Pachet, F. and Roy, P. “Hit Song Science is Not yet a Science”, Proc. of Ismir ‘08, Philadelphia, pp. 355-360

HIT SONG SCIENCE IS NOT YET A SCIENCE

François Pachet
Sony CSL
pachet@csl.sony.fr

Pierre Roy
Sony CSL
roy@csl.sony.fr

ABSTRACT

We describe a large-scale experiment aiming at validating the hypothesis that the popularity of music titles can be predicted from global acoustic or human features. We use a 32.000 title database with 632 manually-entered labels per title including 3 related to the popularity of the title. Our experiment uses two audio feature sets, as well as the set of all the manually-entered labels but the popularity ones. The experiment shows that some subjective labels may indeed be reasonably well-learned by these techniques, but not popularity. This contradicts recent and sustained claims made in the MIR community and in the media about the existence of “Hit Song Science”.

1. INTRODUCTION

Claims have recently been formulated about the possibility of a “Hit Science” that aims at predicting whether a given cultural item, e.g. a song or a movie, will be a hit, prior to its distribution. Such claims have been made in the domains of music [4] as well as movie [7], and are the basis of hit counseling businesses [9], [17].

More precisely, the claim is that cultural items would have specific, technical features that make them preferred by a majority of people, explaining the non uniform distribution of preferences [6]. These features could be extracted by algorithms to entirely automate the prediction process from a given, arbitrary new item (a song or a movie scenario).

A study showed the inherent unpredictability of cultural markets [19]. The unpredictability was shown to stem from a cumulative advantage or rich-get-richer effect. The study did not conclude, however, that there was no objective substrate to user preferences, but demonstrated the existence of a preference bias introduced when users are exposed to judgments of their pairs.

This study assesses to which extent this claim is scientifically grounded in the domain of music, i.e. can we extract automatically features accounting for song popularity, regardless of the cultural biases evidenced by [19].

In particular, [4] describe an experiment in which a system is trained to learn a mapping between various musical features extracted from the acoustic signal and from the lyrics, and the popularity of the song. They conclude from this experiment that their system learns something about popularity, and so that Hit Song Science is indeed possible.

However, the idea that popularity can be inferred from such technical features contradicts the natural intuitions of any musically-trained composer.

In this paper, we describe a larger-scale and more complete experiment designed to further validate this claim. We use a 32.000 song database of popular music titles, associated with fine-grained human metadata, in the spirit of the Pandora effort [16]. To ensure that the experiment is not biased, we use three sets of different features. We describe the various experiments conducted and conclude that popularity is basically not learned by any of these feature sets.

2. EXTRACTING GLOBAL DESCRIPTORS

The most widely used approach to extract global information from acoustic signals is to identify feature sets supposed to be representative of musical information contained in the signal, and to train classifiers such as SVMs (Support Vector Machines) on manually annotated data (*Train* set). These classifiers are then tested, typically on other data sets (the *Test* set), and their performance is evaluated. If the experiment is performed without biases, a good performance of the classifier means that the feature set considered does carry some information pertaining to the classification problem at hand.

In this paper we describe an experiment similar in sprit to that of [4] on a 32,000 song database. We use three different feature sets to train our classifiers: a *generic* acoustic set *à la* MPEG-7, a *specific* acoustic set using proprietary algorithms, and a set of high-level metadata produced by humans. These feature sets are described in the next sections.

2.1. Generic Audio Features

The first feature set we consider is related to the so-called *bag-of-frame* (BOF) approach. The BOF approach owns

his success to its simplicity and generality, as it can be, and has been, used for virtually all possible global descriptor problems. The BOF approach consists in modelling the audio signal as the statistical distribution of audio features computed on individual, short segments. Technically, the signal is segmented into successive, possibly overlapping frames, from which a feature vector is computed. The features are then aggregated together using various statistical methods, varying from computing the means/variance of the features across all frames to more complex modelling such as Gaussian Mixture Models (GMMs). In a supervised classification context, these aggregated features are used to train a classifier. The BOF approach can be parameterized in many ways: frame length and overlap, choice of features and feature vector dimension, choice of statistical reduction methods (statistical moments or Gaussian Mixture Models), and choice of the classifier (Decision Trees, Support Vector Machines, GMM classifiers, etc.). Many papers in the MIR literature report experiments with variations on BOF parameters on varied audio classification problems [1], [5], [12], [15]. Although perfect results are rarely reported, these works demonstrate that the BOF approach is relevant for extracting a wide range of global music descriptors.

