Genre Classification for Brazilian Music using Independent and Discriminant Features

Eduardo F. Simas Filho, Elmo A. Borges Jr. and Antonio C. L. Fernandes Jr.

***Abstract*—Digital music files are largely available both online and in private local collections. These databases may comprise hundreds or thousands of files, which in some cases may not carry tagged information about their content, making the search for the desired audio files very time consuming. An important task in this context is to organize the available database according to the prevailing musical genre. The purpose of this work is to develop an automatic music genre classification system able to identify international music genres (i.e. pop, rock, classic, soul, funk) and also typical Brazilian rhythms such as Samba, Forró and Brazilian Popular Music. The proposed signal processing chain comprises two stages. Initially, audio signal features are computed and their relevance for music genre identification estimated. Independent component analysis is applied to reduce mutual redundancy among the audio attributes. In the following, different classifiers based on neural networks and support vector machines are applied for music genre identification. The proposed system efficiency is evaluated using an experimental dataset.**

***Index Terms*—Neural Networks, Support Vector Machines, Signal Processing, Music Information Retrieval, Independent Component Analysis.**

I. Introduction

CONSIDERING the large amount of audio data files available, both online and in personal collections, along with an increasing availability of mobile digital audio playing devices which are equipped with high capacity storage drives, the search for the desired information may become tedious and time consuming. In this context, the use of an automatic system for efficient managing these large datasets is important for the final user.

Music information retrieval (MIR) [1] is an important and very active research field which combines aspects from signal processing, machine learning and musicology in the search for computational systems able to automatically access and identify the information contained in music data files. When dealing with a musical excerpt, different aspects such as the prevailing genre, the singer and the used instruments are relevant for classification purposes [2].

For audio signals automatic classification, the initial step usually comprises the extraction of relevant features (or attributes) from the digital files. After that, hypothesis testing (classification) is performed to assign to each audio signal a given class.

Eduardo F. Simas Filho, Elmo A. Borges Jr and Antonio C. L. Fernandes Jr. are with the Department of Electrical and Computer Engineering, Federal University of Bahia, Salvador, BA, 40210630 Brazil e-mail: eduardo.simas@ufba.br

This work was partially supported by CNPq and FAPESB.
Digital Object Identifier: 10.14209/jcis.2018.11

Some studies has been carried on in the literature to achieve content-based audio signal classification. For example, the work [3] quantified the relevance of locally estimated parameters for musical instrument recognition. In [2], temporal segmentation was proposed for audio signals analysis by selecting equal time-length segments from different parts of the audio signal file.

The musical genre classification problem was addressed in [4], where Gaussian mixture model (GMM) and k-nearest neighbor (KNN) classifiers were applied for this purpose. The work [5] proposed a feature extraction method for music genre classification based on wavelet analysis. For automatic classification were used both, support vector machines and linear discriminants. In [6] musical genre definitions and hierarchies were discussed; and it was presented techniques for extracting meaningful information from audio data aiming at the characterization of musical excerpts. In [7], signals were classified according to the prevailing audio content into three classes: speech, music, and background noise. Hierarchical approaches for feature selection and classification were also evaluated. The analysis of the bass tracks was proposed in [8] for music genre classification by exploring stylistic similarities.

We noticed that not only traditional and popular rhythms received the attention of the community. For example, in [9] the influence of repeated patterns was considered for Dutch folk song classification. Computational techniques were also used in [10] to discover patterns in Native American music and identify musical differences between indigenous groups.

In this work, the genre classification problem is addressed considering 12 different classes, including typical Brazilian genres such as Samba, Brazilian Popular Music (MPB) and Forró [11]. Brazilian culture is very rich, comprising European, African and Native-American influences. This aspect together with the continental dimensions of the country allowed the appearance of several local (and very particular) music genres, which were properly considered only in few previous works, such as [12], in which visibility network features were used to describe the audio signals.

The main contributions of this paper include the study of audio features relevance for genre classification and a performance comparison between different classifier architectures based on both single layer feedforward neural networks (SLFN) and support vector machines (SVM) [13]. Independent component analysis (ICA) [14] is proposed here for efficient feature transformation, removing information redundancy. It is important to mention that the inclusion of typical Brazilian genres represents itself a relevant novelty of this work. Addi-

tionally, results from a preliminary embedded implementation of the proposed system are presented.

This document is organized as follows. In Section II the proposed system is presented. The audio database and the system validation methodology are described in Section III. The experimental results are presented in Section IV and, the conclusions are derived in Section V.

II. The Proposed System

The proposed system architecture is detailed in this section. Initially, a brief overview is used to present the main signal processing chain. In the following, the proposed audio descriptors are introduced. The applied feature selection and redundancy removal techniques are considered in the next sub-section. In the final sub-section, the used classification systems are presented.

A. System Overview

As illustrated in Fig. 1, the proposed music genre classification system comprises a signal processing chain that initiates with temporal segmentation of full-length audio files. Three data segments of 30 seconds each are selected from the audio files. The first one initiates after 15 seconds from signal beginning, the second one initiates exactly in half of the total signal length, and the last one ends 15 seconds before signal end. The adopted 15 seconds shift from the begin and end of the files intends to avoids selecting audio segments composed mostly of noise or silence, that may occur during the recording process. By selecting segments from different time locations, the estimated features are expected to better represent the complete audio signal characteristics.

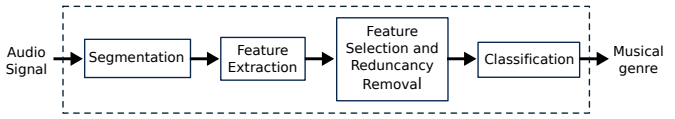


Fig. 1. Diagram of the proposed signal processing chain.

