Improving Multilabel Analysis of Music Titles: A Large-Scale Validation of the Correction Approach

François Pachet and Pierre Roy

***Abstract*—This paper addresses the problem of automatically extracting perceptive information from acoustic signals, in a supervised classification context. Global labels, i.e., atomic information describing a music title in its entirety, such as its genre, mood, main instruments, or type of vocals, are entered by humans. Classifiers are trained to map audio features to these labels. However, the performances of these classifiers on individual labels are rarely satisfactory. In the case we have to predict several labels simultaneously, we introduce a correction scheme to improve these performances. In this scheme—an instance of the classifier fusion paradigm—an extra layer of classifiers is built to exploit redundancies between labels and correct some of the errors coming from the individual acoustic classifiers. We describe a series of experiments aiming at validating this approach on a large-scale database of music and metadata (about 30 000 titles and 600 labels per title). The experiments show that the approach brings statistically significant improvements.**

***Index Terms*—Feature extraction, learning systems, music, pattern classification.**

I. Introduction

DATABASES of digital content of ever growing size are now increasingly available in our digital world. This situation creates economically and culturally desirable phenomena such as the long tail [1], which enhance the accessibility of hitherto poorly distributed content. Consequently, this mass of data creates a need for accurate descriptions of contents. Metadata can help users search and find content, but only precise, robust metadata can turn the presence of an item in a large database in its actual availability.

In the field of music, many methods have addressed the problem of metadata creation (see, e.g., [2] for a review). Three main categories of approaches can be distinguished [3]. First, purely manual approaches consist in building and maintaining online databases of metadata, associated with content using identification technology, such as audio fingerprinting. This is the case of the AllMusicGuide and Pandora endeavours. Second, so-called *cultural* metadata can be collected from the analysis of user usage, such as buying profiles or web pages, the most widespread approach being collaborative filtering [4]. Lastly, *acoustic* metadata attempt to automate the description process from the analysis of the acoustic signal. This last approach is very popular in the music information retrieval (MIR) research community. However, to our knowledge, this approach has not yet been used commercially, due to unsatisfactory performance of these acoustic analysis systems.

The work presented here is an attempt to improve these automatic approaches in the domain of music, i.e., produce automatically robust and fine-grained music metadata using only the acoustic signal as input, and metadata entered by humans as training and testing data.

More precisely, we are interested in the case when several, possibly many, labels have to be predicted simultaneously. The method we propose consists in exploiting possible correlations between labels, which can be exhibited in the case we dispose of sufficiently large metadata databases. In this paper, we describe a comprehensive study aiming at demonstrating the validity of the correction approach on a large-scale database, with minimum bias. Following a general trend in MIR research, we considered a bag-of-frames (*BOF*) approach based on a set of general audio features inspired by the MPEG-7 audio standard.

The following section briefly reviews the BOF approach and its limitations and introduces the feature set used in this study. Section II introduces the large-scale multilabel database used for the experiment. Section III details the implementation of the correction approach, and Section IV presents the results and a discussion.

A. BOF Approach to Global Description

The performance of acoustic analysis systems applied on the extraction of global musical features has so far consistently shown limitations, sometimes referred to as *glass-ceiling* effects. These limitations seem impossible to overcome by simply tuning the various parameters at hand. These limitations are often observed for works based on the so-called BOF approach, which we describe in the next section.

1) BOF Approach: The BOF approach owns its success to its simplicity and generality, as it can be used for virtually all possible global descriptor problems. The BOF approach consists of modeling 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 modeling such as

Manuscript received March 27, 2008; revised October 08, 2008. Current version published February 11, 2009. This work was supported in part by the TAGora project funded by the Future and Emerging Technologies program (IST-FET) of the European Commission under Contract IST-34721. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Sylvain Marchand.

The authors are with the SONY Computer Science Laboratory, 75005 Paris, France (e-mail: pachet@csl.sony.fr).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TASL.2008.2008734

1558-7916/$25.00 © 2009 IEEE

Gaussian mixture models (GMMs). In a supervised classification context, these aggregated features are used to train a classifier. Once trained, the classifier can be used to classify new data, from which features are extracted and aggregated, or tested against a new data set, to assess its performance.

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.). In the remainder of this paper we refer to these parameters as *BOF parameters*.

2) Limitations of the BOF Approach: Many papers in the MIR literature report experiments with variations of BOF parameters on varied audio classification problems. It can be noted that these results are never fully satisfactory. Reference [5] exhibited performance limitations for timbre similarity or genre classification tasks using a BOF approach with Mel-frequency cepstrum coefficients (MFCCs) as features, GMMs as a statistical reduction method and a *kNN* classifier (exploiting a distance produced by the GMMs). Further works slightly improved these results by adding nonspectral features [6], but their approach still exhibited strong limitations. Reference [7] compares two BOF variations for singer identification in polyphonic audio with specific voice features. They also report unsatisfactory performance. In the domain of instrument recognition, [8] investigated a BOF approach with specific features and tuning parameters, notably a clever feature selection mechanism, but, again, report far-from-perfect results. Reference [9] describes interesting experiments in *mood classification* using a BOF approach and standard audio features, on an 800-title database. They report performances of about 80%, which, although still not satisfactory for industrial applications, is above the performances obtained using timbre similarity. However, we claim in this paper that these results are biased by the small size of the database used for training and testing, and our large-scale evaluation with approximately the same approach does not confirm these results, as described in the following sections.

