Classification Accuracy Is Not Enough

On the Evaluation of Music Genre Recognition Systems

Sturm, Bob L.

Published in:
Journal of Intelligent Information Systems

DOI (link to publication from Publisher):
10.1007/s10844-013-0250-y

Publication date:
2013

Document Version
Early version, also known as pre-print

Link to publication from Aalborg University

Citation for published version (APA):
Sturm, B. L. (2013). Classification Accuracy Is Not Enough: On the Evaluation of Music Genre Recognition Systems. *Journal of Intelligent Information Systems*, *41*(3), 371-406. https://doi.org/10.1007/s10844-013-0250-y

Journal of Intelligent Information Systems manuscript No.
(will be inserted by the editor)

Classification Accuracy Is Not Enough

On the Analysis of Music Genre Recognition Systems

Bob L. Sturm

Received: date / Accepted: date

Abstract A recent review of the research literature evaluating music genre recognition (MGR) systems over the past two decades shows that most works (81%) measure the capacity of a system to recognize genre by its classification accuracy. We show here, by implementing and testing three categorically different state-of-the-art MGR systems, that classification accuracy does not necessarily reflect the capacity of a system to recognize genre in musical signals. We argue that a more comprehensive analysis of behavior at the level of the music is needed to address the problem of MGR, and that measuring classification accuracy obscures the aim of MGR: to select labels indistinguishable from those a person would choose.

Keywords Music · Genre · Classification · Evaluation

1 Introduction

For over fifty years, research in information technology has advanced the field of machine learning to reach almost human level performance in discriminating and categorizing the content of text, images, sounds, movies, and other media. For music in particular, the problem of identifying, discriminating between, and learning the criteria of music genres or styles — music genre recognition (MGR) — has motivated much work over the past 28 years [83]. Indeed, a recent review of MGR [34] writes, "Genre classification is the most widely studied area in MIR." There are a few reviews of the variety of features and approaches to MGR by machine listening [7, 34, 73]. MGR research is also making its appearance in textbooks [48].

Most published studies of MGR systems report classification performance significantly better than chance, and sometimes as well as or better than humans. For a

BLS is supported in part by Independent Postdoc Grant 11-105218 from Det Frie Forskningsråd; and the Danish Council for Strategic Research of the Danish Agency for Science Technology and Innovation in the CoSound project, case number 11-115328.

B. L. Sturm
Audio Analysis Lab, Department of Architecture, Design and Media Technology
Aalborg University Copenhagen, A.C. Meyers Vænge 15, DK-2450
Copenhagen SV, Denmark
E-mail: bst@create.aau.dk Tel.: +45-9940-7633

benchmark dataset of music excerpts singly-labeled in ten genres (GTZAN [82,89]), classification accuracies are now reported above 90%, e.g., [19,38,65–67]. Indeed, as [14] writes, "Given the steady and significant improvement in [genre] classification performance since 1997, we wonder if automatic methods are not already more efficient at learning genres than some people." This increase in performance not only merits a closer look at what works so well in these particular systems, but also motivates a re-evaluation of the argument that music genre exists to a large extent outside of the acoustic signal itself [28,58,93]. It might also, most excitingly, reveal fundamental aspects of how people hear and conceptualize the complex and mysterious phenomenon of "music." We might be getting ahead of ourselves, however.

The work in [85] casts doubt on the high classification accuracies reported in [65–67] — results that actually stem from a flaw in the simulations (private correspondence with Y. Panagakis). Another work [83] provides a comprehensive review of the approaches so far used for *evaluating* MGR systems. We see that over 92% of 375 papers approach evaluation of MGR systems by classifying several music excerpts and comparing the labels to the "true" ones. Nearly all of this work (334 papers) uses the classification accuracy as a figure of merit. Also shown is that the most used publicly available benchmark dataset is GTZAN — a dataset that has integrity problems for genre recognition [82]. And the work in [84] shows that, even with high classification accuracy, an MGR system can act as if music genre is not what it is recognizing. Thus, the advances we see in MGR might be misleading: a system with high classification accuracy might not be addressing the problem at all.

In this paper, we show that classification accuracy does not reliably reflect the capacity of an MGR system to recognize music genre. Indeed, recall, precision and confusion tables are still not enough. We claim that these figures of merit — which have been used in the past decade to rank MGR systems, e.g., [7,11,15,17,18,25, 29,34,67,71,88,89] citing one publication from each year since 2001 — do not reliably rank MGR systems. While this claim has not been made in any work surveyed in [83], shades of it appear in [22,23,52,77,84,93]. Those works argue for measuring performance in ways that take into account the ambiguity of genre being in part a cultural and subjective construction. We, however, argue that the evaluation of MGR systems — the experimental designs, the datasets, and the figures of merit — and indeed, the development of future systems, must embrace the fact that the problem of recognizing genre is a *musical* one, and must be evaluated as such. In short, *classification accuracy is not enough* to gauge the success of any MGR system.

In the next section, we distill the variety of MGR evaluation approaches used over the past two decades along three dimensions: experimental design, datasets, and figures of merit. This shows how most work reports classification accuracy of supervised approaches to machine learning using private datasets. The third section reviews three state-of-the-art MGR systems that show high classification accuracy in the most-used publicly-available music genre dataset GTZAN. In the fourth section, we analyze the behaviors of these three systems, from the high-level figures of merit classification accuracy, recall and precision, to mid-level class confusions, and finally to low-level excerpt misclassifications. At this lowest level, we show the pathological misclassifications of these systems argue against the claim that any of them have a capacity to discriminate between and recognize genre based upon musicological principles.

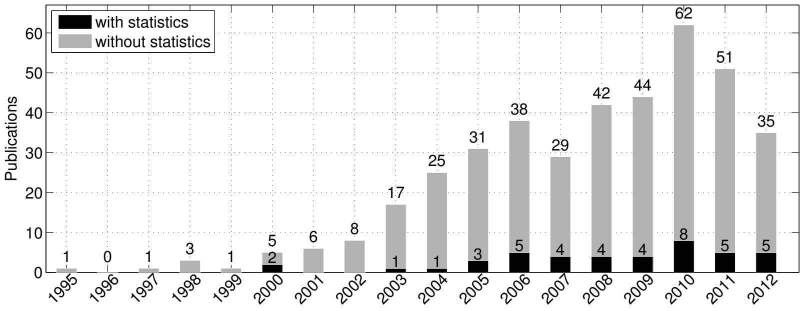


Fig. 1 Annual numbers of publications in MGR separated by which use any form of statistical testing for making comparisons [83]. Overall, about 12% of the MGR literature uses a statistical test

2 Evaluation in Music Genre Recognition Research

Over the past 23 years of MGR research, surprisingly little has been written about evaluation, i.e., experimental design, data, and figures of merit. An experimental design is a method for testing a hypothesis. Data is the material on which a system is tested. A figure of merit describes the confidence in the hypothesis after conducting an experiment. Of three review articles devoted in large part to MGR [7,34,73], only [7] contains a brief paragraph on evaluation. The work in [92] provides a comparison of various figures of merit for music classification. Other works [12,22,23,52,77,93] argue for measuring performance in ways that take into account the natural ambiguity of music genre and similarity. The work in [22,23,84] argues for richer experimental designs than having a system apply a single label to music with a possibly problematic "ground truth." And Flexer [29] notes and criticizes the absence of formal statistical testing in music information retrieval research, and provides an excellent tutorial based upon MGR for how to apply statistical tests. The review in [83] compiles a near-complete bibliography of MGR (surveying over 400 published works), and focuses specifically upon MGR evaluation. Derived from this review, Fig. 1 shows the annual number of publications concerning MGR, and that formal statistical testing in comparing MGR systems remains absent [29].

Table 1 summarizes the ten experimental designs in the MGR literature, all of which address shades of the hypothesis, "system A recognizes genre X." (Some

Table 1 Experimental designs of the music genre recognition literature [83]

Design	**Description**	**% Work**
Classify	system classifies music; researcher compares against "ground truth"	92
Generalize	Classify with two or more datasets, and/or various amounts of training data	20
Features	system ranks and/or selects features; researcher inspects features	18
Cluster	system creates clusters or trees of dataset; researcher inspects these	6
Eyeball	system derives parameters from music; researcher visually compares	3
Robust	system classifies music that researcher modifies or transforms in ways that do not harm its genre identification by a human	3
Scale	Classify with varying numbers of genres	3
Retrieve	system retrieves music similar to query; researcher compares against query	2
Rules	researcher inspects rules used by a system to identify genres	1
Compose	system creates music in specific genres; researcher analyzes representativeness	0.4

Table 2 Datasets of the music genre recognition literature [83]

Dataset	Description	% Work
GTZAN [89]	Audio (`http://marsyas.info/download/data_sets`)	23
ISMIR2004 [41]	Audio (`http://ismir2004.ismir.net/genre_contest/index.htm#genre`)	16
Latin [79]	Features (`http://www.ppgia.pucpr.br/~silla/lmd/`)	4
Homburg [40]	Audio (`http://www-ai.cs.uni-dortmund.de/audio.html`)	3
Bodhidharma [57]	Symbolic (`http://jmir.sourceforge.net/Codaich.html`)	2
RWC [37]	Audio (`http://staff.aist.go.jp/m.goto/RWC-MDB/`)	1
SLAC [59]	Audio and Symbolic (`http://jmir.sourceforge.net/Codaich.html`)	1
USPOP2002 [12]	Audio (`http://labrosa.ee.columbia.edu/projects/musicsim/uspop2002.html`)	1
SOMeJB [49]	Features (`http://www.ifs.tuwien.ac.at/~andi/somejb/experiments/`)	0.9
1517-artists [76]	Audio (`http://www.seyerlehner.info/index.php?p=1_3_Download`)	0.7
Million Song [16,75]	Features (`http://labrosa.ee.columbia.edu/millionsong/`)	0.7
ISMIS2011 [42]	Features (`http://tunedit.org/challenge/music-retrieval/genres`)	0.4
Private	Datasets that are not publicly available	52

works use more than one experimental design.) Here we see that the most widely used design by far, Classify, is that of comparing to a "ground truth" the class(es) selected by a system for particular instances of music. The next most-used experimental design is Generalize. The least-used experimental design, appearing in only two papers [24,84], is having a system compose music that is exemplary of the genres in which it is trained, and testing the representativeness.

Table 2 shows the most used datasets. Overall, 78% of the papers use audio data or features derived from audio data, and 14% use symbolic data. About 20% of work tests MGR systems with two or more datasets (which is the experimental design Generalize). Only 9% of work makes use of an artist or album filter [30,31,43,63]. While more than 50% of the papers use datasets that are not publicly available, the most used public dataset is GTZAN [89] — which has recently been formally shown to have replicas, mislabelings, and distortions [82].

Table 3 shows the figures of merit appearing most in the MGR literature. Consider a single-label classifier trained on M classes, and define the $M \times M$ confusion matrix $\mathbf{Y}$ produced from N observations. Its ijth element $\mathbf{Y}_{ij}$ is the number of elements with true label i assigned label j by the system. For a multilabel system [54], define $\mathscr{L}$ as the set of all possible labels, and so the ith element of N observations has labels

Table 3 Figures of merit of the music genre recognition literature [83]. For a single-label system of M classes, $\mathbf{Y}$ is the $M \times M$ confusion matrix, and N is the number of observations. For a multilabel system, $\mathscr{X}_n$ is the set of true labels of the nth observation, and $\mathscr{Y}_n$ is the set of applied labels

Dataset	Description	% Work
Accuracy	$A = \text{trace}(\mathbf{Y})/N$	81
Confusion table	$\mathbf{Y}$	31
Recall	single-label: $R(i) = \mathbf{Y}_{ii}/\sum_{j=1}^{M} \mathbf{Y}_{ij}$; multilabel: $R = \frac{1}{N}\sum_{n=1}^{N} \lvert \mathscr{Y}_n \cap \mathscr{X}_n \rvert / \lvert \mathscr{X}_n \rvert$	26
Precision	single-label: $P(i) = \mathbf{Y}_{ii}/\sum_{j=1}^{M} \mathbf{Y}_{ji}$; multilabel: $\frac{1}{N}\sum_{n=1}^{N} \lvert \mathscr{Y}_n \cap \mathscr{X}_n \rvert / \lvert \mathscr{Y}_n \rvert$	7
F-measure	single-label: $F(i) = 2R(i)P(i)/(R(i)+P(i))$; multilabel: $F = \frac{1}{N}\sum_{n=1}^{N} 2\lvert \mathscr{Y}_n \cap \mathscr{X}_n \rvert / (\lvert \mathscr{Y}_n \rvert + \lvert \mathscr{X}_n \rvert)$	3

$\mathscr{Y}_i \subseteq \mathscr{L}$, whereas the system applies $\mathscr{X}_i \subseteq \mathscr{L}$. When accuracy appears as a figure of merit, only 22% of the time is it accompanied by variance, standard deviation, or the standard error of the mean. When a confusion table appears as a figure of merit, about 60% of the time it is not accompanied by any kind of musicological reflection.

3 Three State-of-the-art Systems for Music Genre Recognition

We now present three MGR systems. Two of these (AdaBFFs and SRCAM) are used in [84]; but we adjust each one here. We also introduce a new approach (MAPsCAT).

3.1 AdaBFFs

AdaBoost with decision trees and bags of frames of features (AdaBFFs) [14, 84], combines weak classifiers trained by multiclass AdaBoost [32,74] on bags of frames of features. This approach performed the best in the 2005 MIREX music genre classification task [62]. Multiclass AdaBoost [32,74] creates a strong classifier by counting "votes" cast by weak classifiers given an observation $\mathbf{x}$. Its use for MGR is detailed in [14, 84]. Given the labeled features in a training set, iteration l adds a new weak classifier $\mathbf{v}_l(\mathbf{x})$ and weight $w_l \in [0,1]$ to minimize the total prediction error. The weak classifier $\mathbf{v}_l(\mathbf{x})$ produces a length-K vector with elements in $\{\pm w_l\}$. A positive element means it favors a class, whereas a negative means the opposite. After L training steps, our classifier produces the vote vector $\mathbf{f}(\mathbf{x}) \in [-1,1]^K$

$$\mathbf{f}(\mathbf{x}) := \frac{\sum_{l=1}^{L} w_l \mathbf{v}_l(\mathbf{x})}{\sum_{l=1}^{L} w_l}. \tag{1}$$

For an excerpt of recorded music consisting of a set of features $\mathscr{X} := \{\mathbf{x}_i\}$, we pick the class associated with the maximum element in the sum of weighted votes:

$$f_k(\mathscr{X}) := \sum_{i=1}^{|\mathscr{X}|} [\mathbf{f}(\mathbf{x}_i)]_k. \tag{2}$$

We use the "multiboost package" [10] with decision trees as the weak learners, AdaBoost.MH [74] as the strong learner. The features we use are computed using a sliding Hann window of 46.4 ms and 50% overlap: 40 Mel-frequency cepstral coefficients (MFCCs) [80], zero crossings, mean and variance of the magnitude Fourier transform, 16 quantiles of the magnitude Fourier transform, and the error of a 32-order linear predictor. We disjointly partition the set of features into groups of 130 consecutive frames, and then compute for each the means and variances of each dimension. For a 30-s music excerpt, this produces 9 feature vectors of 120 dimensions.

