SIGNAL + CONTEXT = BETTER CLASSIFICATION

Jean-Julien Aucouturier
Grad. School of Arts and Sciences
The University of Tokyo, Japan

François Pachet, Pierre Roy, Anthony Beurivé
SONY CSL Paris
6 rue Amyot, 75005 Paris, France

ABSTRACT

Typical signal-based approaches to extract musical descriptions from audio only have limited precision. A possible explanation is that they do not exploit context, which provides important cues in human cognitive processing of music: e.g. electric guitar is unlikely in 1930s music, children choirs rarely perform heavy metal, etc. We propose an architecture to train a large set of binary classifiers simultaneously, for many different musical metadata (genre, instrument, mood, etc.), in such a way that correlation between metadata is used to reinforce each individual classifier. The system is iterative: it uses classification decisions it made on some classification problems as new features for new, harder problems; and hybrid: it uses a signal classifier based on timbre similarity to bootstrap symbolic inference with decision trees. While further work is needed, the approach seems to outperform signal-only algorithms by 5% precision on average, and sometimes up to 15% for traditionally difficult problems such as cultural and subjective categories.

1 INTRODUCTION: BOOTSTRAPPING SYMBOLIC REASONING WITH ACOUSTIC ANALYSIS

People routinely use many varied high-level descriptions to talk and think about music. Songs are commonly said to be “energetic”, to make us “sad” or ”nostalgic”, to sound “like film music” and to be perfect to “drive a car on the highway” among a possible infinity of similar metaphors. The Electronic Music Distribution industry is in demand of robust computational techniques to extract such descriptions from musical audio signals. The majority of existing systems to this aim rely on a common model of the signal as the long-term accumulative distribution of frame-based spectral features. Musical audio signals are typically cut into short overlapping frames (e.g. 50ms with a 50% overlap), and for each frame, a feature vector is computed. Features usually consists of generic, all-purpose spectral representations such as mel-frequency cepstrum coefficients (MFCCs), but can also be e.g. rhythmic features [1]. The features are then fed to a statistical model, such as a Gaussian mixture model (GMM), which estimates their global distribution over the total length of the extract. Global distributions can then be used to compute decision boundaries between classes (to build e.g. a genre classification system such as [2]) or directly compared to one another to yield a measure of acoustic similarity [3].

While such signal-based approaches are by far the most dominant paradigm currently, recent research increasingly suggests they are plagued with important intrinsic limitations [3, 5]. One possible explanation is that they take an auditory-only approach to music classification. However, many of our musical judgements are not low-level immediate perceptions, but rather high-level cognitive reasoning which accounts for the evidence found in the signal, but also depends on cultural expectations, a priori knowledge, interestingness and “remarkability” of an event, etc. Typical musical descriptions only have a weak and ambiguous mapping to intrinsic acoustic properties of the signal. In [6], subjects were asked to rate the similarity between pairs of 60 sounds and 60 words. The study concludes that there is no immediately obvious correspondence between single acoustic attributes and single semantic dimensions, and go as far as suggesting that the sound/word similarity judgment is a forced comparison (“to what extent would a sound spontaneously evoke the concepts that it is judged to be similar to?”). Similarly, we studied in [4] the performance of a typical classifier on a heterogeneous set of more than 800 high-level musical symbols, manually annotated for more than 4,000 songs. We observed that surprisingly few of such descriptions can be mapped with reasonable precision to acoustic properties of the corresponding signals. Only 6% of the attributes in the database are estimated with more than 80% precision, and more than a half of the database’s attributes are estimated with less that 65% precision (which hardly better than a binary random choice, i.e. 50%). The technique provides very precise estimates for attributes such as homogeneous genre categories or extreme moods like “aggressive” or “warm”, but typically fails on more cultural or subjective attributes which bear little correlation with the actual sound of the music being described, such as “Lyric Content”, or complex moods or genres (such as “Mysterious” or “Electronica”).

