Evaluating automated classification techniques for folk music genres from the Brazilian Northeast

Jeronimo Barbosa[1], **Cory McKay**[1], **Ichiro Fujinaga**[1]

[1]CIRMMT, Schulich School of Music, McGill University
555 Sherbrooke St. West, Montreal, Canada

{jeronimo.costa,cory.mckay}@mail.mcgill.ca, ich@music.mcgill.ca

Abstract. *In this exploratory study, we investigate the problem of automated genre classification focusing on the particular context of folk music genres from the Brazilian Northeast—an issue still under-explored in the literature. As a contribution, we have: a) created a new public dataset with 75 samples equally distributed over 5 different genres (i.e.,* Cavalo Marinho*,* Ciranda*,* Coco*,* Maracatu de Baque Solto *and* Maracatu de Baque Virado*); and b) evaluated 68 features and 10 classifiers, aiming to find those well-suited to this particular classification task. Our results demonstrated high classification accuracy rates: 60% before feature selection when using either the Naive Bayes or Support Vector Machine algorithms, and a very impressive increase to 100% classification after feature selection with our best-performing feature selection methodology.*

1. Introduction

"Brazilian cultural heritage, like racial identification in the country, is mixed, complex and diverse. Nowhere is this situation more evident than in the Brazilian Northeast" [Crook, 2005]. This heritage can also be observed in music. In Pernambuco (a state from the Brazilian Northeast), examples like 'maracatu', 'cavalo marinho', and 'coco' show us how these influences were incorporated over the years into unique musical genres, playing an important role on what is known today as *'Música Popular Brasileira'* (MPB).

However, these genres are fairly unknown outside the "broadcast, recording, print, and electronic media that circulate" about Brazil [Crook, 2005]. The difference between them can be subtle and confusing for non-specialists. This problem is enhanced by the existing oral tradition in the Brazilian Northeast and by the relative lack of educational books to guide newcomers [Souza, 2011]. Perhaps those reasons might explain: a) the lack of studies investigating the suitability of traditional Music Information Retrieval (MIR) techniques in the context of folk music from northeastern Brazil; and b) the lack of public databases of folk music from northeastern Brazil, which could be used by MIR systems.

In order to address these issues, this exploratory study focuses on the usage of automatic genre classification in the context of folk music genres from the northeastern Brazil. As a contribution, we have: a) created a new public dataset consisting of 75 samples equally distributed over 5 different genres (i.e., *Cavalo Marinho*, *Ciranda*, *Coco*, *Maracatu de Baque Solto* and *Maracatu de Baque Virado*); and b) evaluated different features and classifiers, aiming to find which ones would be more suitable for the classification task. As tools, we used jMIR/jAudio [McKay, 2010], for the feature extraction, and Weka [Witten et al., 2011], for the classification and feature selection.

2. Background

Audio-based genre classification is a traditional task for MIR systems, included in the MIREX (Music Information Retrieval Evaluation eXchange) since its beginning in 2005 [1]. As a result, we have today several different approaches—see [Scaringella et al., 2006] for a comprehensive survey—and some high-level support tools designed for the task, such as the jMIR [McKay, 2010] and the Marysas [2].

Although there are no studies concerning the application of these techniques for the folk musical genres from Northeast Brazil, they have already been used in other very specific musical contexts. Recent examples include: a) 7 heavy metal music sub-genres [Tsatsishvili, 2011], in which the author achieved 45.7% accuracy with AdaBost algorithm; b) 9 Latin American Music genres [Völkel et al., 2010], in which the authors achieved 86.7% accuracy using a new technique based on the extraction of rhythmic patterns and template matching; and c) 8 electronic dance music sub-genres [Leimeister et al., 2014], in which the authors achieved 71% accuracy using a similar technique.

3. Methodology

In order to achieve the goals specified, we used the following methodology:

1. We listed and analyzed which musical genres could be considered for this project;
2. We created the necessary dataset for this project;
3. We extracted features from the audio files using jAudio;
4. We compared performances of different classifiers with 5-fold cross validation using Weka; and
5. We investigated different feature selection techniques and their impact on the classification using Weka.

Each one of these steps will be further discussed in the following sections.

3.1. Choosing the genres

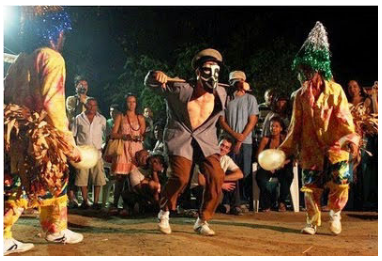
Folk music genres from northeastern Brazil might share some common characteristics. Examples include rigid musical structure (e.g., instrumentation and rhythm could have little variation throughout songs), predominant use of percussion instruments, where rhythm plays an important role, and close relationships to traditional celebrations or rituals—as shown in Figure 1. They also have a great component of oral tradition [Souza, 2011] and—although a simple task for people experienced in music or those traditions—there is not much material dedicated to teaching how to identify these genres.