The *generic* feature set we consider here consists of 49 audio features taken mostly from the MPEG-7 audio standard [11]. This set includes spectral characteristics (Spectral Centroid, Kurtosis and Skewness, HFC, Mel Frequency Cepstrum Coefficients), temporal (ZCR, Inter-Quartile-Range), and harmonic (Chroma). These features are intentionally chosen for their generality, i.e. they do not contain specific musical information nor musically *ad hoc* algorithms.

Various experiments [14] were performed to yield the optimal BOF parameters for this feature set: localization and duration of the signal, statistical aggregation operators used to reduce dimensionality, frame size and overlap. The best trade-off between accuracy and computation time is achieved with the following parameters: 2048 sample frames (at 44,100 Hz) with a 50% overlap computed on a 2-minute signal extracted from the middle part of the title, the features are the two first statistical moments of this distribution, i.e. the mean and variance, are considered, yielding a total feature vector of dimension 98 (49 means + 49 variances).

2.2. Specific Audio Features

The *specific* approach consists in training the same (SVM) classifier with a set of "black-box" acoustic features developed especially for popular music analysis tasks by Sony Corporation. These proprietary features have been used in commercial applications such as hard disk based Hi-Fi systems. Altogether, the specific feature set also yields a feature vector of dimension 98, which guaranties a fair comparison with the generic feature set. As opposed to the generic set, the specific set does not use the *BOF* approach: each feature is computed on the whole signal, possibly integrating specific musical information. For instance, one feature describes the proportion of perfect cadences (i.e. resolutions in the main tonality) in the whole title. Another one represents the proportion of percussive sounds to harmonic sounds. We cannot provide here a detailed description of these features as we are mostly interested in comparing the performances of acoustic classifiers on two reasonable, but different feature sets.

2.3. Human Features

Lastly, we trained a classifier with human-generated features. We use the 632 Boolean labels provided by our manually annotated database (see following section) to train the classifiers. This is not directly comparable to the 98 audio features as these labels are Boolean (and not float values). However, as we will see, these features are good candidate for carrying high-level and precise musical information that are typically not well learnt from features extracted from the acoustic signal.

3. THE HIFIND DATABASE

3.1. A Controlled Categorization Process

Several databases of annotated music have been proposed in the MIR community, such as the RWC database [8], the various databases created for the MIREX tests [3]. However, none of them has the scale and number of labels needed to test our hypothesis. For this study we have used a music and metadata database provided by the HiFind Company [10]. This database is a part of an effort to create and maintain a large repository of fine-grained musical metadata to be used in various music distribution systems, such as playlist generation, recommendation, advanced music browsing, etc. The HiFind labels are binary (0/1 valued) for each song. They are grouped in 16 categories, representing a specific dimension of music: *Style*, *Genre*, *Musical setup*, *Main instruments*, *Variant*, *Dynamics*, *Tempo*, *Era/Epoch*, *Metric*, *Country*, *Situation*, *Mood*, *Character*, *Language*, *Rhythm* and *Popularity*. Labels describe a large range of musical information: objective information such as the "presence of acoustic guitar", or the "tempo range" of the song, as well as more subjective characteristics such as "style", "character" or "mood" of the song. The *Popularity* category contains three (Boolean) labels, *low*, *medium* and *high*. It represents the popularity of the title, as observed e.g. from hit charts and records of music history. These three labels are, in principle, mutually exclusive.

The categorization process at work at HiFind is highly controlled. Each title is listened to entirely by one

categorizer. Labels to be set to *true* are selected using an *ad'hoc* categorization software. Label categories are considered in some specific order. Within a category, some rules may apply that prevent some combinations of labels to be selected. The time taken, for a trained categorizer, to categorize a single title is about 6 minutes. The categorized titles are then considered by a categorization supervisor, who checks, among other things, aspects such as consistency and coherence to ensure that the description ontologies are well-understood and utilized consistently across the categorization team. Although errors and inconsistencies can be made during this process, it nevertheless guaranties a relative good "quality" and consistency of the metadata, as opposed for instance to collaborative tagging approaches in which there is no supervision. Additionally the metadata produced is extremely precise (up to 948 labels can be considered per title), a precision which is difficult to achieve with collaborative tagging approaches.