This does not mean human musical judgements are beyond computational approximation, naturally. The study in [4] shows that there are large amounts of correlation between musical descriptions at the symbolic level. Table 1 shows a selection of pairs of musical metadata items (from a large manually-annotated set), which were found

Table 1. Selected pairs of musical metadata with their Φ score (χ^2 normalized to the size of the population), between 0 (corresponding to statistical independence between the variables) and 1 (complete deteministic association). Data analysed on a a set of 800 metadata values manually annotated for more than 4,000 songs, used in previous study [4]

Attribute1	Attribute2	Φ
Music-independant		
Textcategory Christmas	Genre Special Occasions	0.89
Mood strong	Character powerful	0.68
Mood harmonious	Character well-balanced	0.60
Character robotic	Mood technical	0.55
Mood negative	Character mean	0.51
Music-dependant		
Main Instruments Spoken Vocals	Style Rap	0.75
Style Reggae	Country Jamaica	0.62
Musical Setup Rock Band	Main Instruments Guitar (distortion)	0.54
Character Mean	Style Metal	0.53
Musical Setup Big Band	Aera/Epoch 1940-1950	0.52
Main Instruments transverse flute	Character Warm	0.51

to particularly fail a Pearson's χ^2-test ([7]) of statistical independence. χ^2 tests the hypothesis that the relative frequencies of occurrence of observed events follow a flat random distribution (e.g. that hard rock songs are not significantly more likely to talk about violence than non hard-rock songs). On the one hand, we observe considerable correlative relations between metadata, which have little to do with the actual musical usage of the words. For instance, the analysis reveals common-sense relations such as "Christmas" and "Special occasions" or "Well-known" and "Popular". This illustrates that the process of categorizing music is consistent with psycholinguistics evidences of semantic associations, and that the specific usage of words that describe music is largely consistent with their generic usage: it is difficult to think of music that is e.g. both strong and not powerful. On the other hand, we also find important correlations which are not intrinsic properties of the words used to describe music, but rather extrinsic properties of the music domain being described. Some of these relations capture historical ("ragtime is music from the 1930's") or cultural knowledge ("rock uses guitars"), but also more subjective aspects linked to perception of timbre ("flute sounds warm", "heavy metal sounds mean").

Hence, we are facing a situation where:

1. Traditional signal-based approaches (e.g. nearest-neighbor classification with timbre similarity) work for only a few well-defined categories, which have a clear and unambiguous sound signature (e.g. Heavy metal).

2. Correlations at the symbolic level are potentially useful for many categories, and can be easily exploited by machine learning techniques such as Decision Trees [8]. However, these require the availability of values for non-categorical attributes, to be used as features for prediction: we have to first know that "this has distorted guitar", to infer that "it's probably rock".

This paper quite logically proposes to use the former to bootstrap the latter. First, we use a timbre-based classifier to estimate the values of a few timbre-correlated attributes. Then we use decision trees to make further predictions of cultural attributes on the basis of the pool of timbre-correlated attributes. This results in an iterative system which tries to solve simultaneously a *set* of classification problems, by using classification decisions it made on some problems as new features for new, harder problems.

2 ALGORITHM

This section describes the hybrid classification algorithm, starting with its 2 sub-components: an acoustic classifier based on timbre similarity, and a decision-tree classifier to exploit symbolic-level correlations between metadata. In the following, metadata items are notated as *attributes* $\mathcal{A}_i$, which take boolean values $\mathcal{A}_i(\mathcal{S})$ for a given song $\mathcal{S}$ (e.g. has_guitar($\mathcal{S}$) $\in \{true, false\}$).

2.1 Sub-component1: Signal-based classifier

The acoustic component of the system is a nearest neighbor classifier based on timbre similarity. We use the timbre similarity algorithm described in [3]: 20-coefficient MFCCs, modelled with 50-state GMMs, compared with Monte-Carlo approximation of the Kullback-Leibler distance. The classifier infers the value of a given attribute $\mathcal{A}$ for a given song $\mathcal{S}$ by looking at the values of $\mathcal{A}$ for songs that are timbrally similar to $\mathcal{S}$. For instance, if 9 out of the 10 nearest neighbors of a given song are "Hard Rock" songs, then it is very likely that the seed song be a "Hard Rock" song itself.