One of the few exceptions is the "Manual dos ritmos pernambucanos" [Souza, 2011], in which the author presents rhythmic characteristics of 9 genres from Pernambuco, a state in the Brazilian Northeast. Based on this material, we decided to focus our study on a subset of 5 genres. We believe these examples covers part of the cultural diversity on the Brazilian Northeast, without making our scope too broad. They are:

- Maracatu de baque solto;
- Maracatu de baque virado;
- Cavalo marinho;
- Ciranda; and
- Coco.

(a) Maracatu de Baque Solto

(b) Maracatu de Baque Virado



(c) Cavalo Marinho

(d) Ciranda

(e) Coco

Figure 1: The chosen genres [3]

Audio examples for each of these genres can be found online [4].

3.2. Our dataset

The lack of public databases focused on the selected genres motivated us to create a new one. Our dataset was composed of audio extracted from public videos available on Youtube. We gave preference to audio extracted from amateur videos recorded in the context/conditions those genres are often performed (i.e., in the streets, during rehearsals or presentations, with background noise). We did this in order to create a standard audio quality across the different genres, as we could not find recordings with good audio quality for all genres. In addition, public videos allows other researchers to have access to our dataset.

In total, 15 videos were selected for each genre. For each audio extracted, a small excerpt—30 seconds duration—was randomly selected by using a custom script [5], resulting in 75 labeled samples. The 30-second duration was chosen for two reasons: a) Since the duration of the video files ranged from 40 seconds to 3 minutes, the 30-second selection process would cover all cases; b) We believe that 30-second is sufficient for a human specialist to recognize the genre.

The list of selected videos [6] and the final dataset are publicly available online.

3.3. Feature Extraction

Features were extracted using the jAudio tool, provided by jMIR framework [McKay, 2010]. As settings, we used the default window size of 512 with no overlap. In total, 13 low-level features were used. Except for the 'Method of

[1]http://www.music-ir.org/mirex/wiki/2005

[2]http://marsyas.info/

[3]Image sources: (a) http://goo.gl/0Rxwie ; (b) http://goo.gl/3BtIIZ ; (c) http://goo.gl/NWZsup ; (d) http://goo.gl/UmNEfo ; (e) http://goo.gl/2ymyvP .

[4]https://github.com/jeraman/sbcm2015-dataset/tree/master/database

[5]https://github.com/jeraman/sbcm2015-dataset/blob/master/mp3randomchopper.py

[6]https://github.com/jeraman/sbcm2015-dataset/blob/master/database%20youtube%20list.txt

Moments' (excluded from our study due to instability), these are the default features suggested by jAudio. They are (name followed by its respective jAudio description):

Spectral Centroid: "The centre of mass of the power spectrum";
Spectral Rolloff Point: "The fraction of bins in the power spectrum at which 85% of the power is at lower frequencies. This is a measure of the right-skewness of the power spectrum";
Spectral Flux: "A measure of the amount of spectral change in a signal. Found by calculating the change in the magnitude spectrum from frame to frame";
Compactness: "A measure of the noisiness of a signal. Found by comparing the components of a window's magnitude spectrum with the magnitude spectrum of its neighbouring windows";
Spectral Variability: "The standard deviation of the magnitude spectrum. This is a measure of the variance of a signal's magnitude spectrum";
Root Mean Square: "A measure of the power of a signal";
Fraction Of Low Energy Windows: "The fraction of the last 100 windows that has an RMS less than the mean RMS in the last 100 windows. This can indicate how much of a signal is quiet relative to the rest of the signal";
Zero Crossings: "The number of times the waveform changed sign. An indication of frequency as well as noisiness";
Strongest Beat: "The strongest beat in a signal, in beats per minute, found by finding the strongest bin in the beat histogram";
Beat Sum: "The sum of all entries in the beat histogram. This is a good measure of the importance of regular beats in a signal";
Strength Of Strongest Beat: "How strong the strongest beat in the beat histogram is compared to other potential beats";
MFCC (split into 13 different features): "MFCC calculations based upon Orange Cow code"; and
LPC (split into 10 different features): - 'Linear Prediction Coefficients calculated using autocorrelation and Levinson-Durbin recursion".