3.2 SRCAM

Sparse representation classification with auditory temporal modulations (SRCAM) [67, 84, 85], uses sparse representation classification of long-duration auditory features. This approach is reported to have mean accuracies above 90% [65–67], but those results arise from a flaw in the experiment (private correspondence with Y. Panagakis). Here, as in [84], we modify the approach to produce classification accuracies above 80%. Each feature comes from a modulation analysis of a time-frequency representation, and for a 30-s sound excerpt with sampling rate 22,050 Hz, the feature dimensionality is 768. One can create dictionary atoms by normalizing each feature (mapping all values in each dimension to $[0,1]$ by subtracting the minimum value and dividing by the largest difference). One can also *standardize* them, i.e., making all dimensions have zero mean and unit variance.

Given a matrix of "feature atoms" $\mathbf{D} := [\mathbf{d}_1|\mathbf{d}_2|\cdots|\mathbf{d}_N]$, and the set of class identities $\cup_{k=1}^{K} \mathscr{I}_k = \{1,\ldots,N\}$, where $\mathscr{I}_k$ specifies the columns of $\mathbf{D}$ belonging to class k, sparse representation classification (SRC) [94] first finds for a feature vector $\mathbf{x}'$ (which is the feature $\mathbf{x}$ transformed by the same normalization or standardization approach to create the dictionary) a sparse representation $\mathbf{s}$ by

$$\min \|\mathbf{s}\|_1 \text{ subject to } \|\mathbf{x}' - \mathbf{D}\mathbf{s}\|_2^2 \leq \varepsilon^2 \tag{3}$$

for $\varepsilon^2 > 0$. SRC then defines the set of weights $\mathscr{S} := \{\mathbf{s}_k \in \mathbb{R}^N : \forall n \in \mathscr{I}_k([\mathbf{s}_k]_n = a_n), \forall n \notin \mathscr{I}_k([\mathbf{s}_k]_n = 0), k \in \{1,\ldots,K\}\}$, where $a_n = [\mathbf{s}]_n$, the nth row of $\mathbf{s}$. Thus, $\mathbf{s}_k$ are the weights in $\mathbf{s}$ specific to class k. Finally, SRC classifies $\mathbf{x}$ by solving

$$\hat{k}(\mathbf{x}') := \arg \min_{k \in \{1,\ldots,K\}} \|\mathbf{x}' - \mathbf{D}\mathbf{s}_k\|_2^2. \tag{4}$$

We gauge the confidence of SRC by comparing the class-dependent errors. To this end, we define the "confidence" of SRCAM for assigning class k to $\mathbf{x}'$ as

$$C(k|\mathbf{x}) := \frac{\max_{k'} J_{k'} - J_k}{\sum_l [\max_{k'} J_{k'} - J_l]} \tag{5}$$

where $J_k := \|\mathbf{x}' - \mathbf{D}\mathbf{s}_k\|_2^2$. Thus, $C(k|\mathbf{x}') \in [0,1]$ where 1 is certainty.

3.3 MAPsCAT

Maximum a posteriori classification of scattering coefficients (MAPsCAT) uses the novel features proposed in [56]. The use of these features for MGR is first proposed in [4]. We use scattering coefficients within a Bayesian framework, and achieve accuracies on par with those reported in [4], and quite close to those of SRCAM. Bayesian classification seeks to minimize expected risk given the observation $\mathbf{x}$. Assuming the cost of all misclassifications are the same, and that all classes are equally likely, the Bayesian classifier becomes the maximum a posteriori (MAP) classifier [87]:

$$k^* = \arg \max_{k \in \{1,\ldots,K\}} P[\mathbf{x}|k]P(k) \tag{6}$$

where $P[\mathbf{x}|k]$ models the observations for class k, and $P(k)$ is the prior of class k. We assume $P[\mathbf{x}|k] \sim \mathscr{N}(\mu_k, \mathbf{C}_k)$, i.e., the observations from class k are distributed multivariate Gaussian with mean μ_k and covariance $\mathbf{C}_k$. We may also assume every class is distributed with the same covariance, i.e., $P[\mathbf{x}|k] \sim \mathscr{N}(\mu_k, \mathbf{C})$. With several features from a music excerpt $\mathscr{X} := \{\mathbf{x}_i\}$, we assume independence between the features, and pick the class of $\mathscr{X}$ that maximizes the log posterior:

$$p_k(\mathscr{X}) := \log P(k) + \sum_{i=1}^{|\mathscr{X}|} \log P[\mathbf{x}_i|k]. \tag{7}$$

Scattering coefficients are attractive features because they are designed to be invariant to particular transformations, such as translation and rotation [56]. They also preserve distances between stationary processes, and embody both large- and short-scale structures. One computes these features by convolving the modulus of successive wavelet decompositions with the scaling wavelet. We use the scatterbox implementation [5] with a second-order decomposition, filter q-factor of 16, and a maximum scale of 160. For a 30-s sound excerpt with sampling rate 22,050 Hz, this produces 40 feature vectors of dimension 469. We estimate each class mean and covariance using unbiased minimum mean-squared error estimators on the training set.

4 Analyzing the Behaviors of MGR Systems from High to Low Specificities

As seen in Tables 1 and 3, at least 92% of the published evaluations of MGR systems uses Classify as the experimental design, and at least 81% uses accuracy as the figure of merit. With MGR system accuracies reportedly above 80%, and some over 90% [19,38,65–68], it appears that something must be working — but is that "something" genre recognition? In this section, we evaluate each system above using the Classify experimental design, but unlike most work we analyze the behaviors of the three systems down to the music excerpts themselves. We use the GTZAN dataset [82, 89] for three reasons: 1) it is the publicly available dataset most used in MGR research [83]; 2) it is used in the works proposing AdaBFFs [14], SRCAM [67, 84], and the features of MAPsCAT [4]; and 3) because its contents and faults are now known [82], we can address its problems on a case-by-case basis.

We test each system with stratified 10-fold cross-validation (equal priors), but conduct 10 independent trials to measure the variability of results due to random partitioning of the dataset. For each test fold, we test the systems using the same training and testing data. Every music excerpt is thus classified ten times by each system trained on the same data. For AdaBFFs, we run AdaBoost for 4000 iterations, and test both decision trees of 1 node or no node (stumps). For SRCAM, we test both standardized and normalized features, and solve its inequality-constrained optimization problem (3) for $\varepsilon^2 = 0.01$ using SPGL1 [13] with at most 200 iterations. For MAPsCAT, we test systems trained with either class-dependent covariances or total covariance (covariance of the training data).

4.1 Analyzing Classification Accuracy

As discussed in Section 2, classification accuracy appears in 81% of the MGR literature as a figure of merit for the performance of systems in recognizing genres. In their review of several MGR systems, Fu et al. [34] compare performance using only classification accuracy. The work proposing AdaBFFs [14], SRCAM [67], and the features of MAPsCAT [4], present only classification accuracy. Furthermore, Seyerlehner et al. [77] argue that the gap between classification by MGR systems and humans is narrowing based only on classification accuracy.

For each of these systems (reviewed in Section 3), Table 4 shows the mean classification accuracies with 95% confidence intervals, and the p-values of paired t-tests between the two settings of each system. We see we can reject the null hypothesis of differences between mean accuracies being due to chance. The differences in mean accuracies for SRCAM with normalized features and MAPsCAT with total covariance is also statistically significant ($p < 0.001$). The low mean accuracy for

Table 4 Mean accuracies in GTZAN for each system and design specifics

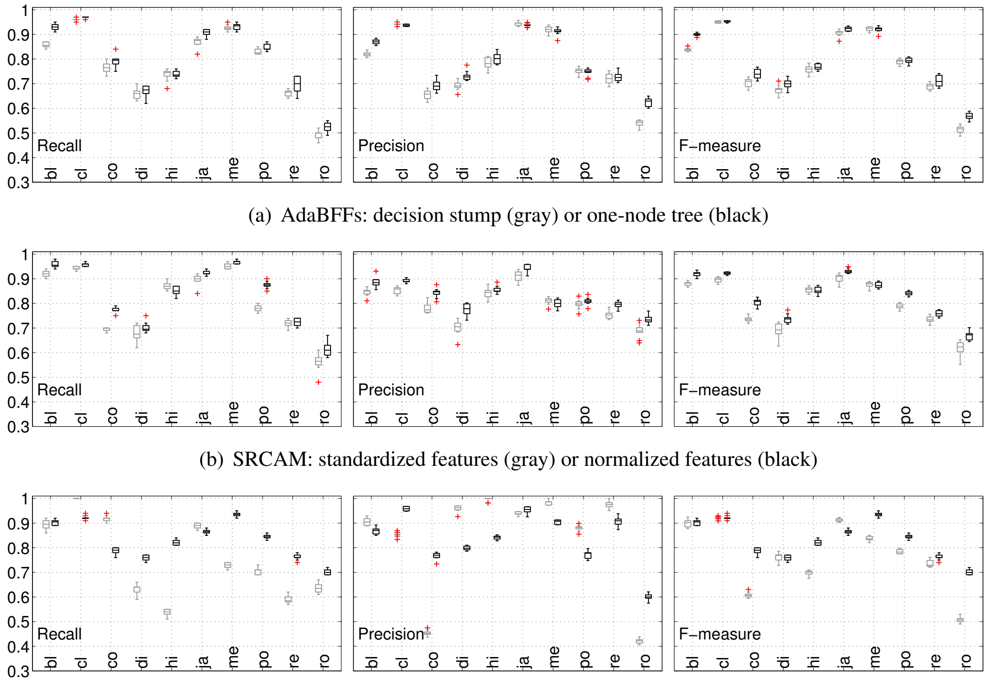
System	System Settings	Mean Accuracy w/ 95%	*p*-value
AdaBFFs	Decision Stump	0.776 ± 0.001	$p < 10^{-7}$
	One-node Tree	0.800 ± 0.002	
SRCAM	Normalized Features	0.835 ± 0.002	$p < 2 \cdot 10^{-7}$
	Standardized Features	0.802 ± 0.002	
MAPsCAT	Class-dependent Covariance	0.754 ± 0.001	$p < 4 \cdot 10^{-11}$
	Total Covariance	0.830 ± 0.001	

MAPsCAT with class-dependent covariance is due to a lack of training data for some classes for estimating covariance matrices from high-dimensional features.

4.2 Analyzing Recall, Precision, and F-measure

The figures of merit recall, precision and the F-measure (see Table 3) are more specific than accuracy, and appear infrequently in the MGR literature. From the observation that their experimental recalls for the Classical- and Rock-labeled excerpts of GTZAN are above that expected from guessing randomly, Wu et al. [95] concludes on the relevance of their features to MGR. With respect to precision, Lin et al. [51] concludes their system is better than another. When it is reported, the F-measure often only accompanies other figures of merit, e.g., recall, precision and accuracy [50].

Figure 2 shows the recalls, precisions, and F-measures for AdaBFFs, SRCAM, and MAPsCAT. We see for the GTZAN Disco excerpts (those excerpts labeled Disco regardless if they are Disco or not) that MAPsCAT using total covariance has the



(a) AdaBFFs: decision stump (gray) or one-node tree (black)

(b) SRCAM: standardized features (gray) or normalized features (black)

(c) MAPsCAT: class-dependent (gray) or total covariance (black)

Fig. 2 Boxplot of recalls (left), precisions (middle), and F-measure (right) from 10 trials of 10-fold stratified cross validation in GTZAN. Classes: Blues (bl), Classical (cl), Country (co), Disco (di), Hip hop (hi), Jazz (ja), Metal (me), Pop (po), Reggae (re), Rock (ro)

highest mean recall (0.76 ± 0.01, standard deviation) of all systems ($p < 2 \cdot 10^{-5}$). Since high recall can come at the price of many false positives, we can look at the precision. Since it shows a high recall but very low precision, we see this is the case for GTZAN Country excerpts for MAPsCAT with class-dependent covariance. However, we see that for GTZAN Disco excerpts, MAPsCAT has the two highest mean precisions: 0.96 ± 0.01 for class-dependent covariance ($p < 3 \cdot 10^{-9}$), and 0.80 ± 0.01 with total covariance ($p < 0.01$). When it comes to GTZAN Classical excerpts, MAPsCAT using class-dependent covariance has perfect recall; and using class-dependent covariance it shows quite high mean precision (0.85 ± 0.01). The F-measure combines recall and precision to reflect class accuracy, where 1 is perfect. We see that AdaBFFs is the most accurate at classifying GTZAN Classical ($p < 8 \cdot 10^{-7}$), and one of the least accurate at classifying GTZAN Disco excerpts.