More precisely, we define as our observation $\mathcal{O}_{\mathcal{A}}(S)$ the number of songs among the set $\mathcal{N}_S$ of the 10 nearest neighbors of $\mathcal{S}$ for which $\mathcal{A}$ is *true*, i.e.

$$\mathcal{O}_{\mathcal{A}}(S) = card\{\mathcal{S}_i \setminus \mathcal{S}_i \in \mathcal{N}_S \wedge \mathcal{A}(\mathcal{S}_i)\} \qquad (1)$$

We make a maximum-likelihood decision (with flat prior) on the value of the attribute $\mathcal{A}$ based on $\mathcal{O}_{\mathcal{A}}(S)$:

$$\mathcal{A}(\mathcal{S}) = p(\mathcal{O}_{\mathcal{A}}(S)/\mathcal{A}(\mathcal{S})) > p(\mathcal{O}_{\mathcal{A}}(S)/\overline{\mathcal{A}(\mathcal{S})}) \qquad (2)$$

where $p(\mathcal{O}_{\mathcal{A}}(S)/\mathcal{A}(\mathcal{S}))$ is the probability to observe a number $\mathcal{O}_{\mathcal{A}}(S)$ of true values in the set of nearest neighbors of S, given that $\mathcal{A}$ is true, and $p(\mathcal{O}_{\mathcal{A}}(S)/\overline{\mathcal{A}(\mathcal{S})})$ is the probability to make the same observation given that $\mathcal{A}$ is false. The likelihood distribution $p(\mathcal{O}_{\mathcal{A}}(S)/\mathcal{A}(\mathcal{S}))$ is estimated on a training database by the histogram of the empirical frequencies of the number of positive neighbors for all songs having $\mathcal{A}(\mathcal{S}) = true$ (similarly for $P(\mathcal{O}_{\mathcal{A}}(S)/\overline{\mathcal{A}(\mathcal{S})})$).

2.2 Sub-component2: Decision-tree classifier

The symbolic component in our system is a decision-tree classifier [8]. It predicts the value of a given attribute (the *category* attribute) on the basis of the values of some *non-category* attributes, in a hierarchical manner. For instance, a typical tree could classify a song as “natural/acoustic”

- if it is not aggressive
- else, if it is from the 50’s (where little amplification was used)
- else, if it’s a folk or a jazz band that performs it,
- else, if it doesn’t use guitar with distortion, etc.

Decision rules are learned on a training database with the implementation of C4.5 provided by the Weka library [8]. As mentionned above, a decision tree for a given attribute is only able to predict its value for a given song if we have access to all the values of the other non-categorical attributes for that same song. Therefore, it is of little use as such. The algorithm described in the next section uses timbre similarity inference to bootstrap the automatic categorization with estimates of a few timbre-grounded attributes, and then use these estimates in decision trees to predict non-timbre correlated attributes.

2.3 Training procedure

The algorithm is a training procedure to generate a set of classifiers for N attributes $\{A_k; k \in [1, N]\}$. Training is iterative, and requires a database of musical signals with annotated values for all A_k. At each iteration i, we produce a set of classifiers $\{\widetilde{A_k}^i; k \in [1, N]\}$, which each estimates the attribute A_k at iteration i. Each classifier is associated with a precision $p(\widetilde{A_k}^i)$. At each iteration i, we define as $best(\widetilde{A_k}^i)$ the best classifier of A_k so far, i.e.

$$best(\widetilde{A_k}^i) = \widetilde{A_k}^m, m = \arg\max_{j \leq i} p(\widetilde{A_k}^j) \qquad (3)$$

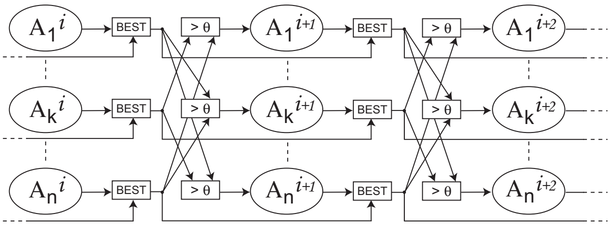

Figure 1. The algorithm constructs successive classifiers for the attributes A_k. At each iteration, the features of the new classifiers is the set of the previous classifiers (in the set of $\{A_l\}_{l \neq k}$) with precision greater than threshold θ. At each iteration, for each attribute, only the best estimate so far is stored for future use. The A_k^i are estimated by timbre inference for $i = 1$, and by decision trees for $i \geq 2$

More precisely, at each iteration i, each of the newly-built classifiers takes as input (aka features) the output of classifiers from the previous generations, based on their precision. Hence, each iteration in this overall training procedure requires some training (to build the classifiers) and some testing (to select the input of the next iteration’s classifiers).

- $i = 1$: Bootstrap with timbre inference

  Classifiers: The classifiers $\widetilde{A_k}^1$ are based on timbre inference, as described in Section 2.1. Each of these classifiers estimates the value of an attribute A_k for a song S based on the audio signal only, and doesn’t require any attribute value in input.