Several factors involved in the processing chain can explain these performance limitations:

1) *Features*. The acoustic features considered may not contain sufficient information for the given classification problem.
2) *Statistical reduction*. The necessity to reduce the quantity of information extracted in individual frames to feature vectors of reasonable dimension using statistical means may introduce a loss of information since perceptually important, but not statistical significant information may be discarded in the process. This statistical erosion has been hypothesized to cause the presence of hubs when distances are computed in this manner [10].
3) *Classifier*. The algorithm that learns the mapping between features and classes may have limitations, especially in the case of nonconvex training sets.
4) *Lack of information*. A limitation may come from an intrinsic difficulty of global music description, possibly related to some unconscious information processing at stake when listening to music, or simply to the fact that the problem has no perfect solution.

The goal of this paper is not to discuss the limitations of the BOF approach *per se*, but to demonstrate the validity of the correction approach with as little bias as possible. To this aim, we conducted our correction experiment using the most general feature set available: a BOF approach with generic audio features, described in the next section.

B. Details of the BOF Approach

We exploit the acoustic signals as provided by the HiFind database (described in Section II). These signals were given in the form of a *wma* file at 128 kb/s. They were converted to raw wav files sampled at 44 100 Hz and mixed down to mono.

In a preliminary study, we tried several classification algorithms with different parameters, to find the best performer for this problem. We evaluated the performance of J48, a decision tree algorithm, of kNN with $k = 1, 2, 3$ and of support vector machine (SVM) with both linear and radial basis function (RBF) kernels. SVM + RBF outperformed every other candidate algorithm, so we chose it to create the acoustic classifiers of our experiment.

More precisely, we used SVM with a set of general audio features. This set 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, MFCC coefficients), temporal (ZCR, inter-quartile-range), and harmonic (chroma).

We performed several experiments 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. More precisely, we considered the following parameter values:

1) signal duration: 30 s, 1 m, 2 m, whole title;
2) frame size: 512, 1024, 2048, 4096 sample frames;
3) frame overlap: 0%, 25%, 50%, and 75%;
4) statistical aggregation: mean variance, skewness, kurtosis, and combinations thereof.

The best tradeoff between accuracy and computation time is achieved with the following parameters: 2048 sample frames (46 ms) with a 50% overlap computed on a 2-min 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, yielding a total feature vector of dimension 98 (49 means + 49 variances). The 49 features are the following.

- Harmonic Spectral Centroid, Harmonic Spectral Deviation, Harmonic Spectral Spread, Spectral Centroid, Spectral Flatness, Spectral Spread, Spectral Kurtosis, Spectral Skewness, Spectral Rolloff, ZCR, RMS, RHF, HFC, IQR, Centroid, Harmonic Spectral Variation.
- 20 first MFCC coefficients.
- A naive Pitch feature is computed using the Harmonic spectral product [12]. This method is applied to each frame, with no attempt at source or voice separation.
- Chroma: the spectrum is divided in pitch-wide bands. Each band is filtered by a triangular filter. Bands an integral number of octaves apart are superimposed, yielding 12 bands corresponding to the 12 pitch classes. The Chroma feature returns the 12-dimensional vector made up of the average value in each class.

II. HiFind: A Large Database of Fine-Grained Music Description

As mentioned above, the need for describing music stems from organizational purposes. Descriptions should be both machine-readable, and kept as simple as possible. Consequently, the most popular form of musical metadata are so-called *global labels*, i.e., descriptors of the music title as a whole, as opposed to labels that would apply only to a specific segment, a musical passage or event, in the musical piece.

The notion of global label is natural as we all need to classify music using simple terms. It is also a controversial notion as music listening, a process typically during three minutes, calls up different cognitive tasks during time, and may involve different affects depending on the structural evolution of the music piece. Some moments can be exciting, boring, sources of tension or relaxation, etc. Models such as the implication-realization model of [13] precisely emphasize the nonglobal nature of music, and the inherent difficulty in reducing a music title to a single word, even a grounded one. Global music labels are therefore intrinsically problematic, and constitute as such a debatable "ground truth." However, the possibility to describe music globally, using precise ontologies for several musical dimensions is acknowledged to be not only possible but useful, as illustrated by the Pandora and HiFind efforts, as well as by the success of the collaborative music tagging projects such as LastFM.

A. *HiFind Database and the Metadata Creation Process*

Several databases of annotated music have been proposed in the MIR community, such as the RWC database [14] or the various databases created for the *Mirex* tasks [15]. However, none of them has the scale and number of labels needed to test our approach. For this study, we have used a music and metadata database provided by the HiFind Company (a subsidiary of Real Networks). 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. Since its founding in 1999, HiFind has categorized 450 000 tracks. For this work, HiFind supplied us with a subset of 49 620 tracks from 2677 artists, selected at random. The database contains popular music produced between 1920 and 2006 and belonging to 39 different genres.

B. *Controlled Categorization Process*

The label annotation is done by a team of about 25 categorizers, mainly trained musicians and music journalists. A categorizer is able to choose from 935 possible labels to describe the piece of music (hereof 340 are from the genre group). On average, a fully categorized track takes about 8 min to be categorized. The HiFind labels are binary (0/1-valued) indicating the validity of the label for a song.

Labels are grouped in the following 16 categories, representing a specific dimension of music description: *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 less objective characteristics, e.g., "style," "character" or "mood" of the song.

The categorization process at work at HiFind is highly controlled. In a first phase, each title is entirely listened to by one categorizer. Categorizers use an *ad hoc* categorization software to set the value (i.e., *true* or *false*) for each label. In practice, about 37 labels in average are set to *true* for a given title, the remaining ones being set to *false* by default. The 16 label *categories* are considered in some specific order to ease and speed up the categorization process. Within a category, some rules may apply that prevent some combinations of labels to be selected. For instance, for the genre category, the categorizer can select as many genre labels as appropriate per song, with at least one. On the average each song has two genre labels (for example "Blues-rock" and "Contemporary Blues"). Concerning the *Main instruments* category, the categorizer selects instruments which dominate the recording. A typical selection for a Metal song would be "Vocals" + "Male" + "Guitar (distorted)." Instruments which play the major role in a solo of a song although not constantly featured in the recording (e.g., a Saxophone) are also selected as *true*.

