J Intell Inf Syst (2013) 41:371–406
DOI 10.1007/s10844-013-0250-y

Classification accuracy is not enough

On the evaluation of music genre recognition systems

Bob L. Sturm

Received: 12 November 2012 / Revised: 10 March 2013 / Accepted: 14 May 2013 /
Published online: 14 July 2013

Abstract We argue that an evaluation of system behavior at the level of the music is required to usefully address the fundamental problems of music genre recognition (MGR), and indeed other tasks of music information retrieval, such as autotagging. A recent review of works in MGR since 1995 shows that most (82 %) measure the capacity of a system to recognize genre by its classification accuracy. After reviewing evaluation in MGR, we show that neither classification accuracy, nor recall and precision, nor confusion tables, necessarily reflect the capacity of a system to recognize genre in musical signals. Hence, such figures of merit cannot be used to reliably rank, promote or discount the genre recognition performance of MGR systems *if* genre recognition (rather than identification by irrelevant confounding factors) is the objective. This motivates the development of a richer experimental toolbox for evaluating any system designed to intelligently extract information from music signals.

Keywords Music · Evaluation · Classification · Genre

1 Introduction

The problem of identifying, discriminating between, and learning the criteria of music genres or styles—music genre recognition (MGR)—has motivated much work since 1995 (Matityaho and Furst 1995), and even earlier, e.g., Porter and Neuringer (1984). Indeed, a recent review of MGR by Fu et al. (2011) writes,

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, AD:MT, Aalborg University Copenhagen,
A.C. Meyers Vænge 15, 2450, Copenhagen SV, Denmark
e-mail: bst@create.aau.dk

"Genre classification is the most widely studied area in [music information retrieval]." MGR research is now making an appearance in textbooks (Lerch 2012). Most published studies of MGR systems report classification performance significantly better than chance, and sometimes as well as or better than humans. For a benchmark dataset of music excerpts singly-labeled in ten genres (GTZAN, Tzanetakis and Cook 2002; Sturm 2013b), reported classification accuracies have risen from 61 % (Tzanetakis and Cook 2002) to above 90 %, e.g., Guaus (2009), Panagakis et al. (2009a, b), Panagakis and Kotropoulos (2010) and Chang et al. (2010). Indeed, as Bergstra et al. (2006a) write, "Given the steady and significant improvement in classification performance since 1997, we wonder if automatic methods are not already more efficient at learning genres than some people." This performance increase merits a closer look at what is working in these systems, and motivates re-evaluating the argument that genre exists to a large extent outside of the acoustic signal itself (Fabbri 1982; McKay and Fujinaga 2006; Wiggins 2009). Most exciting, it might also illuminate how people hear and conceptualize the complex phenomenon of "music" (Aucouturier and Bigand 2013). It might be too soon to ask such questions, however.

Recent work (Sturm 2012b; Marques et al. 2010, 2011a) shows that an MGR system can act as if genre is not what it is recognizing, even if it shows high classification accuracy. In a comprehensive review of the MGR literature (Sturm 2012a), we find that over 91 % of papers with an experimental component (397 of 435 papers) evaluate MGR systems by classifying music excerpts and comparing the labels to the "ground truth," and over 82 % of 467 published works cite classification accuracy as a figure of merit (FoM). Of those that employ this approach to evaluation, 47 % employ only this approach. Furthermore, we find several cases of methodological errors leading to inflated accuracies: those of Panagakis et al. (2009a, b) and Panagakis and Kotropoulos (2010) come from accidentally using the true labels in classification (private correspondence with Y. Panagakis) (Sturm and Noorzad 2012); those of Chang et al. (2010), are irreproducible, and contradict results seen in other areas applying the same technique (Sturm 2013a); and those of Bağci and Erzin (2007) are highly unlikely with an analysis of their approach (Sturm and Gouyon 2013, unpublished). One must wonder if the "progress" in MGR seen since 1995 is not due to solving the problem: can a system have a high classification accuracy in some datasets yet not even address the problem at all?

We show here that classification accuracy does not reliably reflect the capacity of an MGR system to recognize genre. Furthermore, recall, precision and confusion tables are still not enough. We show these FoMs—all of which have been used in the past to rank MGR systems, e.g., Chai and Vercoe (2001), Tzanetakis and Cook (2002), Aucouturier and Pachet (2003), Burred and Lerch (2004), Turnbull and Elkan (2005), Flexer (2006), DeCoro et al. (2007), Benetos and Kotropoulos (2008), Panagakis et al. (2009b), Bergstra et al. (2010), Fu et al. (2011) and Ren and Jang (2012) citing one work from each year since 2001—do not reliably reflect the capacity of an MGR system to recognize genre. While these claims have not been made overt in any of the 467 references we survey (Sturm 2012a), shades of it have appeared before (Craft et al. 2007; Craft 2007; Lippens et al. 2004; Wiggins 2009; Seyerlehner et al. 2010; Sturm 2012b), which argue for evaluating performance in ways that account for the ambiguity of genre being in large part a subjective construction

(Fabbri 1982; Frow 2005). We go further and argue that the evaluation of MGR systems—the experimental designs, the datasets, and the FoMs—and indeed, the development of future systems, must embrace the fact that the recognition of genre is to a large extent a *musical* problem, and must be evaluated as such. In short, *classification accuracy is not enough* to evaluate the extent to which an MGR system addresses what appears to be one of its principal goals: to produce genre labels indistinguishable from those humans would produce.