For feature extraction, audio descriptors are estimated from the previously selected music excerpts. The used descriptors are zero-crossing rate (ZCR), mel-frequency cepstral coefficients (MFCC), spectral power concentration (SPC), spectral centroid (SC), loudness (L) [15], and beat histogram (BH) [4].

Prior to be used to feed the classifiers, the input features relevance for genre identification is estimated. In this stage some non-relevant features are discarded. Additionally, independent component analysis is applied as a pre-processing step for the classification module to reduce features redundancy, which, in some cases, may prevent the proper training of the classifiers, by causing slow convergence and sub-optimal results. Different algorithms were used for pattern recognition and their results compared considering aspects such as the discrimination efficiency and the computational complexity.

B. Used audio descriptors

In this work, feature extraction was performed in short time windows of approximately 30 ms. Hamming windows with 30% overlap were used [2]. For proper characterization of the music files the following audio descriptors were estimated: zero-crossing rate (ZCR), mel-frequency cepstral coefficients (MFCC), spectral power concentration (SPC), beat histogram (BH), spectral centroid (SC) and loudness (L). These descriptors are briefly presented in the following (for more details see, for example [15]).

The ZCR [2] is commonly used as an estimator for the fundamental (pitch) frequency and may be computed by counting the number of times the signal amplitude crosses the zero axis (N_{cross}) during a fixed time interval ΔT:

$$\text{ZCR} = \frac{N_{\text{cross}}}{\Delta T}. \tag{1}$$

The mel-frequency cepstrum coefficients (MFCC) are widely used for audio description (especially in speech processing applications) [16], [17], as they attempt to model the perception of the human ear. For this, a nonlinear frequency scale (the mel scale) is defined as:

$$f_{\text{m}} = 1127 \times \ln\left(1 + \frac{f_{\text{Hz}}}{700}\right), \tag{2}$$

where f_{Hz} is the frequency in Hz.

In order to obtain the mel-frequency cepstrum (see Fig. 2), the discrete Fourier transform (DFT) is applied to each audio signal frame. In the following, the logarithm amplitude spectrum is mapped to the mel-frequency scale and filtered using triangular overlapping filters. Finally the discrete cosine transform (DCT) is applied to produce the MFCC.

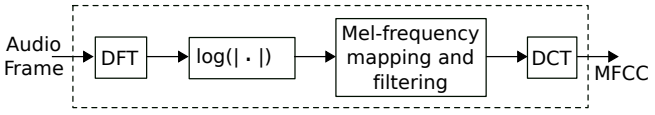


Fig. 2. MFCC estimation diagram.

The spectral power concentration ($\mathbf{S}_{PC}$) vector and the spectral centroid (S_C) are parameters used to evaluate the distribution of signal power throughout the frequency range of interest ($0 \leq f \leq F_S/2$, where F_S is the sampling frequency). The $\mathbf{S}_{PC}$ consists on the power spectral density ($S(f)$) sum within three frequency bands ($\mathbf{S}_{PC} = [S_{PC}^{(1)}, S_{PC}^{(2)}, S_{PC}^{(3)}]$):

$$S_{PC}^{(i)} = \sum_{f=F_L^{(i)}}^{F_H^{(i)}} S(f), \quad i = 1,2,3, \tag{3}$$

where the limit frequencies are: $F_L^{(1)} = 0$, $F_H^{(1)} = 600$ Hz; $F_L^{(2)} = 600$ Hz, $F_H^{(2)} = 2400$ Hz; and $F_L^{(3)} = 2400$ Hz, $F_H^{(3)} = F_S/2$ Hz.

The S_C [18] estimates power spectrum "center of mass":

$$S_C = \sum_{f=0}^{F_S/2} f \cdot |S(f)|^2 \Big/ \sum_{f=0}^{F_S/2} |S(f)|^2. \tag{4}$$

The loudness (L) [19] is an audio descriptor which aims to approximate the human perception of an audio signal intensity. To such end, an approximation of the human ear frequency response is used. As proposed in [20], a frequency-dependent weight factor ($W(f_{kHz})$)) is defined for the outer ear:

$$W(f_{kHz}) = -2.2f_{kHz}^{-0.8} - 6.5e^{-0.6(f_{kHz}-3.3)^2} + 0.004f_{kHz}, \quad (5)$$

where f_{kHz} is the frequency in kHz. The outer ear weighted FFT module coefficients are defined as:

$$X_e(f) = |X(f)|10^{W(f)/20}, \quad (6)$$

where $X(f)$ is the discrete Fourier transform of the audio signal. In this work, a simple estimative for the loudness is obtained by summing the weighted FFT components:

$$L = \sum_{f=0}^{Fs/2} X_e(f). \quad (7)$$

The temporal features such as the tempo and the rhythm are important musical properties. A building block of these parameters is the onset, which may be defined as the beginning of a musical sound event (i.e. a stroke on a percussive instrument). The novelty function is usually applied to estimate the amount of audio signal changes over time and is an important step for the automatic detection of onsets [15].

The Beat Histogram (BH, also called beat spectrum) [4] is used to estimate the amplitude and frequency of the most relevant beats of a song. The BH can be interpreted as the frequency domain representation of the novelty function, as the result, there is a plot of the beat frequency (in BPM) *vs* its respective relevance (number of repetitions). The occurrence of multiple peaks indicates a intense rhythmic content. There are multiple ways of computing the beat histogram. In this work was used the procedure described in [4], as illustrated in Fig. 3.

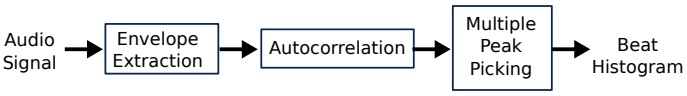


Fig. 3. Beat histogram estimation diagram.