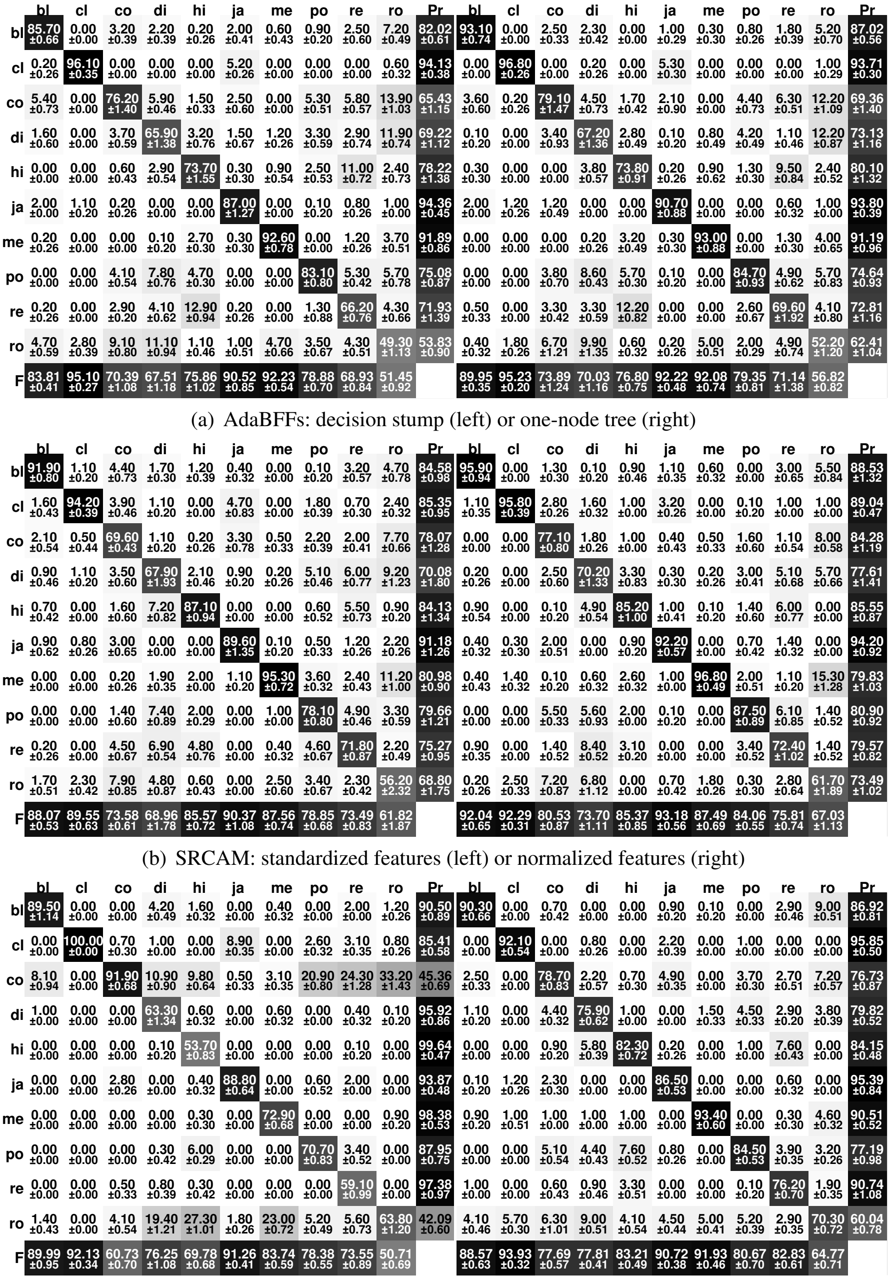
4.3 Analyzing Class-specific Confusions

Confusion tables are reported in 31% of MGR work, of which only 40% discuss them in ways other than repeating what is shown by the table [83]. Sometimes, a confusion table is accompanied by a discussion of how a system appears to perform in ways that makes sense with respect to what experience and musicology say about the variety of influences and commonalities between particular genres, e.g., [1, 2, 21, 39, 40, 70–72, 86, 89–91, 96]. For instance, Tzanetakis and Cook [89] writes that the misclassifications of their system "... are similar to what a human would do. For example, classical music is misclassified as jazz music for pieces with strong rhythm from composers like Leonard Bernstein and George Gershwin. Rock music has the worst classification accuracy and is easily confused with other genres which is expected because of its broad nature." Of their confusion results, Holzapfel and Stylianou [39] writes, "In most cases, misclassifications have musical sense. For example, the genre Rock ... was confused most of the time with Country, while a Disco track is quite possible to be classified as a Pop music piece. ... [The] Rock/Pop genre was mostly misclassified as Metal/Punk. Genres which are assumed to be very different, like Metal and Classic, were never confused."

Figure 3 shows the mean confusions, recalls (diagonal), precisions (right), and F-measures (bottom), all with 95% confidence intervals, for AdaBFFs, SRCAM, and MAPsCAT. We see that all systems confuse the GTZAN Rock excerpts most with other genres: for Country and Disco using AdaBFFs, for Metal using SRCAM, and for Blues using MAPsCAT. It is clear that MAPsCAT with total covariance confuses no pairs of classes over $9 \pm 0.51\%$ than Rock as Blues and Disco as Rock, while the largest confusion for SRCAM with normalized features is $15.3 \pm 1.28\%$ for Rock as Metal, and for AdaBFFs with one-node trees is $12.20 \pm 0.82\%$ for Hip hop as Reggae.

4.4 Analyzing Excerpt-Specific Confusions

Some MGR evaluations describe particular misclassifications, e.g., [26, 45, 47, 73]. Of their experiments, Deshpande et al. [26] writes "... at least in some cases, the classifiers seemed to be making the right mistakes. There was a [classical] song clip that was classified by all classifiers as rock ... When we listened to it, we realized that the clip was the final part of an opera with a significant element of rock in it. As such, even a normal person would also have made such an erroneous classification." Of the confusion table in their review of MGR research, Scaringella et al. [73] finds "... it is



	bl	cl	co	di	hi	ja	me	po	re	ro	Pr	bl	cl	co	di	hi	ja	me	po	re	ro	Pr
bl	85.70 ±0.66	0.00 ±0.00	3.20 ±0.39	2.20 ±0.39	0.20 ±0.26	2.00 ±0.41	0.60 ±0.43	0.90 ±0.20	2.50 ±0.60	7.20 ±0.49	82.02 ±0.61	93.10 ±0.74	0.00 ±0.00	2.50 ±0.33	2.30 ±0.42	0.00 ±0.00	1.00 ±0.29	0.30 ±0.30	0.80 ±0.26	1.80 ±0.39	5.20 ±0.70	87.02 ±0.56
cl	0.20 ±0.26	96.10 ±0.35	0.00 ±0.00	0.00 ±0.00	0.00 ±0.00	5.20 ±0.26	0.00 ±0.00	0.00 ±0.00	0.00 ±0.00	0.60 ±0.32	94.13 ±0.38	0.00 ±0.00	96.80 ±0.26	0.00 ±0.00	0.20 ±0.26	0.00 ±0.00	5.30 ±0.30	0.00 ±0.00	0.00 ±0.00	0.00 ±0.00	1.00 ±0.29	93.71 ±0.30
co	5.40 ±0.73	0.00 ±0.00	76.20 ±1.40	5.90 ±0.46	1.50 ±0.33	2.50 ±0.60	0.00 ±0.00	5.30 ±0.51	5.80 ±0.57	13.90 ±1.03	65.43 ±1.15	3.60 ±0.60	0.20 ±0.26	79.10 ±1.47	4.50 ±0.73	1.70 ±0.42	2.10 ±0.90	0.00 ±0.00	4.40 ±0.73	6.30 ±0.51	12.20 ±1.09	69.36 ±1.40
di	1.60 ±0.60	0.00 ±0.00	3.70 ±0.59	65.90 ±1.38	3.20 ±0.76	1.50 ±0.67	1.20 ±0.26	3.30 ±0.59	2.90 ±0.74	11.90 ±0.74	69.22 ±1.12	0.10 ±0.20	0.00 ±0.00	3.40 ±0.93	67.20 ±1.36	2.80 ±0.49	0.10 ±0.20	0.80 ±0.49	4.20 ±0.49	1.10 ±0.46	12.20 ±0.87	73.13 ±1.16
hi	0.00 ±0.00	0.00 ±0.00	0.60 ±0.43	2.90 ±0.54	73.70 ±1.55	0.30 ±0.30	0.90 ±0.54	2.50 ±0.53	11.00 ±0.72	2.40 ±0.73	78.22 ±1.38	0.30 ±0.30	0.00 ±0.00	0.00 ±0.00	3.80 ±0.57	73.80 ±0.91	0.20 ±0.26	0.90 ±0.62	1.30 ±0.30	9.50 ±0.84	2.40 ±0.52	80.10 ±1.32
ja	2.00 ±0.00	1.10 ±0.20	0.20 ±0.26	0.00 ±0.00	0.00 ±0.00	87.00 ±1.27	0.00 ±0.00	0.10 ±0.20	0.80 ±0.26	1.00 ±0.00	94.36 ±0.45	2.00 ±0.00	1.20 ±0.26	1.20 ±0.49	0.00 ±0.00	0.00 ±0.00	90.70 ±0.88	0.00 ±0.00	0.00 ±0.00	0.60 ±0.32	1.00 ±0.00	93.80 ±0.39
me	0.20 ±0.26	0.00 ±0.00	0.00 ±0.00	0.10 ±0.20	2.70 ±0.30	0.30 ±0.30	92.60 ±0.78	0.00 ±0.00	1.20 ±0.26	3.70 ±0.51	91.89 ±0.86	0.00 ±0.00	0.00 ±0.00	0.00 ±0.00	0.20 ±0.26	3.20 ±0.49	0.30 ±0.30	93.00 ±0.88	0.00 ±0.00	1.30 ±0.30	4.00 ±0.65	91.19 ±0.96
po	0.00 ±0.00	0.00 ±0.00	4.10 ±0.54	7.80 ±0.76	4.70 ±0.30	0.00 ±0.00	0.00 ±0.00	83.10 ±0.80	5.30 ±0.42	5.70 ±0.78	75.08 ±0.87	0.00 ±0.00	0.00 ±0.00	3.80 ±0.70	8.60 ±0.43	5.70 ±0.30	0.10 ±0.20	0.00 ±0.00	84.70 ±0.93	4.90 ±0.62	5.70 ±0.83	74.64 ±0.93
re	0.20 ±0.26	0.00 ±0.00	2.90 ±0.20	4.10 ±0.62	12.90 ±0.94	0.20 ±0.26	0.00 ±0.00	1.30 ±0.88	66.20 ±0.76	4.30 ±0.66	71.93 ±1.39	0.50 ±0.33	0.00 ±0.00	3.30 ±0.42	3.30 ±0.59	12.20 ±0.82	0.00 ±0.00	0.00 ±0.00	2.60 ±0.67	69.60 ±1.92	4.10 ±0.80	72.81 ±1.16
ro	4.70 ±0.59	2.80 ±0.39	9.10 ±0.80	11.10 ±0.94	1.10 ±0.46	1.00 ±0.51	4.70 ±0.66	3.50 ±0.67	4.30 ±0.51	49.30 ±1.13	53.83 ±0.90	0.40 ±0.32	1.80 ±0.26	6.70 ±1.21	9.90 ±1.35	0.60 ±0.32	0.20 ±0.26	5.00 ±0.51	2.00 ±0.29	4.90 ±0.74	52.20 ±1.20	62.41 ±1.04
F	83.81 ±0.41	95.10 ±0.27	70.39 ±1.08	67.51 ±1.18	75.86 ±1.02	90.52 ±0.85	92.23 ±0.54	78.88 ±0.70	68.93 ±0.84	51.45 ±0.92		89.95 ±0.35	95.23 ±0.20	73.89 ±1.24	70.03 ±1.16	76.80 ±0.75	92.22 ±0.48	92.08 ±0.74	79.35 ±0.81	71.14 ±1.38	56.82 ±0.82	

(a) AdaBFFs: decision stump (left) or one-node tree (right)