These low-level features generated in total 68 summary features (i.e., overall standard deviation and average for each low-level-feature), which were used for the classification. A file containing the values of all extracted features, for each sample in our database, is available online [7].

3.4. Classification

For the classification task, we have used Weka. No default parameters were changed (except for the k-NN, for each we used k=3). The classifiers tested were:

- k-NN;
- NaiveBayes;
- j48;
- NBTree;
- MultiClassClassifier;
- Classification via regression;
- Support Vector Machine (SMO);
- Neural Network (Multilayer Perceptron);
- AdaBoost; and
- DecisionTable.

[7] https://github.com/jeraman/sbcm2015-database/tree/master/extracted%20features

It is important to stress that there is a large diversity of approaches towards classification in literature, and some might have been omitted due to scope constraints. We reasoned that the algorithms selected are among the most basic and popular, covering at the same time part of the diversity.

The classifiers were tested with all 68 summary features using 5-fold cross-validation. We chose 5-fold cross validation due to the limited size of our dataset. We emphasize that there was no overlap in terms of the samples in matching training and testing folds during cross validation. In order to provide basis for comparison, we also highlight that the success rate for a random classifier is 20%. Results are summarized in Table 1.

Table 1: Results for each classifier tested with all 68 features, using 5-fold cross validation. The best result was achieved with the Naive Bayes and the Support Vector Machine (60%).

Classifier	Correctly Classified Instances
k-NN (k=3)	58.6667%
NaiveBayes	**60%**
j48	45.33%
NBTree	50.67%
MultiClassClassifier	50.67%
NBTree	54.67%
Classification via regression	46.67%
Support Vector Machine	**60%**
Multilayer Perceptron	57.33%
AdaBoost	33.33%
DecisionTable	44%

The best results were achieved with the Naive Bayes and the Support Vector Machine (both 60%). A complete report for each one of the classifiers can be found online [8]. Considering these cases, we present the detailed accuracy and confusion matrixes respectively in Tables 2, 3, 4, and 5.

Table 2: Detailed accuracy for the *Naive Bayes* classifier using all 68 features.

Class	Precis.	Recall	F-Meas.
Cavalo Marinho	0.692	0.6	0.643
Coco	0.417	0.333	0.37
Ciranda	0.545	0.4	0.462
Maracatu Solto	0.6	0.8	0.686
Maracatu Virado	0.684	0.867	0.765
Weig. Avg.	0.588	0.6	0.585

By analysing these results, we can notice that Coco was the genre which received the larger number of misclassifications (10) throughout the scenarios (i.e., Coco samples were wrongly classified as another genre), followed by Ciranda (9 misclassifications with

[8] https://github.com/jeraman/sbcm2015-database/tree/master/stage%201%20-%20classification

Table 3: Confusion Matrix for the *Naive Bayes* classifier using all 68 features.

Cav. Marinho	Coco	Ciranda	Mar. Solto	Mar. Virado	*<-classified as*
9	2	0	4	0	**Cav. Marinho**
2	**5**	4	1	3	**Coco**
1	2	**6**	3	3	**Ciranda**
1	1	1	**12**	0	**Mar. Solto**
0	2	0	0	**13**	**Mar. Virado**

Table 4: Detailed accuracy for the *Support Vector Machine* classifier using all 68 features.

Class	Precis.	Recall	F-Meas.
Cavalo Marinho	0.9	0.6	0.72
Coco	0.357	0.333	0.345
Ciranda	0.5	0.6	0.545
Maracatu Solto	0.625	0.667	0.645
Maracatu Virado	0.706	0.8	0.75
Weig. Avg.	0.618	0.6	0.601

Table 5: Confusion Matrix for the *Support Vector Machine* classifier using all 68 features.

Cav. Marinho	Coco	Ciranda	Mar. Solto	Mar. Virado	*<-classified as*
9	3	1	2	0	**Cav. Marinho**
0	**5**	5	2	3	**Coco**
0	2	**9**	2	2	**Ciranda**
1	2	2	**10**	0	**Mar. Solto**
0	2	1	0	**12**	**Mar. Virado**

Naive Bayes, 6 with Support Vector Machines). Their accuracy are considerably below the average. In the case of Coco, for example, variables such as 'Precision', 'Recall' and 'F-Measure' are almost half of the average. Further studies are needed in order to address this issue.