1.1 Arguments

Some argue that since MGR is now replaced by, or is a subproblem of, the more general problem of automatic tagging (Aucouturier and Pampalk 2008; Bertin-Mahieux et al. 2010), work in MGR is irrelevant. However, genre is one of the most used descriptors of music (Aucouturier and Pachet 2003; Scaringella et al. 2006; McKay and Fujinaga 2006): in 2007, nearly 70 % of the tags on `last.fm` are genre labels (Bertin-Mahieux et al. 2010); and a not insignificant portion of the tags in the Million Song Dataset are genre (Bertin-Mahieux et al. 2011; Schindler et al. 2012). Some argue that automatic tagging is more realistic than MGR because multiple tags can be given rather than the single one in MGR, e.g., Panagakis et al. (2010b), Marques et al. (2011a), Fu et al. (2011) and Seyerlehner et al. (2012). This claim and its origins are mysterious because nothing about MGR—the problem of identifying, discriminating between, and learning the criteria of music genres or styles—naturally restricts the number of genre labels people use to describe a piece of music. Perhaps this imagined limitation of MGR comes from the fact that of 435 works with an experimental component we survey (Sturm 2012a), we find only ten that use a multilabel approach (Barbedo and Lopes 2008; Lukashevich et al. 2009; Mace et al. 2011; McKay 2004; Sanden 2010; Sanden and Zhang 2011a, b; Scaringella et al. 2006; Tacchini and Damiani 2011; Wang et al. 2009). Perhaps it comes from the fact that most of the private and public datasets so far used in MGR assume a model of one genre per musical excerpt (Sturm 2012a). Perhaps it comes from the assumption that genre works in such a way that an object *belongs* to a genre, rather than *uses* a genre (Frow 2005).

Some argue that, given the ambiguity of genre and the observed lack of human consensus about such matters, MGR is an ill-posed problem (McKay and Fujinaga 2006). However, people often do agree, even under surprising constraints (Gjerdingen and Perrott 2008; Krumhansl 2010; Mace et al. 2011). Researchers have compiled MGR datasets with validation from listening tests, e.g., (Lippens et al. 2004; Meng et al. 2005); and very few researchers have overtly argued against any of the genre assignments of the most-used public dataset for MGR (Sturm 2012a, 2013b). Hence, MGR does not always appear to be an ill-posed problem since people often use genre to describe and discuss music in consistent ways, and that, not to forget, MGR makes no restriction on the number of genres relevant for describing a particular piece of music. Some argue that though people show some consistency in using genre, they are making decisions based on information not present in the audio signal, such as composer intention or marketing strategies (McKay and Fujinaga 2006; Bergstra et al. 2006b; Wiggins 2009). However, there exist some genres or styles that appear distinguishable and identifiable from the sound, e.g., musicological criteria like tempo (Gouyon and Dixon 2004), chord progressions (Anglade et al. 2010), instrumentation (McKay and Fujinaga 2005), lyrics (Li and Ogihara 2004), and so on.

Some argue that MGR is really just a proxy problem that has little value in and of itself; and that the purpose of MGR is really to provide an efficient means to gauge the performance of features and algorithms solving the problem of measuring music similarity (Pampalk 2006; Schedl and Flexer 2012). This point of view, however, is not evident in much of the MGR literature, e.g., the three reviews devoted specifically to MGR (Aucouturier and Pachet 2003; Scaringella et al. 2006; Fu et al. 2011), the work of Tzanetakis and Cook (2002), Barbedo and Lopes (2008), Bergstra et al. (2006a), Holzapfel and Stylianou (2008), Marques et al. (2011b), Panagakis et al. (2010a), Benetos and Kotropoulos (2010), and so on. It is thus not idiosyncratic to claim that one purpose of MGR could be to identify, discriminate between, and learn the criteria of music genres in order to produce genre labels that are indistinguishable from those humans would produce. One might argue, "MGR does not have much value since most tracks today are already annotated with genre." However, genre is not a fixed attribute like artist or instrumentation (Fabbri 1982; Frow 2005); and it is certainly not an attribute of only commercial music infallibly ordained by composers, producers, and/or consumers using perfect historical and musicological reflection. One cannot assume such metadata are static and unquestionable, or that even such information is useful, e.g., for computational musicology (Collins 2012).

Some might argue that the reasons MGR work is still published is that: 1) it provides a way to evaluate new features; and 2) it provides a way to evaluate new approaches to machine learning. While such a claim about publication is tenuous, we argue that it makes little sense to evaluate features or machine learning approaches without considering for what they are to be used, and then designing and using appropriate procedures for evaluation. We show in this paper that the typical ways in which new features and machine learning methods are evaluated *for* MGR provide little information about the extents to which the features and machine learning *for* MGR address the fundamental problem of *recognizing* music genre.