There is no systematic way to ensure that the categorization produces absolutely correct and consistent information, so we had to consider the database as it was provided as ground truth. Some minor "clean up" was performed before use, by discarding titles with metadata of obviously of low quality. For instance, we discarded songs having much less labels set to "true" than the average (37). Additionally, we kept only those labels for which we had a significant amount of titles (above 20) with the *true* and *false* values, to build training and testing sets of sufficient size. As a result of this cleanup, the total number of titles considered in this study is 32978, and the number of labels 632. (Note that those labels correspond to the 632 human features for the experiment described in Section 2.3) Acoustic signals were given in the form of a *wma* file at 128 kbps. This database was used both for training our classifiers and for testing them, as described in Section 4.1.

3.2. Database Redundancy

The HiFind database is *sparse*: the mean number of labels set to true per song (occupation factor) is 5.8% (i.e. 37 on a total of 632). Sparseness suggests the dominant role of the true-valued labels compared to false-valued labels for a given song. It is also *redundant*. For instance, labels 'Country Greece' and 'Language Greek' are well correlated. This redundancy has been analyzed and exploited for performing statistical inference, e.g. to infer unknown attributes from a partial description of a music title, or for suggesting modifications [18].

3.3. Assessing Classifiers

To avoid the problems inherent to the sole use of precision or recall, the traditional approach is to use *F-Measure* to assess the performance of classifiers. For a given label, the *recall R* is the proportion of positive examples (i.e. the titles that are *true* for this label) that were correctly predicted. The *precision P* is the proportion of the predicted positive examples that were correct. When the proportion of positive examples is high compared to that of negative examples, the precision will usually be artificially very high and the recall very low, regardless of the actual quality of the classifier. The F-measure addresses this issue and is defined as:

$$F = 2 \times {R \times P}/{R + P}$$

However, in our case, we have to cope with a particularly unbalanced 2-class (*True* and *False*) database. So the mean value of the F-measure for each class (*True* and *False*) can still be artificially good. To avoid this bias, we assess the performance of our classifiers with the more demanding *min F-measure*, defined as the minimum value of the F-measure for the positive and negative cases. A *min-F-measure* near 1 for a given label really means that the two classes (*True* and *False*) are well predicted.

4. EXPERIMENT

4.1. Experiment Design

We first split the *HiFind* database in two "balanced" parts *Train* and *Test*, so that *Train* contains approximately the same proportion of examples and counter-examples for each labels as *Test*. We obtained this state by performing repeated random splits until a balanced partition was observed. We trained three classifiers, one for each feature set (*generic*, *specific* and *human*). These classifiers all used a Support Vector Machine (SVM) algorithm with a Radial-Basis Function (RBF) kernel, and were trained and tested using *Train* and *Test*. More precisely, each classifier, for a given label, is trained on a maximally "balanced" subset of *Train*, i.e. the largest subset of *Train* with the same number of "True" and "False" titles for this label (popularity *Low*, *Medium* and *High*). In practice the size of these individual train databases varies from 20 to 16320. This train database size somehow represents the "grounding" of the corresponding label. The classifiers are then tested on the whole *Test* base. Note that the *Test* base is usually not balanced with regards to a particular label, which justifies the use of the *min-F-measure* to assess the performance of each classifier.

4.2. Random Oracles

To assess the performance of our classifiers, we compare them to that of *random oracles* defined as follows: given an label with *p* positive examples (and therefore *N-p* negative ones, with *N* the size of the test set), this oracle returns *true* with a probability *p/N*.

By definition, the min-F-measure of a random oracle only depends on the proportion of positive and negative examples in the test database.

For instance, for a label with balanced positive and negative examples, the random oracle defined as above has a *min-F-measure* of 50%. A label with 200 positive examples (and therefore around 16,000 negative examples) leads to a random oracle with a min-F-measure of 2.3%. So the performance of the random oracle is a good indicator of the size of the train set, and can therefore be used for comparing classifiers as we will see below.