  Training: Timbre inference requires training to evaluate the likelihood distributions $p(\mathcal{O}_{\mathcal{A}_k}(S)/\mathcal{A}_k(\mathcal{S}))$, and hence a training set $\mathcal{L}^1(\mathcal{A}_k)$ for each attribute $\mathcal{A}_k$

  Testing: Each $\widetilde{A_k}^1$ is tested on a testing set $\mathcal{T}^1(\mathcal{A}_k)$. For each song $S \in \mathcal{T}^1(\mathcal{A}_k)$, estimates $\widetilde{A_k}^1(S)$ are compared to groundtruth $A_k(S)$, to yield a precision value $p(\widetilde{A_k}^1)$.

- $\forall i \geq 2$: Iterative improvement by decision trees

  Classifiers: The classifiers $\widetilde{A_k}^i$ are decision trees, as described in Section 2.2. Each of them estimates the value of an attribute A_k based on the output of the best classifiers from previous generations. More precisely, the non-category attributes (aka features) used in the $\widetilde{A_k}^i$ are the attribute estimates generated by a subset $\mathcal{F}_k{}^i$ of all previous classifiers $\{\widetilde{A_l}^j; l \in [1, N], j < i\}$, defined as:

  $$\mathcal{F}_k{}^i = \{best(\widetilde{A_l}^{i-1}); l \neq k, p(best(\widetilde{A_l}^{i-1})) \geq \theta\} \qquad (4)$$

  where $0 \leq \theta \leq 1$ is a precision threshold. $\mathcal{F}_k{}^i$ contains the estimate generated by the best classifier so far (up to iteration $i - 1$) for every attribute other than A_k, provided that its precision be greater than θ. This is illustrated in Figure 1.

  Training: Decision trees require training to build and trim decision rules, and hence a training set $\mathcal{L}^i(\mathcal{A}_k)$ for each category attribute $\mathcal{A}_k$ and each iteration $i \geq 2$: new trees have to be trained for every new set of features attributes ${\mathcal{F}_k}^i$, which are selected based on their precision at previous iterations. Trees are trained using the *true* values (groundtruth) of the non-categorical attributes (but they will be tested using *estimated* values for these same attributes, see below).

Testing: Each $\widetilde{{A_k}^i}$ is tested on a testing set $\mathcal{T}^i(\mathcal{A}_k)$. For each song $S \in \mathcal{T}^i(\mathcal{A}_k)$, estimates $\widetilde{{A_k}^i(S)}$ are computed using the *estimated* values $best(\widetilde{{A_l}^{i-1}})(S)$ of the non-categorical attributes A_l, i.e. values computed by the corresponding best classifier, and compared to the *true* value $A_k(S)$, to yield a precision value $p(\widetilde{{A_k}^i})$.

- Stop condition: The training procedure terminates when there is no more improvement of precision between successive classifiers for any attribute, i.e. the set of all $best(\widetilde{{A_k}^i})$ reaches a fixed point.

2.4 Output

The output of the above training procedure is a final set of classifiers, containing the best classifiers for each A_k, i.e. $\{best(\widetilde{{A_k}^{i_f}}), k \in [1, N]\}$, where i_f is the iteration where stop condition is reached. For a given attribute A_k, the final classifier is a set of $1 \leq n \leq N.i_f$ component classifiers, arranged in a tree where parent classifiers use results of their children. The top-level node is a decision tree[1] for A_k, the intermediate nodes are decision trees for A_k but also part of the other attributes $A_l, l \in [1, N]$, and the leaves are timbre classifiers also for part of the $A_l, l \in [1, N]$. Each component classifier has fixed parameters (likelihood distributions for timbre classifiers and rules for decision trees) and fixed features (the ${\mathcal{F}_k}^i$), as determined by the above training process. Therefore, they are standalone algorithms which take as input an audio signal S, and outputs an estimate $\widetilde{A_k}(S)$.