The Popularity category is particular and consists of three labels: *high*, *medium* and *low*. The selection of a label from this group is done by a categorizer based on 1) chart success, 2) artist popularity, 3) genre specific popularity. A "popularity_high" is given to a song which was in the global/local charts or has a key role within its specific genre (for instance Jazz tracks might not be chart hits but still be popular among its subset of listeners). A "popularity_medium" is given to a song which is a non-hit from a highly popular artist, a low chart entry song or a popular song among a specific genre. For instance, a track from the genre "Detroit Techno" is neither widely popular nor is the artist very known. Still, there are levels of popularity for this subgroup of listeners. A "popularity_low" is given to a track from an unknown artist with no chart history.

In a second phase, the categorized titles are checked 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, e.g., to collaborative tagging approaches in which there is no supervision by definition. Additionally the metadata produced is extremely precise (a maximum of 948 labels per title), a precision which, again, 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, however, 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 32 978, and the number of labels

632. This database was used both for training our classifiers and for testing them, as described in Section III-D.

C. Database Sparseness and Redundancy

As observed in [16], this database is quite *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.

Another feature of the database is its *redundancy*. For instance, labels like "Country Greece" and "Language Greek" are well correlated. There are correlations between pairs of labels, as described in [17], typically between *language* and *country* labels. There are more complex correlations between arbitrary numbers of labels. This redundancy is a sign of the presence of many inter-label dependencies that justifies the deployment of a statistical approach to label inference, either for correcting acoustic classifiers as we will describe here, or as a stand-alone process as described in [16]. The correction approach attempts to exploit them all.

III. Correction Approach

In this section, we describe the experiment conducted in training and testing acoustic classifiers on the HiFind database presented above. In the first section, we introduce the correction approach. We then describe the experiment to validate the approach.

A. Classifier Fusion

The idea of combining the decisions of different classifiers to improve pattern recognition systems emerged in the early 1990s, in particular, in the domain of handwritten character recognition [18]. The idea has since received different names, e.g., *classifier* (or *data*) *fusion*, *mixture of experts*, *classifier ensembles* [19].

Several strategies have been devised to combine the decisions of a set of classifiers, like fuzzy aggregation methods [20] or probabilistic methods, such as the Dempster–Shafer fusion [21]. Wolpert [22] proposes to train a higher-lever classifier on the outputs of a set of original classifiers trained on multiple partitioning of the learning set. The higher-lever classifier, called *stacked classifier*, estimates and corrects the generalization biases of the original classifiers.

Our approach is close to that of Wolpert as we use the outputs of a set of original (acoustic) classifiers to train a higher-level (correction) classifier. The fundamental difference is that all the original classifiers are trained on the same data (i.e., same acoustic features extracted from the same music titles) but are predictors for different classes. The *stacked classifiers* of Wolpert's approach estimate and correct the biases of the original classifiers, whereas *correction classifiers* learn the relationships between the different classes, thus exploiting redundancy in the database's metadata structure.

B. Correction Approach

Our correction approach aims at exploiting statistical correlations between the labels we attempt to model, in the training corpus. These correlations can be used to correct the individual classifiers obtained by using only acoustic information. In the rest of this paper, those classifiers are referred to as *acoustic classifiers*.

This approach is based on a hypothesis. First, the information produced by the acoustic classifiers as well as the information about the correlations is fuzzy and approximate by nature. In practice, however, it is hoped that these approximations can somehow be compensated by an appropriate feature extraction scheme. Second, the classifiers we build (acoustic and correction) exploit the information extracted from the same set of acoustic features, albeit in different ways. From a signal viewpoint, there is strictly speaking no "additional information" introduced at the correction stage. However, the correction targets inter-labels dependencies that are typically not exploited explicitly in traditional approaches, in which labels are predicted individually. This hypothesis justifies the need for a large-scale validation.

A previous study, described in [17], proposed to use inter-label correlation to improve multilabel music classification, considered as "contextual information." The cited paper reported an average of 15% of improvement in precision over the purely acoustic classifiers. In this paper, successive generations of classifiers are created. The first generation consists of acoustic classifiers trained on timbre features. The second generation consists of classifiers trained on the output of the acoustic classifiers having a precision greater that a predefined threshold. In general, the $(N+1)$th generation classifiers are trained on the output of the best classifiers found before the Nth generation.

However, this approach revealed flawed for several reasons, which motivated in part the present study. The most important problem was that the approach, iterative by nature, was biased by a cumulated "contamination" of train and test samples, eventually producing overfitting. The performance of the "first-generation" classifiers was taken into account to compute the next-generation classifiers. This performance was computed using the *Test* databases which were then reused to evaluate the performance of the next-generation classifiers, thereby creating a contamination. The second source of contamination stemmed from the way *Train* and *Test* databases were constructed for each label. No separation was enforced across labels, i.e., some training samples for a label could be test samples for another one. This created another source of contamination in the case of a classifier X based on classifier Y, such as some test samples for X were included in the *Train* set of Y. Lastly, the scheme did prevent direct reuse of a classifier for the next generation, but not its indirect use after more than two generations.

The encouraging results of this study were used as the basis of an argumentation to sustain the hypothesis that the limitations observed in timbre similarity studies would be caused primarily by some high-level processing involving the brain, supposedly requiring other information than the signal [23]. However interesting, these arguments were based on a heavily biased experiment so no serious conclusion could be drawn at this stage concerning the correction approach.

