A HIERARCHICAL MUSIC GENRE CLASSIFIER BASED ON USER-DEFINED TAXONOMIES

Juan José Burred
Communication Systems Group
Technical University of Berlin
burred@nue.tu-berlin.de

ABSTRACT

A system for classifying audio files according to music genre has been thoroughly evaluated within the MIREX 2005 Audio Description Contest. The system is based on a hierarchical classifier and on automatic feature selection. The results of the contest evaluation are presented here and compared with a previous evaluation performed by the authors.

Keywords: music genre classification, hierarchical classifier, music content analysis

1 INTRODUCTION

The music genre classification system submitted to the MIREX 2005 Audio Description Contest was originally presented in Burred and Lerch (2003) and thouroughly explained and evaluated in Burred and Lerch (2004). The main characteristics of the system are the following:

- **Long-term feature processing:** Classification is based on feature vectors whose elements are either statistical measures of short-time, frame-based features across long-term windows (with lengths of several seconds), or measures describing the rhythmic properties of the excerpt. The following short-term features are extracted: zero crossings, spectral centroid, spectral rolloff, spectral flux, Mel Frequency Cepstral Coefficients (MFCC), rms energy, time envelope, low energy rate, loudness, skewness, kurtosis, predictivity ratio and following MPEG-7 Audio features: AudioSpectrumCentroid, AudioSpectrumSpread, AudioSpectrumFlatness and HarmonicRatio. From each of these short-time features, the mean and standard deviation, as well as the mean and standard deviation of their derivatives, are extracted to form the feature vectors. This is the principle of *texture window* processing introduced by Tzanetakis and Cook (2002). The rhythm is described by means of the beat strength and rhythmic regularity features, which are based on the beat histogram of the analysis window beeing classified.

- **Automatic feature selection:** Given a subset of genres with their corresponding training data, the algorithm selects out of the full list of available features the ones that maximize genre separability. To this end, a sequential forward feature selection algorithm based on an objective measure of class separability is used. It should be noted that the separability measure used is solely based on the examination of the training examples, and not on the classification results.

- **Hierarchical feature selection and classification:** The system has been designed to work with hierarchical, multi-level taxonomies. The feature selection is repeated for each subset of classes of the taxonomy tree, so that only the features that are most suitable for separating that particular subset are retained. When classifying an unknown input signal, the appropriate features are selected and computed at each level of the hierarchy.

- **Parametric classification:** The classes are modelled as 3-cluster Gaussian Mixture Models. The classification is performed on a Maximum Likelihood basis.

The original system, as described in the above references, was designed and implemented using a fixed, predefined genre taxonomy (see fig. 1). Here, the algorithm has been enhanced in order to work with any taxonomy given as an input by the user. The taxonomy structure is extracted from the input text file listing the labeled training examples.

Also, the program has been wrapped into a M2K framework (using MATLAB integration modules), and its input and output requirements adapted to the standard agreed by MIREX participants.

Apart from that, the audio-specific part of the program submitted to MIREX 2005 is essentially the same as the original one in Burred and Lerch (2003), except for two slight modifications. Firstly, the tests on robustness to irrelevancies (specifically, added noise and reduced signal bandwidth) have been left out, because all the files on the contest database were expected to be of high quality. Secondly, the modified harmonic ratio feature has also been ignored due to its high computational requirements, and to the fact that it mostly contributed to improve performance in the case of chamber music and orchestral subgenres, which were not present in the contest.