Figure 2 illustrates a possible outcome scenario of the above process, using a set of attributes including “Style Metal”, “Character Warm”, “Style Rap” (which are attributes well correlated with timbre) and “TextCategory Love” and “Setup Female Singer”, which are poor timbre estimates (the former being arguably “too cultural”, and the latter apparently too complex to be precisely described by timbre). The first set of classifiers $\widetilde{S_A}^0$ is built using timbre inference, and logically performs well for the timbre-correlated attributes, and poorly for the others. Classifiers at iteration 2 estimate each of the attributes using a decision tree on the output of the timbre classifiers (only keeping classifiers above $\theta = 0.75$, which appear in gray). For instance, the classifier for “Style Metal” uses a decision tree on the output of the classifiers for “Character Warm” and “Style Rap”, and achieves poorer classification precision that the original timbre classifier. Similarly, the classifier for “Setup Female Singer” uses a decision tree on “Style Metal”, “Character Warm” and “Style Rap”, which results on better precision than the original timbre classifier. At the next iteration, the just-produced classifier for “Setup Female Singer” (which happens to be above threshold θ) is used in a decision tree to give a good estimate of “TextCategory Love” (as e.g. the knowledge of whether the singer is a female may give some information about the lyric contents of the song). At the next iteration, all best classifiers so far may be used in a decision tree to yield a classifier of “Style Metal” which is even better than the original timbre classifier (as it uses some additional cultural information).

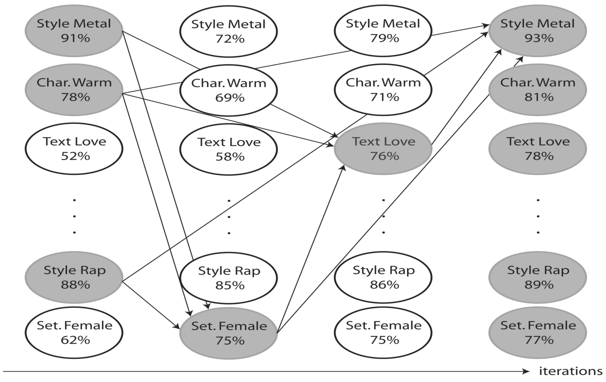

Figure 2. An example scenario of iterative attribute estimation

3 RESULTS

3.1 Database

We report here on preliminary results of the above algorithm, using a database of human-made judgments of high-level musical descriptions, collected for a large quantity of commercial music pieces. The data is proprietary, and made available to the authors by research partnerships. The database contains 4936 songs, each described by a set of 801 boolean attributes (e.g. “Mood happy”= *true*). These attributes are grouped in 18 categories, some of which being correlated with some acoustic aspect of the sound (“Main Instrument”,“Dynamics”), while others seem to result from a more cultural take on the music object (“Genre”, “Mood”, “Situation[2]”). Attribute values were filled in manually by human listeners, under a process related to Collaborative Tagging, in a business initiative comparable to the Pandora project[3].

[1] or a timbre classifier for A_k if $i_f = 1$

[2] i.e. in which everyday situation would the user like to listen to a given song, e.g. “this is music for a birthday party”

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

3.2 About the isolation between training and testing

As seen above, there are several distinct training and testing stages in the training procedure described here. For a joint optimisation of N attributes over i_f iterations, as many as $N.i_f$ training sets $\mathcal{L}^i(\mathcal{A}_k)$ and testing sets $\mathcal{T}^i(\mathcal{A}_k)$ have to be constructed dynamically. Additionally, for the purpose of evaluation when this training procedure is finished, final algorithms for all $\mathcal{A}_k$ have to be tested on separate testing sets $\mathcal{W}(\mathcal{A}_k)$.

The construction of these various datasets has to respect several constraints to ensure isolation between training and testing data.

- No overlap between $\mathcal{T}^i(\mathcal{A}_k)$ and $\mathcal{L}^i(\mathcal{A}_k)$
- No overlap between $\mathcal{T}^i(\mathcal{A}_k)$ and $\cup_{j<i,\forall l}\mathcal{L}^j(\mathcal{A}_l)$ (since the classifier $\widetilde{A_k}^i$ uses component classifiers from previous iterations, for possibly all attributes)
- No overlap between $\mathcal{W}(\mathcal{A}_k)$ and all other training and testing sets, i.e. $\{\mathcal{L}^i(\mathcal{A}_l); 1 \leq i \leq i_f, 1 \leq l \leq N\} \cup \{\mathcal{T}^i(\mathcal{A}_l); 1 \leq i \leq i_f, 1 \leq l \leq N\}$.