A subsequent study addressed and solved the contamination issue, by proposing a clean correction scheme with no use of test information whatsoever in the train phase [24]. This preliminary study reported interesting and comparable improvements in classifier performance, but on a more solid experimental setting. However, this study was also based on the sole use of timbre similarity as an acoustic seed, and could only be applied to a subset of the same, small, 5000-title reference database, for reason of time and memory limitations (the timbre similarity matrix cannot be computed for large datasets). These results highlighted the potential interest of the correction approach, but were not fully convincing to the present authors, as the observed improvements in classification performance could not clearly be attributed to the correction effect *per se*, so no real insight was gained.

C. Min-F-Measure as Evaluation Criteria

To avoid the problems inherent to the sole use of precision or recall, the traditional approach to evaluation classifiers is to use *F-Measure*. 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.

D. Description of the Experiment

In this section, we describe the correction experiment. Due to the nature of the correction process (see below), we first had to ensure that *Train* and *Test* are structurally equivalent. So we split the database in two "balanced" parts *Train* and *Test*: for each label, *Train* contains approximately the same proportion of examples and counter-examples as *Test*. We obtained this state by performing repeated random splits until a balanced partition was observed. Then, we built the acoustic and correction classifiers as follows.

1) Acoustic Classifiers: First, we build a set of 632 *acoustic* classifiers, one for each label. As described in Section I-B, we chose a SVM + RBF kernel classifier.

The acoustic classifiers are trained and tested using the *Train* and *Test* databases described above. 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. The most frequent items are selected randomly.

In practice, the size of these individual train databases varies from 20 to 16 320, with an average of 1700. This train database size somehow represents the "grounding" of the labels. 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.

2) Correction Classifiers: In a second step, we build a new set of 632 classifiers of second generation, referred to as *correction* classifiers. The correction classifiers also use a SVM algorithm but with a *linear kernel*, which turned out to perform slightly better than the RBF in this configuration (632 Boolean inputs, instead of the 98 floats for the acoustic classifiers).

Each correction classifier takes as input the vector of Boolean predictions from the 632 acoustic classifiers.

Correction classifiers are then, again, trained on *Train* and tested on *Test*. In this case, we consider the whole *Train* database for training, as opposed to maximally balanced sets used in the previous step. This is justified by the fact that, at this step, we try to learn the inter-label dependencies, so a set which is balanced for a given label may not be representative of these dependencies.

It is important to note that in this scheme there is no contamination between the two phases, as opposed to the approach described in [17]: the performance of acoustic classifiers on *Test* is not exploited in the correction phase. This performance is indicated here only for comparison purposes. The results are described in the next section.

IV. Results

The main results of this experiment are the following.

Acoustic classifiers perform much *better than random*.

The *correction approach is efficient*: we observe significant improvements in the second-generation classifiers. This improvement is uniform and does not depend on the grounding of labels.

There is no clear dependency between the nature of the labels, as represented by their category (see Section IV-B) and their "acoustic classifiability." Categories exhibit a high variance with regards to acoustic classifiability. Strikingly, *a priori* "subjective" labels are not necessarily more difficult to learn than *a priori* "objective" ones.

These points are detailed and discussed in the next sections.

A. Compared Performance of Acoustic Classifiers

In this section, we report on the performances of the acoustic classifiers

1) Acoustic Classifiers Outperform Random Oracles: A random oracle is a classifier that yields a random but systematic answer, solely based on the distribution of examples in the training set. A naive random oracle that would always draw the most represented class could have a nonzero (mean) F-measure, but its min -F-measure would be 0, by definition.

For our comparison, we defined a less naive random oracle as follows: given a label with p positive examples (and therefore $N-p$ negative ones, with N the size of the dataset), this oracle returns *true* with probability p/N. Note that a simple random oracle that would return *true* and *false* with probability $1/2$ for every label would have a lower performance.

Note also that the min-F-measure of this random oracle only depends on the proportion of positive and negative examples in the test dataset. Roughly speaking, when using our random oracle, a label with balanced positive and negative instances has a min-F-measure of approximately 50%, whereas a label with 200 positive examples (and therefore around 16 000 negative examples) has a min-F-measure of 2.3%.

The comparison of the performance of acoustic classifiers with random oracles shows that the classifiers do indeed learn something. For more than 450 out of 632 labels, the acoustic classifier outperforms the corresponding random oracle. Table I indicates the distribution of the relative performance of acoustic

TABLE I

NUMBER OF LABELS FOR WHICH AN ACOUSTIC CLASSIFIER IMPROVES OVER A RANDOM CLASSIFIER BY A CERTAIN AMOUNT. COLUMN "IMPROVEMENT" INDICATES PERFORMANCE IMPROVEMENT. THERE ARE 79 LABELS FOR WHICH AN ACOUSTIC CLASSIFIER OUTPERFORMS A RANDOM CLASSIFIER BY $+20$ (IN MIN-F-MEASURE)

Improvement	# of Labels
50	0
40	15
30	20
20	79
10	360
0	158

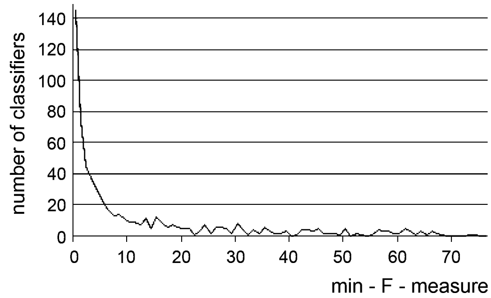


Fig. 1. Distribution of the performance of acoustic classifiers. There are many labels for which the corresponding acoustic classifier performs poorly. As performance increases, the number of corresponding labels decreases.

classifiers compared to random oracles.Table I shows that 158 labels lead to low-performance classifiers, i.e., they do not perform much better than a random oracle; half of the labels lead to classifiers that outperform a random oracle by less than 10; the remaining ones clearly outperform a random oracle, i.e., they are well-modeled by acoustic classifiers.

