THREE CURRENT ISSUES IN MUSIC AUTOTAGGING

Gonçalo Marques[1], **Marcos Aurélio Domingues**[2], **Thibault Langlois**[3], **Fabien Gouyon**[4]
[1]gmarques@isel.pt DEETC-ISEL Lisboa, [2]maddomingues@gmail.com INESC Porto,
[3]tl@di.fc.ul.pt DI-FCUL Lisboa, [4]fgouyon@inescporto.pt INESC Porto

ABSTRACT

The purpose of this paper is to address several aspects of music autotagging. We start by presenting autotagging experiments conducted with two different systems and show performances on a par with a method representative of the state-of-the-art. Beyond that, we illustrate via systematic experiments the importance of a number of issues relevant to autotagging, yet seldom reported in the literature. First, we show that the evaluation of autotagging techniques is fragile in the sense that small alterations to the set of tags to be learned, or in the set of music pieces may lead to dramatically different results. Hence we stress a set of methodological recommendations regarding data and evaluation metrics. Second, we conduct experiments on the generality of autotagging models, showing that a number of different methods at a similar performance level to the state-of-the-art fail to learn tag models able to generalize to datasets from different origins. Third we show that current performance level of a direct mapping between audio features and tags still appears insufficient to enable the possibility of exploiting natural tag correlations as a second stage to improve performance.

1. INTRODUCTION

Music autotagging refers to the task of automatically classifying music audio excerpts with respect to a number of high-level concepts (the "tags") from potentially very diverse music facets such as Emotion, Musical instruments, Genre, Usage, etc. In the literature, a number of approaches to the task have been proposed that build upon previous work in genre and artist classification, where a direct mapping is sought via machine learning models between low-level features computed on short audio signal frames and tags [2, 4, 10, 11]. These approaches are tailored to the fact that the task is more difficult than genre classification in that the number of classes is usually much higher (genres correspond in fact to one among many facets), and models must account for the possibility that multiple labels usually apply to a given excerpt. Music tags are often correlated (for instance, Genre tags often co-occur with Instruments or Emotion tags), this is often the rationale behind implementing a 2-stage architecture, where a second stage of processing, modeling tag co-occurence relationships, can "correct" [8] the imperfect tag predictions of the first stage (see illustration in figure 1). A number of authors report on performance improvements with this procedure over the one-stage approach [1, 6–9].

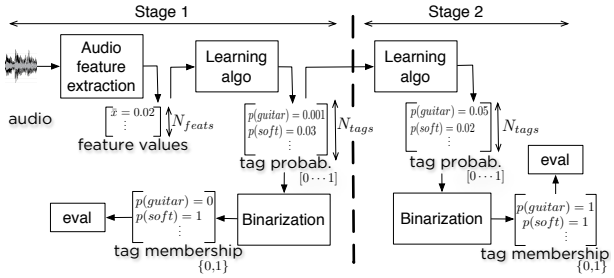

Figure 1. Generic 2-stage music autotagging framework (training of learning algorithms not represented; audio feature extraction can be statistics or time series).

This paper aims at demonstrating via systematic experiments the relevance of a number of music autotagging issues that we believe are, to the best of our knowledge, only addressed superficially in current literature. After presenting the data and systems used and reporting on initial experiments in sections 2 and 3, we address in section 4 the notion of "fragility" of evaluation methodologies and stress a number of methodological recommendations. In section 5, we address the issue of generality of autotagging models, and in section 6, we address limitations of exploiting tag correlations in a second processing stage. We finally propose a discussion on these issues and directions for future work in section 7.

2. DATA AND SIGNAL FEATURES

In this paper we use two datasets with tag annotations made available publicly to the community by fellow researchers

and on which a number of papers have reported results.

CAL500. The Computer Audition Lab 500 (CAL500) dataset (http://cosmal.ucsd.edu/cal/projects/AnnRet/) is made up of 500 Western popular song excerpts of different lengths. Excerpts annotations are among a set of 174 tags.

Magnatagatune. The Magnatagatune dataset (http://tagatune.org/Magnatagatune.html) consists of 21642 excerpts of length 30 s from 230 different artists. Excerpts annotations are among a set of 188 tags. Some pre-processing was applied to yield a cleaner dataset, referred to as Magtag5k (see section 4.2 for more details), on which we ran most of the experiments below.