In practice, these set constraints are very difficult to enforce when one requires balanced datasets (roughly as many positive and negative examples for all attributes in all training and testing sets): it is a complex combinatorial problem, all the more so as the number of attributes N increases (which is a desirable feature as seen in Section 3.3). Unbalanced datasets create additional learning problems which we found were also difficult to handle in the current iterative framework, notably because cross-validation cannot be conducted at every iteration [9].

Therefore, we opted for an approximation strategy where datasets were taken as:

- All $\mathcal{T}^i(\mathcal{A}_k)$ equal $\forall i$; all $\mathcal{L}^i(\mathcal{A}_k)$ equal $\forall i$
- $\mathcal{T}^i(\mathcal{A}_k)$ and $\mathcal{L}^i(\mathcal{A}_k)$ contain as many positive and negative examples for $\mathcal{A}_k$
- No overlap between $\mathcal{W}(\mathcal{A}_k)$ and $\{\mathcal{L}^i(\mathcal{A}_k)\} \cup \{\mathcal{T}^i(\mathcal{A}_k)\}$.

Such datasets cannot guarantee complete isolation between training ($\mathcal{L}$) and testing ($\mathcal{T}$) data *during the training procedure*. This doesn't affect the reliability of the final testing stage, as the $\mathcal{W}$ sets are properly independent from the $\mathcal{L}$ and $\mathcal{T}$ data used during training. However, interactions between the sets used for training probably leads to over-estimations of the performance on training data ($\mathcal{L}$ and $\mathcal{T}$ sets), as well as the high variance observed on test performance ($\mathcal{W}$ sets) (see below). On the whole, this is a consequence of the rather unconventional learning architecture investigated here, and is clearly subjected to further work and clarification.

3.3 Evaluation

Table 2 shows the test performance of the above algorithm on a set of 45 randomly chosen attributes, using $\theta = 0.7$. 30 out of the 45 attributes see their classification precision improved by the iterative process (the remaining 15 do not appear in the table). We observe that, for 10 classifiers, the precision improves by more than 10% (absolute), and that 15 classifiers have a final precision greater than 70%. Cultural attributes such as “Situation Sailing” or “Situation Love” can be estimated with reasonable precision, whereas their initial timbre estimate was poor. It also appears that two “Main Instrument” attributes (guitar and choir), that were surprinsingly bad timbre correlates, have been refined using correlations between cultural attributes. This is consistent with the example scenario in Figure 2.

Table 2. Set Optimization of 45 attribute estimates

Attribute	$p(\widetilde{A_k}^0)$	$p(\widetilde{A_k}^{i_f})$	i_f	$\Delta(p)$
Situation Sailing	0.48	0.71	10	0.23
Situation Flying	0.49	0.64	3	0.15
Situation Rain	0.50	0.64	9	0.14
Instrument Guitar	0.60	0.69	4	0.09
Situation Sex	0.59	0.68	11	0.09
Situation Love	0.63	0.70	3	0.07
Lyrics Love	0.61	0.67	11	0.06
Situation Party	0.60	0.66	6	0.06
Tempo medium	0.59	0.64	4	0.05
Character slick	0.65	0.69	11	0.04
Aera/Epoch 90s	0.71	0.75	13	0.04
Character harmony	0.62	0.66	6	0.04
Rhythmics rhythmic	0.64	0.68	4	0.04
Genre Dancemusic	0.65	0.68	12	0.03
Mood dreamy	0.64	0.67	2	0.03
Style Pop	0.71	0.74	6	0.03
Mood positive	0.58	0.61	6	0.03
Mood harmonious	0.62	0.65	4	0.03
Instrument Choir	0.60	0.63	13	0.03
Dynamics up+down	0.61	0.63	5	0.02
Lyrics Associations	0.57	0.59	10	0.02
Variant expressive	0.62	0.64	2	0.02
Setup Pop Band	0.72	0.74	7	0.02
Lyrics Poetry	0.57	0.59	10	0.02
Character friendly	0.65	0.67	6	0.02
Character repeating	0.63	0.64	9	0.01
Rhythmics groovy	0.63	0.64	4	0.01
Mood romantic	0.69	0.70	9	0.01
Lyrics Wisdom	0.58	0.59	4	0.01
Lyrics Romantics	0.65	0.66	14	0.01