1.2 Organization and conventions

We organize this article as follows. Section 2 distills along three dimensions the variety of approaches that have been used to evaluate MGR: experimental design, datasets, and FoMs. We delimit our study to work specifically addressing the recognition of music genre and style, and not tags in general, i.e., the 467 works we survey (Sturm 2012a). We show most work in MGR reports classification accuracy from a comparison of predicted labels to "ground truths" of private datasets. The third section reviews three state-of-the-art MGR systems that show high classification accuracy in the most-used public music genre dataset GTZAN (Tzanetakis and Cook 2002; Sturm 2013b). In the fourth section, we evaluate the performance statistics of these three systems, starting from high-level FoMs such as classification accuracy, recall and precision, continuing to mid-level class confusions. In the fifth section, we evaluate the behaviors of these systems by inspecting low-level excerpt misclassifications, and performing a listening test that proves the behaviors of all three systems are highly distinguishable from those of humans. We conclude by discussing our results and further criticisms, and a look forward to the development and practice of better means for evaluation, not only in MGR, but also the more general problem of music description.

We use the following conventions throughout. When we refer to Disco, we are referring to those 100 excerpts in the GTZAN category named "Disco" without advocating that they are exemplary of the genre disco. The same applies for the excerpts of the other nine categories of GTZAN. We capitalize the *categories* of GTZAN, e.g., Disco, capitalize and quote *labels*, e.g., "Disco," but do not capitalize *genres*, e.g., disco. A number following a category in GTZAN refers to the identifying number of its excerpt filename. All together, "it appears this system does not recognize disco because it classifies Disco 72 as 'Metal'."

2 Evaluation in music genre recognition research

Surprisingly little has been written about evaluation, i.e., experimental design, data, and FoMs, with respect to MGR (Sturm 2012a). An experimental design is a method for testing a hypothesis. Data is the material on which a system is tested. A FoM reflects the confidence in the hypothesis after conducting an experiment. Of three review articles devoted in large part to MGR (Aucouturier and Pachet 2003; Scaringella et al. 2006; Fu et al. 2011), only Aucouturier and Pachet (2003) give a brief paragraph on evaluation. The work by Vatolkin (2012) provides a comparison of various performance statistics for music classification. Other works (Berenzweig et al. 2004; Craft et al. 2007; Craft 2007; Lippens et al. 2004; Wiggins 2009; Seyerlehner et al. 2010; Sturm 2012b) argue for measuring performance in ways that take into account the natural ambiguity of music genre and similarity. For instance, we Sturm (2012b), Craft et al. (2007) and Craft (2007) argue for richer experimental designs than having a system apply a single label to music with a possibly problematic "ground truth." Flexer (2006) criticizes the absence of formal statistical testing in music information research, and provides an excellent tutorial based upon MGR for how to apply statistical tests. Derived from our survey (Sturm 2012a), Fig. 1 shows the annual number of publications in MGR, and the proportion that use formal statistical testing in comparing MGR systems.

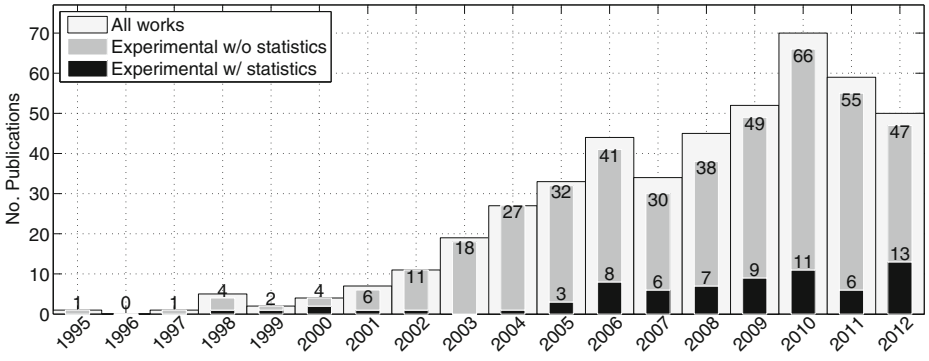


Fig. 1 Annual numbers of references in MGR divided by which use and do not use formal statistical tests for making comparisons (Sturm 2012a). Only about 12 % of references in MGR employ formal statistical testing; and only 19.4 % of the work (91 papers) appears at the Conference of the International Society for Music Information Retrieval

Table 1 Ten experimental designs of MGR, and the percentage of references having an experimental component (435 references) in our survey (Sturm 2012a) that employ them

Design	Description	%
Classify	To answer the question, "How well does the system predict the genres used by music?" The system applies genre labels to music, which researcher then compares to a "ground truth"	91
Features	To answer the question, "At what is the system looking to identify the genres used by music?" The system ranks and/or selects features, which researcher then inspects	33
Generalize	To answer the question, "How well does the system identify genre in varied datasets?" *Classify* with two or more datasets having different genres, and/or various amounts of training data	16
Robust	To answer the question, "To what extent is the system invariant to aspects inconsequential for identifying genre?" The system classifies music that researcher modifies or transforms in ways that do not harm its genre identification by a human	7
Eyeball	To answer the question, "How well do the parameters make sense with respect to identifying genre?" The system derives parameters from music; researcher visually compares	7
Cluster	To answer the question, "How well does the system group together music using the same genres?" The system creates clusters of dataset, which researcher then inspect	7
Scale	To answer the question, "How well does the system identify music genre with varying numbers of genres?" *Classify* with varying numbers of genres	7
Retrieve	To answer the question, "How well does the system identify music using the same genres used by the query?" The system retrieves music similar to query, which researcher then inspects	4
Rules	To answer the question, "What are the decisions the system is making to identify genres?" The researcher inspects rules used by a system to identify genres	4
Compose	To answer the question, "What are the internal genre models of the system?" The system creates music in specific genres, which the researcher then inspects	0.7