The feature vector used to feed the classifiers is composed by the mean and variance estimates of the following audio descriptors: ZCR, first five MFCC, S_{PC} (in three frequency bands), S_C, L, and four beat histogram measures: the relative amplitude of the first and second peaks; the sum of the histogram; and the period of the first peak. Considering this, the feature vector comprises a total of 26 parameters (see Table I).

C. Feature Selection and Redundancy Removal

In classification systems, preprocessing the inputs is important in order to feed the classifiers from a compact and discriminant set of features. There are multiple ways to do such feature ranking, in this work it is used the sequential backward elimination procedure: the classifier system is initially trained from the full set of features and further re-trained after eliminating each feature individually [15]. Comparing the efficiency results it is possible to determine if the discrimination performance changes after removing each feature.

TABLE I
USED AUDIO DESCRIPTORS AND ESTIMATED FEATURES.

Audio Descriptor	Estimated Parameters	Number of Features
ZCR	Mean and variance	2
First 5 MFCC	Mean and variance	10
S_{PC} (in 3 bands)	Mean and variance	6
BH	Relative amplitude of first and second peaks; histogram sum; and first peak period	4
S_C	Mean and variance	2
L	Mean and variance	2
Total number of features:		**26**

Another issue that may me observed is the mutual redundancy among the input features. The classifier training process may be hampered if redundant features are used. To avoid this problem, independent component analysis (ICA) [21] is proposed in this work as a preprocessing step.

Considering that a set of N observed variables $\mathbf{x} = [x_1,...,x_N]^T$ is generated from a linear combination of unknown sources $\mathbf{s} = [s_1,...,s_N]^T$, such that:

$$\mathbf{x} = \mathbf{W} \cdot \mathbf{s}, \quad (8)$$

where $\mathbf{W}$ is the $N \times N$ mixing matrix [14], ICA deals with the problem of finding an estimate $\mathbf{y}$ of $\mathbf{s}$ considering that the components y_i are mutually independent.

As the exact inverse mixing matrix is ill-conditioned (it is not possible to guarantee the correct multiplying factor in the estimated sources s_i) [14], a solution may be obtained if it is possible to find an approximation for the inverse of the mixing matrix $\mathbf{B} \approx \mathbf{W}^{-1}$ and so:

$$\mathbf{s} \approx \mathbf{y} = \mathbf{Bx}. \quad (9)$$

In this work the FastICA algorithm [14] is applied for independent components estimation. Among the advantages of the method we can mention fast and more reliable convergence, computational simplicity and low memory requirements.

ICA is closely related do principal component analysis (PCA) [22], which is used in this work to estimate the level of redundancy in the feature vector. PCA explores second-order statistics, removes signal correlation and produces linear projections (principal components) ordered by the amount of retained energy. Indeed, some ICA algorithms use PCA as a preprocessing step. Considering that after PCA all second order dependence is removed, the ICA problem is reduced to deal with the higher order statistics information.

D. The Proposed Classifier System

Automatic classifier systems have been successfully applied in different problem such as detection of partial discharges in electrical power systems [23], location of faults in electrical power lines [24], detection of broken bars in induction

motors [25], and nondestructive evaluation of materials and structures [26].

In this work two different types of classifiers are applied. One based on a single-hidden layer feedforward neural network (SLFN) and another based on support vector machine (SVM). The obtained results are compared considering the discrimination efficiency. A preliminary implementation of the proposed system in dedicated electronics is used to estimate the computational complexity of each signal processing step.

SLFN are widely applied for classification problems (see for example [26], [27], [4]) and in this work, two different neural network architectures were used for the SLFN classifiers: (i) one comprises in the output layer one neuron associated to each musical genre (in this case, 12 neurons); and (ii) other uses one SLFN classifier specialized for each class of interest, in an one-against-all (OAA) approach. For both cases, the number of neurons in the hidden layer is chosen using a network growing procedure (starting from a small number of hidden neurons and adding hidden units until the desired discrimination performance is achieved). The hyperbolic tangent is used as activation function for all neurons and the standard error back propagation algorithm is applied for training.

SVM algorithm comprises a three-layered feedforward network structure, which initially projects the input data in a high-dimensional space and, in the following, uses kernel functions (usually nonlinear) to generate a low-dimensional feature space. In this work, the multi-class SVM classifier was implemented using an one-against-all approach.

Some example of popular kernel functions for SVM are the q-th order polynomial kernel, the radial basis function (RBF) kernel and the sigmoid kernel, respectively given by [28]:

$$K_{Pol}(\mathbf{z}_1,\mathbf{z}_2) = (\mathbf{z}_1^T\mathbf{z}_2 + 1)^q, \qquad q > 0; \tag{10}$$

$$K_{RBF}(\mathbf{z}_1,\mathbf{z}_2) = \exp\left(-\frac{||\mathbf{z}_1 - \mathbf{z}_2||^2}{\sigma^2}\right); \tag{11}$$

$$K_{Sig}(\mathbf{z}_1,\mathbf{z}_2) = \tanh\left(\beta\mathbf{z}_1^T\mathbf{z}_2 + \gamma\right); \tag{12}$$

where $\mathbf{z}_1$ and $\mathbf{z}_2$ are vectors in the input space, $||\cdot||$ and T are the vector norm and the transpose operators, respectively, q (polynomial kernel order), σ^2 (RFB kernel variance), β and γ (sigmoid kernel gain and bias, respectively) are constants used to adjust each kernel function [28].

Applications of SVM include voice activity detection [29] and facial expression recognition [30].

III. Database and System Validation Methodology

The database comprises 1008 music files assigned by five expert listeners to twelve different musical genres: Blues, Classical, Country, Forró, Hip Hop, Jazz, Brazilian Popular Music (MPB), Pop, Reggae, Rock, Soul, and Samba (see Table II). It is important to mention that, during the class assignment procedure for the used dataset (required for supervised training), the expert listeners did not always agree in the music genre classification. In these cases, the class which received a larger number of indications was assigned to the music file.