4.3. Evaluation of the Performance of Acoustic Classifiers

4.3.1. Comparison with random oracles

The comparison of the performance of acoustic classifiers with random oracles shows that the classifiers do indeed learn something about many of the HiFind labels. More than 450, out of 632, are better learned with the acoustic classifiers than with our random oracle. Table 1 indicates, for each feature set, the distribution of the relative performances of acoustic classifiers with regards to random oracles.

Improvement	Specific	Generic
50	8	0
40	12	15
30	43	20
20	111	79
10	330	360
0	128	158

Table 1. Number of labels for which an acoustic classifier improves over a random classifier by a certain amount. Column "Improvement" reads as follows: there are 111 labels for which a specific acoustic classifier outperforms a random classifier by +20 (in min-F-measure).

Table 1 also shows that around 130 to 150 labels lead to low-performance classifiers, i.e. acoustic classifiers that do not perform significantly better than a random oracle (the last row of the table); approximately half of the labels lead to classifiers that improve over the performance of a random classifier by less than 10; the rest (top rows) clearly outperform a random oracle, i.e. they are well-modeled by acoustic classifiers.

4.3.2. Distribution of performances for acoustic classifiers

At this point, it is interesting to look at the distribution of the performances of these acoustic classifiers. These performances vary from 0% for both feature sets to 74% for the generic features and 76% for the specific ones. The statistical distribution of the performances is close to a power law distribution, as illustrated by the log-log graph of **Figure 1**.

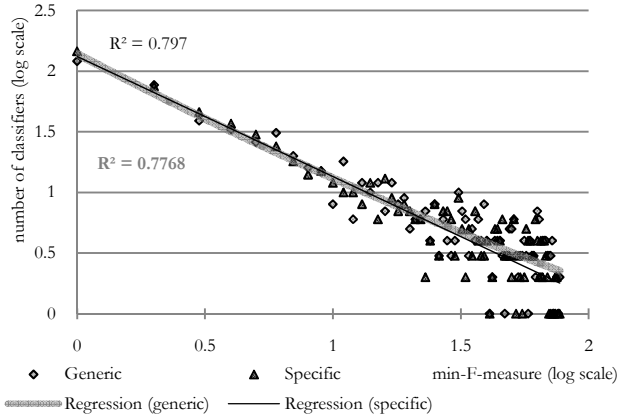

Figure 1. Log-log graph of the distribution of the performance of acoustic classifiers for both feature sets. This distribution of the performance of classifiers is close to a power law.

These power laws suggest that a natural organization process is taking place in the representation of human musical knowledge, and that the process of automatic audio classification maintains this organization.

4.3.3. Specific features slightly outperform generic features

Not surprisingly, we can see that specific features perform always better than the generic ones. This is illustrated by **Figure 2**. Since the classifiers are both based on the same SVM/kernel, the difference can only come from 1) the actual features extracted or 2) the aggregation method. For the generic features, the aggregation is based on means and averages over all the segments of the song. For the specific features, the aggregation is *ad hoc*.

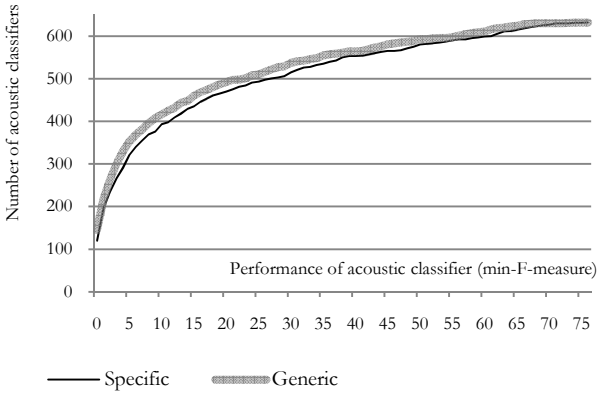

Figure 2. Cumulated distribution of the performance of acoustic classifiers for the generic and specific feature sets. There are more classifiers with low performance for the generic feature sets (leftmost side of the graph).