Some references use more than one design

Table 1 summarizes ten experimental designs we find in our survey (Sturm 2012a). Here we see that the most widely used design by far is *Classify*. The experimental design used the least is *Compose*, and appears in only three works (Cruz and Vidal 2003, 2008; Sturm 2012b). Almost half of the works we survey (213 references), uses only one experimental design; and of these, 47 % employ *Classify*. We find only 36 works explicitly mention evaluating with an artist or album filter (Pampalk et al. 2005; Flexer 2007; Flexer and Schnitzer 2009, 2010). We find only 12 works using human evaluation for gauging the success of a system.

Typically, formally justifying a misclassification as an error is a task research in MGR often defers to the "ground truth" of a dataset, whether created by a listener (Tzanetakis and Cook 2002), the artist (Seyerlehner et al. 2010), music vendors (Gjerdingen and Perrott 2008; Ariyaratne and Zhang 2012), the collective agreement of several listeners (Lippens et al. 2004; García et al. 2007) professional musicologists (Abeßer et al. 2012), or multiple tags given by an online community (Law 2011). Table 2 shows the datasets used by references in our survey (Sturm 2012a). Overall, 79 % of this work uses audio data or features derived from audio data, about 19 %

Table 2 Datasets used in MGR, the type of data they contain, and the percentage of experimental work (435 references) in our survey (Sturm 2012a) that use them

Dataset	Description	%
Private	Constructed for research but not made available	58
GTZAN	Audio; http://marsyas.info/download/data_sets	23
ISMIR2004	Audio; http://ismir2004.ismir.net/genre_contest	17
Latin (Silla et al. 2008)	Features; http://www.ppgia.pucpr.br/~silla/lmd/	5
Ballroom	Audio; http://mtg.upf.edu/ismir2004/contest/tempoContest/	3
Homburg (Homburg et al. 2005)	Audio; http://www-ai.cs.uni-dortmund.de/audio.html	3
Bodhidharma	Symbolic; http://jmir.sourceforge.net/Codaich.html	3
USPOP2002 (Berenzweig et al. 2004)	Audio; http://labrosa.ee.columbia.edu/projects/musicsim/uspop2002.html	2
1517-artists	Audio; http://www.seyerlehner.info	1
RWC (Goto et al. 2003)	Audio; http://staff.aist.go.jp/m.goto/RWC-MDB/	1
SOMeJB	Features; http://www.ifs.tuwien.ac.at/~andi/somejb/	1
SLAC	Audio & symbols; http://jmir.sourceforge.net/Codaich.html	1
SALAMI (Smith et al. 2011)	Features; http://ddmal.music.mcgill.ca/research/salami	0.7
Unique	Features; http://www.seyerlehner.info	0.7
Million song (Bertin-Mahieux et al. 2011)	Features; http://labrosa.ee.columbia.edu/millionsong/	0.7
ISMIS2011	Features; http://tunedit.org/challenge/music-retrieval	0.4

All datasets listed after *Private* are public

uses symbolic music data, and 6 % uses features derived from other sources, e.g., lyrics, the WWW, and album art. (Some works use more than one type of data.) About 27 % of work evaluates MGR systems using two or more datasets. While more than 58 % of the works uses datasets that are not publicly available, the most-used public dataset is GTZAN (Tzanetakis and Cook 2002; Sturm 2013b).

Table 3 shows the FoMs used in the works we survey (Sturm 2012a). Given *Classify* is the most-used design, it is not surprising to find mean accuracy appears

Table 3 Figures of merit (FoMs) of MGR, their description, and the percentage of work (467 references) in our survey (Sturm 2012a) that use them

FoM	Description	%
Mean accuracy	Proportion of the number of correct trials to the total number of trials	82
Confusion table	Counts of labeling outcomes for each labeled input	32
Recall	For a specific input label, proportion of the number of correct trials to the total number of trials	25
Confusions	Discussion of confusions of the system in general or with specifics	24
Precision	For a specific output label, proportion of the number of correct trials to the total number of trials	10
F-measure	Twice the product of *Recall* and *Precision* divided by their sum	4
Composition	Observations of the composition of clusters created by the system, distances within and between	4
Precision@k	Proportion of the number of correct items of a specific label in the k items retrieved	3
ROC	*Precision* vs. *Recall* (true positives vs. false positives) for several systems, parameters, etc.	1

the most often. When it appears, only about 25 % of the time is it accompanied by standard deviation (or equivalent). We find 6 % of the references report mean accuracy as well as recall and precision. Confusion tables are the next most prevalent FoM; and when one appears, it is not accompanied by any kind of musicological reflection about half the time. Of the works that use *Classify*, we find about 44 % of them report one FoM only, and about 53 % report more than one FoM. At least six works report human-weighted ratings of classification and/or clustering results.

