EXPLORING NEW FEATURES FOR MUSIC CLASSIFICATION

Rémi Foucard[1], *Slim Essid*[1], *Gaël Richard*[1] *and Mathieu Lagrange*[2]

[1] Institut Mines-Télécom - TELECOM ParisTech, CNRS-LTCI
37, rue Dareau
75014 Paris, France

[2] Ircam - STMS, CNRS, UPMC
1, place Igor Stravinsky
75004 Paris, France

remi.foucard@telecom-paristech.fr

ABSTRACT

Automatic music classification aims at grouping unknown songs in predefined categories such as music genre or induced emotion. To obtain perceptually relevant results, it is needed to design appropriate features that carry important information for semantic inference. In this paper, we explore novel features and evaluate them in a task of music automatic tagging. The proposed features span various aspects of the music: timbre, textual metadata, visual descriptors of cover art, and features characterizing the lyrics of sung music. The merit of these novel features is then evaluated using a classification system based on a boosting algorithm on binary decision trees. Their effectiveness for the task at hand is discussed with reference to the very common Mel Frequency Cepstral Coefficients features. We show that some of these features alone bring useful information, and that the classification system takes great advantage of a description covering such diverse aspects of songs.

Index Terms— Music information retrieval, Features, Autotagging, Boosting, Missing features

1. INTRODUCTION

Automatic music classification is the task of predicting categories for previously unknown pieces of music. These categories can be as diverse as musical genre, vocalist gender or emotional content. They can be organized into labels (*i.e. Emotion-Happy* or *Genre-Rock*), often called *tags*. In this configuration, the two considered categories are: presence *vs.* absence of any particular tag.

Many works have been published in the recent years to deal with the problem of automatic tagging (*autotagging*) [1]. They usually include machine learning techniques to learn decision rules from an annotated corpus. The most common learning algorithms are: Support-Vector Machines (SVM) [2], Gaussian Mixture Models (GMM) [3], and Boosting [4].

The representation of the song by features to be used by the learning algorithm is also a critical aspect of the system. As far as music is concerned, these features can be extracted not only from the audio recordings but also from a variety of information sources surrounding the musical content, in particular different types of metadata usually attached to it, especially cover art and lyrics (in sung music).

Features extracted from the audio content appear to be the most popular ones. They often rely on some kind of energy levels over different frequency ranges. Those levels are computed over a short time interval in order to observe processes that are quasi-stationary. Because this stationarity assumption will only hold over a short time scale, the issue is then to compute meaningful statistics on several features to discard noise and to focus on meaningful characteristics, during this step called integration, or pooling [5]. This integration can go from the simplest operators (average, max) to more complex ones like GMMs. In order to take temporality into account, integration using Auto-Regressive (AR) modelling [6] or tokenization using Hidden Markov Models (HMM)s [7] have been attempted. More recently, bags of model parameters, extracted over several temporal horizons, have been proposed as a representation which accounts for some temporal dynamics [8].

In addition to audio-content descriptors, some researchers have considered features relating to textual metadata or social tags [9].

The quest for a set of efficient features is indeed a vibrant field of research. The present paper falls within this research on the features, and proposes new (*e.g.* drum energy, instrument probabilities) or rarely used (*e.g.* psychoacoustic) descriptors, which cover a wider range of semantic levels, and are useful in certain classification cases. In particular, we explore new ideas of descriptors relating to editorial metadata, especially features extracted from song lyrics and cover art. Owing to the fact that such features are not always available (for instance lyrics are only valid for sung music), we resort to a particular boosting algorithm that can cope with missing features. In the following section, we describe the explored features. Then in Section 3, we present a classification algorithm adapted to their learning. In Section 4, we present our experiment and comment the results, and we conclude in Section 5.

2. EXPLORED FEATURES

Many aspects of music are potentially useful for a classification algorithm, namely timbral, tonal or rhythmical aspects. We choose to retain only timbral features, which have consistently proven to be superior to the two others in previous works (which was also verified in our experiments), and to focus the study on the usefulness of original editorial and social metadata.

2.1. Timbral features

The following descriptors are extracted from the audio signal of each song.

MFCC-HMM

The timbre of a song can be commonly represented by the first 13 Mel Frequency Cepstral Coefficients (MFCC). In order to represent longer portions of the signal, one can compute the average and standard deviation of the MFCC. But this appears quite simplistic as the

This work was realized as part of the Quaero Programme, funded by OSEO, French State agency for innovation

temporal evolution and dynamics of the timbre are not taken into account.

