cepstrum Coefficients. Cepstral mean normalization is not used because it removes important genre attributes that characterize the piece style (see [2]). Since one uses a classifier based on time averaged measures, we assume that the influence of delta and delta-delta MFCC coefficients is not of major importance.

### 4.3. The reference system

For comparison purposes, Gaussian Mixture Models (GMM) of the MFCC is used and 16 mixtures and diagonal covariance matrices are estimated via the Expectation Maximization (EM) algorithm.

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

		$\mu_G$	$\mu_{Sc}$	$I_{ij}$	$J_{ij}$
Categories (12)	$\mu(T, R)$	95	92	96	97
	$\mu(R, T)$	96	96	95	97
Subcategories (40)	$\mu(T, R)$	94	87	89	93
	$\mu(R, T)$	89	87	93	93

## 5. EXPERIMENTS

### 5.1. Prescriptive taxonomies

We found that, in the RWC database, the number of musical pieces corresponding to each subcategory are not the same. This suggests the use of two procedures: *i*) Keep only the subcategories which number of musical pieces is three. In that case we obtain 9 categories and 29 subcategories(see [5]). *ii*) Use all the existent 12 categories and 40 subcategories.

### 5.2. Experiments

For each musical piece, the first half of the signal is used for training only, then long or short testing is performed on the remaining half (long) or on one minute from the second half (short). Matched and mismatched conditions are also reported for each strategy. Matched conditions refer to experiments where each musical piece is used in training (first half piece) and testing sessions (second half piece). Mismatched conditions refer to experiments where the musical test piece was never used or presented during the training session. This can be useful to simulate new musical pieces and to evaluate the systems performance for new genre classification.

## 6. 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 is used to design the reference. Similarly, during the testing session, the measures between the test file and all reference prototypes of musical genre are evaluated. The reference prototype style with the minimal distance to the test is assigned to the recognized style.

## 7. RESULTS AND DISCUSSION

All measures presented in section 3 have been used. They are tested in their asymmetrical form  $U(X, Y)$  and  $U(Y, X)$ , except for the divergence distance that is originally symmetrical. All results presented here are based on supervised learning techniques.

Two structures were defined for the RWC database, each comprising two levels with a different node number on each level. We are interested in automatically reproduce the proposed taxonomy of the RWC musical database. For each category and subcategory, results are reported on tables 1 to 8. Table 9 illustrates the recognition scores of the GMM reference system for each strategy and specific experimental conditions. Tables 1 to 8 report the scores for the two

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

		$\mu_G$	$\mu_{Sc}$	$I_{ij}$	$J_{ij}$
Categories (11)	$\mu(T, R)$	96	94	96	97
	$\mu(R, T)$	96	96	96	97
Subcategories (29)	$\mu(T, R)$	95	89	88	92
	$\mu(R, T)$	88	87	93	92

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

		$\mu_G$	$\mu_{Sc}$	$I_{ij}$	$J_{ij}$
Categories (12)	$\mu(T, R)$	65	60	73	68
	$\mu(R, T)$	70	58	65	68
Subcategories (40)	$\mu(T, R)$	52	43	53	58
	$\mu(R, T)$	43	43	53	58

second-order statistical and two information theoretical measures for all strategies and experimental conditions. Each measure  $\mu_G$ ,  $\mu_{Sc}$ ,  $I_{i,j}$ ,  $J_{i,j}$  and  $GMM$  models, 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 carried out based on the subcategory label (see [5]).

### 7.1. Matched conditions

For all experiments the best performance (for every measures and training/testing conditions) is observed for the matched conditions. Precisely, with the  $\mu_G$  measure, we obtain recognition scores from 94% to 98% for genre classification by categories and from 88% to 97% when genre classification was addressed by subcategories. With the  $\mu_{Sc}$  measure, we obtain recognition scores from 94% to 97% for genre classification by categories and from 87% to 94% when genre classification was addressed by subcategories. With the  $I_{i,j}$  measure, we obtain recognition scores from 95% to 98% for genre classification by categories and from 88% to 97% when genre classification was addressed by subcategories. With the  $J_{i,j}$  measure, we obtain recognition scores from 97% to 99% for genre classification by categories and from 92% to 97% when genre classification was

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

		$\mu_G$	$\mu_{Sc}$	$I_{ij}$	$J_{ij}$
Categories (11)	$\mu(T, R)$	58	55	67	60
	$\mu(R, T)$	64	58	55	60
Subcategories (29)	$\mu(T, R)$	42	36	42	48
	$\mu(R, T)$	43	36	40	48

addressed by subcategories. Scores for these different measures and strategies remain similar with the matched conditions. It is also observed that the long and short testing does not have many influence on the scores. The good performance obtained with the proposed measures confirm that previous theoretical assumptions are probably verified. We recall that the parameters are assumed to be Gaussian. This assumption would be too restrictive when analyzing speech and better fits the music distribution.