One might argue that the evaluation above does not clearly reflect that most papers on automatic music tagging report recall, precision, and F-measures, and not mean accuracy. However, in our survey we do not consider work in automatic tagging unless part of the evaluation specifically considers the resulting genre tags. Hence, we see that most work in MGR uses classification accuracy (the experimental design *Classify* with mean accuracy as a FoM) in private datasets, or GTZAN (Tzanetakis and Cook 2002; Sturm 2013b).

3 Three state-of-the-art systems for music genre recognition

We now discuss three MGR systems that appear to perform well with respect to state of the art classification accuracy in GTZAN (Tzanetakis and Cook 2002; Sturm 2013b), and which we evaluate in later sections.

3.1 AdaBoost with decision trees and bags of frames of features (AdaBFFs)

AdaBFFs was proposed by Bergstra et al. (2006a), and performed the best in the 2005 MIREX MGR task (MIREX 2005). It combines weak classifiers trained by multiclass AdaBoost (Freund and Schapire 1997; Schapire and Singer 1999), which creates a strong classifier by counting "votes" of weak classifiers given observation $\mathbf{x}$. With the features in $\mathbb{R}^M$ of a training set labeled in K classes, iteration l adds a weak classifier $\mathbf{v}_l(\mathbf{x})$: $\mathbb{R}^M \to \{-1, 1\}^K$ and weight $w_l \in [0, 1]$ to minimize the total prediction error. A positive element means it favors a class, whereas negative means the opposite. After L training iterations, the classifier is the function $\mathbf{f}(\mathbf{x})$: $\mathbb{R}^M \to [-1, 1]^K$ defined

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

For an excerpt of recorded music consisting of a set of features $\mathscr{X} := \{\mathbf{x}_i\}$, AdaBFFs picks the class $k \in \{1, \ldots, K\}$ associated with the maximum element in the sum of weighted votes:

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

where $[\mathbf{a}]_k$ is the kth element of the vector $\mathbf{a}$.

We use the "multiboost package" (Benbouzid et al. 2012) with decision trees as the weak learners, and AdaBoost.MH (Schapire and Singer 1999) as the strong learner. The features we use are computed from a sliding Hann window of 46.4 ms and 50 % overlap: 40 Mel-frequency cepstral coefficients (MFCCs) (Slaney 1998), zero crossings, mean and variance of the magnitude Fourier transform (centroid and

spread), 16 quantiles of the magnitude Fourier transform (rolloff), 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 group the means and variances of each dimension. For a 30-s music excerpt, this produces 9 feature vectors of 120 dimensions. Bergstra et al. (2006a) report this approach obtains a classification accuracy of up to 83 % in GTZAN. In our reproduction of the approach (Sturm 2012b), we achieve using stumps (single node decision trees) as weak classifiers a classification accuracy of up to 77.6 % in GTZAN. We increase this to about 80 % by using two-node decision trees.

3.2 Sparse representation classification with auditory temporal modulations (SRCAM)

SRCAM (Panagakis et al. 2009b; Sturm and Noorzad 2012) uses sparse representation classification (Wright et al. 2009) in a dictionary composed of auditory features. This approach is reported to have classification accuracies above 90 % (Panagakis et al. 2009a, b; Panagakis and Kotropoulos 2010), but those results arise from a flaw in the experiment inflating accuracies from around 60 % (Sturm and Noorzad 2012) (private correspondence with Y. Panagakis). We modify the approach to produce classification accuracies above 80 % (Sturm 2012b). 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. To create a dictionary, we either *normalize* the set of features (mapping all values in each dimension to [0,1] by subtracting the minimum value and dividing by the largest difference), or *standardize* them (making all dimensions have zero mean and unit variance).

With the dictionary $\mathbf{D} := [\mathbf{d}_1|\mathbf{d}_2|\cdots|\mathbf{d}_N]$, and a mapping of columns to class identities $\bigcup_{k=1}^{K} \mathscr{I}_k = \{1,\ldots,N\}$, where $\mathscr{I}_k$ specifies the columns of $\mathbf{D}$ belonging to class k, SRCAM finds for a feature vector $\mathbf{x}'$ (which is the feature $\mathbf{x}$ we transform by the same normalization or standardization approach used to create the dictionary) a sparse representation $\mathbf{s}$ by solving

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

for a $\varepsilon^2 > 0$ we specify. SRCAM then defines the set of class-restricted weights $\{\mathbf{s}_k \in \mathbb{R}^N\}_{k\in\{1,\ldots,K\}}$

$$[\mathbf{s}_k]_n := \begin{cases} [\mathbf{s}]_n, & n \in \mathscr{I}_k \\ 0, & \text{else.} \end{cases} \tag{4}$$

Thus, $\mathbf{s}_k$ are the weights in $\mathbf{s}$ specific to class k. Finally, SRCAM classifies $\mathbf{x}$ by finding the class-dependent weights giving the smallest error

$$\hat{k}(\mathbf{x}') := \arg \min_{k\in\{1,\ldots,K\}} \|\mathbf{x}' - \mathbf{D}\mathbf{s}_k\|_2^2. \tag{5}$$