2) Distribution of Acoustic Classifiers Performance: It is interesting to look at the distribution of the performances (*min-F-measure*) of the acoustic classifiers. These performances vary from 0% to 74%. Fig. 1 shows that the statistical distribution of the performances is close to a power law distribution, which is confirmed by the log-log graph of Fig. 2.

More precisely, if y is the number of acoustic classifiers with performance x expressed as a *min-F-measure*, the distribution of the performance of acoustic classifiers is best approximated by $y = 263 \cdot x^{-1.34}$.

3) Acoustic Classifiers Trained on Large Datasets Perform Better: Lastly, we can observe the relationship between the performance and the size of the training set. The trend line in Fig. 3 shows that the performance of acoustic classifiers increases with the size of the training dataset. This is consistent with the acknowledged fact that machine-learning algorithms require large numbers of training samples, especially for high-dimensional feature sets.

These experiments show that acoustic classifiers definitely learn some musical information, with varying degrees of performance. The next section is devoted to the problem of improving these performances using the correction approach.

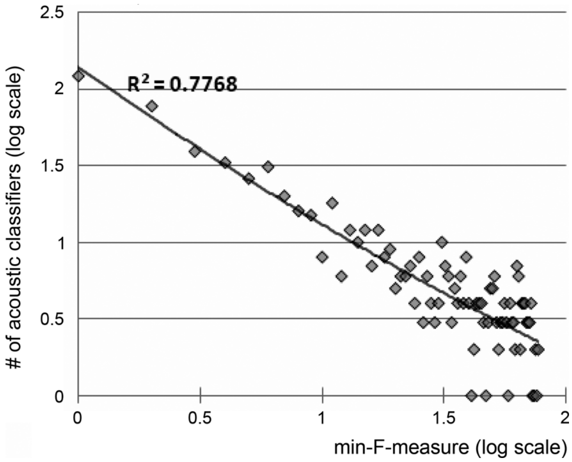


Fig. 2. Log-log graph of the distribution of the performance of acoustic classifiers. The distribution of the performance of acoustic classifiers is close to a power law (with more data fluctuation as we reach high performance, which can be due to the small number of labels considered, i.e., labels well-modeled by an acoustic classifier).

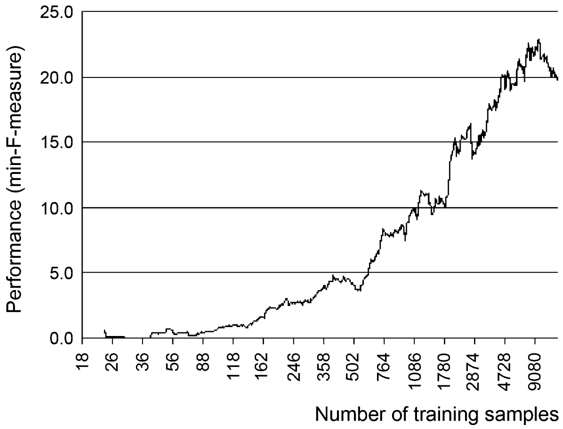


Fig. 3. Relative performances of the 632 acoustic classifiers (i.e., the difference between the min-F-measures of the acoustic classifier and of the corresponding random classifier), as a function of the training database size (moving average over 30 labels). This graph shows that the performance of the acoustic classifiers increases with the size of the training database.

B. Validity of the Correction Approach

In this section, we compare the performance of the acoustic classifiers with their corrected counterparts, as described in Section III-D.

The experiment shows that the correction approach is valid. Fig. 4 shows a comparison of the acoustic versus correction classifiers. One can observe that the correction curve is generally above the acoustic one, except for the few labels listed as follows:

Popularity low (-14%);
Tempo/Moderato (-11%);
Situation/Night (-5%);
Character/smokey (-4%);
Mood/positive (-4%).

This result is highlighted by the "mean" value which is clearly separated and in favor of the corrected classifiers.

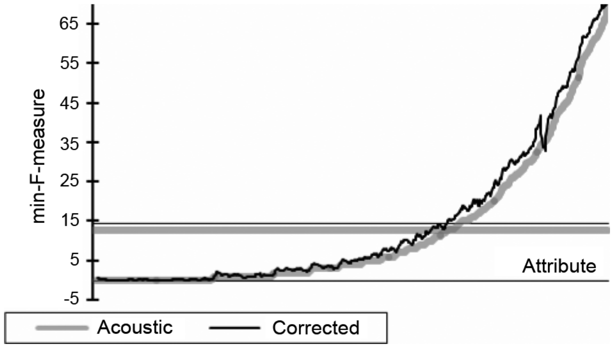


Fig. 4. Global improvement in classifier performance (min-F-measure). The x-axis represents label indices, sorted by increasing performance of the corresponding acoustic classifier. The y-axis represents the min-F-measure of the classifiers. Consequently the performance of the acoustic classifiers is the "smooth" curve, and the other one shows the performance of the corrected classifiers. The two horizontal lines in each graph represent the average performance of the acoustic (thick gray line) and corrected classifier (thin black line).

TABLE II
RANKING OF THE MEAN OF THE MIN-F-MEASURES FOR ALL LABELS IN THE FOUR EXPERIMENTS

Experiment	Mean of the min-F-measures for all labels
Corrected	14.29
Acoustic	12.58

As a conclusion on the validity of the correction approach, we rank the classifier performances in Table II. The average improvement is 1.71. The relative performance improvement is 16.4%.