Truth → Classification ↓	Ambient	Blues	Classical	Electr.	Ethnic	Folk	Jazz	New Age	Punk	Rock
Ambient	**55.88%**	0.00%	1.27%	1.22%	1.20%	0.00%	4.55%	26.47%	0.00%	4.76%
Blues	0.00%	**67.65%**	0.00%	1.22%	1.20%	4.17%	18.18%	2.94%	0.00%	3.57%
Classical	2.94%	5.88%	**78.48%**	0.00%	6.02%	0.00%	0.00%	8.82%	0.00%	0.00%
Electr.	5.88%	0.00%	0.00%	**45.12%**	7.23%	4.17%	4.55%	2.94%	5.88%	14.29%
Ethnic	11.76%	11.76%	12.66%	7.32%	**44.58%**	8.33%	0.00%	8.82%	0.00%	11.90%
Folk	0.00%	2.94%	1.27%	7.32%	7.23%	**58.33%**	0.00%	5.88%	0.00%	11.90%
Jazz	0.00%	5.88%	0.00%	4.88%	8.43%	0.00%	**54.55%**	2.94%	0.00%	1.19%
New Age	20.59%	0.00%	3.80%	12.20%	12.05%	8.33%	4.55%	**29.41%**	0.00%	7.14%
Punk	0.00%	0.00%	0.00%	0.00%	0.00%	4.17%	0.00%	0.00%	**82.35%**	4.76%
Rock	2.94%	5.88%	2.53%	20.73%	12.05%	12.50%	13.64%	11.76%	11.76%	**40.48%**

Table 1: Confusion matrix for the first MIREX 2005 evaluation (Magnatune dataset).

Truth → Classification ↓	Country	Electr. & Dance	New Age	Rap & Hip Hop	Reggae	Rock
Country	**86.90%**	7.46%	0.00%	0.00%	0.00%	24.55%
Electr. & Dance	2.38%	**65.67%**	4.76%	14.53%	11.11%	9.58%
New Age	1.19%	4.48%	**90.48%**	0.00%	0.00%	7.78%
Rap & Hip Hop	0.00%	8.96%	0.00%	**71.79%**	11.11%	8.98%
Reggae	0.00%	4.48%	0.00%	8.55%	**72.22%**	1.20%
Rock	9.52%	8.96%	4.76%	5.13%	5.56%	**47.90%**

Table 2: Confusion matrix for the second MIREX 2005 evaluation (USPOP dataset).

2 PREVIOUS EVALUATION

In the original paper, the system was evaluated using an audio database compiled by the author and consisting of 850 audio excerpts of about 30 seconds, sampled at 44.1 kHz and equally distributed into 17 audio classes (i.e., 50 files per class). The classes included 13 music genres as well as 3 speech classes and one background noise class, as can be seen in the taxonomy (fig. 1).

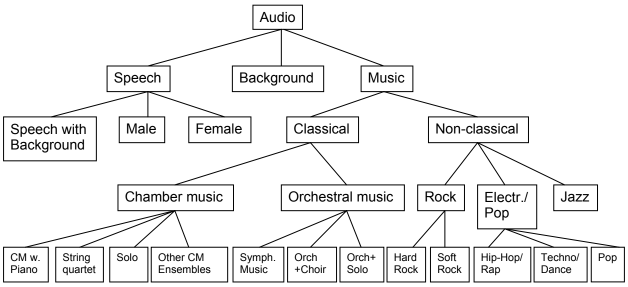

Figure 1: Audio taxonomy for the previous evaluation.

The system was evaluated using 10-fold cross-validation, obtaining an overall classification rate for all 17 classes of 58.71% (with standard deviation of 2.85%) using the hierarchical approach and of 59.76% (standard deviation 5.23%) using a flat, direct classification approach without hierarchy. It can be seen that the performance was similar for both approaches. Nevertheless, the hierarchical approach features the additional advantages of making the errors concentrate on their subgenre branches, a higher flexibility for future expansions, and the fact that it allows the design of genre-dependent features (Aucouturier and Pachet (2003)).

In the course of this evaluation, it was also obtained that analysis windows longer than 15 seconds did not significantly improve performance.

3 MIREX 2005 EVALUATION

For the MIREX 2005 contest, the system was evaluated two times with two different audio databases: Magnatune and USPOP, each time with a single-fold validation (just one iteration was run for each database).

3.1 Magnatune dataset

The Magnatune dataset consisted of 1515 full-length audio files, organized into 10 genres using the hierarchy of fig. 2. 66.4% of the files (1005) were used for training, the other 33.6% (510) for testing. The present system used the files sampled at 44.1 kHz and downmixed to mono. For the training and classification, the 30 central seconds of each piece were analyzed, instead of the whole file. This was motivated, in part, by the previous observation of an optimal window length of 15 seconds, as mentioned above, and also to comply with the timing requirements of the contest (less than 24 hours for one iteration).