Other datasets. We made use of two other publicly available datasets with only genre annotations: the Latin Music Dataset (LMD, http://www.ppgia.pucpr.br/~silla/lmd/index.html) and the ISMIR04 dataset (http://ismir2004.ismir.net/genre_contest/index.htm) to evaluate the generalization capacity of our autotagging systems (see section 5).

Features. We used MARSYAS to extract 16 audio features from 46ms frames of the audio signals with no overlap. The features are: the spectral centroid, rolloff frequency, spectral flux, and 13 MFCCs, including MFCC0. These features as the same ones used in [7].

3. AUTOTAGGING SYSTEMS

3.1 Benchmark

In order to better compare our experiments with previous literature and to facilitate the reproducibility of our experiments, we use as a benchmark the system proposed in [7], which is available under GPL in MARSYAS. [1] Performance of the Benchmark have been reported in the 2010 MIREX evaluation. In this system, frame features are collapsed in a two steps process (texture windowing and computation of global mean and standard deviation) into a 64-dimensional feature vector for the whole audio excerpt [7]. This system implements an architecture with two stages of processing, illustrated in figure 1. A multiclass SVM classifier is used in both stages. We report below on the performance of using just the first stage of processing alone, or the whole system.

3.2 Alternative systems

1. **External multiclass SVM in both stages**: This system (referred to as **Sys1**) is a 2-stage system similar to the Benchmark, with the difference that it externalizes the learning algorithm and directly uses the libSVM software package (http://www.csie.ntu.edu.tw/~cjlin/libsvm). The other difference is that normalization of the data is done via the libSVM package and not in the MARSYAS code.

2. **One-stage Markov models-based classifier**: This approach consists of using the method for genre classification based on Markov models previously described in [5]. In the context of autotagging, for each tag a pair of models are estimated and used to assign a tag to a piece of audio. This approach is referred to as **Sys2**.

3.3 Autotagging performance

	CAL500	Magtag5k
Benchmark	0.452|0.245	0.312|0.083
Sys1	0.464|0.269	0.423|0.176
Sys2	0.480|0.246	0.411|0.171

Table 1. F-score$_g$ | F-score$_{pt}$ for Benchmark, Sys1, and Sys2 on CAL500 and Magtag5k. Evaluation methodology described in section 4.2.

Table 1 presents a comparison of the performance achieved with the methods described previously and the performance obtained with the Benchmark. The performance measure is the F-score computed on global classification rates (denoted F-score$_g$) and the F-score based on the average per-tag classification rates (denoted F-score$_{pt}$, see section 4.1 for further methodological considerations). For both datasets Sys1 and Sys2 perform better than the Benchmark albeit in small proportions in some cases. The Benchmark was chosen in order to have a fair point of comparison to evaluate our approaches: it is a recent contribution that rates among the best in the latest MIREX evaluation (2010).

Other examples using the same datasets can be found in the literature: Using CAL500, Turnbull et al. [11], Hoffman et al. [4] and Mahieux et al. [2] obtain F-scores$_{pt}$ equal to 0.20, 0.21 and 0.14 respectively but the evaluation is based on a ranking of the first 10 most probable tags and thus not comparable with our results. Seyerlehner et al. [9] obtains F-score$_g = 0.50$ and F-score$_{pt} = 0.30$ on CAL500 and 0.42 | 0.22 with the Magnatagatune dataset thus slightly above our results. Zhao et al. [12] achieve F-score$_{pt} = 0.31$ on CAL500 but tags that were not recognized in the dataset were ignored in the evaluation (using this metric we were able to achieve F-score$_{pt} = 0.33$ using Sys2). Similarly, Miotto et al. [6] obtain a F-Score$_{pt} = 0.30$ on CAL500 but less frequent tags were removed which, as we will see in the next section, affects significantly the results. To summarize, we claim that the approaches presented in this paper are on a par with the state-of-the-art as described in the recent literature.

[1] The authors are grateful to Ness & Tzakenakis for kindly providing and commenting the code used for these experiments.