The statistical significance of this result is assessed using the χ^2 test. Note that this test cannot be computed on the F-measure since it represents a frequency. We therefore compute χ^2 on data from the whole confusion matrices. More precisely, we compute χ^2 on the following.

1) The well-classified titles (True positive + True negative titles for the acoustic and corrected classifiers).
2) The misclassified titles (False positive + False negative).

We obtain the following results: $\chi^2_{631} = 99\,402$, with $p < 0.001$ for the well-classifier examples; $\chi^2_{631} = 61\,631$, with $p < 0.001$ for the misclassified examples.

Fig. 5 shows that the average $\mathrm{TP} + \mathrm{TN}$ is higher for the corrected classifiers, and that the average $\mathrm{FP}+\mathrm{FN}$ is smaller for the corrected classifiers. This shows that correction significantly improves the performance of the acoustic classifiers.

C. Categories are not Grounded

In this section, we analyze the relation between the performance of the classifiers and their "semantics," as given by their categories (there are 16 categories, see Section II-A).

1) General Case: First, we observe that some categories are, on average, better learned than others by the acoustic classifiers. Fig. 6 shows the minimum, maximum, and mean performance for each category. We can see, for instance, that the "Style" category is in average poorly modeled, as compared to, e.g., The "Dynamics" category.

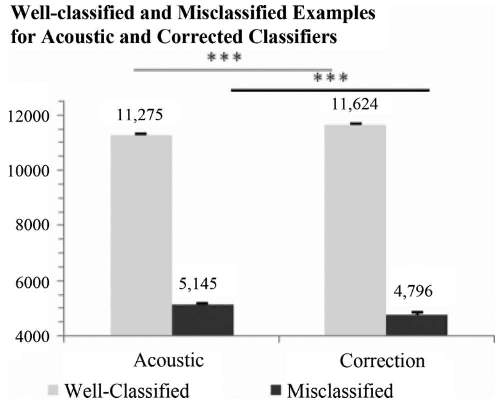


Fig. 5. Statistical significance of the results shown in Table II. The light gray bars represent the average number of well-classified titles $(\mathrm{TP} + \mathrm{TN})$ for the acoustic (left) and corrected classifiers (right). The dark bars represent the average number of misclassified titles.

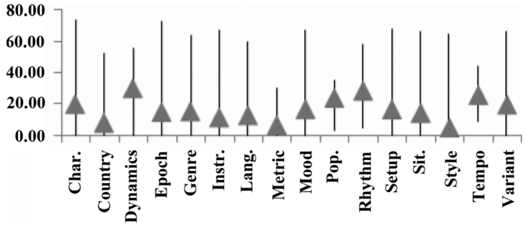


Fig. 6. Performance of acoustic classifiers on each of the 16 categories. The graph shows, for each category, the lowest performance (bottom of the vertical line), the best performance (top of the line) and the mean performance across labels of this category (triangles). It is clear that categories exhibit a high variance with regards to "acoustic classifiability."

However, each category contains both very good and very bad labels, as shown in Fig. 6. For instance, the Style category contains the "Urban" and "Rock" labels which are in the top list of the acoustic classifiers (68% and 66% in min-F-measure, respectively), and also "Folk Rock" which yields a min-F-measure of 2.8% (with a random oracle yielding 2.2%), so extremely bad. This high variance of category is illustrated in Fig. 6. In other words, no category is intrinsically easy or difficult to model using acoustic features.

Things are somehow different with the correction. In Table III and Fig. 7, we see that the improvement of correction classifiers does depend on the category. For categories *Style*, *Country*, *Genre*, *Language*, *Musical Setup*, and *Epoch*, correction classifiers perform significantly better than acoustic classifiers. Conversely, categories *Main Instruments*, *Mood*, *Variant*, *Popularity*, *Tempo*, *Metric*, *Situation*, *Character, Rhythm*, and *Dynamics* do not benefit from the correction approach (the performance even degrades for *Popularity* and *Tempo*).

However, the performance of corrected classifiers (not shown here) follows the same pattern than the performance of acoustic classifiers (see Fig. 6). This confirms that categories are not related to the capacity of being learned. In other words, labels are somehow grounded, but not categories.

It is important to stress the counter-intuitive nature of these results. Seemingly « high-level" labels, e.g., "Situation Landscape Panning Shot" (min-F-measure 69) or "Mood Sentimental" (17), can be better learned than seemingly low-level

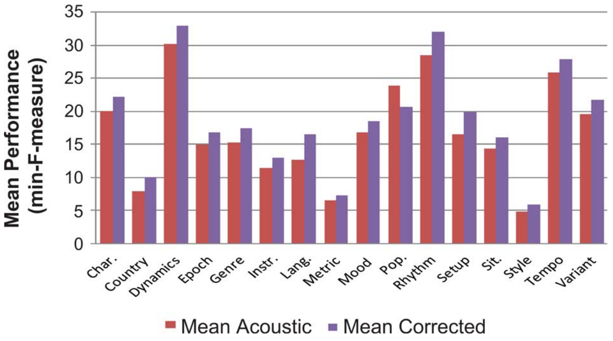


Fig. 7. Graphical representation of the performance of acoustic and corrected classifiers for each label category.

TABLE III
PERFORMANCE OF ACOUSTIC AND CORRECTED CLASSIFIERS FOR EACH CATEGORY

Dimension	MeanAcoustic	Mean Correction
Char.	20.14	22.32
Country	8.04	10.11
Dynamics	30.25	33.00
Epoch	15.00	16.88
Genre	15.35	17.48
Instr.	11.50	13.03
Lang.	12.81	16.63
Metric	6.60	7.40
Mood	16.81	18.61
Period	0.00	0.50
Pop.	24.00	20.67
Rhythm	28.60	32.10
Setup	16.64	19.88
Sit.	14.45	16.16
Style	4.94	5.95
Tempo	25.88	28.00
Variant	19.67	21.77

ones, e.g., "Character Metallic" (18) or "Instrument Accordion" (13). This can be partly explained by the same reasons that make the correction approach work: some high-level labels are correlated with lower-level ones which are easier to model. The corresponding acoustic classifier may then reach indirectly a good performance.