We define the confidence of SRCAM in assigning class k to $\mathbf{x}$ by comparing the errors:

$$C(k|\mathbf{x}) := \frac{\max_{k'} J_{k'} - J_k}{\sum_l [\max_{k'} J_{k'} - J_l]} \tag{6}$$

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 Maximum a posteriori classification of scattering coefficients (MAPsCAT)

MAPsCAT uses the novel features proposed in Mallat (2012), the use of which for MGR was first proposed by Andén and Mallat (2011). MAPsCAT applies these features within a Bayesian framework, which seeks to choose the class with minimum expected risk given 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 (Theodoridis and Koutroumbas 2009):

$$k^* = \arg \max_{k \in \{1,\ldots,K\}} P[\mathbf{x}|k] P(k) \tag{7}$$

where $P[\mathbf{x}|k]$ is the conditional model of the observations for class k, and $P(k)$ is a prior. MAPsCAT assumes $P[\mathbf{x}|k] \sim \mathcal{N}(\mu_k, \mathbf{C}_k)$, i.e., the observations from class k are distributed multivariate Gaussian with mean μ_k and covariance $\mathbf{C}_k$. MAPsCAT estimates these parameters using unbiased minimum mean-squared error estimation and the training set. When a music excerpt produces several features $\mathscr{X} := \{\mathbf{x}_i\}$, MAPsCAT assumes independence between them, and picks the class maximizing the sum of the log posteriors:

$$p_k(\mathscr{X}) := \log P(k) + \sum_{i=1}^{|\mathscr{X}|} \log P[\mathbf{x}_i|k]. \tag{8}$$

Scattering coefficients are attractive features for classification because they are designed to be invariant to particular transformations, such as translation and rotation, to preserve distances between stationary processes, and to embody both large- and short-scale structures (Mallat 2012). One computes these features by convolving the modulus of successive wavelet decompositions with the scaling wavelet. We use the "scatterbox" implementation (Andén and Mallat 2012) 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. Andén and Mallat (2011) report these features used with a support vector machine obtains a classification accuracy of 82 % in GTZAN. We obtain comparable results.

4 Evaluating the performance statistics of MGR systems

We now evaluate the performance of AdaBFFs, SRCAM and MAPsCAT using *Classify* and mean accuracy in GTZAN (Tzanetakis and Cook 2002). Despite the fact that GTZAN is a problematic dataset—it has many repetitions, mislabelings, and distortions (Sturm 2013b)—we use it for four reasons: 1) it is the public benchmark dataset most used in MGR research (Table 2); 2) it was used in the initial evaluation of AdaBFFs (Bergstra et al. 2006a), SRCAM (Panagakis et al. 2009b), and the features of MAPsCAT (Andén and Mallat 2011); 3) evaluations of MGR systems using GTZAN and other datasets show comparable performance, e.g., Moerchen et al. (2006), Ren and Jang (2012), Dixon et al. (2010), Schindler and Rauber (2012); and 4) since its contents and faults are now well-studied (Sturm 2013b), we can appropriately handle its problems, and in fact use them to our advantage.

We test each system by 10 trials of stratified 10-fold cross-validation (10×10 fCV). For each fold, we test all systems using the same training and testing data.

Every music excerpt is thus classified ten times by each system trained with the same data. For AdaBFFs, we run AdaBoost for 4000 iterations, and test both decision trees of two nodes or one 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 (van den Berg and Friedlander 2008) with at most 200 iterations. For MAPsCAT, we test systems trained with class-dependent covariances (each $\mathbf{C}_k$ can be different) or total covariance (all $\mathbf{C}_k$ the same). We define all priors to be equal. It might be that the size of this dataset is too small for some approaches. For instance, since for SRCAM one excerpt produces a 768-dimensional feature, we might not expect it to learn a good model from only 90 excerpts. However, we start as many have before: assume GTZAN is large enough and has enough integrity for evaluating an MGR system.

4.1 Evaluating classification accuracy

Table 4 shows classification accuracy statistics for two configurations of each system presented above. In their review of several MGR systems, Fu et al. (2011) compare the performance of several algorithms using only classification accuracy in GTZAN. The work proposing AdaBFFs (Bergstra et al. 2006a), SRCAM (Panagakis et al. 2009b), and the features of MAPsCAT (Andén and Mallat 2011), present only classification accuracy. Furthermore, based on classification accuracy, Seyerlehner et al. (2010) argue that the performance gap between MGR systems and humans is narrowing; and in this issue, Humphrey et al. conclude "progress in content-based music informatics is plateauing" (Humphrey et al. 2013). Figure 2 shows that with respect to the classification accuracies in GTZAN reported in 83 published works (Sturm 2013b), those of AdaBFFs, SRCAM, and MAPsCAT lie above what is reported best in half of this work. It is thus tempting to conclude from these that, with respect to the mean accuracy and its standard deviation, some configurations of these systems are better than others, that AdaBFFs is not as good as SRCAM and MAPsCAT, and that AdaBFFs, SRCAM, and MAPsCAT are recognizing genre better than at least half of the "competition".