	bl	cl	co	di	hi	ja	me	po	re	ro	Pr	bl	cl	co	di	hi	ja	me	po	re	ro	Pr
bl	91.90 ±0.80	1.10 ±0.20	4.40 ±0.73	1.70 ±0.30	1.20 ±0.39	0.40 ±0.32	0.00 ±0.00	0.10 ±0.20	3.20 ±0.57	4.70 ±0.78	84.58 ±0.98	95.90 ±0.94	0.00 ±0.00	1.30 ±0.30	0.10 ±0.20	0.90 ±0.46	1.10 ±0.35	0.60 ±0.32	0.00 ±0.00	3.00 ±0.65	5.50 ±0.84	88.53 ±1.32
cl	1.60 ±0.43	94.20 ±0.39	3.90 ±0.46	1.10 ±0.20	0.00 ±0.00	4.70 ±0.83	0.00 ±0.00	1.80 ±0.39	0.70 ±0.30	2.40 ±0.32	85.35 ±0.95	1.10 ±0.35	95.80 ±0.39	2.80 ±0.26	1.60 ±0.32	1.00 ±0.00	3.20 ±0.26	0.00 ±0.00	0.10 ±0.20	1.00 ±0.00	1.00 ±0.00	89.04 ±0.47
co	2.10 ±0.54	0.50 ±0.44	69.60 ±0.43	1.10 ±0.20	0.20 ±0.26	3.30 ±0.78	0.50 ±0.33	2.20 ±0.39	2.00 ±0.41	7.70 ±0.66	78.07 ±1.28	0.00 ±0.00	0.00 ±0.00	77.10 ±0.80	1.80 ±0.26	1.00 ±0.00	0.40 ±0.43	0.50 ±0.33	1.60 ±0.60	1.10 ±0.54	8.00 ±0.58	84.28 ±1.19
di	0.90 ±0.46	1.10 ±0.20	3.50 ±0.60	67.90 ±1.93	2.10 ±0.46	0.90 ±0.20	0.20 ±0.26	5.10 ±0.46	6.00 ±0.77	9.20 ±1.23	70.08 ±1.80	0.20 ±0.26	0.00 ±0.00	2.50 ±0.60	70.20 ±1.33	3.30 ±0.83	0.30 ±0.30	0.20 ±0.26	3.00 ±0.41	5.10 ±0.68	5.70 ±0.66	77.61 ±1.41
hi	0.70 ±0.42	0.00 ±0.00	1.60 ±0.60	7.20 ±0.82	87.10 ±0.94	0.00 ±0.00	0.00 ±0.00	0.60 ±0.52	5.50 ±0.73	0.90 ±0.20	84.13 ±1.34	0.90 ±0.54	0.00 ±0.00	0.10 ±0.20	4.90 ±0.54	85.20 ±1.00	1.00 ±0.41	0.10 ±0.20	1.40 ±0.60	6.00 ±0.77	0.00 ±0.00	85.55 ±0.87
ja	0.90 ±0.62	0.80 ±0.26	3.00 ±0.65	0.00 ±0.00	0.00 ±0.00	89.60 ±1.35	0.10 ±0.20	0.50 ±0.33	1.20 ±0.26	2.20 ±0.26	91.18 ±1.26	0.40 ±0.32	0.30 ±0.30	2.00 ±0.51	0.00 ±0.00	0.90 ±0.20	92.20 ±0.57	0.00 ±0.00	0.70 ±0.42	1.40 ±0.32	0.00 ±0.00	94.20 ±0.92
me	0.00 ±0.00	0.00 ±0.00	0.20 ±0.26	1.90 ±0.35	2.00 ±0.00	1.10 ±0.20	95.30 ±0.72	3.60 ±0.32	2.40 ±0.43	11.20 ±1.00	80.98 ±0.90	0.40 ±0.43	1.40 ±0.32	0.10 ±0.20	0.60 ±0.32	2.60 ±0.32	1.00 ±0.00	96.80 ±0.49	2.00 ±0.51	1.10 ±0.20	15.30 ±1.28	79.83 ±1.03
po	0.00 ±0.00	0.00 ±0.00	1.40 ±0.60	7.40 ±0.89	2.00 ±0.29	0.00 ±0.00	1.00 ±0.00	78.10 ±0.80	4.90 ±0.46	3.30 ±0.59	79.66 ±1.21	0.00 ±0.00	0.00 ±0.00	5.50 ±0.33	5.60 ±0.93	2.00 ±0.00	0.10 ±0.20	0.00 ±0.00	87.50 ±0.89	6.10 ±0.85	1.40 ±0.52	80.90 ±0.92
re	0.20 ±0.26	0.00 ±0.00	4.50 ±0.67	6.90 ±0.54	4.80 ±0.76	0.00 ±0.00	0.40 ±0.32	4.60 ±0.67	71.80 ±0.87	2.20 ±0.49	75.27 ±0.95	0.90 ±0.35	0.00 ±0.00	1.40 ±0.52	8.40 ±0.52	3.10 ±0.20	0.00 ±0.00	0.00 ±0.00	3.40 ±0.52	72.40 ±1.02	1.40 ±0.52	79.57 ±0.82
ro	1.70 ±0.51	2.30 ±0.42	7.90 ±0.85	4.80 ±0.87	0.60 ±0.43	0.00 ±0.00	2.50 ±0.60	3.40 ±0.67	2.30 ±0.42	56.20 ±2.32	68.80 ±1.75	0.20 ±0.26	2.50 ±0.33	7.20 ±0.87	6.80 ±1.12	0.00 ±0.00	0.70 ±0.42	1.80 ±0.26	0.30 ±0.30	2.80 ±0.64	61.70 ±1.89	73.49 ±1.02
F	88.07 ±0.53	89.55 ±0.63	73.58 ±0.61	68.96 ±1.78	85.57 ±0.72	90.37 ±1.08	87.56 ±0.74	78.85 ±0.68	73.49 ±0.83	61.82 ±1.87		92.04 ±0.65	92.29 ±0.31	80.53 ±0.87	73.70 ±1.11	85.37 ±0.85	93.18 ±0.56	87.49 ±0.69	84.06 ±0.55	75.81 ±0.74	67.03 ±1.13	

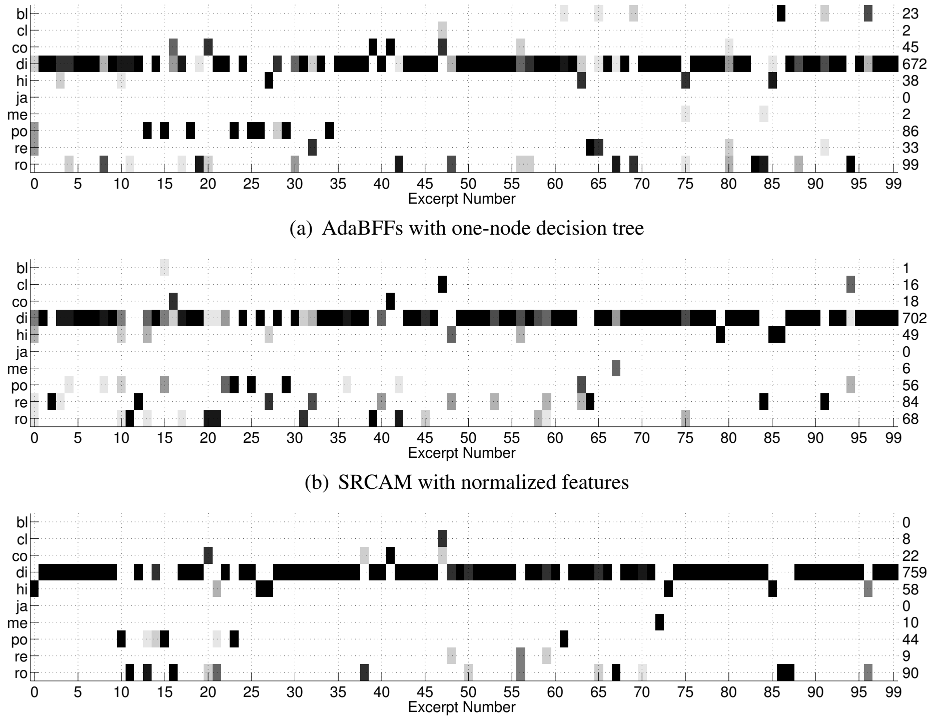
(b) SRCAM: standardized features (left) or normalized features (right)

	bl	cl	co	di	hi	ja	me	po	re	ro	Pr	bl	cl	co	di	hi	ja	me	po	re	ro	Pr
bl	89.50 ±1.14	0.00 ±0.00	0.00 ±0.00	4.20 ±0.49	1.60 ±0.32	0.00 ±0.00	0.40 ±0.32	0.00 ±0.00	2.00 ±0.00	1.20 ±0.26	90.50 ±0.89	90.30 ±0.66	0.00 ±0.00	0.70 ±0.42	0.00 ±0.00	0.00 ±0.00	0.90 ±0.20	0.10 ±0.20	0.00 ±0.00	2.90 ±0.46	9.00 ±0.51	86.92 ±0.81
cl	0.00 ±0.00	100.00 ±0.00	0.70 ±0.30	1.00 ±0.00	0.00 ±0.00	8.90 ±0.35	0.00 ±0.00	2.60 ±0.32	3.10 ±0.35	0.80 ±0.26	85.41 ±0.58	0.00 ±0.00	92.10 ±0.54	0.00 ±0.00	0.80 ±0.26	0.00 ±0.00	2.20 ±0.39	0.00 ±0.00	1.00 ±0.00	0.00 ±0.00	0.00 ±0.00	95.85 ±0.50
co	8.10 ±0.94	0.00 ±0.00	91.90 ±0.68	10.90 ±0.90	9.80 ±0.64	0.50 ±0.33	3.10 ±0.35	20.90 ±0.80	24.30 ±1.28	33.20 ±1.43	45.36 ±0.69	2.50 ±0.33	0.00 ±0.00	78.70 ±0.83	2.20 ±0.57	0.70 ±0.30	4.90 ±0.35	0.00 ±0.00	3.70 ±0.30	2.70 ±0.51	7.20 ±0.57	76.73 ±0.87
di	1.00 ±0.00	0.00 ±0.00	0.00 ±0.00	63.30 ±1.34	0.60 ±0.32	0.00 ±0.00	0.60 ±0.32	0.00 ±0.00	0.40 ±0.32	0.10 ±0.20	95.92 ±0.86	1.10 ±0.20	0.00 ±0.00	4.40 ±0.32	75.90 ±0.62	1.00 ±0.00	0.00 ±0.00	1.50 ±0.33	4.50 ±0.33	2.90 ±0.20	3.80 ±0.39	79.82 ±0.52
hi	0.00 ±0.00	0.00 ±0.00	0.00 ±0.00	0.10 ±0.20	53.70 ±0.83	0.00 ±0.00	0.00 ±0.00	0.00 ±0.00	0.10 ±0.20	0.00 ±0.00	99.64 ±0.47	0.00 ±0.00	0.00 ±0.00	0.90 ±0.20	5.80 ±0.39	82.30 ±0.72	0.20 ±0.26	0.00 ±0.00	1.00 ±0.00	7.60 ±0.43	0.00 ±0.00	84.15 ±0.48
ja	0.00 ±0.00	0.00 ±0.00	2.80 ±0.26	0.00 ±0.00	0.40 ±0.32	88.80 ±0.64	0.00 ±0.00	0.60 ±0.52	2.00 ±0.00	0.00 ±0.00	93.87 ±0.48	0.10 ±0.20	1.20 ±0.26	2.30 ±0.30	0.00 ±0.00	0.00 ±0.00	86.50 ±0.53	0.00 ±0.00	0.00 ±0.00	0.60 ±0.32	0.00 ±0.00	95.39 ±0.84
me	0.00 ±0.00	0.00 ±0.00	0.00 ±0.00	0.00 ±0.00	0.30 ±0.30	0.00 ±0.00	72.90 ±0.68	0.00 ±0.00	0.00 ±0.00	0.90 ±0.20	98.38 ±0.53	0.90 ±0.20	1.00 ±0.51	1.00 ±0.00	1.00 ±0.00	1.00 ±0.00	0.00 ±0.00	93.40 ±0.60	0.00 ±0.00	0.30 ±0.30	4.60 ±0.32	90.51 ±0.52
po	0.00 ±0.00	0.00 ±0.00	0.00 ±0.00	0.30 ±0.42	6.00 ±0.29	0.00 ±0.00	0.00 ±0.00	70.70 ±0.83	3.40 ±0.52	0.00 ±0.00	87.95 ±0.75	0.00 ±0.00	0.00 ±0.00	5.10 ±0.54	4.40 ±0.43	7.60 ±0.52	0.80 ±0.26	0.00 ±0.00	84.50 ±0.53	3.90 ±0.35	3.20 ±0.26	77.19 ±0.98
re	0.00 ±0.00	0.00 ±0.00	0.50 ±0.33	0.80 ±0.39	0.30 ±0.42	0.00 ±0.00	0.00 ±0.00	0.00 ±0.00	59.10 ±0.99	0.00 ±0.00	97.38 ±0.97	1.00 ±0.00	0.00 ±0.00	0.60 ±0.43	0.90 ±0.46	3.30 ±0.51	0.00 ±0.00	0.00 ±0.00	0.10 ±0.20	76.20 ±0.70	1.90 ±0.35	90.74 ±1.08
ro	1.40 ±0.43	0.00 ±0.00	4.10 ±0.54	19.40 ±1.21	27.30 ±1.01	1.80 ±0.26	23.00 ±0.72	5.20 ±0.49	5.60 ±0.73	63.80 ±1.20	42.09 ±0.60	4.10 ±0.46	5.70 ±0.30	6.30 ±1.01	9.00 ±0.51	4.10 ±0.54	4.50 ±0.44	5.00 ±0.41	5.20 ±0.39	2.90 ±0.35	70.30 ±0.72	60.04 ±0.78
F	89.99 ±0.95	92.13 ±0.34	60.73 ±0.70	76.25 ±1.08	69.78 ±0.68	91.26 ±0.41	83.74 ±0.59	78.38 ±0.55	73.55 ±0.89	50.71 ±0.69		88.57 ±0.63	93.93 ±0.32	77.69 ±0.57	77.81 ±0.41	83.21 ±0.49	90.72 ±0.38	91.93 ±0.46	80.67 ±0.70	82.83 ±0.61	64.77 ±0.71	

(c) MAPsCAT: class-dependent covariance (left) or total covariance (right)

Fig. 3 Mean confusions with 95% confidence intervals for each system. Columns are true genres, with mean precision (Pr) shown in last column. Rows are predicted genres, with mean F-measure (F ×100) in last row. Mean recalls are on diagonal. Classes as in Fig. 2

noticeable that classification errors make sense. For example, 29.41% of the ambient songs were misclassified as new-age, and these two classes seem to clearly overlap



(a) AdaBFFs with one-node decision tree

(b) SRCAM with normalized features

(c) MAPsCAT with total covariance

Fig. 4 GTZAN Disco excerpt confusions for each system, with number of classifications in each genre labeled at right. Classes as in Fig. 2

when listening to the audio files. In the same way, 14.71% of the blues examples were considered as rock by the algorithm."

Figure 4 shows how the GTZAN Disco excerpts are classified by AdaBFFs, SRCAM, and MAPsCAT over all trials. For lack of space, we only look at these excerpts, and herein only consider the setting that shows the best classification accuracy (Table 4): AdaBFFs with one-node decision tree; SRCAM with normalized features; and MAPsCAT with total covariance. Unlike in Fig. 3, we can see here the specific excerpts that AdaBFFs most often misclassifies as Pop and Rock, that SRCAM most often misclassifies as Reggae and Rock, and that MAPsCAT most often misclassifies as Rock and Hip hop. We can also see particular excerpts that are misclassified by the systems in all trials, i.e., GTZAN Disco excerpts 20, 27, 41, 47, and 85.

4.5 Analyzing System Proclivity

Very few works analyze the proclivity of an MGR system and its pathological behaviors, e.g., when an MGR system always favors the same wrong class. The analysis in [84] introduces the idea of studying the pathological misclassifications of MGR systems. Related to this is the MGR system proposed in [53], which selects from the training data only those instances of music that are easily separable, i.e., those instances for which a classifier rarely chooses the wrong genre. Also related is the work in [30,31,64], which analyzes the effects on system performance of artist/album replication across training and test sets.