Thus, following an idea developed in [7] for music segmentation, we use an integrated representation of MFCC, in which the MFCC sequence is seen as a realization of a Hidden Markov Model (HMM). Beforehand, MFCCs describing one eighth of a second of audio are extracted, by temporally averaging instantaneous MFCC (covering 23 ms, during which the signal is considered as stationary). This gives a MFCC sequence for each song. Then, an HMM is estimated, using all the sequences from the training data as examples. Subsequently, to describe a song, its HMM state sequence is decoded, and the feature description will be the histogram of these decoded states. In this paper, HMM have simple Gaussian probabilities and full covariance matrix.

Drum Energy

We have also supposed that the relative power of the drums could be useful for assessing some characteristics of a song, such as genre or emotion. To this end, we have used a drum separation algorithm [10], which yields two separated signals for each song. This algorithm runs fast and its code is freely available[1]. The relative power of the drums, compared to the other components, is first computed on frames of 200 ms. Then, we keep as descriptor the mean and standard deviation of this value over the whole song.

Instrument Presence Probabilities

The instrumentation constitutes a high-semantic-level information, that can be correlated to musical genre, or some other categories of tags. This leads us to represent it with a feature deduced from the output of an automatic musical instrument recognition system. Prior to the song analysis, we build rough SVM instruments classifiers, using a real-music solo performance database (similar to the one in [4]). The instruments are: double bass, drums, guitar, piano and voice. They use the following features: MFCC, Octave band signal intensity, and Line Spectral Frequency coefficients[2]. Then, the final feature consists in the predicted probability of presence for each instrument on the song.

Psychoacoustic Descriptors

We have also experimented a psychoacoustically-motivated representation of the signal content, which has been used in [11], but, to our knowledge, never separately evaluated on an automatic tagging task. It consists of 13 features representing timbral or perceptual properties of the sound. Some are already widely used (spectral centroid, loudness), whereas most are very rare: tonal dissonance, spectral dissonance, perceived size of the sound, number of perceived tones... They are extracted using Psysound[3], and temporally integrated (mean) on the whole song. As for MFCC and Drum Energy, these descriptors are computed over short frames, and then integrated over the whole song by mean and variance.

2.2. Editorial and social metadata

In the recent years, a vast amount of music metadata has become freely available online. All this data may be very useful for automatic classification.

[1] `http://perso.telecom-paristech.fr/~liutkus/`

[2] These features are extracted using Yaafe. For more details, visit `http://yaafe.sourceforge.net/`

[3] `http://www.psysound.org/)`

Echo Nest Tags

User tags give a high-level description of the user perception of music. Even if the free vocabulary used by users may not be always relevant, user tags can still be used as a feature for automatically predicting other, more relevant and more reliable, tags. This kind of feature has already been used [9], but remains quite new in the field, and would then benefit from being evaluated in the present work.

To this end, we collect the tags given by The Echo Nest[4]. This service provides a set of labels for each song, with their relative relevance ($0 \leq r < 1$). We only keep the 20 tags that are most frequently given on our set of songs. All these tags are represented by their relevance.

Lyrics Topics

Another aspect of a song is the lyrics. Most pop songs contain some, and they can consitute indications on the genre, or emotion for instance. But they are quite difficult to extract accurately from the signal. Fortunately, a simple search on the Web with the title of the song and the name of the artist is likely to bring all the lyrics (given that the song is popular enough). Based on this observation, we have automatically queried ChartLyrics[5] to fetch the lyrics of the analyzed songs. The texts are only queried by with the song title and artist name. Such a simple request is particularly attractive since it allows building fully automatic search systems. Thus, when this request does not bring any response, we do not reformulate it and we consider the title to be the lyrics.