3.5. Feature Selection

To improve the classification rates, we explored two different feature selection approaches: the wrapper method and filter method. Again, we reasoned that those are among the most basic and popular approaches employed in literature. For the wrapper method-based approaches, we tested:

- **'ClassifierSubsetEval'** as attribute method, and **'BestFirst'** as search method;
- **'ClassifierSubsetEval'** as attribute method, and **'GreedyStepWise'** as search method;
- **'WrapperSubsetEval'** as attribute method, and **'BestFirst'** as search method; and
- **'WrapperSubsetEval'** as attribute method, and **'GreedyStepWise'** as search method.

Regarding the filter method-based approaches, we tested:

- **'InfoGainAttributeEval'** as attribute method, and **'Ranker'** as search method.

In both cases, we used default parameters as provided in Weka. The only exception was the default classifier used to guide the feature selection, which was replaced by the Support Vector Machine classifier. The reason was its performance, as described in the previous subsection (*Classification*).

We trained these approaches with 80% of our original dataset (our training set), validating their performance with the remaining 20% (our validation set). Our intention was to "use an independent set to evaluate feature selection's efficacy", avoiding previously mentioned pitfalls in music classification [Fiebrink and Fujinaga, 2006]. Thus, the validation set did not include samples also found in the training set. The Support Vector Machine classifier was used again as a benchmark to compare improvement on accuracy. Without feature selection, its accuracy was 60%.

It is important to highlight the reduced size of our validation set (i.e., 15 samples, or 20% of the original dataset). Such small size may cause uncertainty about the precision of the results achieved. In order to minimize this issue, we have:

1. Randomly repartitioned the our training set into 5 new subsets. Each subset was composed of 48 (i.e., 80% of our training set) randomly selected samples from our training set. This step was done by using a custom script [9];
2. For each subset created, we performed an independent feature selection. All selected features were then independently validated (with our validation set). In the particular case of the 'InfoGainAttributeEval' based approach (which ranks all features by giving it a score), we only selected the features with scores greater than zero. Both selection and validation were performed in Weka;
3. For each tested approach, we first calculated the average and the standard error over the results achieved in each subset [10], as presented in Table 6. Finally, we have analysed the occurrence of the selected features over the subsets—summarized in Table 7. Features were listed if they appeared at least in half of the subsets [11].

Table 6: Average accuracy and standard error for each feature selection approach tested. Average calculated over the results achieved in each subset.

Feature Selection Approach	Avg. Accuracy	Avg. Standard Error
ClassifierSubsetEval + BestFirst	86.67%	3.4%
ClassifierSubsetEval + GreedyStepWise	88%	3.27%
InfoGainAttributeEval + Ranker	100%	0%
WrapperSubsetEval + BestFirst	86.67%	3.65%
WrapperSubsetEval + GreedyStepWise	85.33%	2.49%

As presented in Table 6, we noticed a significant improvement when using feature selection—no matter the approach chosen. The best result was achieved when using the 'InfoGainAttributeEval' (100% average accuracy with the Ranker search method). Even the worst scenario ('WrapperSubsetEval' plus 'GreedyStepWise') presented a significant improvement on accuracy (around 25.33% increase, with 2.49% of standard error).

[9] https://goo.gl/uJpQTH
[10] https://goo.gl/uI9b52
[11] https://goo.gl/KFNhmJ

Regarding the most common features—among the 68 tested— for the classification task, results are presented in Table 7. We note the predominance of 'Mel-frequency cepstral coefficients' (MFCC) related features. Considering the 15 features suggested by 'InfoGainAttributeEval', for example, 12 are MFCC related. Further studies are needed in order to address this issue.

Table 7: The most common features selected over the subsets used for training. Features presented below appeared at least in half of the subsets. Approaches are represented as: (a) *ClassifierSubsetEval + BestFirst*; (b) *ClassifierSubsetEval + GreedyStepWise*; (c) *InfoGainAttributeEval + Ranker*; (d) *WrapperSubsetEval + BestFirst*; (e) *WrapperSubsetEval + GreedyStepWise*.

Feature Selection Approach	Selected features	Occurrence over the folds
(a)	MFCC Overall Standard Deviation9	80%
	Fraction of Low Energy Wind. Overall Standard Deviation0	60%
	MFCC Overall Standard Deviation0	60%
	MFCC Overall Standard Deviation10	60%
	Spectral Variability Overall Average0	60%
	Beat Sum Overall Average0	60%
	MFCC Overall Average3	60%
	MFCC Overall Average8	60%
(b)	MFCC Overall Standard Deviation0	60%
	MFCC Overall Standard Deviation9	60%
	Spectral Variability Overall Average0	60%
	MFCC Overall Average3	60%
(c)	MFCC Overall Standard Deviation7	100%
	MFCC Overall Standard Deviation8	100%
	MFCC Overall Standard Deviation9	100%
	MFCC Overall Standard Deviation10	100%
	MFCC Overall Standard Deviation11	100%
	MFCC Overall Standard Deviation12	100%
	LPC Overall Standard Deviation0	100%
	Spectral Variability Overall Average0	100%
	MFCC Overall Average3	100%
	MFCC Overall Average4	100%
	MFCC Overall Average5	100%
	MFCC Overall Average9	100%
	MFCC Overall Average7	80%
	LPC Overall Standard Deviation8	60%
	MFCC Overall Average8	60%
(d)	MFCC Overall Average5	80%
	LPC Overall Standard Deviation3	60%
(e)	LPC Overall Standard Deviation3	60%
	MFCC Overall Average9	60%