Table 5 Classification type results for each system on GTZAN. The column "CM as" signifies the number of excerpts consistently misclassified as the genre of the relevant row

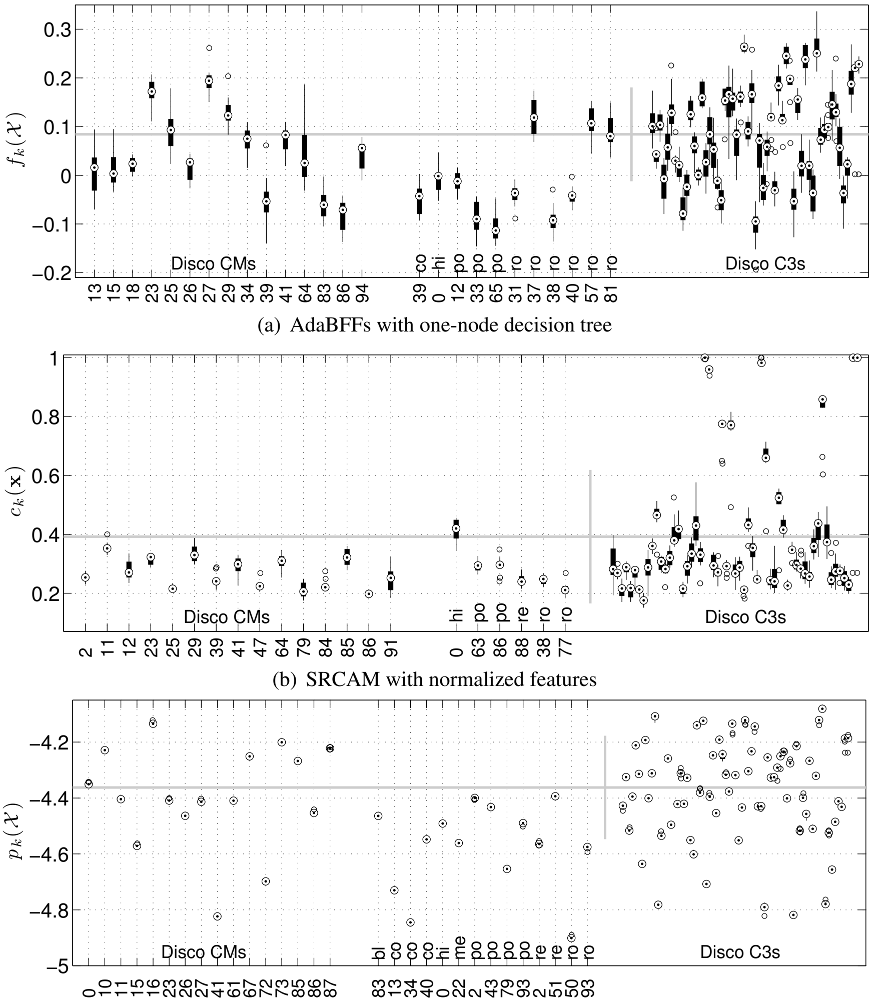
	AdaBFFs				SRCAM				MAPsCAT			
Genre	C3	CM	PM	CM as	C3	CM	PM	CM as	C3	CM	PM	CM as
Blues	88	2	1	5	89	0	1	4	86	6	0	8
Classical	95	2	1	5	95	2	1	9	90	6	0	2
Country	67	10	3	15	66	10	8	5	69	12	2	15
Disco	55	15	5	11	57	15	4	6	71	16	7	14
Hip hop	64	11	5	4	78	9	0	4	77	10	2	13
Jazz	80	5	1	3	85	4	2	1	82	6	3	3
Metal	87	2	0	6	94	0	0	16	89	5	0	5
Pop	81	6	3	19	77	5	3	12	81	13	0	19
Reggae	58	11	6	7	60	10	7	9	73	17	4	3
Rock	34	21	10	10	42	18	6	7	64	19	1	28
Total	709	85	35	85	743	73	73	73	855	110	19	110

Building upon the work in [84], we define three types of system proclivity. When in all trials a system selects the "correct" class for an excerpt (the label in GTZAN), we call it a *consistently correct classification* (C3). When in all trials a system selects the same but "wrong" class for an excerpt, we call it a *consistent misclassification* (CM). When in all trials a system selects different "wrong" classes for an excerpt, we call it a *persistent misclassification* (PM). Table 5 summarizes the numbers of these classification types for all of GTZAN for AdaBFFs, SRCAM, and MAPsCAT. Here we see that of the GTZAN Disco excerpts, AdaBFFs produced 55 C3s, 15 CMs, 5 PMs, and it consistently misclassified 11 excerpts as Disco. We see that, in total, MAPsCAT has the highest and AdaBFFs the lowest number of C3s and CMs.

We can ask about the relative confidence of a system betwen a CM and C3. For instance, is for AdaBFFs the value (2) larger for its CMs than for its C3s? This amounts to comparing the votes (2) for the CMs and C3s in the GTZAN excerpts. We plot in Fig. 5 the statistics of (2) for AdaBFFs, (5) for SRCAM, and (7) for MAPsCAT, for only the GTZAN Disco excerpts. The left-most portion of each subfigure is of the CMs of Table 5; and the right-most portion is from the C3s. The middle portion is of those GTZAN excerpts not labeled Disco, but that each system consistently misclassifies as Disco (CMs as Disco). The gray horizontal line is the mean value of the C3s for a system; and the vertical gray line marks one standard deviation above and below the mean. Figure 5(a) shows that for AdaBFFs the votes (2) of most Disco CMs and CMs as Disco are indistinguishable from those of the Disco C3s, even though the majority of them lie under the mean of the C3s. They all are within two standard deviations of the mean. Figure 5(b) shows that the mean confidence (5) of all Disco CMs and CMs as Disco are indistinguishable from those of the Disco C3s. Most exist below the mean, but all are well within one standard deviation. Figure 5(c) shows that the mean log posteriors (7) of most Disco CMs and CMs as Disco are indistinguishable from those of the Disco C3s. About half lie above the mean than below it, and all but one are within two standard deviations. These results point to the idea that AdaBFFs, SRCAM, and MAPsCAT are as confident in their C3s as they are in their pathological misclassifications.

4.6 Analyzing Consistently Misclassified Excerpts

So far, our evaluation has used statistical tests, rough discussions of genre labels, and mentioned specific excerpt numbers, but has yet to make mention of and use



(a) AdaBFFs with one-node decision tree

(b) SRCAM with normalized features

(c) MAPsCAT with total covariance

Fig. 5 Figures of merit for GTZAN Disco labeled excerpts. Disco CMs labeled on left. CMs as Disco labeled in center. Disco C3s on right. Mean for Disco C3s shown as gray line with one standard deviation above and below. Classes as in Fig. 2

the actual music embodied by any excerpts. It is in fact quite extraordinary to find in the MGR literature any identification of the music behind problematic classifications. Langlois and Marques [45] notice in their system evaluation that all tracks from an album by Bossa Nova artist João Gilberto are PMs. They attribute this to the tracks coming from a live recording with speaking and applause. We see that the system by Lee et al. [47] misclassifies John Denver's "Rocky Mountain High" as Techno, but they do not discuss this problematic result.

In Table 6, we list the specific excerpts of each pathological classification of AdaBFFs, SRCAM, and MAPsCAT for only the Disco class. We take into account that GTZAN has among its Disco excerpts: six replicas, two coming from the same recording, and seven conspicuous and three contentious mislabelings [84]. Further-

Table 6 Classification type results for each system: C3s, CMs and PMs for GTZAN Disco excerpts; and CMs as Disco. We take into account the problems of GTZAN [84], and strike-though particular excerpts

	Classification Type			
System	C3	CM *excerpts*	PM *excerpts*	CM as Disco *label and excerpts*
AdaBFFs	45	*13, 15, 18, ~~23~~, 25, ~~26~~, ~~27~~, ~~29~~, 34, 39, ~~41~~, 64, 83, 86, ~~94~~*	*~~20~~, ~~47~~, 69, 75, 84*	*~~co 39~~; hi 00; po 12, 33, 65; ro 31, 37, 38, 40, 57, 81*
SRCAM	47	*02, 11, 12, ~~23~~, 25, ~~29~~, 39, ~~41~~, ~~47~~, 64, 79, 84, ~~85~~, 86, 91*	*~~27~~, 42, 48, 63*	*hi 00; po 63, 86; re 88; ro 38, 77*
MAPsCAT	58	*00, 10, 11, 15, 16, ~~23~~, ~~26~~, ~~27~~, ~~41~~, 61, 67, 72, 73, ~~85~~, 86, 87*	*13, ~~20~~, ~~21~~, 38, ~~47~~, 56, 96*	*bl 83; co 13, 34, 40; hi 00; me 22; po 02, 43, 79, 93; re 02, 51; ro 50, 93*
in common	32	*~~23~~, ~~41~~, 86*		*hi 00*

more, the genre of Country excerpt 39 is none of the 10 labels in GTZAN, and so we do not consider it here [84]. We strike-through these problematic excerpts. We see that, as in Table 5, MAPsCAT has the highest number of C3s and CMs, and AdaBFFs has the lowest. All systems share a common CM and CM as Disco.

From the GTZAN track listing produced in [84], Tables 7 and 8 list the music artist and title of the GTZAN Disco excerpt CMs and the CMs as Disco of AdaBFFs, SRCAM, and MAPsCAT. Even though each system commits CMs in almost all the other genres of GTZAN (Table 5), we show only the CMs of the GTZAN Disco excerpts, and CMs as Disco, for lack of space. We list the statistics associated with each

Table 7 Details of Disco CMs from Table 6. An excerpt number marked by * means the classifier "confidence" is larger than that of the Disco CM marked by † ($p < 0.041$)

				Assigned	$\{f_k(\mathscr{X}) - f_{Disco}(\mathscr{X})\}$ (2)			
	No.	Artist	Title of Work	Class k	min	max	mean ± 95%	Top `last.fm` Tags
AdaBFFs	13*	Donna Summer	Back Off Boogaloo	Pop	0.08	0.24	0.15 ± 0.03	artist: disco, pop, 70s
	15*	Heatwave	Boogie Nights	Pop	0.06	0.30	0.17 ± 0.05	disco, funk, 70s
	18*	?	?	Pop	0.11	0.25	0.17 ± 0.03	
	25	Evelyn Thomas	High Energy	Pop	0.05	0.24	0.14 ± 0.03	Disco, 80s, dance
	34†	Evelyn Thomas	Reflections	Pop	0.002	0.17	0.09 ± 0.03	80s, disco; artist: disco
	39	George McCrae	I Can't Leave You Alone	Country	0.01	0.21	0.10 ± 0.04	Disco, 70s, pop
	64*	Lipps, Inc.	Funkytown	Reggae	0.07	0.41	0.20 ± 0.06	Disco, 80s, 70s
	83*	Rick Dees	Disco Duck	Rock	0.12	0.22	0.18 ± 0.02	70s, pop, Disco
	86*	Alicia Bridges	I Love the Night Life	Blues	0.13	0.28	0.22 ± 0.03	70s; artist: disco, dance

				Assigned	$\{c_k(\mathbf{x}) - c_{Disco}(\mathbf{x})\}$ (5)			
	No.	Artist	Title of Work	Class k	min	max	mean ± 95%	Top `last.fm` Tags
SRCAM	02*	Archie Bell and The Drells	Look Back Over Your Shoulder	Reggae	0.12	0.19	0.16 ± 0.02	northern soul, soul
	11*	Billy Ocean	Can You Feel It	Rock	0.26	0.37	0.29 ± 0.02	rock, pop; artist: 80s
	12	Carl Carlton	She's A Bad Mama Jama	Reggae	0.002	0.16	0.07 ± 0.03	funk, disco, 70s
	25	Evelyn Thomas	High Energy	Pop	0.03	0.07	0.05 ± 0.01	disco, 80s, dance
	39*	George McCrae	I Can't Leave You Alone	Rock	0.08	0.18	0.12 ± 0.02	disco, 70s, pop
	64*	Lipps, Inc.	Funky Town	Reggae	0.12	0.21	0.16 ± 0.02	Disco, 80s, 70s
	79	Peter Brown	Love Is Just The Game	Hip hop	0.02	0.12	0.07 ± 0.02	70s, disco; artist: funk
	84*	?	?	Reggae	0.10	0.18	0.12 ± 0.01	
	86†	Alicia Bridges	I Love the Night Life	Hip hop	0.04	0.13	0.08 ± 0.02	70s; artist: disco, dance
	91	Silver Convention	Fly Robin Fly	Reggae	0.02	0.22	0.08 ± 0.05	Disco, pop

				Assigned	$\frac{1}{100}\{p_k(\mathscr{X}) - p_{Disco}(\mathscr{X})\}$ (7)			
	No.	Artist	Title of Work	Class k	min	max	mean ± 95%	Top `last.fm` Tags
MAPsCAT	00*	Boz Scaggs	Lowdown	Hip hop	1.29	1.97	1.68 ± 0.11	70s, classic rock
	10	?	?	Pop	0.56	0.97	0.71 ± 0.08	
	11*	Billy Ocean	Can You Feel It	Rock	0.82	1.19	1.03 ± 0.07	rock, pop; artist: 80s
	15†	Heatwave	Boogie Nights	Pop	0.45	0.89	0.65 ± 0.10	disco, funk, 70s
	16	?	?	Rock	0.15	0.60	0.35 ± 0.09	
	61	Anita Ward	Ring My Bell	Pop	0.43	1.17	0.75 ± 0.18	Disco, 70s, dance
	67*	ABBA	Dancing Queen	Rock	0.53	1.15	0.87 ± 0.11	pop, Disco, 70s
	72*	ABBA	Mamma Mia	Metal	0.81	1.18	1.00 ± 0.09	pop, 70s, disco
	73	KC & Sunshine Band	I'm Your Boogie Man	Hip hop	0.18	0.61	0.40 ± 0.08	Disco, 70s, funk
	86*	Alicia Bridges	I Love the Night Life	Rock	0.95	1.47	1.21 ± 0.10	70s; artist: disco, dance
	87	The Supremes	He's My Man	Rock	0.04	0.58	0.31 ± 0.11	soul, vocalization

Table 8 Details of GTZAN excerpts consistently misclassified as Disco from Table 6

	Genre			$\{f_{Disco}(\mathscr{X}) - f_k(\mathscr{X})\}$ (2)			
	& No.	Artist	Title of Work	min	max	mean ± 95%	Top `last.fm` Tags
AdaBFFs	hi 00	Afrika Bambaataa	Looking for the Perfect Beat	0.02	0.26	0.16 ± 0.04	electro, Hip-Hop, old school
	po 12	Aretha Franklin, et al.	You Make Me Feel Like A Natural Woman	0.09	0.22	0.16 ± 0.02	pop, Ballad; artist: soul
	po 33	Britney Spears	Pepsi Now and Then	0.13	0.26	0.19 ± 0.03	artist: pop, dance
	po 65	Prince	The Beautiful Ones	0.15	0.27	0.21 ± 0.02	80s, funk, pop
	ro 31	The Rolling Stones	Honky Tonk Women	0.09	0.21	0.15 ± 0.02	classic rock, rock, 60s
	ro 37	The Rolling Stones	Brown Sugar	0.02	0.17	0.08 ± 0.03	classic rock, rock, 70s
	ro 38	Guns 'N Roses	Knockin' On Heaven's Door	0.14	0.28	0.19 ± 0.03	rock, hard rock, classic rock
	ro 40	Led Zeppelin	The Crunge	0.24	0.33	0.28 ± 0.02	classic rock, hard rock, rock
	ro 57	Sting	If you love somebody set them free	0.14	0.31	0.23 ± 0.04	rock, 80s, pop
	ro 81	Survivor	Poor Man's Son	0.29	0.46	0.34 ± 0.03	80s, rock, melodic rock