These conclusions are unwarranted for at least three reasons. First, we cannot compare mean classification accuracies computed from 10×10 fCV because the samples are highly dependent (Dietterich 1996; Salzberg 1997). Hence, we cannot test a hypothesis of one system being better than another by using, e.g., a *t*-test, as we have erroneously done before (Sturm 2012b). Second, *Classify* is answering the question, "How well does the system predict a label assigned to a piece of data?"

Table 4 Mean accuracies in GTZAN for each system, and the maximum $\{p_i\}$ (9) over all 10 CV runs

System	System configuation	Mean acc., std. dev.	Max $\{p_i\}$
AdaBFFs	Decision stumps	0.776 ± 0.004	>0.024
	Two-node trees	0.800 ± 0.006	>0.024
SRCAM	Normalized features	0.835 ± 0.005	>0.024
	Standardized features	0.802 ± 0.006	>0.024
MAPsCAT	Class-dependent covariances	0.754 ± 0.004	$<10^{-6}$
	Total covariance	0.830 ± 0.004	$<10^{-6}$

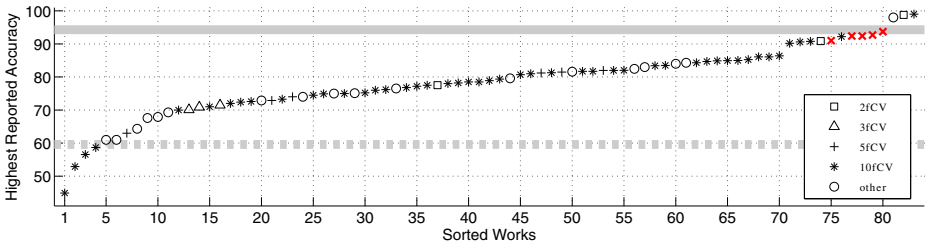


Fig. 2 Highest reported classification accuracies in GTZAN (Sturm 2013b). The legend shows evaluation parameters. *Top gray line* is the estimated maximum accuracy possible in GTZAN given its repetitions and mislabelings. The five "x" are results that are disputed, or known to be invalid. The *dashed gray line* is the accuracy we observe for SRCAM with normalized features and 2 fCV using an artist-filtered GTZAN without repetitions

Since many independent variables change between genre labels in GTZAN, and since *Classify* does nothing to deal with that, we cannot guard against confounds (Sturm 2012b; Urbano et al. 2013). This becomes clear when we see that artist filtering (Pampalk et al. 2005; Flexer 2007; Flexer and Schnitzer 2009, 2010) drops classification accuracy 20 points in GTZAN (Sturm 2013b). Thus, even if a system obtains the highest possible classification accuracy in GTZAN of 94.3 % (Sturm 2013b), we cannot reasonably say it is due to a capacity to recognize genre. Finally, since none of the results shown in Fig. 2 come from procedures that guard against confounds, we still cannot make meaningful comparisons between them.

We can, however, make some supportable conclusions about our systems and their configurations. For each of the systems in Table 4, we test for a significant difference in performance using a binomial hypothesis test (Salzberg 1997). For one run of 10 fCV, we define $c_{h>l}$ as the number of times the system with the high mean accuracy is correct and the other wrong; and $c_{h<l}$ as the number of times the system with the low mean accuracy is correct, and the other wrong. We define the random variable $C_{h>l}$ as that from which $c_{h>l}$ is a sample; and similarly for $C_{h<l}$. When the two systems perform equally well, we expect $C_{h>l} = C_{h<l}$, and each to be distributed Binomial with parameters $S = c_{h>l} + c_{h<l}$ and $q = 0.5$—assuming each of S trials is iid Bernoulli. Then, the probability the system with high mean accuracy performs better than the other given that they actually perform the same is

$$p = P[C_{h>l} \geq c_{h>l} | q = 0.5] = \sum_{s=c_{h>l}}^{c_{h>l}+c_{h<l}} \binom{c_{h>l} + c_{h<l}}{s} (0.5)^{c_{h>l}+c_{h<l}}. \quad (9)$$

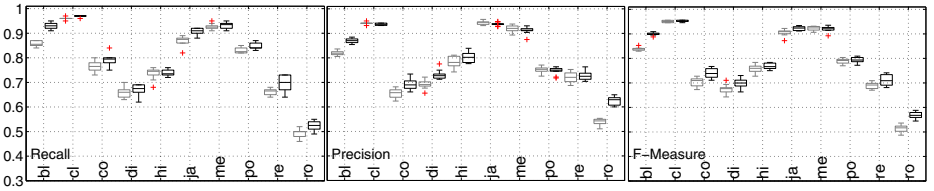
We define statistical significance as $\alpha = 0.025$ (one-tailed test). With the Bonferroni correction, we consider a result statistically significant if over all 10 CV runs, $\max\{p_i\} < \alpha/10$.

The last column of Table 4 shows that we can only reject the null hypothesis for the two configurations of MAPsCAT. In the same way, we test for significant differences between pairs of the systems showing the highest mean accuracy, e.g., SRCAM with normalized features and MAPsCAT with total covariance. Since we

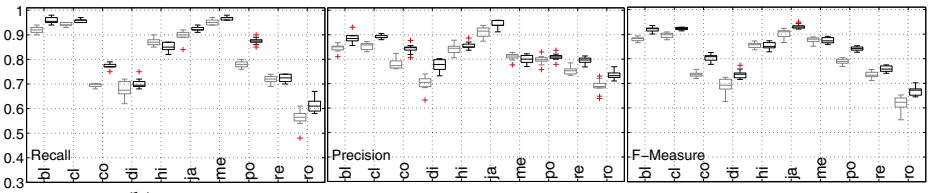
find no significant differences between any of these pairs, we fail to reject the null hypothesis that any performs better than another.