4. ISSUE 1: METHODOLOGICAL ISSUES IN EVALUATING AUTOTAGGING SYSTEMS

4.1 On evaluation measures

Evaluation for autotagging systems is mostly based on Information Retrieval measures, such as accuracy, precision, recall, F-score, etc. These measures are generally computed on a per-tag basis, separately for each tag and then averaged across tags, or globally across the whole dataset. Music datasets typically have a strong imbalance in tag distributions, and results on a per-tag or global basis can differ significantly. This imbalance drives global scores artificially high. The reason is simple: since the most common tags account for a large percentage of all annotations, classifiers that predict these tags well start off with high global scores. Figure 2 shows the F-scores on CAL500 for Sys1 and Sys2, when the most frequent tags (left) or the least frequent tags (right) are removed from the dataset (tests with Magtag5k had a similar outcome). Results confirm the dependence of global scores on the most common tags [2, 6, 11]: the left plot shows a sharp decrease in F-scores$_g$ when the top tags are removed (F-scores$_{pt}$ also decrease, albeit relatively less). This indicates that the most frequent tags are on average better classified and have a substantial effect on the overall performance. This is also seen in figure 6, where the most common tags (the ones represented by larger circles) have high scores, and the least frequent tags low scores. On the other hand, figure 2 (right plot) also shows that the least frequent tags, with lower classification rates, have little impact on global scores but have a dramatic effect on per-tag scores, a fact that is most often ignored. We therefore stress the importance when reporting results on reference data to include both global and per-tag metrics, and to consider the influence of both the least and most frequent tags. For instance, in [6] the evaluation is obtained excluding the 77 least frequent tags, which in our systems would result in a increase in the F-scores$_{pt}$ above 10%.

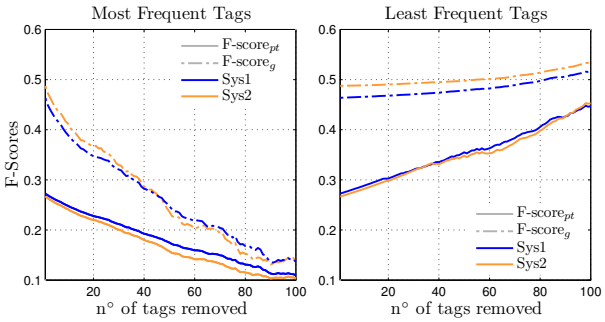

Figure 2. F-score$_g$ and F-score$_{pt}$ on CAL500 for Sys1 and Sys2 autotaggers, as the most frequent (left) or the least frequent tags (right) are removed.

Another important factor that can influence performance scores is how thoroughly the songs in the dataset are annotated. CAL500 has a high number of tags per song (an average of 26 tags per song): a trivial classifier (i.e. always predicting all tags) has a precision of $\approx 15\%$ (with 100% recall). This "starting point" yields a F-score$_{pt}$ of 26%, which is misleadingly high, and almost on a par with other results reported in the literature (see the F-scores$_{pt}$ reported in section 3.3). Note that in this case the F-score$_g$ equals F-score$_{pt}$ and is much lower than what is reported in the literature, hence a good indicator of the system's sub-optimal performance.

The choice of evaluation measure can hinder comparisons between different methods and can also conceal sub-optimal performances. It is therefore important to report both per-tag and global scores, and ideally, also document how the individual tag performances are related to the a priori tag frequencies in the datasets used.

4.2 On data and evaluation methodology

Depending on the data gathering method, tag-annotated datasets can present several problems [11] such as misspelling, impossible combinations of values, diverse types of noise, etc. However, only few papers consider these potential problems when reporting on autotagging experiments with the CAL500 or Magnatagatune datasets.

The Magnatagatune dataset reveals a significant number of problems with annotation: (1) **synonymy**: we merged a number of tags (e.g. "clasical", "classical" and "classic"), (2) **trivial cases**: we removed excerpts with tags such as e.g. "silence", (3) **antonymy**: we removed tag attributions of an excerpt when they were not compatible (e.g. having both "drums" and "no-drums" tags, or "fast" and "slow"), (4) **extreme sparseness**: we removed excerpts with no tags, and (5) **duplication**: many excerpts in the Magnatagatune dataset are segments of the same original piece and have different tag annotations, we kept those segments with the maximum number of tags and removed the other segments. After pre-processing the Magnatagatune dataset as detailed above, the remaining data, referred henceforth as *Magtag5k*, consists of 137 tags, 5259 excerpts from 230 artists. CAL500 did not require such pre-processing.