These results suggest that feature selection plays an important role in improving accuracy on automated classification of folk music genres from the northeastern Brazil. Details (for both training and validation stages) are available online [12].

[12]https://github.com/jeraman/sbcm2015-database/tree/master/stage%202%20-%20features%20selection

4. Conclusion & Future Work

In this exploratory work, we have investigated the problem of automated genre classification considering the particular context of folk music genres from northeastern Brazil. As a contribution, we have: a) created a new public dataset with 75 samples equally distributed over 5 different genres (i.e., *Cavalo Marinho*, *Ciranda*, *Coco*, *Maracatu de Baque Solto* and *Maracatu de Baque Virado*); and b) evaluated different features and classifiers, aiming to find ones well-suited to this particular classification task.

Classification results—evaluated using 5-fold cross validation—showed high accuracy rates (e.g., 60% for both Naive Bayes and the Support Vector Machine classifier with all features) when compared to a random classifier (around 20% accuracy). We also highlight that Coco was the genre which received the larger number of misclassifications (10) throughout the best two scenarios. Further studies need to be performed in order to assess this issue.

Regarding the feature selection, a significant improvement (tested with an average accuracy of 89.33%) was found in all five approaches tested. As a highlight, 100% accuracy was achieved with the Support Vector Machine using 'InfoGainAttributeEval' as attribute method, and 'Ranker' as search method. The most common features were the ones related to the MFCC.

Finally, concerning the size of our dataset, it is important to stress its limitations. Previous works [McKay et al., 2006] have suggested that MIR databases should "include many thousands of recordings". The authors argue that this would "allow sufficient variety", and also "avoid research overuse of a relatively small number of recordings, which can result in overtraining". Such considerations must be taken into account when analyzing our results—especially the feature selection, as already pointed out. However, given the lack of other open databases of folk music from the Brazilian Northeast, we believe that the dataset we present is a valuable and relevant initial contribution to the research community.

As future work, we plan to compare the results achieved in this work with the ones achieved by human evaluators (both specialists and non specialists). In addition, over the long term we plan to include other genres in our analysis, such as Frevo and Baião.

References

Crook, L. N. (2005). *Brazilian Music: Northeastern Traditions and the Heartbeat of a Modern Nation*. ABC-CLIO, Santa Barbara, California.

Fiebrink, R. and Fujinaga, I. (2006). Feature Selection Pitfalls and Music Classification. In *International Conference on Music Information Retrieval*, pages 340–341.

Leimeister, M., Gaertner, D., and Dittmar, C. (2014). Rhythmic Classification of Electronic Dance Music. In *Audio Engineering Society Conference: 53rd International Conference: Semantic Audio*.

McKay, C. (2010). *Automatic music classification with jMIR*. PhD thesis, McGill University.

McKay, C., McEnnis, D., and Fujinaga, I. (2006). A large publicly accessible prototype audio database for music research. In *Proceedings of the International Conference on Music Information Retrieval*, pages 160–163.

Scaringella, N., Zoia, G., and Mlynek, D. (2006). Automatic genre classification of music content. *IEEE Signal Processing Magazine*, 23(2):133–141.

Souza, F. (2011). *Esquentando tambores: manual de percussao dos ritmos pernambucanos - escrita e tecnica*. Funcultura, CEL, Recife, Brazil.

Tsatsishvili, V. (2011). *Automatic subgenre classification of heavy metal music*. Master's thesis on music, mind and technology, University of Jyväskylä.

Völkel, T., Abeßer, J., Dittmar, C., and Großmann, H. (2010). Automatic genre classification of Latin American music using characteristic rhythmic patterns. *Proceedings of the 5th Audio Mostly Conference on A Conference on Interaction with Sound - AM '10*, pages 1–7.

Witten, I. H., Frank, E., and Hall, M. A. (2011). *Data Mining: Practical Machine Learning Tools and Techniques*. Morgan Kaufmann Publishers, 3rd edition.