Of course, these raw lyrics are not expected to be meaningful enough for an automatic classifier. This is why we exploit a representation based on "latent lyrics topics". We begin by computing a matrix M of dimension $N_{songs} \times N_{words}$, indicating the *term frequency-inverse document frequency* (TFxIDF) of each word in each song. TFxIDF takes into account the global frequency of the words, in order to represent the relative strength of a word to a particular song, compared to other songs. The word vocabulary is built on all songs (excluding the *stop words*). Then, we perform a Non-negative Matrix Factorization (NMF) on this matrix, similarly to [12]. This technique consists in factorizing matrix M into the product of two new matrices: $M \approx R \cdot T$, to prompt the emergence of the topics. The first matrix R (of dimension $N_{songs} \times N_{topics}$) gives the relevance of each topic for each song. And T (dimension $N_{topics} \times N_{words}$) describes the contribution of each word to each topic. The rows of R constitute our song "topic" feature.

Lyrics Emotions

Topics can be informative for automatic learning, but lyrics are also very likely to evocate particular emotions. Consequently, we have used a corpus, made by Bradley & Lang [13], which places 2477 common english words in an emotional space, with dimensions: *valence*, *arousal* and *dominance*. All words w in the text are replaced by their emotional value found in this $\mathcal{D}$ dictionary: $\mathbf{v} = \mathcal{D}(w)$. When no exact matching entry is found in the dictionary , we use *Lancaster stemming* [14] to match a stemmed entry with the stemmed word: $\mathbf{v} = \mathcal{D}_{stemmed}(\mathrm{stem}(w))$. Stemming is an operation, calculating a base that is shared by different words which derive from the same root (*i.e.* love, loving, loved ...).Then, we represent each song by the mean and variance of the three emotion dimensions, for all words found in $\mathcal{D}$ or $\mathcal{D}_{stemmed}$.

[4] `http://www.echonest.com/`

[5] `http://www.chartlyrics.com/`

Cover Image Descriptors

All commercial records have a cover image. This image is supposed to be related to the music it illustrates. This is why we have retrieved cover images on Discogs[6] and last.fm[7]. These images are represented by many MPEG-7 descriptors : scalable color, dominant color, color layout, color structure, homogeneous texture, edge histogram, contour shape and region shape. To these features, we add an estimated number of visible human faces, and color histograms.

3. CLASSIFICATION ALGORITHM

All our featured are processed by a boosting algorithm. Boosting is a learning technique, that iteratively trains several complementary versions of a "weak" (performing slightly better than chance) classifier h. In this work, the weak classifier is a decision stump (decision tree with only two leaves), which boils down to a simple threshold on an optimal dimension. At the end, the trained versions of the weak classifier are gathered by a weighted sum, which can be thresholded for binary decision. This configuration has proven efficient [4] and can be trained fast.

Because our descriptors cannot be computed for all songs (for instance when the cover image was not retrieved, or when the song contains no lyrics), we have to handle missing features. Thanks to the flexibility of the boosting algorithm, it can be easily adapted to this configuration [15]. In this case, the weak classifier may abstain from answering, and the boosting algorithm minimizes a loss function which takes this abstention into account.

4. EXPERIMENTS AND RESULTS

4.1. Experimental framework

The experiment is done on the CAL500 database [3]. It contains 500 pop songs, with tags describing emotion, genre, instrumentation, general song characteristics (energy, major/minor tonality), the ideal listening context ("*Usage*"), and characteristics of the singing voice. This dataset, although of modest size, presents the advantage of having highly reliable ground-truth annotations. We use the same 61 tags as in [2], and aim at predicting their presence or absence. The tests are conducted with 10-fold cross-validation, keeping 450 songs for training, and 50 for testing. For complexity reduction, we only use 30 s of each song: between instants 30 s and 60 s.

To compare the prediction accuracy of the different systems on the test set, we use two different ranking metrics. Ranking metrics evaluate the list of examples ranked by predicted score, compared against the binary ground truth. Our first measure is the Mean Average Precision (MAP). It can be interpreted as the average precision for each possible recall value. We also use the Area Under the Receiver Operating Characteristic curve (AROC). This curve represents the correct detection rate with respect to the false alarm rate, computed at each element in the ranking. More information about these metrics can be found in [16].

4.2. Results and discussion

The global performance of the different descriptors on all tags is presented in Table 1, and compared to the common first 13 MFCC, integrated to the whole song duration by mean and variance. First, we can see that the psychoacoustic description is the best timbral one, and EchoNest tags constitute the best feature based on metadata. These two representations even lead to better performance than MFCC. We also see that the Drum energy performs surprisingly well, given that it consists in only two coefficients (mean,variance). Indeed, even though it has the lowest performance among timbre descriptors, it sustains comparison with all of them. A cross-validated Student t-test [17] on the whole set of tags confirmed that the 95% confidence interval in AROC is $[58.9, 61.5]$. We can also notice that the MFCC-HMM yield performance close to those of mean/var-MFCC, but do not beat the latter. This may be due to the small size of the training corpus.