To avoid overfitting the data in building autotagging models, the literature fosters a number of evaluation methodologies, e.g. holdout validation, S-fold cross-validation, etc. However, it seldom takes into account artist filtering in the definition of the training and test datasets, a method whose importance has been demonstrated in music similarity research [3] (over-optimistic results can be achieved when the same artists are present in both sets). Taking this additional factor into account, the evaluation methodology should agree with a number of constraints related to the statistics of the data, i.e. the number of folds should not be higher than the

number of artists per tag, nor than the number of excerpts per tag. For instance, constraints from CAL500 favors a 2-fold cross-validation or holdout validation (instead of 10-fold cross-validation [11]). We report results with the latter (with a 50% split). In Magtag5k, some tags have few instances, from few artists (e.g. tag "water" has 16 songs from 6 artists). Thus, we chose to set the maximum number of folds to 3 (ensuring at least 2 different artists per tag per fold) and report on results with 3-fold cross-validation. We can clearly see in table 2 that very different results are obtained when considering data and methodology issues discussed here and when not. To facilitate reproducible research, the whole Magtag5k data pre-processing and resulting data are available [2] .

5. ISSUE 2: WHAT ARE WE REALLY LEARNING?

In this section we present results of a set of experiments that were conducted in order to evaluate the extent of the results obtained with the various systems. The objective was to evaluate models' ability to generalize when used with data from different origins. We selected songs annotated with 35 tags common to both Magtag5k and CAL500. [3] Figure 3 shows for both the Benchmark (left) and Sys2 (right) two F-scores for each of the 35 tags, these F-scores are obtained with Magtag5k as *test* set, but with two different *training* sets for building models, either Magtag5k or CAL500. [4] The F-score obtained with CAL500 is shown on the horizontal axis while the F-score obtained with is Magtag5k shown on the vertical axis. On these plots a model that perform equally well when trained with either datasets would be on the diagonal, those performing worse when trained with CAL500 data are above the diagonal.

When comparing performance obtained on the same test set (Magtag5k) we observe much lower performance for models based on CAL500 training than for those trained with Magtag5k. This observation is valid for the Benchmark, Sys1 (not shown here) and Sys2. Nearly every point is above the diagonal. Sys2 seems to perform slightly better than other systems in terms of generalization but still the performance is much lower for models trained with CAL500: only three tags obtain a relatively high F-score for both training sets (*man.singing*, *electro*, and *female.singing*).

Models were also tested on the LMD and the ISMIR04 genre classification datasets. These two datasets were not created for autotagging tasks therefore no ground truth is available so our analysis is based on tag assignment frequency. We processed the music pieces from these datasets with Sys1 models trained with CAL500 and Magtag5k. Figure 4 shows the proportion of songs from a given test set to which each tag was assigned. Each color/shade corresponds to a training set and each row to a test set. We can see for example that when testing with CAL500 (first row) and training with Magtag5k (orange, light shade) nine tags are assigned to all songs. When testing with Magtag5k (second row), models trained with CAL500 (blue, dark shade) recognize very few tags. When testing on LMD and ISMIR04 we observe a strange phenomenon: the proportion of music per tag is almost the same for both datasets and for all tags. This indicates a strong bias on the models side and a weak power of generalization.