## 7.2. Mismatched conditions

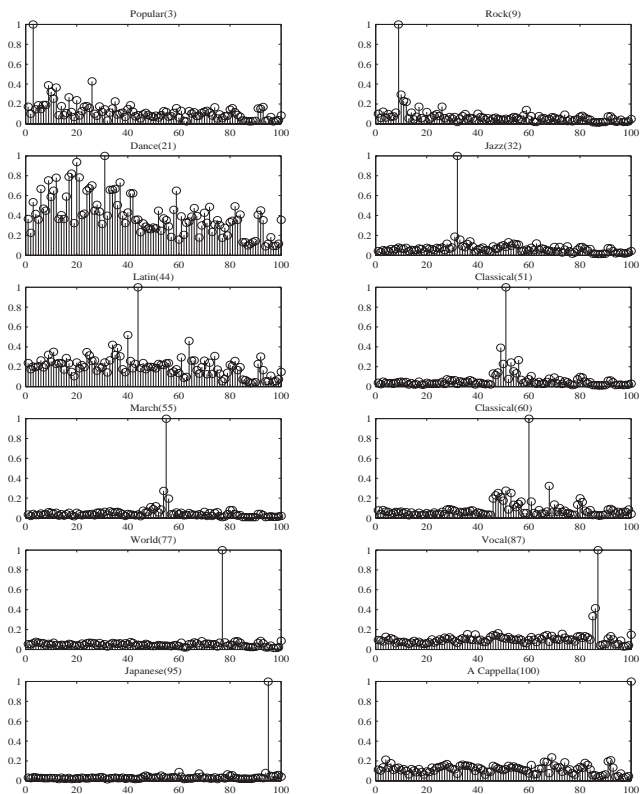
For mismatched conditions, where tests were never seen during training sessions, the scores drop significantly. The recognition scores varied from 50% to 75% for the genre classification by categories, and from 35% to 62% by subcategories. The information theoretical measures seems to be more interesting and yield the best scores in comparison to the other measures for all experimental strategies. Particularly, it is observed that the discrimination measure  $I_{i,j}$  yields a better score when the classification is addressed by category. When the problem is addressed to subcategories, the best score is obtained with the divergence measure  $J_{i,j}$ . However, the decrease in performance with mismatched conditions can partially be explained by the fact that human experts themselves do not always agree on the category or sub-category for a specific musical. Furthermore, the auditory system is able to perceive subtle features (likes tremolo, hangs of rhythm,etc..) that are not encoded in the MFCC parameters.

## 7.3. Reference system

The reference classification system yields the same recognition rates in matched conditions comparatively to the proposed measures, except with short testing utterances where the score drops significantly to 30%. With mismatched conditions, recognition scores of categories are similar for all measures. But with subcategories, the best performance of the reference system is globally 10% lower than  $\mu_G$ ,  $I_{ij}$  and  $J_{ij}$ .

## 7.4. Graphical analysis

Figure 1 reports results of the divergence measure for the classification of 100 musical pieces into 12 categories for the matched conditions. The same analysis can be carried out for other scenarios as classification into 33 subcategories. It also predicts the performance and behavior of the system against style of musics not presented in training (never seen before). The inverse is used on the figure for an easier interpretation even if the minimization of the divergence is the classification criterion. Each subplot corresponds to ones



**Fig. 1.** Results of the divergence measure for the classification of 100 musical pieces into 12 categories. The x axis is the label (number from 1 to 100) associated to each musical file of the RWC database. The file numbers are the same than the ones used in the description file of the RWC database (see [5]). For example, the first category includes 6 musical pieces which are labeled 1 to 6 on the x axis. Rock category contains also 6 musical pieces that are labeled from 7 to 12 and so on. Vertical axis corresponds to the inverse of the divergence measure. For each subplot, the title specifies the name of the category of the file presented in test and between parenthesis its order (label).