Feature	MAP (%)	AROC (%)
MFCC (mean,var)	44.4	63.3
Instrument proba	43.4	62.2
Drum energy	41.1	59.6
Psychoacoustic	47.1	64.9
MFCC-HMM 32 states	44.3	62.7
MFCC-HMM 64 states	44.1	62.6
Echo Nest tags	46.0	65.0
Lyrics 16 topics	38.6	56.3
Lyrics 32 topics	38.0	55.9
Lyrics emotions	36.5	54.4
Visual descriptors	35.2	51.8
Best features concatenated	50.1	69.1

Table 1. Global performance of the different descriptors

Figure 1 represents the AROC of the different songs, grouped by tag category. We can see that the performance of the HMM-MFCC is closely correlated to the one of mean/var-MFCC. We can also see that the EchoNest tags are particularly suitable for predicting *Genre*, *Instrument* or *Usage* tags. Indeed, user tags are very often related to these categories. The lyrics topics, although yielding modest results on the whole tag set, are the best performing representation on the *Vocals* tags. The observation that the lyrics influence the singing, do not appear surprising. This processing of the lyrics performs better than Lyrics emotions, which are still quite good for the genre tags.

The performance of the Instrument presence probabilities can seem surprising, since they are beaten by MFCC on *Instrument* tags, but give more accurate predictions on the three last categories, which we expect to be less closely correlated to the audio signal. This may be due to the fact that the presence of instrument is a higher-semantic level description, and is more suitable for predicting tags less correlated to the audio. This result is still encouraging, because these instrument predictors are quite simple, and can be greatly improved. Also, the visual descriptors give the least accurate predictions among all representations. However, they perform better than random[8], which shows that some useful information is indeed contained in the cover image. Overall, Figure 1 shows also, that the representations give information of different nature, because their relative performance is not constant among tags. This observation leads us to make a last experiment, in which we concatenate good and complementary descriptors: mean/var-MFCC, Drum energy, Psychoacoustic, EchoNest tags and Lyrics topics. The result of this experiment is shown at the bottom of Table 1, and shows that, indeed, these descriptors give complementary information, with limited redundancy,

[6] `http://www.discogs.com/`
[7] `http://www.lastfm.fr/`

[8] This difference has also been proved significant by a cross-validated Student t-test [17], which yielded a p-value of $p = 2.9\%$.

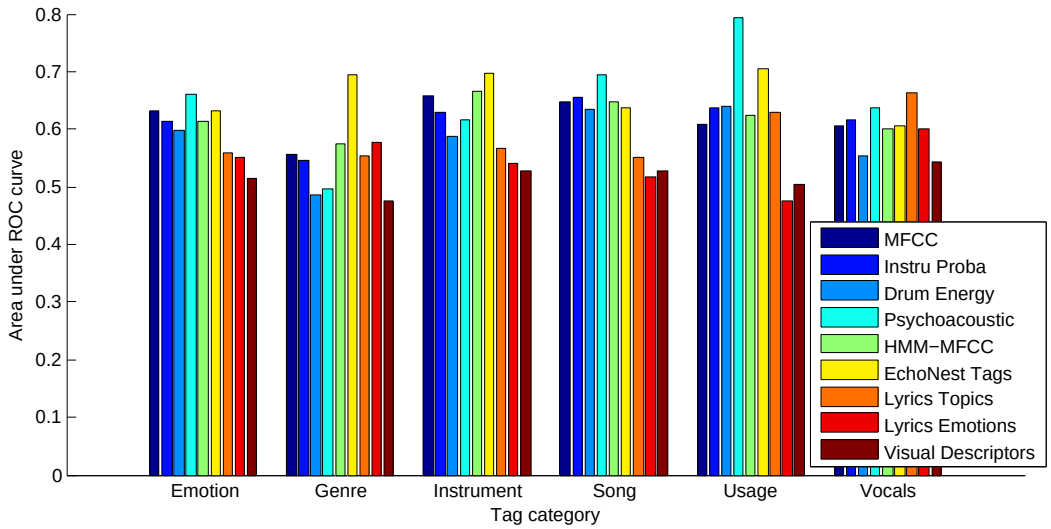

Fig. 1. AROC obtained with the different features, grouped by tag category.

and benefit from being exploited together.