The confusion matrix corresponding to this evaluation is shown in table 1. The total hierarchical accuracy obtained was of 59.22%, and the raw accuracy of 54.12%. When normalized by the number of files in each class (which, in contrast with the previous evaluation, were not equally distributed), these performances improve slightly (see table 3).

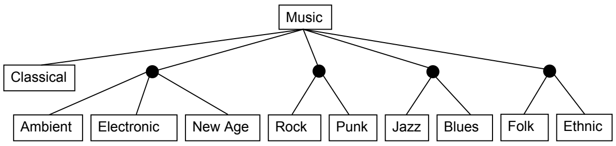

Figure 2: Audio taxonomy for the first MIREX 2005 evaluation (Magnatune dataset).

	Number of classes	Hierarchical accuracy	Normalized hierarchical accuracy	Raw accuracy	Normalized raw accuracy
2003 evaluation	17	-	-	58.71% ± 2.85%	58.71% ± 2.85%
MIREX 2005 (Magnatune)	10	59.22%	61.96%	54.12%	55.68%
MIREX 2005 (USPOP)	6	-	-	66.03%	72.50%

Table 3: Summary of results.

3.2 USPOP dataset

The USPOP dataset consisted of 1414 full-length examples belonging to 6 genres (see fig. 3), 940 (66.4%) for training and 474 (33.6%) for testing. This time, no hierarchical organization was used, resulting in a flat or direct classification approach. Again, the 30 central seconds of 44.1 kHz mono files were analyzed. Table 2 shows the confusion matrix. The raw accuracy obtained was of 66.03%.

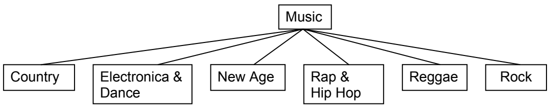

Figure 3: Audio taxonomy for the second MIREX 2005 evaluation (USPOP dataset).

3.3 Computation time

The total computation time in seconds for each of the MIREX iterations is shown in table 4. In average, the algorithm needed 7.35 seconds computation time for each processed training or testing audio file. The algorithm was run on a Dual AMD Opteron 64 1.6 GHz processor with 4 GB RAM.

Dataset	Total runtime (s)	Total runtime (h)	Runtime per file (s)
Magnatune	12483	3.47	8.2
USPOP	9233	2.56	6.5

Table 4: Computation time for the MIREX evaluations.

4 CONCLUSIONS

The MIREX 2005 evaluation with the Magnatune database has resulted in a worse raw classification accuracy than in the previous 2003 evaluation, even if in the latter case less classes have been used (10 instead of 17). This can be due to the high mutual similarity between the classes defined in the Magnatune taxonomy, a fact that can be observed in the confusion matrix (e.g., many new age examples have been misclassified as ambient, and the same is valid for rock classified as electronic). In contrast, the original taxonomy included very dissimilar classes like speech, noise, chamber music, hard rock, etc.

In contrast, the USPOP evaluation was a simpler task, with only 6 classes that do not follow a hierarchy. Accordingly, the best results were obtained in this case.

When the first evaluation was performed (sec. 2), the classifier obtained accuracies that were similar to other systems at that time. However, the results of the MIREX 2005 evaluation have shown that its performance is no longer comparable to that of more recent systems, and that significant improvement is needed.

References

J.-J. Aucouturier and F. Pachet. Representing musical genre: A state of the art. *Journal of New Music Research*, 32 (1), 2003.

J.J. Burred and A. Lerch. A hierarchical approach to automatic musical genre classification. *Proc. Int. Conference on Digital Audio Effects (DAFX), London, UK*, 2003.

J.J. Burred and A. Lerch. Hierarchical automatic audio signal classification. *Journal of the Audio Engineering Society*, 52, No. 7/8, 2004.

G. Tzanetakis and P. Cook. Musical genre classification of audio signals. *IEEE Trans. on Speech and Audio Processing*, 10, No. 5, 2002.