Figure 5 shows the proportion of music pieces for which each tag was selected when trained with Magtag5k and tested with both LMD and ISMIR04 datasets (different colors) for two modeling techniques (different rows). The first row confirms what was seen on figure 4: with Sys1 the proportion of songs per tag is almost the same independently of the test set. When Sys2 is used, a different anomaly is observed: very few tags are recognized and these tags are over-represented. Moreover the same tags seem to be over-

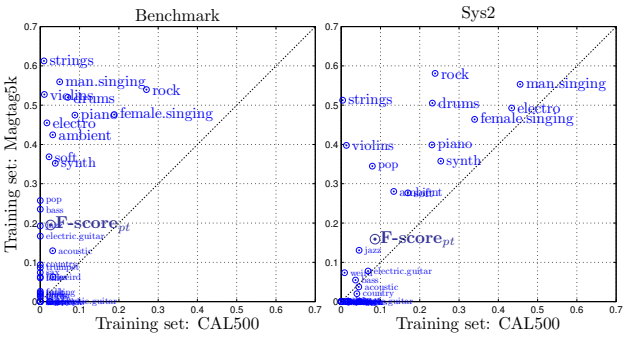

Figure 3. F-score on Magtag5k *test* set for Sys2 (right) and Benchmark (left) autotaggers, either *trained* with CAL500 (x-axis) or Magtag5k (y-axis).

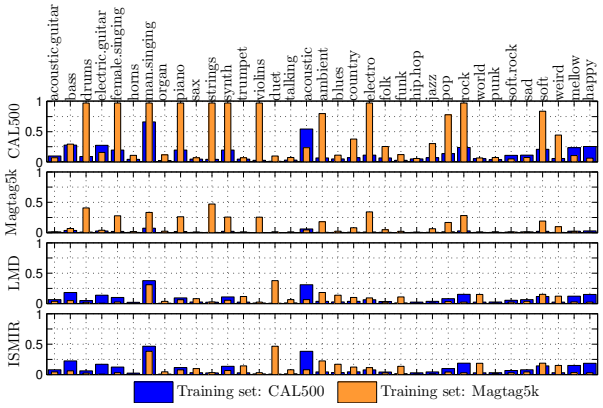

Figure 4. Proportion of music pieces for which each tag was assigned in the corresponding test set (rows). Sys1.

[2] Please follow this link: `http://tl.di.fc.ul.pt/t/magtag5k.zip`.

[3] Hence reducing Magtag5k to 4549 songs.

[4] Note that artist filtering and non-overlap of training and test data are observed for Magtag5k.

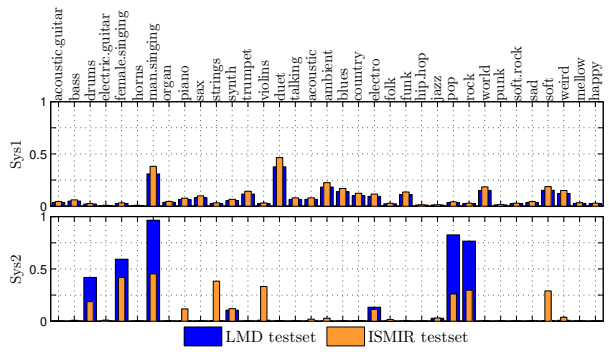

Figure 5. Proportion of music pieces for which each tag was assigned for two kinds of models (rows) and two test sets (colors).

represented in both datasets. When comparing the two rows of the plot, we can see that the two autotagging techniques have a very low level of agreement, for both test sets.

These experiments show that models obtained with autotagging techniques at the level of the state-of-the-art show very limited ability to generalize to new datasets and that the level of performance observed on a single finite dataset is somewhat misleading. Current autotagging techniques are still far from the long-term goal that is to allow automatic tagging of sounds independently of their origin.

6. ISSUE 3: EXPLOITING TAG CORRELATIONS IN A SECOND PROCESSING STAGE

	Magtag5k	2-fold
Bench. stage 1	0.409|0.164	0.342|0.126
Bench. both stages	0.312|0.083	0.347|0.136
Sys1 stage 1	0.411|0.165	0.341|0.127
Sys1 both stages	0.423|0.176	0.347|0.136

Table 2. Comparison of F-score$_g$|F-score$_{pt}$ for different configurations of the Magnatagatune dataset: Magtag5k, and 2-fold cross-validation over unprocessed Magnatagatune dataset (no artist filter).