4.3.4. Acoustic classifiers perform better for large training sets

Lastly, we can observe the relationship between the performance and the size of the training set. The trend lines in Figure 3 show that the performances of acoustic classifiers increase with the training dataset size, regardless of the feature set. This is consistent with the acknowledged fact that machine-learning algorithms require large numbers of training samples, especially for high-dimensional feature sets.

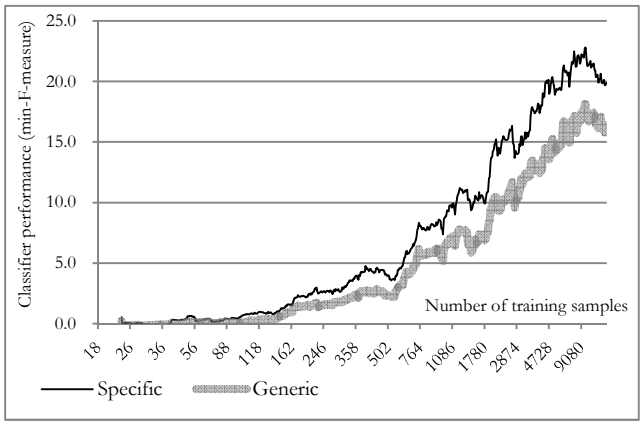

Figure 3. The relative performances of the 632 acoustic classifiers (i.e. the difference between the min-F-measures of the classifier and of the corresponding random oracle) for specific and generic features, as a function of the training database size. The performance of the acoustic classifiers increases with the size of the training database.

These experiments show that acoustic classifiers definitely learn some musical information, with varying degrees of performance. It also shows that the subjective nature of the label do not seem to influence their capacity to be learned by audio features. For instance, the label "Mood nostalgic" is learnt with the performances of 48% (specific features), and 43% (generic features), to be compared to the 6% of the random oracle. Similarly, label "Situation evening mood" is learnt with 62% and 56% respectively, against 36% for random. So popularity is, *a priori*, a possible candidate for this task.

4.4. Inference from Human Data

This double feature experiment is complemented by another experiment in which we train a classifier using all the HiFind labels but the Popularity ones. This is justified by the low entropy of the database as discussed in Section 3.2. Contrarily to the acoustic classifiers, we do not present here the performances of the classifiers for all HiFind labels. Indeed, some pairs of HiFind labels are perfectly well correlated so this scheme works perfectly for those, but this result is not necessarily meaningful (e.g. to infer the country from the language). The same *Train / Test* procedure described above applied with the 629 non-popularity labels as input yields the following result (*min-F-measure*): 41% (*Popularity-Low*), 37% (*Popularity-Medium*) and 3% (*Popularity-High*).

4.5. Summary of Results for Inferring Popularity

The results concerning the Popularity labels are summarized in Table 2.

Popularity label	Generic features	Specific features	Corrected specific	Human features	Dumb features	Random oracle
Low	36	35	31	41	32	27
Medium	36	34	38	37	28	22
High	4	3	3	3	3	3

Table 2. The performances (min-F-measures) of the various classifiers for the three Popularity labels. No significant improvement on the random oracle is observed.

These results show clearly that the Popularity category is not well-modeled by acoustic classifiers: its mean performance is ranked fourth out of 16 categories considered, but with the second lowest maximum value among categories.

Although these performances appear to be not so bad at least for the "Low" label, the comparison with the associated random classifiers shows that popularity is in fact practically not learnt. Incidentally, these performances are not improved with the *correction scheme*, a method that exploits inter-relations between labels to correct the results [14], in the spirit of the contextual approach described in [2].

Interestingly, the use of human features (all HiFind labels) does not show either any significant performance.

Lastly, we also considered *a priori* irrelevant information to train our classifiers: the letters of the song title, i.e. a feature vector of size 26, containing the number of occurrences of each letter in the song title. The performances of the corresponding classifiers are respectively 32% 28% and 3% (for the low, medium and high popularity labels, see Table 2). This shows that even dumb classifiers can slightly improve the performances of random classifiers (by 5% in this case for the medium and low popularity labels), but that this information does not teach us anything about the nature of hits.

These results suggest that there are no significant statistical patterns concerning popularity using these features sets.