To evaluate the performance of the proposed classifiers, the confusion matrix and the efficiencies geometric mean ($\overline{EF}$) are computed. The confusion matrix presents the discrimination efficiencies (in the main diagonal) and classification errors (in off-diagonal positions) for each class of interest. $\overline{EF}$ provides a measure for the classifier overall performance:

$$\overline{EF} = \sqrt[M]{\prod_{m=1}^{M} EF_m}, \tag{13}$$

where EF_M is the classification efficiency obtained for class m and $M = 12$ is the number of classes. The geometric mean is preferred here instead of the simple mean as it tends faster to small values when there occurs a low efficiency for a single class.

In order to account for statistical fluctuations in the dataset, for each classifier architecture the training procedure was restarted 10 times using different samples for the training, testing and validation sets. In this cross-validation procedure the amount of examples in each set are kept fixed into 50%, 30% and 20% of available signals, respectively. After that, the maximum value $\overline{EF}_{\max}$ and standard deviation $\sigma_{\overline{EF}}$ are computed.

TABLE II
Detailed information about the used database.

Genre	Files	Genre	Files
Blues (BL)	81	Soul (SO)	75
Classical (CL)	83	Samba (SA)	96
Country (CO)	68	Rock (RO)	87
Forro (FO)	84	Reggae (RE)	98
Hip-hop (HH)	77	MPB	102
Jazz (JA)	74	Pop	93
		Total	**1008**

IV. Experimental Results

This section is divided into three parts, initially the features selection and preprocessing results are presented. In the following, the proposed classifier systems are presented and compared. Finally, the proposed system is compared to previous works.

A. Features selection and preprocessing

It is possible to observe from Fig. 4 typical audio attributes (MFCCs and beat histogram) for different music genres. It is interesting to note that the patterns are not easily distinguished, and indeed in some cases are quite similar (e.g. MPB and Samba).

In order to evaluate the effects of information compaction, the principal component analysis (PCA) load curve was estimated. From Fig. 5 it can be observed that the first 22 more energetic principal components (from the 26 original features) retain approximately 99.9% of the total energy. This indicates that probably some features present high mutual redundancy.

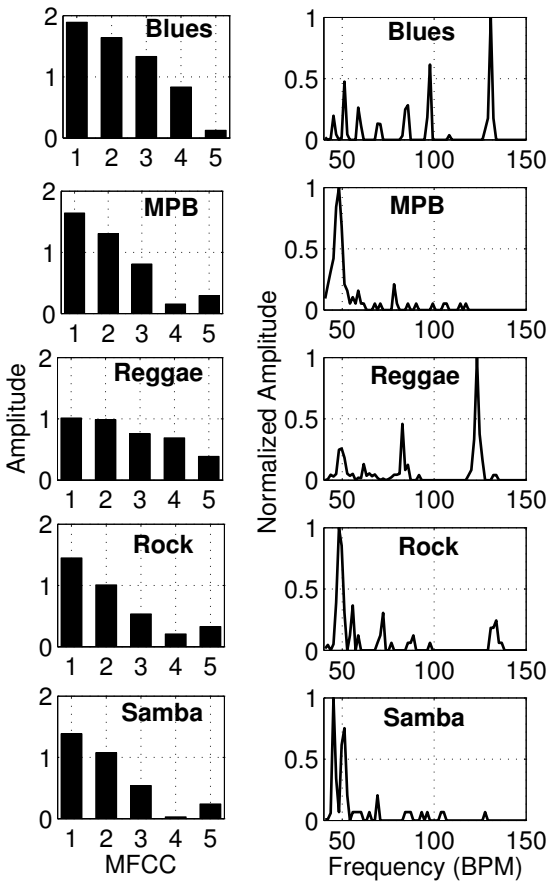


Fig. 4. Typical MFCC and Beat Histograms for five different music genres.

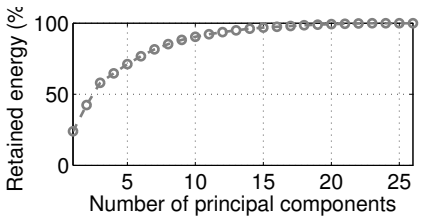
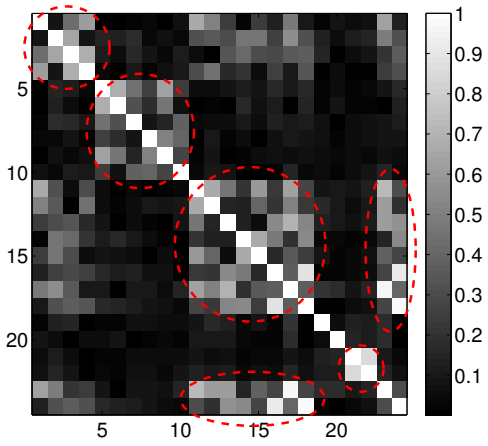


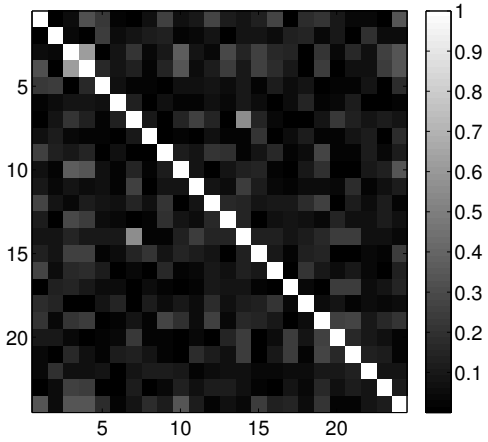
Fig. 5. PCA load curve for the used input features.