2) Case of Popularity: The case of the popularity category is interesting to point out. The corresponding result contradicts recent claims about the possibility of a "Hit Song Science" that aims at predicting whether a given song will be a hit, prior to its distribution. More precisely, Hit Song Science claims that cultural items have specific, technical features that make them preferred by a majority of people, explaining the nonuniform distribution of preferences [25]. These features could be extracted by algorithms to entirely automate the prediction process from a given, arbitrary new song. The idea that popularity can be inferred from technical features contradicts the natural intuitions of any musically trained listener. A study showed the inherent unpredictability of cultural markets [26]. 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 merely exposed to judgments of their pairs. The "popularity feature" claims have been made in the domains of music [27] as well as movie [28], leading to the appearance of hit counseling businesses [29], [30]. In particular, Dhanaraj and Logan describe an experiment [27] 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 indeed something about popularity, and so that Hit Song Science is possible. However, the experiment described by Dhanaraj and Logan was performed on a relatively small database (1700 songs), with rudimentary features, mostly based on timbre.

As a side-effect of our core experiment, we can revisit this claim on a larger database. The *Popularity* category in our HiFind database contains three labels, *low*, *medium*, and *high*. It represents the popularity of the title, as observed, e.g., from hit charts and records of music history. It is interesting to observe that this *Popularity* category is not well modeled by acoustic classifiers: its mean performance is ranked fourth, but with the second lowest maximum value among categories. Moreover the acoustic performances are respectively of 36%, 36%, and 4% for the three classes. This is to be compared to the performances of the associated random classifiers for these classes, which are 27%, 22%, and 3%, respectively. This means that popularity is practically not learned by acoustic classifiers. Furthermore, these performances are not improved with correction, as shown on Fig. 7 and Table III. This further suggests that there are no significant statistical patterns concerning popularity.

This result, obtained on a large-scale evaluation with the best techniques available to our knowledge, contradicts recent claims of so-called "Hit Song Science" that the popularity of a music title can be learned effectively from acoustic features. We suggest that these claims are either based on spurious data or on biased experiments.

V. IMPLEMENTATION

All the data used in this experiment is available online at http://www.csl.sony.fr/~pachet/correction/.

The implementation of the feature extraction was done in C (Windows dll). The evaluation was done using a Java-based framework, using the Weka library [31]. Feature extraction tasks were parallelized and distributed on two machines (a 1.7-GHz dual-core Xeon and a 2 dual-core Xeon at 2.66 GHz). The training and testing phases (series of 632 SVMs) were distributed on eight processors. The overall computation took two weeks.

VI. CONCLUSION

We introduced the correction approach, an attempt to use statistical redundancies in the training database to correct errors of individual classifiers by learning how to correct errors using statistical redundancy in the training database. The experiments described here confirm the validity of the approach in general. For some labels, the improvements in classification are substantial. For other labels the performances obtained are still not sufficient to automate the metadata extraction process. However, the correction approach is worth considering in the case of multilabel classification, as it is compatible with other known mechanisms for improving individual classification tasks (e.g., feature selection, boosting, bagging, parameter tuning).

The correction approach can be applied to other types of metadata databases, in particular social tagging systems which also provide multilabel descriptions of items. It can also be applied to other fields than music, in which signal-based analysis shows limitations, such as picture or video collections.

Acknowledgment

The authors would like to thank N. Friedenberger, H. Ross, and B. Steffens from HiFind for providing a lot of support during this project. A. Beurivé implemented the features library and participated in the implementation of the evaluation framework. J.-J. Aucouturier designed and implemented a preliminary version of the correction approach. R. Campana designed the version of the correction approach used in this paper and performed an initial validation using a timbre distance. The second author designed and implemented the framework for the computation and evaluation of the classifiers.

References

[1] C. Anderson*, The Long Tail: Why the Future of Business is Selling Less of More Rev*, Updated ed. New York: Hyperion, 2008.

[2] G. Widmer, S. Dixon, P. Knees, E. Pampalk, and T. Pohle, "From sound to sense via feature extraction and machine learning: Deriving high-level descriptors for characterising music," in *Sound to Sense, Sense to Sound—A State of the Art in Sound and Music Computing*, P. Polotti and D. Rocchesso, Eds. Berlin, Germany: Logos Verlag, 2007.

[3] F. Pachet, "Knowledge management and musical metadata," Encyclopedia of Knowledge Management Idea Group, 2005.

[4] U. Shardanand and P. Maes, "Social information filtering: algorithms for automating word of mouth," in *Proc. CHI*, 1995, pp. 210–217.

[5] J.-J. Aucouturier and F. Pachet, "Improving timbre similarity: How high is the sky?," *J. Negative Results Speech Audio Sci.*, vol. 1, no. 1, 2004.

[6] E. Pampalk, A. Flexer, and G. Widmer, "Improvements of audio-based music similarity and genre classification," in *Proc. ISMIR*, London, U.K., 2005, pp. 628–633.

[7] E. K. Youngmoo and B. Whitman, "Singer identification in popular music recordings using voice coding features," in *Proc. ISMIR*, Paris, France, 2002, pp. 329–336.

[8] E. G. Richard and B. David, "Instrument recognition in polyphonic music based on automatic taxonomies," *IEEE Trans. Speech, Audio, Lang. Process.*, vol. 14, no. 1, pp. 68–80, Jan. 2006.