5. CONCLUSION

We have shown that the popularity of a song cannot be learnt by using state-of-the-art machine learning

techniques with two sets of reasonable audio features. This result is confirmed when using supposedly higher-level human metadata. This large-scale evaluation, using the best machine-learning techniques available to our knowledge, contradicts the claims of "Hit Song Science", i.e. that the popularity of a music title can be learned effectively from known features of music titles, either acoustic or human. We think that these claims are either based on spurious data or on biased experiments. This experiment is all the more convincing that some other subjective labels can indeed be learnt reasonably well using the features sets described here (e.g. the "mood nostalgic" label).

This experiment does not mean, however, that popularity cannot be learnt from the analysis of a music signal or from other features. It rather suggests that the features used commonly for music analysis are not informative enough to grasp anything related to such subjective aesthetic judgments. Current works are in progress to determine "good features" using feature generation techniques [13], which have been shown to outperform manually designed features for specific analysis tasks. However, more work remains to be done to understand the features of subjectivity for even simpler musical objects such as sounds or monophonic melodies. Hit song science is not yet a science, but a wide open field.

6. ACKNOWLEDGEMENT

This research has been partly supported by the TAGora project funded by the Future and Emerging Technologies program(IST-FET) of the European Commission under the contract IST-34721. The information provided is the sole responsibility of the authors and does not reflect the Commission's opinion. The commission is not responsible for any use that may be made of data appearing in this publication.

7. REFERENCES

[1] Aucouturier, J.-J. and Pachet, F. Improving Timbre Similarity: How high is the sky? J. of Negative Results in Speech and Audio Sciences, 1(1), 2004.

[2] Aucouturier, J.-J., Pachet, F., Roy, P. and Beurivé, A. Signal + Context = Better Classification. *Proc. of Ismir 2007*, Vienna, Austria.

[3] Cano, P. Gómez, E. Gouyon, F. Herrera, P. Koppenberger, M. Ong, B. Serra, X. Streich, S. Wack, N. (2006). *ISMIR 2004 Audio Description Contest*, *MTG Technical Report:* MTG-TR-2006-02.

[4] Dhanaraj, R. and Logan, B. Automatic Prediction of Hit Songs, *Proc. of Ismir 2005*, London, UK.

[5] Essid, S. Richard, G. and David, B. Instrument Recognition in Polyphonic Music Based on Automatic Taxonomies, *IEEE Trans. on Speech, Audio and Lang. Proc*., 14(1), 68-80, 2006.

[6] Frank, R. H. Cook, P. J. The Winner-Take-All Society, Free Press, New York, NY, 1995.

[7] Gladwell, M. The Formula. *The New Yorker*, 2006.

[8] Goto, M. Hashigushi, H., Nishimura, T., Oka, R. "RWC Music Database: Popular, Classical and Jazz Music Databases", *Proc. of Ismir 2002*, Paris, France.

[9] http://www.hitsongscience.com

[10] http://www.HiFind.com

[11] Kim, H. G. Moreau, N. Sikora, T. MPEG-7 Audio and Beyond: Audio Content Indexing and Retrieval, Wiley & Sons, 2005.

[12] Liu, D. Lu, L. Zhang, H.-J. Automatic mood detection and tracking of music audio signals. *IEEE Trans. on Speech Audio and Language Processing*, 14(1), pp 5-18, 2006.

[13] Pachet, F. and Roy, P. Exploring billions of audio features. In Eurasip, editor, *Proc. of CBMI 07*, Bordeaux, France.

[14] Pachet, F. and Roy, P. Improving Multi-Label Analysis of Music Titles: A Large-Scale Validation of the Correction Hypothesis, submitted to *IEEE TALSP,* 2008.

[15] Pampalk, E., Flexer, A., Widmer G. Improvements of Audio-Based Music Similarity and Genre Classification, pp. 628-633, *Proc. of Ismir 2005*, London, UK.

[16] http://www.pandora.com

[17] http://www.platinumblueinc.com/

[18] Rabbat, P. and Pachet, F. Statistical Inference in Large-Scale Databases: How to Make a Song Funk? *Proc. of Ismir 2008*, Philadephia, USA.

[19] Salganik, M. J. Dodds, P. S. Watts, D. J. Experimental Study of Inequality and Unpredictability in an Artificial Cultural Market, *Science*, 311, 854-856, 2006.