As it can be seen from the input feature correlation matrix in Fig. 6-(a), there is considerable correlation between some characteristics (see the dashed-circled areas in Fig. 6-(a)). This may contribute to hamper the classifiers training process. To reduce the redundancy between the input features, the attributes are processed using independent component analysis. It can be seen from Fig. 6-(b) that there is a considerable correlation reduction, as evidenced by the quasi-diagonal correlation matrix after ICA. The estimated independent components are used as new inputs for the classifier systems.

For estimation of features relevance the procedure adopted in this work was the sequential backward elimination [31]. In this case, each individual feature is removed from the feature set, the classifier re-trained and the performance index



(a)



(b)

Fig. 6. Input features correlation matrix for the testing set (a) before and (b) after ICA.

computed and compared to the one obtained by using the complete features set. For this procedure a SLFN classifier using 40 hidden neurons was used (the choise of the number of hidden neurons will be explained further). The results are illustrated in Fig 7-(a). It can be observed that by removing two individual features (namely feat-01, the mean of ZCR; and feat-20, the period of beat histogram first peak) the global classification performance improves. This clearly indicates that these are confusing features and thus, they should be removed from the features set. There are also some other features that do not considerably contribute for class discrimination as their removal produces a slight variation on the global discrimination index (namely feat-4, the variance of the first MFCC; and feat-25, the mean of loudness).

A complementary analysis considered the removal of sets of features computed from the same audio descriptor (see Fig 7-(b)). In this case it is observed that the MFCC and the beat-histogram (BH) sets of features are the most relevant ones. This is interesting to note that the BH individual features (feat-19 to feat-22 in Fig 7-(a)) are not highly relevant, but when

they are considered together, they contribute significantly for class discrimination. After this feature relevance analysis it was decided to eliminate features 01 and 20.

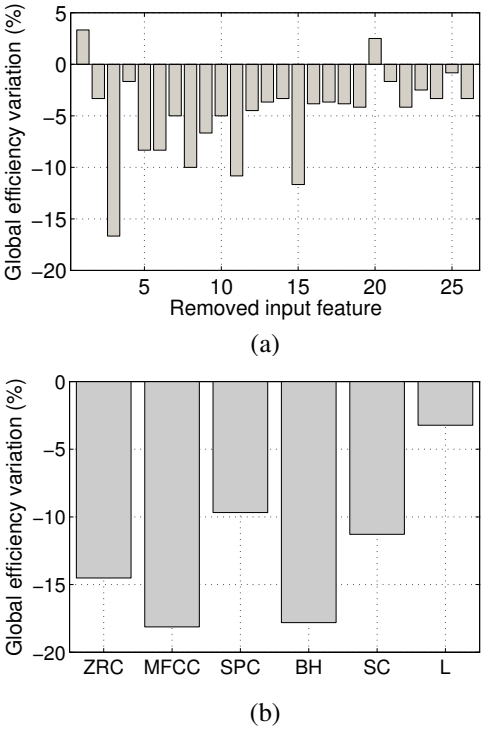


Fig. 7. Global efficiency variation after removing (a) individual features and (b) sets of features.

B. Proposed classifiers efficiency evaluation

In this work four different classifiers were trained for the Brazilian music genre classification problem: (i) a SLFN classifier fed from the 24 more discriminant features (called SLFN); (ii) a SLFN classifier fed from 24 discriminant and independent features (SLFN-I); (iii) SLFN classifiers trained in a one-against-all configuration (called SLFN_{OAA}); and (iv) a SVM classifier, which was also trained in a one-against-all configuration (called SVM).

For training of SLFN classifiers it is important to properly determine the number of neurons in the hidden layer. In this work this was achieved by a network growing procedure. As illustrated in Fig. 8 for SLFN and SLFN-I classifiers, it can be seen that the highest discrimination efficiency was achieved for the SLFN-I classifier with 35 hidden neurons. The best result for the SLFN classifier was achieved for 40 hidden neurons. An interesting aspect also observed in Fig. 8 is that the use of independent features consistently produced higher discrimination efficiencies and smaller statistical fluctuations in the final global performance.

For the SVM classifiers tests were performed using different kernel functions (linear, radial basis, polynomial and sigmoid) and the best discrimination results were obtained for the sigmoid kernel.

The discrimination efficiencies obtained from different classifiers are summarized in Table III. As it can be seen, the

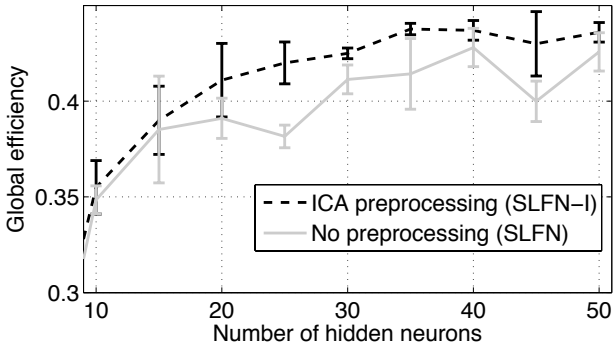


Fig. 8. Global discrimination efficiency of the SLFN classifiers as a function of the number of hidden neurons.

SVM classifier presented the highest global efficiency (62.6 %) among the used classifier designs. For some specific musical genres such as Classical, Forró, Pop and Reggae, different classifiers present higher individual class discrimination accuracy. If a combination of the best-cases was used for genre identification the global efficiency shall be increased to 64.4 %.

TABLE III
DISCRIMINATION EFFICIENCIES (EF_n IN %) FOR DIFFERENT CLASSIFIERS, BEST CASES ARE IN BOLD, GLOBAL PERFORMANCE IS EXPRESSED IN TERMS OF $\overline{EF}_{\text{MAX}}$ ($\sigma_{\overline{EF}}$).