Figure 3 shows the influence of the number of attributes considered for joint classification on the average improvement of precision. It appears that using more attributes leads to larger improvements of (testing) precision over purely signal-based approaches: this allows the decision trees to exploit stronger correlations than in smaller sets. Larger sets also improve the stability of the results: the performance on small sets depends critically on the quality of the original timbre-correlated estimates, which are

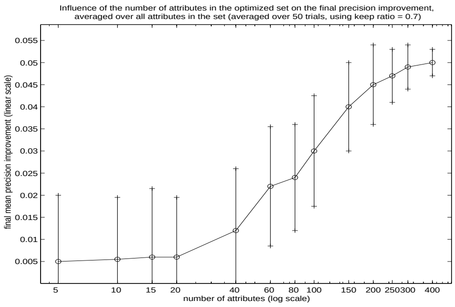

Figure 3. Influence of the number of attributes considered for joint optimization on the average improvement of precision

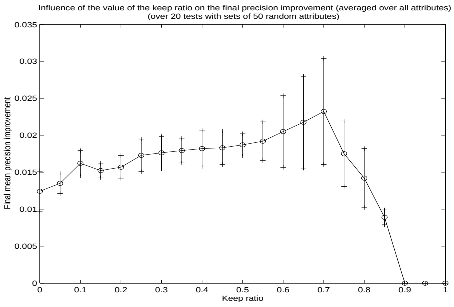

Figure 4. Influence of precision threshold θ on the average improvement of precision

used for bootstrap.

On the whole, it appears that the approach can improve precision over simple signal-based approaches by as much as 5% on average, when considering sets of several hundreds of attributes.

Figure 4 shows the influence of the precision threshold parameter, used at each iteration to select classifiers from previous iterations to be used as features. The parameter is a tradeoff between quantity and quality of the correlations to be exploited in decision trees. The curve has an intuitive inverted-U shape: small θ values lead to selecting too many bad classifiers, whereas large θ values constrain the system to use only high-quality features, which are ultimately too few to boostrap correlation analysis. The optimal value is found around 70% precision, which is consistent with the empirical upper-bound found with signal-only approaches (so-called “glass ceiling”) [3]

4 CONCLUSION

We have described an iterative procedure to train simultaneously a set of classifiers for high-level music metadata. The system exploits correlations between metadata, using decision trees, to reinforce each individual classifier. The approach outperforms signal-only algorithms by 5% precision on average when a sufficient number of metadata are considered jointly. It provides reasonable solutions to traditionally difficult problems, such as complex genres or “situations in which one would like to play the song”. However, the concurrent training and testing of very many classification algorithms makes the task of constructing well-behaved training and testing datasets unusually difficult. Some solutions remain to be found, either in the direction of algorithms to resample balanced datasets (e.g. combinatorial optimisation) or alternative formulations of the learning architecture (e.g. Bayesian belief networks).

5 REFERENCES

[1] A. Flexer, F. Gouyon, S. Dixon, and G. Widmer, “Probabilistic combination of features for music classification,” in *Proceedings of the 5th International Conference on Music Information Retrieval*, Victoria, BC (Canada), 2006.

[2] G. Tzanetakis, G. Essl, and P. Cook, “Automatic musical genre classification of audio signals,” in *proceedings ISMIR*, 2001.

[3] J.-J. Aucouturier and F. Pachet, “Improving timbre similarity: How high’s the sky ?” *Journal of Negative Results in Speech and Audio Sciences*, vol. 1, no. 1, 2004.

[4] ——, “How much audition involved in everyday categorization of music?” *Cognitive Science (submitted)*, 2007.

[5] C. McKay and I. Fujinaga, “Musical genre classification: Is it worth pursuing and how can it be improved?” in *Proceedings International Conference on Music Information Retrieval (ISMIR)*, Vancouver (BC), Canada, 2006.

[6] P. Janata, “Timbre and semantics,” keynote Presentation, Journées fondatrices Perception Sonore, Lyon (France), January 2007. Available: http://www.sfa.asso.fr/fr/gps.

[7] D. Freedman, R. Pisani, and R. Purves, *Statistics, 3rd edition*. W.W. Norton and Co., New York, 1997.

[8] J. Quinlan, *C4.5: Programs for Machine Learning*. Morgan Kauffman, 1993.

[9] J. Zhang and I. Mani, “k-nn approach to unbalanced data distributions,” in *Proceedings of the International Conference on Machine Learning (Workshop on Learning from Unbalanced Datasets)*, Washington DC, USA, 2003.