In table 2, we compare Sys1 against the Benchmark, considering either stage 1 only or both stages. The first column reports results on Magtag5k while the second reports results with the data and evaluation methodology from [7]: 2-fold over the whole Magnatagatune data, without artist filtering. Looking at results for the Benchmark, we can see that although results of the first stage (first row, second column) are very similar to those published in [7], the second stage in fact impairs results from the first stage only, i.e. the opposite phenomenon than [7]. Similar improvements for the second stage as those published can only be found when considering unadapted evaluation methodologies (e.g. no artist filter)

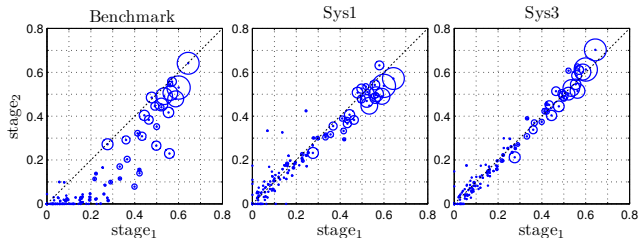

Figure 6. Performance of stage 1 vs both stages, Magtag5k. Individual tag F-scores are represented by circle centers. x-axis are the stage 1 F-scores, and y-axis both stages. Radius are proportional to corresponding tag frequency.

and noisy (see problems 1 to 4 in section 4.2) and redundant data (see problem 5), as illustrated in the second column.

Results also show that the second stage of Sys1 does appear to bring a small improvement on the first stage. However, we can gain more insights on the actual effect of the second stage by looking at figure 6 which illustrates the difference in tag's individual F-scores between using only one stage of processing vs using both stages. For a given data point (i.e. a particular tag) to lie above the diagonal means that the second stage improves results, while below the diagonal means impairing results from stage 1. For the Benchmark (left plot), the decrease in overall performance can be seen on almost all tags individually. For Sys1 (middle plot), if average results are better with both stages, we can see that not all tags are affected in the same way by the second stage: some improve (are above the diagonal) while others do not. In our opinion, this distribution around both sides of the diagonal indicates that no clear pattern of improvement can be identified with the 2-stage procedure.

A possible reason for the inability of the system to take advantage of existing tag correlations may reside in the nature of the second stage classifier. Hence we experimented a different option for the second stage: a pool of binary SVMs (one per tag) [8]. These experiments are restricted to the particular task of tag co-occurrence modeling, i.e. we compare classifiers that process *correct* input (we are *not* evaluating the full system here, only what can serve as its second stage). Results show that binary SVMs are clearly better at the task than a multiclass SVM: in three-fold cross-validation on Magtag5k the former reaches a F-score$_g$ and F-score$_{pt}$ of 0.839 and 0.822 respectively while the latter reaches 0.581 and 0.567. A corollary of the above is that the second stage may fail precisely because it is trained on data that only represents *estimations* of these correlations (and relatively bad ones, as indicated by the performance of stage 1). Hence we modified Sys1 with binary SVMs in stage 2, trained with *true* tag annotations instead of probability estimations from stage 1. We refer to this system as **Sys3**. Overall, Sys3 reaches F-score$_g$ and F-score$_{pt}$ of 0.411

and 0.162, therefore slightly below the performance of Sys1 and comparable to using only stage 1 (see table 2). However, when looking at the case of individual tags, i.e. rightmost plot of figure 6, we can spot an interesting pattern: improvements with stage 2 seem higher for tags with better performance in stage 1. In other words, this seems to indicate that a minimum performance in stage 1 should be expected for a given tag —i.e. for its probability estimation— to be useful in a second stage. Although proving this claim will require more data, we wish to argue here that this pattern appears as a logical and desirable property for an autotagging system, and it indicates clear directions for future work: e.g. improving stage 1; tailoring stage 2 classifier to a selection of particular tags (e.g. the most reliable, the most "influential" [1]) instead of processing all tags the same way.

7. DISCUSSION

The experiments described in this paper show that diverse techniques on a par with the state-of-the art in music autotagging fail to achieve their goal in several aspects. It was shown that autotagging tasks must be evaluated more carefully than what is usually done, that changing the set of tags or altering the evaluation measure (per tag vs global F-score) may dramatically alter the results, sometimes hiding weaknesses. It was also shown that current techniques used for autotagging fail the generalization test. Finally it was shown that the performance achieved with these techniques is not sufficient to be able to take advantage of the correlations between tags. Research in music genre classification and music similarity has seen recent progresses but its adaptation to autotagging shows severe drawbacks. What are the causes of these relatively negative results?