Genre	SLFN	SLFN-I	$\text{SLFN-I}_{(\text{OAA})}$	SVM
BL	49.8	54.1	65.5	**77.6**
CL	**77.6**	75.3	75.7	69.8
CO	50.3	43.5	67.9	**82.0**
FO	51.6	56.2	**59.3**	57.1
HH	60.3	53.2	57.8	**60.4**
JA	33.5	29.8	47.5	**50.1**
MPB	23.1	27.4	43.6	**64.7**
Pop	43.1	**46.5**	44.1	41.3
RE	60.7	**63.5**	60.2	60.2
RO	28.3	33.7	51.0	**53.9**
SA	34.0	38.9	57.5	**86.6**
SO	26.9	29.1	45.1	**63.2**
Global	42.1 (4.3)	43.8 (2.1)	55.4 (2.3)	**62.6 (1.8)**

The confusion matrix for the SVM classifier is presented in Table IV. It can be seen that, in most cases, the cross-confusion between two genres is below 5 % and only in few cases it is above 10 % (highlighted inside boxes in Table IV). Some high confusion rates appeared for genres which present similar characteristics (and in some cases are also confused by human listeners) such as Reggae and Forró, Jazz and Soul, Pop and Reggae.

Considering aspects related to the computational cost, Table V presents the average computational time in system operation (in % of the total time) required for each signal processing step. The used classifier was the SVM (the training phase was not considered, only system operation) and this analysis was performed using a Texas Instruments TMS320C6713 DSP (clock-frequency of 225 MHz, 192 KB of internal memory, 512 KB of flash memory and 16 MB of SDRAM [32]). As it

TABLE IV
CONFUSION MATRIX (IN %) FOR THE SVM CLASSIFIER WITH ICA PREPROCESSING (IN ROWS ARE THE TRUE CLASS AND DETECTED CLASS IN COLUMNS, CLASS EFFICIENCIES ARE IN BOLD).

	BL	**CL**	**CO**	**FO**	**HH**	**JA**
BL	**77.6**	2.3	1.0	0.5	0.0	2.2
CL	0.9	**69.8**	1.0	2.0	0.0	2.2
CO	1.1	0.0	**82.0**	7.0	0.0	4.0
FO	2.7	4.4	1.7	**57.1**	7.5	4.5
HH	2.0	2.4	2.1	5.6	**60.4**	5.0
JA	6.6	12.6	0.8	0.5	0.0	**50.1**
MPB	0.0	0.0	3.4	4.9	7.7	0.0
Pop	4.1	11.2	2.0	4.6	0.0	10.5
RE	0.8	0.0	1.1	10.3	11.9	2.1
RO	4.9	6.2	2.6	2.0	0.0	7.9
SA	0.0	0.0	2.9	4.5	0.0	0.0
SO	3.1	8.2	2.8	0.0	0.0	4.8
	MPB	**Pop**	**RE**	**RO**	**SA**	**SO**
BL	4.5	3.4	1.8	4.5	1.2	1.0
CL	4.4	6.7	1.1	11.5	0.4	0.0
CO	0.0	5.2	0.0	0.0	0.7	0.0
FO	2.9	3.1	8.6	3.6	2.2	1.7
HH	4.8	8.5	2.5	3.2	2.2	1.3
JA	2.0	10.0	1.8	1.5	0.2	13.9
MPB	**64.7**	6.1	9.6	0.6	3.0	0.0
Pop	3.5	**41.3**	10.1	7.9	1.4	3.4
RE	0.4	4.7	**60.2**	3.2	0.5	4.8
RO	2.9	3.0	5.4	**53.9**	1.4	9.8
SA	2.9	0.0	0.0	3.1	**86.6**	0.0
SO	4.8	6.4	0.0	4.2	2.5	**63.2**

can be observed, the classification module requires only 1 % of the total time. The average processing time for each audio file is approximately 300 ms, and the complete dataset may be processed in less than six minutes. These results indicate that it may be possible to produce a version of the proposed system for embedded applications which may present a relatively fast response to the final user.

TABLE V
AVERAGE PROCESSING TIME (IN % OF THE TOTAL TIME) REQUIRED FOR EACH SIGNAL PROCESSING STEP.

MFCC	**SPC**	**BH**	**L**	**SC**	**ZRC**	**Class.**
49	27	15	3	3	2	1

C. Comparison with previous works

In order to allow fair comparison of the proposed system with previous research in automatic music genre classification, were used here the results presented in [33] for different databases such as GTZAN [4], ISMIR 2004 [34], Homburg [35] and 1,517 Artists [36].

Table VI presents a summary comprising some relevant aspects of the proposed classification systems (the cases marked with '*' use short-length audio excerpts, instead of full-length music files). It can be observed that the global discrimination results are usually higher for datasets with smaller number of genres. Considering this, the proposed system, which comprises 12 genres and is the only one that uses discriminant and independent features, present results comparable to a dataset comprising only 9 genres.

Another interesting aspect is that only in [36] Latin music is considered explicitly in the classification problem and none of these works considered the diversity and the particularities of Brazilian music.

TABLE VI
COMPARISON OF THE PROPOSED SYSTEM WITH RELATED PREVIOUS WORKS, THE DATABASE USED IN THIS WORK IS NAMED 'BR_MUS'.

Database	[4]	[34]	[35]	[36]	**BR_Mus**
Genres	10	6	9	19	12
Files	1,000*	1,458	1,886*	3,180	1,008
SVM	86.6 %	83.0 %	62.6 %	53.3 %	62.5 %
SLFN	81.1 %	78.6 %	50.3 %	44.9 %	55.4 %

V. CONCLUSIONS

Music information retrieval from multimedia files is very important in the search for desired contents in large non-tagged databases. This work deals with the identification of the prevailing musical genre for a dataset which includes Brazilian genres. As the Brazilian culture comprises multiple influences (European, African and Native-American), its musical genres present very specific characteristics, which can only be accounted by designing a specific automatic music genre identification system. The experimental results indicates that the combination of relevant and independent input features with SVM classifiers produce a music genre classification system with efficiency compatible to previous results presented in this field. Additionally, a preliminary implementation of the proposed system in embedded electronics indicates that it may be possible to develop a version for mobile devices.