[9] D. Liu, L. Lu, and H.-J. Zhang, "Automatic mood detection and tracking of music audio signals," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 14, no. 1, pp. 5–18, Jan. 2006.

[10] J.-J. Aucouturier and F. Pachet, "A scale-free distribution of false positives for a large class of audio similarity measures," *Pattern Recognition*, vol. 41, no. 1, pp. 272–284, 2007.

[11] H. G. Kim, N. Moreau, and T. Sikora*, MPEG7 Audio and Beyond: Audio Content Indexing and Retrieval*. New York: Wiley, 2005.

[12] M. Alonso, B. David, and G. Richard, "A study of tempo tracking algorithms from polyphonic music signals," in *Proc. 4th COST 276 Workshop*, France, Mar. 2003.

[13] E. Narmour*, The Analysis and Cognition of Basic Melodic Structures*. Chicago, IL: Chicago Univ. Press, 1990.

[14] M. Goto, H. Hashigushi, T. Nishimura, and R. Oka, "RWC music database: Popular, classical and jazz music databases," in *Proc. ISMIR*, Paris, France, 2002, pp. 287–288.

[15] P. Cano, E. Gómez, F. Gouyon, P. Herrera, M. Koppenberger, B. Ong, X. Serra, S. Streich, and N. Wack, "ISMIR 2004 audio description contest," MTG Tech. Rep. MTG-TR-2006-02, 2006.

[16] P. Rabbat and F. Pachet, "Statistical inference in large-scale databases: How to make a title funk?," in *Proc. ISMIR*, Philadelphia, PA, 2008, pp. 589–594.

[17] J.-J. Aucouturier, F. Pachet, P. Roy, and A. Beurivé, "Signal + context = better classification," in *Proc. ISMIR*, Vienna, Austria, 2007, pp. 425–430.

[18] Y. S. Huang and C. Y. Suen, "A method of combining multiple experts for the recognition of unconstrained handwritten numerals," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 17, no. 1, pp. 90–94, Jan. 1995.

[19] T. K. Ho, "Multiple classifier combination: Lessons and next steps," in *Kandel, Bunke, Hybrid Methods in Pattern Recognition*. Singapore: World Scientific, 2002, pp. 171–198.

[20] S.-B. Cho and J. H. Kim, "Combining multiple neural networks by fuzzy integral and robust classification," *IEEE Trans. Syst,, Man, Cybern.*, vol. 25, no. 2, pp. 380–384, Feb. 1995.

[21] Y. Lu, "Knowledge integration in a multiple classifier system," *Appl. Intell.*, vol. 6, pp. 75–86, 1996.

[22] D. H. Wolpert, "Stacked generalization," *Neural Netw.*, vol. 5, no. 2, pp. 241–260, 1992.

[23] J.-J. Aucouturier, "Ten experiments on the modeling of polyphonic timbre," Ph.D. dissertation, Univ. Pierre et Marie Curie, Paris, France, 2006.

[24] R. Campana, "Organisation et catégorisation de la musique populaire par apprentissage statistique," Sony CSL, Paris, France, 2006, Tech. Rep. CSLP-TR-06-01.

[25] R. H. Frank and P. J. Cook*, The Winner-Take-All Society*. New York: Free Press, 1995.

[26] M. J. Salganik, P. S. Dodds, and D. J. Watts, "Experimental study of inequality and unpredictability in an artificial cultural market," *Science*, vol. 311, pp. 854–856, Feb. 2006.

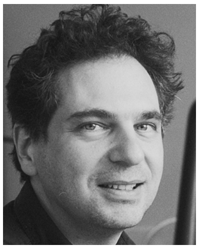
[27] R. Dhanaraj and B. Logan, "Automatic prediction of hit songs," in *Proc. of ISMIR 2005*, London, UK, 2005.

[28] M. G. Gladwell, "The formula, The New Yorker," 2006 [Online]. Available: http://www.gladwell.com/2006/2006_10_16_a_formula.html.

[29] "PlatiniumBlue." [Online]. Available: http://www.platinum-blueinc.com.

[30] "HitSongScience." [Online]. Available: http://www.hitsong-science.com.

[31] I. H. Witten and F. Eibe*, Data Mining: Practical Machine Learning Tools and Techniques*, 2nd ed. San Francisco, CA: Morgan Kaufmann, 2005.



Francois Pachet received the Ph.D. degree and Habilitation à diriger des Recherches from University of Paris 6, Paris, France.

He is a Civil Engineer (Ecole des Ponts and Chaussées) and was an Assistant Professor in Artificial Intelligence and Computer Science, Paris 6 University, until 1997. He then set up the music research team at SONY Computer Science Laboratory, Paris, and developed the vision that metadata can greatly enhance the musical experience in all its dimensions, from listening to performance. His team conducts research in interactive music listening and performance and musical metadata and developed several innovative technologies and award winning systems (MusicSpace, constraint-based spatialization, PathBuilder, intelligent music scheduling using metadata, The Continuator for Interactive Music Improvization). He is the author of over 80 scientific publications in the fields of musical metadata and interactive instruments.

Pierre Roy was born in France in 1970. He studied mathematics and computer science and received the Ph.D. degree from the University Paris 6, Paris, France, in 1998.

He then spent a year in Santa Barbara, CA, as a Postdoctoral Fellow in the Center for Research in Electronic Arts and Technology. He came back to France and worked a few years for a multimedia company in Paris. In 2003, he worked as a consultant for a business loyalty start-up company in Beirut, Lebanon. He joined Sony CSL, Paris, in 2005 as a Research Assistant in the music team where he specializes in intelligent digital signal processing techniques with applications in domains such as music delivery, music interaction, environmental sound analysis, and animal behavior.