	Genre			$\{c_{Disco}(\mathbf{x}) - c_k(\mathbf{x})\}$ (5)			
	& No.	Artist	Title of Work	min	max	mean ± 95%	Top `last.fm` Tags
SRCAM	hi 00	Afrika Bambaataa	Looking for the Perfect Beat	0.21	0.31	0.27 ± 0.02	electro, Hip-Hop, old school
	po 86	Madonna	Cherish	0.02	0.13	0.08 ± 0.02	pop, 80s, dance
	re 88	Marcia Griffiths	Electric Boogie	0.14	0.26	0.20 ± 0.02	funk, reggae, dance
	ro 38	Guns 'N Roses	Knocking On Heaven's Door	0.03	0.13	0.08 ± 0.02	rock, hard rock, classic rock
	ro 77	Simply Red	Freedom	0.01	0.12	0.08 ± 0.02	pop, rock, easy

	Genre			$\frac{1}{100}\{p_{Disco}(\mathscr{X}) - p_k(\mathscr{X})\}$ (7)			
	& No.	Artist	Title of Work	min	max	mean ± 95%	Top `last.fm` Tags
MAPsCAT	co 13	Loretta Lynn	Let Your Love Flow	0.47	1.11	0.96 ± 0.12	country, Traditional Country
	co 34	Merle Haggard	Sally let your bangs hang down	0.37	1.65	0.81 ± 0.22	artist: country, classic country, outlaw country
	co 40	Kentucky Headhunters	Dumas Walker	0.24	1.17	0.67 ± 0.18	country, outlaw country, rock
	hi 00	Afrika Bambaataa	Looking for the Perfect Beat	2.01	2.49	2.25 ± 0.10	electro, Hip-Hop, old school
	me 22	Ozzy Osbourne	Crazy Train	0.49	0.96	0.75 ± 0.09	heavy metal, metal, hard rock
	po 02	Mariah Carey	My All	0.05	0.68	0.26 ± 0.13	pop, rnb, soul
	po 43	Cher	Believe	0.89	1.60	1.17 ± 0.15	pop, dance, 90s
	po 79	Kate Bush	Couldbusting	1.72	2.09	1.93 ± 0.08	80s, pop, alternative
	po 93	Mandy Moore	I Wanna Be With You	0.84	1.34	1.08 ± 0.10	pop, romantic, Love
	re 02	Bob Marley	Could You Be Loved	0.43	0.95	0.70 ± 0.12	reggae, roots reggae, jamaican
	ro 50	Simple Minds	See The Lights	0.87	1.51	1.21 ± 0.14	rock, 80s, new wave
	ro 93	The Stone Roses	Waterfall	0.09	0.69	0.43 ± 0.13	indie, britpop, madchester

system decision, and up to three top `last.fm` tags (ranked by the "count" parameter) of each song or artist (retrieved from `last.fm` on Oct. 15, 2012). We do not include tags that are the artist or song name; and when a song has no tags associated with it, we take the top tags for the artist. We define the "confidence" of a classification as the difference between the score — (2) for AdaBFFs, (5) for SRCAM, and (7) for MAPsCAT — of the selected class with that of the "correct" Disco class, This is different in Fig. 5, where we show the decision statistic but not a difference.

Table 7 lists the class consistently selected by each system. We see the tag "disco" as a top tag for all eight identified Disco CMs of AdaBFFs. We see AdaBFFs consistently misclassifies as Pop excerpts 25 and 34, both by Evelyn Thomas. The mean of the differences between the votes for Pop and those for Disco is larger for excerpt 25 than it is for 34, but we find from a paired t-test that we cannot reject the null hypothesis that one is not larger than another ($p > 0.07$). However, with respect to excerpt 34, we can reject such a null hypothesis for excerpts 13, 15, 18, 64, 83 and 86 ($p < 0.041$), and thus consider these classifications to be "confident." In other words, if we assume excerpt 34 is a borderline classification in all trials, then the other six having larger votes are not borderline. Of those six, only one (13) shares a tag matching the class given by AdaBFFs. The consistent misclassification of “Funkytown” by Lipps, Inc. as Reggae, and Alicia Bridges’ “I Love the Night Life” as Blues seem quite odd, as they do not sound similar to the GTZAN excerpts exemplifying each category, i.e., Bob Marley, Dennis Brown, Burning Spear, and Gregory Isaacs making more than 50% of the Reggae excerpts; and Robert Johnson, John Lee Hooker, Stevie Ray Vaughn, and Magic Slim making more than 50% of the Blues excerpts [82].

Of the nine identified Disco CMs of SRCAM, the tag “disco” exists for seven. Excerpt 25 has smallest mean confidence difference among the CMs, and we find the mean differences for excerpts 12, 79 and 91 are not larger ($p > 0.09$), but that of excerpt 86 is ($p < 0.005$). With respect to excerpt 86 then, we can reject the null hypothesis that its mean difference is not smaller than those of excerpts 2, 11, 39, 64, and 84 ($p < 0.013$), and thus consider these classifications to be confident. Of those five, only one (11) shares a tag matching the class given by SRCAM. SRCAM, like AdaBFFs, consistently misclassifies as Reggae “Funkytown” by Lipps, Inc.

Of the nine identified Disco CMs of MAPsCAT, the tag “disco” exists for six. The smallest log posterior difference occurs for excerpt 87, which we find is not smaller than 16 and 73 ($p > 0.1$). With respect to excerpt 15, we can reject the null hypothesis that its mean log posterior difference is not smaller than those of excerpts 00, 11, 67, 72, and 86 ($p < 0.016$), and thus consider these classifications as confident. Of these five, only one (11) shares a tag matching the class given. MAPsCAT consistently misclassifies “Lowdown” by Boz Scaggs as Hip hop, and ABBA’s “Mamma Mia” as Metal. These are odd considering the composition of the majority of each class in GTZAN, i.e., excerpts by Beastie Boys, A Tribe Called Quest, and Public Enemy make more than 56% of the Hip hop excerpts; and Metallica, Dark Tranquillity, Iron Maiden, Black Sabbath, Anthrax, Dio, Motörhead, Rage Against The Machine, and New Bomb Turks make more than 50% of the Metal excerpts [82].

Of the CMs as Disco by all three systems, Table 8 shows none of the tags of the music or artist contains “disco,” and only a few contain Disco-relatable tags, such as “dance” and “70s.” For those CMs as Disco of AdaBFFs, we see that the mean vote difference for Rock excerpt 37 is the smallest, and all others are significantly larger ($p < 0.015$). For SRCAM, the mean difference in confidence for Pop excerpt 86, and Rock excerpts 38 and 77 are the smallest, but those for Hip hop 00 and Reggae 88 are significantly larger ($p < 2 \cdot 10^{-6}$). For MAPsCAT, the mean log posterior difference for Pop excerpt 02 is the smallest, and all others are significantly larger ($p < 0.013$).

Common to all three systems are one CM and CM as Disco. All systems consistently misclassify Disco excerpt 86, “I Love the Night Life” by Alicia Bridges from 1978: AdaBFFs confidently labels it Blues, SRCAM labels it Hip hop, and MAPsCAT confidently labels it Rock. These labels do not agree well with the tags for the song or the artist. The common CM as Disco is Hip hop 00, which is Afrika Bambaataa’s “Looking for the Perfect Beat” from 1982. The insistence of all systems that this excerpt is Disco might be seen as forgivable, since early Hip hop used Disco records as background for rapping [78] — but such a claim assumes these systems learn that fact from the 90 GTZAN Hip hop excerpts in every cross-validation fold.

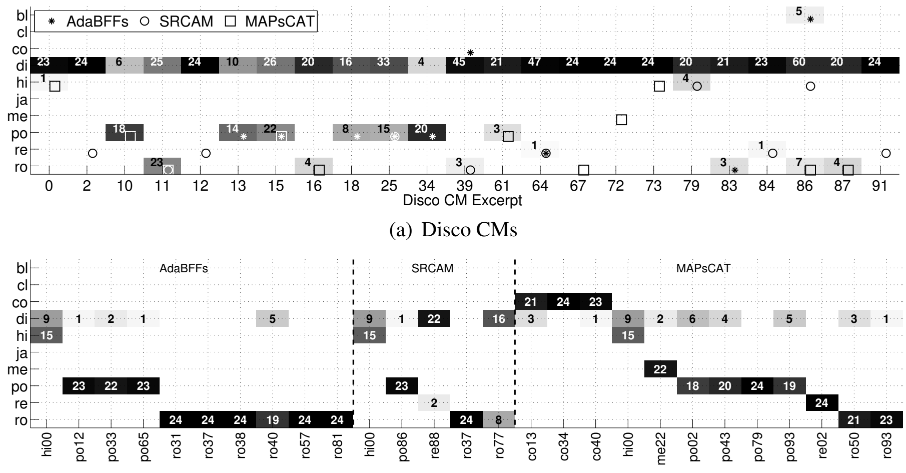
We also see more generally in Tables 7 and 8 that all systems misclassify several GTZAN Disco excerpts as Pop, and GTZAN Pop excerpts as Disco. We might also see these as forgivable as much music we now call “disco” was in fact part of the "popular" charts in the late 1970s [78]. Furthermore, the two excerpts of "High Energy" and "Reflections" by Evelyn Thomas come from 1984 and 1985, respectively, which is five years after "disco died" in 1979 in the USA [78]. Hence, a better single label for these particular excerpts is Pop. Aside from these, some consistent misclassifications appear quite unsatisfactory. For instance, AdaBFFs and SRCAM consistently misclassifies as Disco "Knockin' on Heaven's Door" by Guns 'N Roses; and MAPsCAT consistently misclassifies as Disco "Sally Let Your Bangs Hang Down" sung by Merle Haggard, and misclassifies as Metal "Mamma Mia" by ABBA.

4.7 Analyzing by Listening Tests

The question thus arises: to what extent do humans show the same kinds of classification behavior as AdaBFFs, SRCAM, and MAPsCAT? By and large, the MGR literature is concerned with the performance of algorithms, pigeons [69], and fish [20]. There have, however, been a few studies made of human music genre classification. One of the most widely cited [8] is the work of Gjerdingen and Perrott [36], which studies human genre recognition capacity as a function of excerpt length. Krumhansl [44], and Mace et al. [55] expands upon this in several directions. Ahrendt et al. [2, 3] and Meng et al. [60, 61] both use listening tests to gauge the difficulty of discriminating the genres of their genre datasets, and to evaluate their systems' performance. Lippens et al. [52] use listening tests to produce a music genre dataset that has excerpts more exemplary of single genres; and Craft et al. [22, 23] addresses a fundamental problem of that work, proposing that evaluating MGR systems makes sense only with respect to the generic ambiguity of music. Seyerlehner et al. [77] reproduces and expands upon these works. In a novel direction, Guaus [38] conducts listening tests to determine the relative importance of timbre or rhythm in genre recognition. Finally, Cruz and Vidal [24] and Sturm [84] conduct listening tests to determine if a system can create music using the genres it has learned to identify.

These latter works motivate the use of listening tests to circumvent the need to demarcate the stylistic elements — assuming they exist to a large extent in the acoustic realm [93] — required to formally justify whether it is more appropriate, e.g., to label an excerpt Disco or Metal, neither or both. Our hypothesis is that for some excerpts the difference between the pair of genre labels given by a human and an MGR system will be large enough that it is extremely clear to listeners familiar with the genres which label is given by a human. For other pairs of genre labels for some excerpts, the difference will be small enough that listeners familiar with the genres can make no real distinction. This boils down to something like a Turing test to determine whether the consistent misclassifications of AdaBFFs, SRCAM, or MAPsCAT are appropriate. With this approach, we can circumvent the comparisons between tags, classes, and "ground truth" labels, and also the task of having to define genres in ways that allow us to, e.g., challenge the labeling of Alicia Bridge's "I Love the Night Life" as using the Blues, Hip hop, and/or Rock genres.

To approach our hypothesis, we conduct a listening test in which a subject must choose for each 12 second excerpt which label of two was given by a human (i.e., Tzanetakis); the other label was given by the computer. The experiment has two parts, both facilitated by GUIs built in MATLAB. In the first part, we screen subjects for their ability to distinguish between the ten genres in GTZAN. (The representative



(a) Disco CMs

(b) CMs as Disco

Fig. 6 Distribution of choices from listening test for each consistent misclassification of AdaBFFs, SRCAM, and MAPsCAT. (a) Disco CMs in Table 7. The class selected by each system is marked by the symbol shown in the legend. (b) CMs as Disco in Table 8. Classes as in Fig. 2

GTZAN excerpts we use are: Blues 05, John Lee Hooker, "Sugar Mama"; Classical 96, Vivaldi, "The Four Seasons, Summer, Presto"; Country 12, Billy Joe Shaver, "Music City"; Disco 66, Peaches and Herb, "Shake Your Groove Thing"; Hip hop 47, A Tribe Called Quest, "Award Tour"; Jazz 19, Joe Lovano, "Birds Of Springtimes Gone By"; Metal 11, unknown; Pop 95, Mandy Moore, "Love you for always"; Reggae 71, Dennis Brown, "Big Ships"; Rock 37, The Rolling Stones, "Brown Sugar.") A subject correctly identifying the genre of all excerpts continues to the second part of the test, where s/he must discriminate between the human- and algorithm-given genres for each music excerpt. For instance, the test application presents the excerpt of Donna Summer's "Back Off Boogaloo" along with the two labels "Disco" and "Pop." The subject must select the one s/he thinks is given by a human before proceeding to the next excerpt. We also record the time a subject spends listening to an excerpt before proceeding to the next one. We test all unique Disco CMs and CMs as Disco in Tables 7 and 8. In total, 24 test subjects completed the second part.