5. CONCLUSION AND PERSPECTIVES

We have explored original features for the classification of music, and shown their usefulness for this task. It was also demonstrated that the joint exploitation of these diverse descriptors permits to significantly improve classification performances. This supports the fact that the proposed features are both meaningful and complementary to the classical MFCC.

Future work will be dedicated to the design of more accurate representations for both lyrics emotion and cover image. Indeed, it seems that the simple emotion description adopted here (only based on mean and variance) has shown the interest of this feature. But it is probably too crude to well capture the complexity and richness of this modality. Cover art has also been shown useful, but needs more advanced reflection on its representation.

6. REFERENCES

[1] T. Bertin-Mahieux, D. Eck, and M. Mandel, "Automatic Tagging of Audio: The State-of-the-Art," in *Machine Audition: Principles, Algorithms and Systems*, Wenwu Wang, Ed. IGI Publishing, 2010.

[2] L. Barrington, M. Yazdani, D. Turnbull, and G. Lanckriet, "Combining Feature Kernels for Semantic Music Retrieval," in *ISMIR*, 2008, pp. 614–619.

[3] D. Turnbull, L. Barrington, D. Torres, and G. Lanckriet, "Semantic Annotation and Retrieval of Music and Sound Effects," *TASLP*, vol. 16, no. 2, pp. 467–476, Feb. 2008.

[4] R. Foucard, S. Essid, M. Lagrange, and G. Richard, "Multi-scale temporal fusion by boosting for music classification," in *ISMIR*, 2011.

[5] Y. Boureau, F. Bach, Y. LeCun, and J. Ponce, "Learning mid-level features for recognition," in *CVPR*. 2010, IEEE.

[6] A. Meng, P. Ahrendt, J. Larsen, and L. K. Hansen, "Temporal feature integration for music genre classification," in *TASLP*, 2007, pp. 1654–1664.

[7] M. Levy, M. Sandler, and M. Casey, "Extraction of high-level musical structure from audio data and its application to thumbnail generation," in *ICASSP*. 2006, vol. 5, IEEE.

[8] K. Ellis, E. Coviello, and G. Lanckriet, "Semantic annotation and retrieval of music using a bag of systems representation," in *ISMIR*, 2011.

[9] L. Barrington, D. Turnbull, M. Yazdani, and G. Lanckriet, "Combining audio content and social context for semantic music discovery," in *SIGIR*. 2009, pp. 387–394, ACM.

[10] A. Liutkus, R. Badeau, and G. Richard, "Gaussian processes for underdetermined source separation," *IEEE Transactions on Signal Processing*, vol. 59, no. 7, pp. 3155 –3167, July 2011.

[11] R. Foucard, S. Essid, M. Lgrange, and G. Richard, "A Regressive boosting approach to automatic audio tagging based on soft annotator fusion," in *ICASSP*, 2012.

[12] F. Kleedorfer, P. Knees, and T. Pohle, "Oh Oh Oh Whoah! Towards automatic topic detection in song lyrics," in *Ismir*, 2008, pp. 287–292.

[13] M. Bradley and P. Lang, "Affective norms for English words (ANEW): Instruction manual and affective ratings," Tech. Rep., Center for Research in Psychophysiology, University of Florida, Gainesville, FL, 2010.

[14] C. Paice, "A word stemmer based on the Lancaster stemming algorithm," in *ACM SIGIR*, 1990, pp. 56–61.

[15] F. Smeraldi, M. Defoin-Platel, and M. Saqi, "Handling missing features with boosting algorithms for protein-protein interaction prediction," in *Data Integration in the Life Science*, Aug. 2010, vol. 6254, pp. 132–147.

[16] J. Herlocker, J. Konstan, L. Terveen, and J. Riedl, "Evaluating collaborative filtering recommender systems," *Transactions on Information Systems*, vol. 22, no. 1, pp. 5–53, Jan. 2004.

[17] T.G. Dietterich, "Approximate Statistical Tests for Comparing Supervised Classification Learning Algorithms," *Neural Computation*, vol. 10, no. 7, 1998.