of the 12 categories. For each subplot, one musical testing piece corresponding to the category is randomly chosen. The divergence measure is computed from 100 reference prototypes estimated during the training session as mentioned in section 5.2. The normalized inversion of all measures is reported in the 12 subplots. Most of the tested files were correctly classified. Moreover, a significant difference of the similarity measure between the recognized model and the others can be highlighted. This confirms the good performance obtained with the matched conditions. To simulate the mismatched conditions, the reference prototype corresponding to each test file is ignored. In this context, only the similarity measures estimated from the others prototypes (models) are considered and compared. Only 5 categories are well identified (Rock, Jazz, Classical, March and Vocal). This is because files from the same category as the test file yield the best similarity measure. However, the others styles are badly discriminated. The badly recognized categories are characterized with a small inter-variability. Several prototypes should then be used for each of these categories. From these results, one can see that unsupervised musical genre classification would yield a different taxonomy than the one given on the RWC database.

## 8. CONCLUSION

Automatic musical genre taxonomy has been realized. It is based on four statistical measures. They were already used in the context of text independent speaker identification. Matched/mismatched and long/short testing strategies have been studied. The best results were observed in all matching conditions and yielded score until 99% and 98% recognition for categories and subcategories, respectively. Worst results were observed in all mismatched conditions and decreased to about 75% and 60% for categories and subcategories, respectively. A Gaussian Mixture Model is used as a reference system. In mismatched conditions, the proposed measures yield better scores than the reference system especially for the short testing case. Particularly, it is observed that the averaged discrimination information measure is most appropriate for musical categories classification and on the other hand the divergence measure is most suitable for music subcategories classification. In future works, we propose to further study these measures since they yielded better scores and only two parameters are required, the mean vector ( $x_{12}$ ) and the covariance matrix ( $12 \times 12$ ) for each prototype. We plan to adapt this technique to unsupervised musical genre classification in the same conditions. We will focus our research on the automatic generation of new categories and subcategories.

**Table 9.** GMM score recognition (in%) for matched mismatched conditions.

	time	matched	mismatched
Categories 11	long	100	73
	short	81	60
Categories 12	long	100	75
	short	81	61
Subcategories 34	long	100	44
	short	80	38
Subcategories 40	long	95	50
	short	77	42

## 9. REFERENCES

- [1] Frédéric Bimbot, Ivan Magrin-Chagnolleau and Luc Mathan, Second-Order Statistical Measures for Text-Independent Speaker Identification, In *Speech Communication*, pp. 177-192, Vol. 17, 1995.
- [2] H. Ezzaidi and J. Rouat, Speech, music and songs discrimination in the context of handsets variability, In *proceedings of ICSLP 2002*, 16-20 September 2002.
- [3] FT. Lambrou, P. Kudumakis, M. Sandler, R. Speller and A. Linney, Classification of Audio Signals using Statistical Features on Time and Wavelet Transform Domains, In *IEEE ICASSP 98*, May 1998, Seattle, USA.
- [4] J. Tou and R. Gonzalez. Pattern recognition principles, Addison-Wesley Publishing Company, Reading, Massachusetts, 1974.
- [5] Masataka Goto, Hiroki Hashiguchi, Takuichi Nishimura, and Ryuichi Oka, RWC Music Database: Music Genre Database and Musical Instrument Sound Database, ISMIR, pp. 229-230, October 2003.
- [6] Aucouturier, J.J and Pachet, F. Musical Genre: a Survey. In *Journal of New Music Research*, Vol. 32, No 1, pp.83-93, 2003.
- [7] H. Soltau, T. Schultz, M. Westphal, and A. Waibel. Recognition of Musical Types. In *Proceedings ICASSP*, May 1998, vol. 2, pp. 1137-1140.
- [8] Tzanetakis, G., Essl, G., and Cook, P., Automatic Musical Genre Classification of Audio Signals, In *Proceedings of the 2001 International Symposium on Music Information Retrieval.*, 2001.
- [9] Tzanetakis, G., and P. R. Cook, Musical Genre Classification of Audio Signals, In *IEEE Transactions on Speech and Audio*, July, 2002.
- [10] Li, Tao and Tzanetakis, George, Factors in Automatic Musical Genre Classification of Audio Signals, In *Proc. IEEE WASPAA*, New Paltz, NY October 2003.