Figure 6 shows the choices made by subjects for all Disco CMs, and CMs as Disco for each MGR system. Figure 6(a) shows that for the nine Disco CMs of AdaBFFs, a majority of subjects sided with the non-human class in two cases: excerpts 13 and 34. In one case, excerpt 39, no subject chose the class given by AdaBFFs. In no case for the 10 Disco CMs of SRCAM did a majority of subjects pick the non-human class; and for four excerpts — 2, 12, 86, and 91 — no subject chose the class given by SRCAM. For the eleven Disco CMs of MAPsCAT, only for excerpt 10 did a majority of subjects choose the non-human class; and no subject chose the MAPsCAT class for three excerpts: 67, 72, and 73. In Fig. 6(b), we see that of the ten CMs as Disco of AdaBFFs, and of the twelve of MAPsCAT, in no case did a majority of subjects select "Disco." Of the five CMs as Disco of SRCAM, we see for two excerpts — Reggae 88 and Rock 77 — a majority of subjects chose "Disco".

Now we test the null hypothesis that the subjects are unable to recognize the difference between the genre label given by a human and the class selected by AdaBFFs, SRCAM, or MAPsCAT. We can consider the outcome of each trial as a Bernoulli random variable with parameter x (the probability of a subject selecting the label given by a human). For a given excerpt to which the human label is selected h times by N independent subjects, we can estimate the Bernoulli parameter x using the minimum mean-squared error estimator, assuming x is distributed uniform in $[0,1]$: $\hat{x}(h) = (h+1)/(N+2)$ [81]. The variance of this estimate is given by [81]

$$\hat{\sigma}^2(\hat{x}) = \frac{\hat{x}(1-\hat{x})}{(N-1) + \frac{N+1}{N\hat{x}(1-\hat{x})}}. \quad (8)$$

We test the null hypothesis $x - 0.5 = 0$ by computing $P[T > |\hat{x} - 0.5|/\hat{\sigma}(\hat{x})]$ where T is distributed Student's t with $N-2$ degrees of freedom (two degrees lost in the estimation of the Bernoulli parameter and its variance). For only four Disco CM excerpts — 11, 13, 15, and 18 — do we find that we cannot reject the null hypothesis $(p > 0.1)$. Furthermore, in the case of excerpts 10 and 34, we can reject the null hypothesis in favor of the misclassification of MAPsCAT and AdaBFFs, respectively ($p < 0.012$). For all other 21 Disco excerpts, we reject the null hypothesis in favor of the labels given by Tzanetakis ($p < 0.008$). For only two CMs as Disco excerpts (Hip hop 00 and Rock 77) do we find that we cannot reject the null hypothesis $(p > 0.1)$. Furthermore, only in the case of excerpt Reggae 88 can we reject the null hypothesis in favor of SRCAM $(p < 4 \cdot 10^{-7})$. For all other 20 excerpts, we can reject the null hypothesis in favor of Tzanetakis' labels $(p < 0.012)$.

So, what is it about Disco excerpts 13, 15 and 18 that made subjects divided between the labels "Disco" and "Pop," and choose more often "Pop" for Disco excerpts 10 and 34? Many subjects that passed the screening mentioned in post-test interviews that the most challenging pair of tags was "Disco" and "Pop." When asked what cues they used to make the choice, many could not state specifics, referring instead to the "feel" of the music. Some said they decided based upon whether the excerpt sounded "old" or "more produced." Hence, it is reasonable to believe that whatever makes something Disco but not Pop is unclear without further specification, e.g., "Pop like Britney Spears" and "70s Disco." In these cases then, we might as well conclude that AdaBFFs and MAPsCAT are classifying appropriately. Some subjects were also dissatisfied by some label pairs, e.g., "Metal" and "Disco" for ABBA's "Mamma Mia" because in their opinion ABBA is Pop.

In the case of Disco excerpt 11, subjects were divided between "Disco" and "Rock." When asked in the post-test interview about how quickly they made each selection, many subjects said they were quite quick, e.g., within the first few seconds. Some mentioned that they changed their answers after listening to some of the excerpts longer; and a few subjects said that they made sure to listen beyond what sounded like the introduction. We thus look at the duration each subject spent listening to Disco excerpt 11 before proceeding. We find that the listening time difference between subjects who selected "Rock" (8.5 ± 1.2 s, with 95% confidence interval) versus those who selected "Disco" (7.9 ± 1.1 s), is not statistically significant ($p > 0.48$). However, for Hip hop excerpt 00, the mean listening durations of subjects who selected "Disco" (4.9 ± 1.1 s) versus those who selected "Hip hop" (9.5 ± 1.6 s) is significant ($p < 6 \cdot 10^{-5}$). Apparently, many subjects hastily chose the label "Disco" — which brings up the question of whether the genres used by an entire piece of music applies to its parts [54]. In these cases, then, we can conclude that SRCAM and MAPsCAT are classifying appropriately.

Finally, there are Reggae 88 and Rock 77, for which subjects selected significantly more often the non-human class. In the first case, it is clear that people do not agree with the Tzanetakis label. "Electric Boogie" by Marcia Griffiths is quite unlike the majority of the GTZAN Reggae excerpts, even though one of its top tags is "reggae." Hence, we can consider as appropriate this misclassification by SRCAM. For the Rock 77 excerpt, since we do not find a significant difference in listening times ($p > 0.6$), we can regard SRCAM is classifying appropriately. All other CMs and CMs as Disco, however, are not appropriate: for AdaBFFs, 5 of its 9 CMs and 9 of its 10 CMs as Disco; for SRCAM, 9 of its 10 CMs and 3 of its 6 CMs as Disco; and for MAPsCAT, 8 of its 11 CMs and 13 of its 14 CMs as Disco.

5 Conclusion

While genre is an inescapable result of human communication [33], it is also ambiguous [22,23,93] as humans do not always agree, e.g., [2,3,22,23,36,52,61,77]. A major conundrum in the evaluation of MGR systems is thus the formal justification of why a particular label is better than another. For instance, while I deride the misclassification above, an argument might be made that ABBA's "Mamma Mia" employs some of the same stylistic elements used by Motörhead in "Ace Of Spades" — though it is difficult to imagine the audiences of the two would perceive that to be the case. The matter of evaluating MGR systems would be quite simple if only we had a checklist of essential, or at least important, attributes for each genre. Barbedo and Lopes [9] provides a long list of such attributes in each of several genres and sub-genres, e.g., Light Orchestra Instrument Classical is marked by "light and slow songs ... played by an orchestra" and have no vocal element (like J. S. Bach's "Air on the G String"); and Soft Country Organic Pop/Rock is marked by "slow and soft songs ... typical of southern United States [with] elements both from rock and blues [and where] electric guitars and vocals are [strongly] predominant [but there is little if any] electronic elements" (like "Your Cheating Heart" by Hank Williams Sr.). Some of these attributes are clear and actionable, like "slow," but others are not, like "[with] elements both from rock and blues." Categorically different from this is the expert system devised by Dixon et al. [27], where temporal characteristics of music, e.g., tempo and meter, can sometimes restrict its membership to particular dance styles.

In this work, we have analyzed from multiple perspectives the performance of three MGR systems to measure the extent to which they recognize music genre. From Table 4, we see the classification accuracies of AdaBFFs, SRCAM, and MAPsCAT are significantly higher than chance, and are among the best observed (and reproduced) for the GTZAN dataset. Thus, one might take such a high classification accuracy as evidence that a system is capable of recognizing the genres in a test set. However, from the nature of the Classify experimental design, we are not able to reject the null hypothesis that one of these systems is not able to recognize genre, *no matter the accuracy observed*. In essence, "genre" is not the only independent variable that changes between the excerpts of particular genres. There is also, just to name a few, instrumentation (Disco and Classical may or may not use strings), loudness (Metal and Classical can be listened to at high or low volume), tempo (Blues and Country can be played fast or slow), dynamics (Classical and Jazz can have few or several large changes in dynamics), reverberation (Reggae can involve spring reverberation, and Classical can be performed in small or large halls), production (Hip hop and Rock can be produced in a studio or in a concert), channel bandwidth (Country and Classical can be heard on AM or FM radio), noise (Blues and Jazz can be heard from an old record or a new CD), and so on. To determine if an MGR system has a capacity to recognize any genre, we must look deeper than classification accuracy.

In Fig. 2, we see the recalls, precisions, and F-measures for AdaBFFs, SRCAM, and MAPsCAT. With these figures of merit then, one might be inclined to claim that we can reject the null hypothesis that MAPsCAT cannot recognize Disco, or that AdaBFFs cannot recognize Classical. However, “to recognize” is not equivalent to having high recall, precision, or F-measure; and “recognize Disco” is not equivalent to “recognize as Disco an excerpt labeled Disco” — especially with the problems inherent to the GTZAN dataset [82]. Thus, we still cannot reject the null hypothesis that MAPsCAT cannot recognize Disco, *even with perfect accuracy, and thus precision, recall, and F-measure.* To answer whether any of these MGR systems has a capacity to recognize Disco, we must dig deeper than these figures of merit.

We might claim that the confusion behavior of AdaBFFs, SRCAM, and MAPsCAT “makes musical sense;” but by doing so we implicitly make two critical assumptions: 1) that the dataset being used has integrity for MGR; and 2) that the system is using cues similar to those used by humans when categorizing music, e.g., what instruments are playing, and how are they being played? what is the rhythm, and how fast is the tempo? is it for dancing, moshing, protesting or listening? is someone singing, and if so what is the subject? For the first assumption, though it is the most used dataset in MGR — appearing in 23% of MGR research since 2002 — GTZAN has numerous problems, including repetitions of excerpts and artists, many mislabelings, and distortions [84]. Hence, GTZAN is not a dataset with high integrity. The second assumption is much harder to justify, and requires us again to dig deeper than the confusion behaviors. We thus have to look at the level of the music itself to answer these questions.

Analyzing the pathological behaviors of an MGR system provides insight into whether its internal models of genres make sense with respect to the ambiguous nature of genre. Tables 5 – 8 provide details on persistent kinds of confusions that appear for AdaBFFs, SRCAM, and MAPsCAT. Comparing the classification results with the tags given by a community of listeners show that some behaviors do indeed “make musical sense,” but other appear less rational. In the case of using tags, the implicit assumption is that the tags given by an unspecified population to make their music more useful to them are to be trusted in describing the elements of music that characterize the genre(s) it uses — whether users found these upon genre (“funk” and “soul”), style (“melodic” and “classic”), form (“ballad”), function (“dance”), history (“70s” and “old school”), geography (“jamaican” and “brit pop”), or others (“romantic”). This assumption is thus quite unsatisfying, and one wonders whether tags present a good way to formally evaluate MGR systems.

Analyzing the same pathological behaviors of an MGR system, but by a listening test designed specifically to test the sensibility of its choices, circumvents the need to compare tags, and gets to the heart of whether the system is recognizing and comparing salient characteristics typical to genres, e.g., instrumentation, rhythm, form, and so on. Hence, we finally see through this that though AdaBFFs, SRCAM, and MAPsCAT have classification accuracies that are significantly higher than chance, and though each system has confusion tables that appear reasonable, a closer analysis of their confusions at the level of the music and a listening test measuring the appropriateness of their classifications, reveals that they are not *recognizing* genre since a large majority of their consistent misclassifications are easily detected as artificial.

Typically, formally justifying a misclassification as an error is a task MGR research often defers to the "ground truth" of a dataset, whether created by a listener [89], the artist [77], music vendors [6, 36], the collective agreement of several listeners [35, 52] professional musicologists [1], or multiple tags given by an online community [46]. However, the focus of developing an algorithm to pick the correct or best label actually obscures what should be the goal of any MGR system: to produce labels that are indistinguishable from those humans would produce. Hence, to this end, classification accuracy is not enough.

Acknowledgements Many thanks to: Carla Sturm for her bibliographic prowess; and Geraint Wiggens, Nick Collins, Matthew Davies, Fabien Gouyon, and Mark Plumbley for numerous and insightful conversations.

References

1. Abeßer, J., Lukashevich, H., Bräuer, P.: Classification of music genres based on repetitive basslines. J. New Music Research **41**(3), 239–257 (2012)
2. Ahrendt, P.: Music genre classification systems – a computational approach. Ph.D. thesis, Technical University of Denmark (2006)
3. Ahrendt, P., Larsen, J., Goutte, C.: Co-occurrence models in music genre classification. In: Proc. IEEE Workshop Machine Learning Signal Process. (2005)
4. Andén, J., Mallat, S.: Multiscale scattering for audio classification. In: Proc. ISMIR, pp. 657–662 (2011)
5. Andén, J., Mallat, S.: Scatterbox v. 1.02. http://www.cmap.polytechnique.fr/scattering/ (2012)
6. Ariyaratne, H., Zhang, D.: A novel automatic hierachical approach to music genre classification. In: Proc. ICME, pp. 564 –569 (2012)
7. Aucouturier, J.J., Pachet, F.: Representing music genre: A state of the art. J. New Music Research **32**(1), 83–93 (2003)
8. Aucouturier, J.J., Pampalk, E.: Introduction – from genres to tags: A little epistemology of music information retrieval research. J. New Music Research **37**(2), 87–92 (2008)
9. Barbedo, J.G.A., Lopes, A.: Automatic genre classification of musical signals. EURASIP J. Adv. Sig. Process. (2007)
10. Benbouzid, D., Busa-Fekete, R., Casagrande, N., Collin, F.D., Kégl, B.: Multiboost: a multi-purpose boosting package. J. Machine Learning Res. **13**, 549–553 (2012)
11. Benetos, E., Kotropoulos, C.: A tensor-based approach for automatic music genre classification. In: Proc. EUSIPCO. Lausanne, Switzerland (2008)
12. Berenzweig, A., Logan, B., Ellis, D.P.W., Whitman, B.: A large-scale evaluation of acoustic and subjective music-similarity measures. Computer Music J. **28**(2), 63–76 (2004)
13. van den Berg, E., Friedlander, M.P.: Probing the Pareto frontier for basis pursuit solutions. SIAM J. on Scientific Computing **31**(2), 890–912 (2008)
14. Bergstra, J., Casagrande, N., Erhan, D., Eck, D., Kégl, B.: Aggregate features and AdaBoost for music classification. Machine Learning **65**(2-3), 473–484 (2006)
15. Bergstra, J., Mandel, M., Eck, D.: Scalable genre and tag prediction with spectral covariance. In: Proc. ISMIR (2010)