It is our opinion that some key differences between autotagging and genre classification should be given more emphasis in autotagging research. In particular with regards to data recollection and annotation [10]. Tags can correspond to music facets more subjective than music genre. Or they can have multiple meanings, as in the case of Instrument tags: a song tagged "piano" can mean e.g. that piano is salient all over the song, or that there is a piano accompanying (but it may be relatively quiet), or that some parts have piano (but may have a short temporal span). In autotagging the procedure used to obtain ground truth differs from one dataset to another, which results in a lack of consistency. Public datasets are limited in quantity and in many cases present errors or incompleteness. Also, where datasets for genre classification are usually limited to 10-20 genres, it is common to deal with hundreds of tags. This is not a problem per-se but in these conditions it is much more difficult to achieve good results for every tags and to follow good practices (artist filtering, S-fold cross validation). It is hard to build models based on extremely unbalanced data but it is even harder if the ground truth lacks consistency. Future work will include seeking for improvements in terms of generalization using recently published datasets like the Million Songs or CAL10k datasets.

This paper's results and previous observations lead us to propose some directions regarding future work in music autotagging: Different processing could be applied depending on categories of tags: (1) 2-stage architectures may be beneficial for some tags (e.g. tags with reasonable performance might help build models for other tags) but not for others (discussion in [1] is also insightful on this matter). (2) Tag models could be differentiated according to temporal characteristics: models for tags that correspond to a short time span should be based on local features whereas tags that correspond to whole songs should use global features.

8. ACKNOWLEDGMENTS

Thanks to Alessandro Koerich, Luiz Oliveira and Alceu Brito in Curitiba, this research was supported by FCT and QREN-AdI grant for the project Palco3.0/3121, by FCT through LASIGE Multiannual Funding and VIRUS research project (PTDC/EIAEIA/101012/2008). The first author is supported by PROTEC grant SFRH/BD/50118/2009.

9. REFERENCES

[1] J.-J. Aucouturier. *Language, Evolution and the Brain, Frontiers in Linguistics Series*, chapter Sounds like Teen Spirit: Computational Insights into the Grounding of Everyday Musical Terms. Academia Sinica Press, 2009.

[2] T. Bertin-Mahieux, D. Eck, F. Maillet, and P. Lamere. Autotagger: A model for predicting social tags from acoustic features on large music databases. *JNMR*, 37(2), 2008.

[3] A. Flexer. A closer look on artists filters for musical genre classification. In *ISMIR*, 2007.

[4] M. Hoffman, D. Blei, and P. Cook. Easy as CBA: A simple probabilistic model for tagging music. In *ISMIR*, 2009.

[5] T. Langlois and G. Marques. A music classification method based on timbral features. In *ISMIR*, 2009.

[6] R. Miotto, L. Barrington, and G. Lanckriet. Improving auto-tagging by modeling semantic co-occurrences. In *ISMIR*, 2010.

[7] S. Ness, A. Theocharis, G. Tzanetakis, and L. Martins. Improving automatic music tag annotation using stacked generalization of probabilistic SVM outputs. In *ACM Multimedia*, 2009.

[8] F. Pachet and P. Roy. Improving multilabel analysis of music titles: A large-scale validation of the correction approach. *IEEE TASLP*, 17(2):335 –343, 2009.

[9] K. Seyerlehner, G. Widmer, M. Schedl, and P. Knees. Automatic music tag classification based on block-level features. In *SMC Conference*, 2010.

[10] D. Tingle, Y. E. Kim, and D. Turnbull. Exploring automatic music annotation with "acoustically-objective" tags. In *ACM Int. Conf. on Multimedia Information Retrieval*, 2010.

[11] D. Turnbull, Barrington. L., D. Torres, and G. Lanckriet. Semantic annotation and retrieval of music and sound effects. *IEEE TASLP*, 16(2):467–476, 2008.

[12] Z. Zhao, X. Wang, Q. Xiang, A. M. Sarroff, Z. Li, and Y. Wang. Large-scale music tag recommendation with explicit multiple attributes. In *ACM Multimedia*, 2010.