ACKNOWLEDGEMENTS

Financial support from the Brazilian Research Council (CNPq) and Bahia State Foundation for Research Support (FAPESB) is gratefully acknowledged.

REFERENCES

[1] T. Li, M. Ogihara, and G. Tzanetakis, Eds., *Music Data Mining*, 1st ed., ser. Chapman & Hall/CRC Data Mining and Knowledge Discovery Series. Boca Raton, US: CRC Press, 2011, no. 21. [Online]. Available: http://onlinelibrary.wiley.com/doi/10.1111/j.1751-5823.2012.0017913.x/full

[2] Z. Fu, G. Lu, K. M. Ting, and D. Zhang, "A survey of audio-based music classification and annotation," *Multimedia, IEEE Transactions on*, vol. 13, no. 2, pp. 303–319, April 2011. [Online]. Available: https://doi.org/10.1109/TMM.2010.2098858

[3] F. Fuhrmann and P. Herrera, "Quantifying the revelance of locally extracted information for musical instrument recognition from entire pieces of music," in *Proceedings of the 12th International Society for Music Retrieval Conference*, Miami, US, October 2011, pp. 239–244. [Online]. Available: https://doi.org/10.1.1.352.6584

[4] G. Tzanetakis and P. Cook, "Musical genre classification of audio signals," *Speech and Audio Processing, IEEE Transactions on*, vol. 10, no. 5, pp. 293–302, Jul 2002. [Online]. Available: https://doi.org/10.1109/TSA.2002.800560

[5] T. Li, M. Ogihara, and Q. Li, "A comparative study on content-based music genre classification," in *Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Informaion Retrieval*, ser. SIGIR '03, New York, NY, USA, 2003, pp. 282–289. [Online]. Available: https://doi.org/10.1145/860435.860487

[6] N. Scaringella, G. Zoia, and D. Mlynek, "Automatic genre classification of music content: a survey," *Signal Processing Magazine, IEEE*, vol. 23, no. 2, pp. 133–141, March 2006. [Online]. Available: https://doi.org/10.1109/MSP.2006.1598089

[7] J. J. Burred and A. Lerch, "Hierarchical automatic audio signal classification," *Journal of the Audio Engineering Society (JAES)*, vol. 07, no. 52, pp. 724–739, 2004. [Online]. Available: http://www.aes.org/e-lib/browse.cfm?elib=13015

[8] J. Abeber, H. Lukashevich, and P. Brauer, "Classification of music genres based on repetitive basslines," *Journal of New Music Research*, vol. 41, no. 3, pp. 239–257, 2012. [Online]. Available: http://dx.doi.org/10.1080/09298215.2011.641571

[9] P. Boot, A. Volk, and W. B. de Haas, "Evaluating the role of repeated patterns in folk song classification and compression," *Journal of New Music Research*, vol. 45, no. 3, pp. 223–238, 2016. [Online]. Available: http://dx.doi.org/10.1080/09298215.2016.1208666

[10] K. Neubarth, D. Shanahan, and D. Conklin, "Supervised descriptive pattern discovery in native american music," *Journal of New Music Research*, vol. 46, no. 0, pp. 1–16, 2017. [Online]. Available: http://dx.doi.org/10.1080/09298215.2017.1353637

[11] C. McGowan and R. Pessanha, *The Brazilian Sound: Samba, Bossa Nova, and the Popular Music of Brazil*, Revised ed. Philadelphia, US: Temple University Press, 2008.

[12] D. F. P. Melo, I. S. Fadigas, and H. B. B. Pereira, "Categorisation of polyphonic musical signals by using modularity community detection in audio-associated visibility network," *Applied Network Science*, vol. 2, no. 32, pp. 1–15, 2017. [Online]. Available: http://dx.doi.org/10.1007/s41109-017-0052-1

[13] S. O. Haykin, *Neural Networks and Learning Machines*, 3rd ed. Pearson, 2008.

[14] A. Hyvarinen and E. Oja, "Independent component analysis: Algorithms and applications," *Neural Networks*, vol. 13, no. 4-5, pp. 411–430, 2000. [Online]. Available: https://doi.org/10.1016/S0893-6080(00)00026-5

[15] A. Lerch, *An Introduction to Audio Content Analysis*. Wiley-IEEE Press, 2012. [Online]. Available: http://dx.doi.org/10.1002/9781118393550

[16] M. Sahidullah and G. Saha, "Design, analysis and experimental evaluation of block based transformation in MFCC computation for speaker recognition," *Speech Communication*, vol. 54, no. 4, pp. 543 – 565, 2012. [Online]. Available: https://doi.org/10.1016/j.specom.2011.11.004

[17] J. Nirmal, M. Zaveri, S. Patnaik, and P. Kachare, "Novel approach of MFCC based alignment and wd-residual modification for voice conversion using RBF," *Neurocomputing*, vol. 237, pp. 39 – 49, 2017. [Online]. Available: https://doi.org/10.1016/j.neucom.2016.07.048

[18] E. Schubert, J. Wolfe, and A. Tarnopolsky, "Spectral centroid and timbre in complex, multiple instrumental textures," in *Proceedings of the International Conference on Music Perception and Cognition*, Evanston, IL, August 2004. [Online]. Available: http://icmpc.org/icmpc14/files/ICMPC14_Proceedings.pdf

[19] S. Ferguson, D. Cabrera, and E. Schubert, "Comparing continuous subjective loudness responses and computational models of loudness for temporally varying sounds," in *129th AES Convention*, San Francisco, CA, November 2010. [Online]. Available: http://www.aes.org/e-lib/browse.cfm?elib=15600

[20] International Telecommunication Union, "Recommendation ITU-R BS.1387-1 - method for objective measurements of perceived audio quality," ITU, Tech. Rep., 2001. [Online]. Available: https://www.itu.int/rec/R-REC-BS.1387

[21] D. G. Silva, L. T. Duarte, and R. Attux, "Blind source separation: Fundamentals and perspectives on galois fields and sparse signals," *Journal of Communication and Information Systems*, vol. 31, no. 1, pp. 177–187, 2016. [Online]. Available: http://dx.doi.org/10.14209/jcis.2016.16

[22] I. T. Jolliffe, *Principal Component Analysis*, 2nd ed. New York, USA: Springer, 2002.