16. Bertin-Mahieux, T., Ellis, D.P., Whitman, B., Lamere, P.: The million song dataset. In: Proc. ISMIR (2011). URL `http://labrosa.ee.columbia.edu/millionsong/`
17. Burred, J.: Hierarchical automatic audio signal classification. J. Audio Engineering Society **52**(7), 724–739 (2004)
18. Chai, W., Vercoe, B.: Folk music classification using hidden markov modelts. In: Int. Conf. on Artificial Intel. (2001)
19. Chang, K., Jang, J.S.R., Iliopoulos, C.S.: Music genre classification via compressive sampling. In: Proc. ISMIR, pp. 387–392. Amsterdam, The Netherlands (2010)
20. Chase, A.: Music discriminations by carp "cyprinus carpio". Learning & Behavior **29**, 336–353 (2001)
21. Chen, S.H., Chen, S.H.: Content-based music genre classification using timbral feature vectors and support vector machine. In: Proc. Int. Conf. Interaction Sciences, pp. 1095–1101 (2009)
22. Craft, A.: The role of culture in the music genre classification task: human behaviour and its effect on methodology and evaluation. Tech. rep., Queen Mary University of London (2007)
23. Craft, A., Wiggins, G.A., Crawform, T.: How many beans make five? The consensus problem in music-genre classification and a new evaluation method for single-genre categorisation systems. In: Proc. ISMIR (2007)
24. Cruz, P.P., Vidal, E.: Two grammatical inference applications in music processing. Applied Artificial Intel. **22**(1/2), 53–76 (2008)
25. DeCoro, C., Barutcuoglu, S., Fiebrink, R.: Bayesian aggregation for hierarchical genre classification. In: Proc. ISMIR (2007)
26. Deshpande, H., Singh, R., Nam, U.: Classification of music signals in the visual domain. In: Proc. DAFx. Limerick, Ireland (2001)
27. Dixon, S., Pampalk, E., Widmer, G.: Classification of dance music by periodicity patterns. In: Proc. ISIMIR (2003)
28. Fabbri, F.: A theory of musical genres: Two applications. In: Proc. Int. Conf. Popular Music Studies. Amsterdam, The Netherlands (1980)
29. Flexer, A.: Statistical evaluation of music information retrieval experiments. J. New Music Research **35**(2), 113–120 (2006)
30. Flexer, A.: A closer look on artist filters for musical genre classification. In: Proc. ISMIR. Vienna, Austria (2007)
31. Flexer, A., Schnitzer, D.: Album and artist effects for audio similarity at the scale of the web. In: Proc. SMC, pp. 59–64. Porto, Portugal (2009)
32. Freund, Y., Schapire, R.E.: A decision-theoretic generalization of on-line learning and an application to boosting. J. Computer System Sci. **55**, 119–139 (1997)
33. Frow, J.: Genre. Routledge, New York, NY, USA (2005)
34. Fu, Z., Lu, G., Ting, K.M., Zhang, D.: A survey of audio-based music classification and annotation. IEEE Trans. Multimedia **13**(2), 303–319 (2011)
35. García, J., Hernández, E., Meng, A., Hansen, L.K., Larsen, J.: Discovering music structure via similarity fusion. In: Proc. Music, Brain and Cognition Workshop (2007)
36. Gjerdingen, R.O., Perrott, D.: Scanning the dial: The rapid recognition of music genres. J. New Music Research **37**(2), 93–100 (2008)
37. Goto, M., Hashiguchi, H., Nishimura, T., Oka, R.: RWC music database: Music genre database and musical instrument sound database. In: Proc. ISMIR (2003)
38. Guaus, E.: Audio content processing for automatic music genre classification: descriptors, databases, and classifiers. Ph.D. thesis, Universitat Pompeu Fabra, Barcelona, Spain (2009)
39. Holzapfel, A., Stylianou, Y.: Musical genre classification using nonnegative matrix factorization-based features. IEEE Trans. Audio, Speech and Lang. Process. **16**(2), 424–434 (2008)
40. Homburg, H., Mierswa, I., Möller, B., Morik, K., Wurst, M.: A benchmark dataset for audio classification and clustering. In: Proc. ISMIR. London, U.K. (2005)
41. ISMIR: Genre results. `http://ismir2004.ismir.net/genre_contest/index.htm` (2004)
42. ISMIS: Genre results. `http://tunedit.org/challenge/music-retrieval` (2011)
43. Kim, Y.E., Williamson, D.S., Pilli, S.: Towards quantifying the "album effect" in artist identification. In: ISMIR, pp. 393–394 (2006)
44. Krumhansl, C.L.: Plink: "thin slices" of music. Music Perception: An Interdisciplinary Journal **27**(5), 337–354 (2010)
45. Langlois, T., Marques, G.: Automatic music genre classification using a hierarchical clustering and a language model approach. In: Proc. Int. Conf. Advances in Multimedia (2009)
46. Law, E.: Human computation for music classification. In: T. Li, M. Ogihara, G. Tzanetakis (eds.) Music Data Mining, pp. 281–301. CRC Press (2011)

47. Lee, J.W., Park, S.B., Kim, S.K.: Music genre classification using a time-delay neural network. In: J. Wang, Z. Yi, J. Zurada, B.L. Lu, H. Yin (eds.) Advances in Neural Networks, pp. 178–187. Springer Berlin / Heidelberg (2006). URL `http://dx.doi.org/10.1007/11760023_27`
48. Lerch, A.: An Introduction to Audio Content Analysis: Applications in Signal Processing and Music Informatics. Wiley/IEEE Press, Hoboken, New York (2012)
49. Lidy, T.: Marsyas and rhythm patterns: Evaluation of two music genre classification systems. In: Proc. Workshop Data Anal. (2003)
50. Lidy, T., Rauber, A., Pertusa, A., Iñesta, J.M.: Improving genre classification by combination of audio and symbolic descriptors using a transcription system. In: Proc. ISMIR, pp. 61–66. Vienna, Austria (2007)
51. Lin, C.R., Liu, N.H., Wu, Y.H., Chen, A.: Music classification using significant repeating patterns. In: Y. Lee, J. Li, K.Y. Whang, D. Lee (eds.) Database Systems for Advanced Applications, pp. 27–29. Springer Berlin / Heidelberg (2004)
52. Lippens, S., Martens, J., De Mulder, T.: A comparison of human and automatic musical genre classification. In: Proc. ICASSP, pp. 233–236 (2004)
53. Lopes, M., Gouyon, F., Koerich, A., Oliveira, L.E.S.: Selection of training instances for music genre classification. In: Proc. ICPR. Istanbul, Turkey (2010)
54. Lukashevich, H., Abeßer, J., Dittmar, C., Großmann, H.: From multi-labeling to multi-domain-labeling: A novel two-dimensional approach to music genre classification. In: ISMIR (2009)
55. Mace, S.T., Wagoner, C.L., Teachout, D.J., Hodges, D.A.: Genre identification of very brief musical excerpts. Psychology of Music **40**(1), 112–128 (2011)
56. Mallat, S.: Group invariant scattering. Comm. Pure Appl. Math. (2012 (to appear))
57. McKay, C.: Automatic music classification with jMIR. Ph.D. thesis, McGill University, Montréal, Canada (2010)
58. McKay, C., Fujinaga, I.: Music genre classification: Is it worth pursuing and how can it be improved? In: Proc. ISMIR. Victoria, Canada (2006)
59. McKay, C., Fujinaga, I.: Combining features extracted from audio, symbolic and cultural sources. In: Proc. ISMIR, pp. 597–602 (2008)
60. Meng, A., Ahrendt, P., Larsen, J.: Improving music genre classification by short-time feature integration. In: Proc. ICASSP, pp. 497–500. Philadelphia, PA (2005)
61. Meng, A., Shawe-Taylor, J.: An investigation of feature models for music genre classification using the support vector classifier. In: Proc. ISMIR (2008)
62. MIREX: Genre results. `http://www.music-ir.org/mirex/wiki/2005:MIREX2005_Results` (2005)
63. Pampalk, E.: Computational models of music similarity and their application in music information retrieval. Ph.D. thesis, Vienna University of Tech., Vienna, Austria (2006)
64. Pampalk, E., Flexer, A., Widmer, G.: Improvements of audio-based music similarity and genre classification. In: Proc. ISMIR, pp. 628–233. London, U.K. (2005)
65. Panagakis, Y., Kotropoulos, C.: Music genre classification via topology preserving non-negative tensor factorization and sparse representations. In: Proc. ICASSP, pp. 249–252 (2010)
66. Panagakis, Y., Kotropoulos, C., Arce, G.R.: Music genre classification using locality preserving non-negative tensor factorization and sparse representations. In: Proc. ISMIR, pp. 249–254. Kobe, Japan (2009)
67. Panagakis, Y., Kotropoulos, C., Arce, G.R.: Music genre classification via sparse representations of auditory temporal modulations. In: Proc. EUSIPCO. Glasgow, Scotland (2009)
68. Panagakis, Y., Kotropoulos, C., Arce, G.R.: Non-negative multilinear principal component analysis of auditory temporal modulations for music genre classification. IEEE Trans. Acoustics, Speech, Lang. Process. **18**(3), 576–588 (2010)
69. Porter, D., Neuringer, A.: Music discriminations by pigeons. Experimental Psychology: Animal Behavior Processes **10**(2), 138–148 (1984)
70. Ren, J.M., Jang, J.S.R.: Time-constrained sequential pattern discovery for music genre classification. In: Proc. ICASSP, pp. 173–176 (2011)
71. Ren, J.M., Jang, J.S.R.: Discovering time-constrained sequential patterns for music genre classification. IEEE Trans. Audio, Speech, and Lang. Process. **20**(4), 1134–1144 (2012)
72. Rizzi, A., Buccino, N.M., Panella, M., Uncini, A.: Genre classification of compressed audio data. In: Proc. Int. Workshop on Multimedia Signal Process. (2008)
73. Scaringella, N., Zoia, G., Mlynek, D.: Automatic genre classification of music content: A survey. IEEE Signal Process. Mag. **23**(2), 133–141 (2006)

74. Schapire, R., Singer, Y.: Improved boosting algorithms using confidence-rated predictions. Machine Learning **37**(3), 297–336 (1999)
75. Schindler, A., Mayer, R., Rauber, A.: Facilitating comprehensive benchmarking experiments on the million song dataset. In: Proc. ISMIR (2012)
76. Seyerlehner, K.: Content-based music recommender systems: Beyond simple frame-level audio similarity. Ph.D. thesis, Johannes Kepler University, Linz, Austria (2010)
77. Seyerlehner, K., Widmer, G., Knees, P.: A comparison of human, automatic and collaborative music genre classification and user centric evaluation of genre classification systems. In: M. Detyniecki, P. Knees, A. Nürnberger, M. Schedl, S. Stober (eds.) Adaptive Multimedia Retrieval. Context, Exploration, and Fusion, pp. 118–131. Springer Berlin / Heidelberg (2011)
78. Shapiro, P.: Turn the Beat Around: The Secret History of Disco. Faber & Faber, London, U.K. (2005)
79. Silla, C.N., Koerich, A.L., Kaestner, C.A.A.: The latin music database. In: Proc. ISMIR (2008)
80. Slaney, M.: Auditory toolbox. Tech. rep., Interval Research Corporation (1998)
81. Song, W., Chang, C.J., Liou, S.: Improved confidence intervals on the bernoulli parameter. Communications and Statistics Theory and Methods **38**(19), 3544–3560 (2009)
82. Sturm, B.L.: An analysis of the GTZAN music genre dataset. In: Proc. ACM MIRUM Workshop. Nara, Japan (2012)
83. Sturm, B.L.: A survey of evaluation in music genre recognition. In: Proc. Adaptive Multimedia Retrieval. Copenhagen, Denmark (2012)
84. Sturm, B.L.: Two systems for automatic music genre recognition: What are they really recognizing? In: Proc. ACM MIRUM Workshop. Nara, Japan (2012)
85. Sturm, B.L., Noorzad, P.: On automatic music genre recognition by sparse representation classification using auditory temporal modulations. In: Proc. CMMR. London, UK (2012)
86. Sundaram, S., Narayanan, S.: Experiments in automatic genre classification of full-length music tracks using audio activity rate. In: Proc. IEEE Workshop Multimedia Signal Process. (2007)
87. Theodoridis, S., Koutroumbas, K.: Pattern Recognition, 4 edn. Academic Press, Elsevier, Amsterdam, The Netherlands (2009)
88. Turnbull, D., Elkan, C.: Fast recognition of musical genres using RBF networks. IEEE Trans. Knowl. Data Eng. **17**(4), 580–584 (2005)
89. Tzanetakis, G., Cook, P.: Musical genre classification of audio signals. IEEE Trans. Speech Audio Process. **10**(5), 293–302 (2002)
90. Tzanetakis, G., Ermolinskyi, A., Cook, P.: Pitch histograms in audio and symbolic music information retrieval. J. New Music Research **32**(2), 143–152 (2003)
91. Umapathy, K., Krishnan, S., Jimaa, S.: Multigroup classification of audio signals using time-frequency parameters. IEEE Trans. Multimedia **7**(2), 308–315 (2005)
92. Vatolkin, I.: Multi-objective evaluation of music classification. In: W.A. Gaul, A. Geyer-Schulz, L. Schmidt-Thieme, J. Kunze (eds.) Challenges at the Interface of Data Analysis, Computer Science, and Optimization, pp. 401–410. Springer Berlin (2012)
93. Wiggins, G.A.: Semantic gap?? Schemantic schmap!! Methodological considerations in the scientific study of music. In: Proc. IEEE Int. Symp. Mulitmedia, pp. 477–482 (2009)
94. Wright, J., Yang, A.Y., Ganesh, A., Sastry, S.S., Ma, Y.: Robust face recognition via sparse representation. IEEE Trans. Pattern Anal. Machine Intell. **31**(2), 210–227 (2009)
95. Wu, M.J., Chen, Z.S., Jang, J.S.R., Ren, J.M.: Combining visual and acoustic features for music genre classification. In: Int. Conf. Machine Learning and Applications (2011)
96. Yao, Q., Li, H., Sun, J., Ma, L.: Visualized feature fusion and style evaluation for musical genre analysis. In: Int. Conf. Pervasive Computing, Signal Process. and App. (2010)