[23] Y. Khan, "Partial discharge pattern analysis using PCA and back-propagation artificial neural network for the estimation of size and position of metallic particle adhering to spacer in gis," *Electrical Engineering*, vol. 98, no. 1, pp. 29–42, 2016. [Online]. Available: https://doi.org/10.1007/s00202-015-0343-4

[24] Y. Aslan and Y. E. Yagan, "Artificial neural-network-based fault location for power distribution lines using the frequency spectra of fault data," *Electrical Engineering*, pp. 1–11, 2016. [Online]. Available: https://doi.org/10.1007/s00202-016-0428-8

[25] G. Trejo-Caballero, H. Rostro-Gonzalez, R. J. Romero-Troncoso, C. H. Garcia-Capulin, O. G. Ibarra-Manzano, J. G. Avina-Cervantes, and A. Garcia-Perez, "Multiple signal classification based on automatic order selection method for broken rotor bar detection in induction motors," *Electrical Engineering*, pp. 1–10, 2016. [Online]. Available: https://doi.org/10.1007/s00202-016-0463-5

[26] F. C. Cruz, E. F. Simas Filho, M. C. S. Albuquerque, I. C. Silva, C. T. T. Farias, and L. L. Gouvea, "Efficient feature selection for neural network based detection of flaws in steel welded joints using ultrasound testing," *Ultrasonics*, vol. 73, pp. 1 – 8, 2017. [Online]. Available: https://doi.org/10.1016/j.ultras.2016.08.017

[27] H. Lam, U. Ekong, H. Liu, B. Xiao, H. Araujo, S. H. Ling, and K. Y. Chan, "A study of neural-network-based classifiers for material classification," *Neurocomputing*, vol. 144, pp. 367 – 377, 2014. [Online]. Available: https://doi.org/10.1016/j.neucom.2014.05.019

[28] I. Steinwart and A. Chrinstmann, *Support Vector Machines*. Springer, 2008.

[29] J. Saeedi, S. M. Ahadi, and K. Faez, "Robust voice activity detection directed by noise classification," *Signal, Image and Video Processing*, vol. 9, no. 3, pp. 561–572, 2015. [Online]. Available: https://doi.org/10.1007/s11760-013-0479-5

[30] A. M. Ashir and A. Eleyan, "Facial expression recognition based on image pyramid and single-branch decision tree," *Signal, Image and Video Processing*, pp. 1–8, 2017. [Online]. Available: https://doi.org/10.1007/s11760-016-1052-9

[31] I. Guyon, S. Gunn, M. Nikravesh, and L. A. Zadeh, *Feature Extraction: Foundations and Applications*. Springer, 2006.

[32] Texas Instruments, *TMS320c6713 Datasheet*, 2001. [Online]. Available: http://www.ti.com/lit/ds/symlink/tms320c6713.pdf

[33] Y. Panagakis and C. Kotropoulos, "Music classification by low-rank semantic mappings," *EURASIP Journal on Audio, Speech, and Music Processing*, vol. 2013, no. 1, p. 13, 2013. [Online]. Available: http://dx.doi.org/10.1186/1687-4722-2013-13

[34] ISMIR, "ISMIR2004 audio description contest," 2004, http://ismir2004.ismir.net/.

[35] H. Homburg, I. Mierswa, B. Moller, K. Morik, and M. Wurst, "A benchmark dataset for audio classification and clustering," in *Proc. 6th Int. Conf. Music Information Retrieval*, London, September 2005, pp. 528–531. [Online]. Available: http://ismir2005.ismir.net/proceedings/2117.pdf

[36] K. Seyerlehner, G. Widmer, T. Pohle, and P. Knees, "Fusing block-level features for music similarity estimation," in *Proc. 13th Int. Conf. Digital Audio Effects*, Graz, September 2010, pp. 1–8. [Online]. Available: http://dafx10.iem.at/proceedings

Eduardo F. Simas Filho received the B.Sc. (2001) and M.Sc. (2004) in Electrical Engineering from the Federal University of Bahia, Brazil and the PhD in Electrical Engineering from the Federal University of Rio de Janeiro (COPPE/UFRJ), Brazil (2010). Since 2011 he has been an associate professor with the Federal University of Bahia (Electrical and Computer Engineering Department), where he served as the coordinator of the Computer Engineering undergraduate program (2014-2016). His research interests include the application of digital signal processing and machine learning methods to instrumentation systems.

Elmo A. Borges Jr. received the B.Sc. (2013) in Electrical Engineering from the Federal Institute for Education, Science and Technology of Bahia, Brazil and is currently attempting to achieve his M.Sc. in Electrical Engineering at the Federal University of Bahia, Brazil. Since 2012 he has been a maintenance technician with Petrobras.

Antonio C. L. Fernandes Jr. received the B.Sc. (2000) in Electrical Engineering from the Federal University of Bahia, Brazil and the M.Sc. (2005) and PhD (2015) in Electrical Engineering from the State University of Campinas (UNICAMP), Brazil. Since 2012 he has been with the Federal University of Bahia (Electrical and Computer Engineering Department). He is currently the coordinator of the Computer Engineering undergraduate program. His main research interest include audio signal processing and machine learning.