4.2 Evaluating performance in particular classes

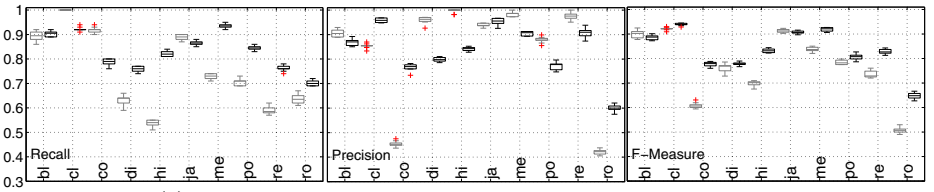
Figure 3 shows the recalls, precisions, and F-measures for AdaBFFs, SRCAM, and MAPsCAT. These FoMs, which appear infrequently in the MGR literature, can be more specific than mean accuracy, and provide a measure of how a system performs for specific classes. Wu et al. (2011) concludes on the relevance of their features to MGR by observing that the empirical recalls for Classical and Rock in GTZAN are above that expected for random. With respect to precision, Lin et al. (2004) concludes their system is better than another. We see in Fig. 3 for Disco that MAPsCAT using total covariance shows the highest recall (0.76 ± 0.01, std. dev.) of all systems. Since high recall can come at the price of many false positives, we look at the precision. MAPsCAT displays this characteristic for Country. When it comes to Classical, we see MAPsCAT using class-dependent covariance has perfect recall; and using class-dependent covariance it shows high precision (0.85 ± 0.01). The F-measure combines recall and precision to reflect class accuracy. We see that AdaBFFs appears to be one of the most accurate for Classical, and one of the least accurate for Disco.



(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. 3 Boxplots of recalls, precisions, and F-measures in 10×10-fold CV in GTZAN. Classes: Blues (*bl*), Classical (*cl*), Country (*co*), Disco (*di*), Hip hop (*hi*), Jazz (*ja*), Metal (*me*), Pop (*po*), Reggae (*re*), Rock (*ro*)

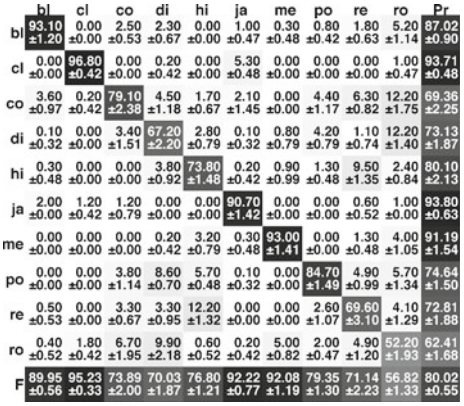
It is tempting to conclude that for these systems classical is quite easy to identify (high F-measure), and that this is reasonable since humans can easily identify such music as well. One might also be tempted to conclude that from looking at the F-measures between each system, AdaBFFs is the worst at identifying rock but the best at identifying classical. All of these are unfounded, however. First, our implementation of *Classify* is not testing whether classical is identifiable by these systems, but rather the extent to which they identify excerpts labeled "Classical" among the others in the dataset. Second, we cannot implicitly assume that whatever aspects humans use to identify classical music are the same used by these systems. Finally, just as the statistics of the accuracy in Table 4 come from dependent samples, we cannot simply compare these statistics over 10×10 fCV.

We can, however, test the null hypothesis that two systems perform equally well for identifying a particular class. We compare the performance of pairs of systems in each of the classes in each run of 10 fCV, i.e., we compute (9) but restricted to each class. We find that MAPsCAT with total covariance performs significantly better than MAPsCAT with class-dependent covariances for Hip hop and Metal ($p < 10^{-4}$). We are unable, however, to reject the null hypothesis for any of the systems having the highest mean accuracy in Table 4.

4.3 Evaluating confusions

Figure 4 shows the mean confusions, recalls (diagonal), precisions (right), F-measures (bottom), and accuracy (bottom right), for three of our systems. (We herein only consider the configuration that shows the highest mean classification accuracy in Table 4.) We find confusion tables reported in 32 % of MGR work (Sturm 2012a). Confusion tables are sometimes accompanied by a discussion of how a system appears to perform in ways that make sense with respect to what experience and musicology say about the variety of influences and characteristics shared between particular genres, e.g., Tzanetakis et al. (2003), Ahrendt (2006), Rizzi et al. (2008), Sundaram and Narayanan (2007), Abeßer et al. (2012), Yao et al. (2010), Ren and Jang (2011, 2012), Umapathy et al. (2005), Homburg et al. (2005), Tzanetakis and Cook (2002), Holzapfel and Stylianou (2008), Chen and Chen (2009). For instance, Tzanetakis and Cook (2002) write 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 results, Holzapfel and Stylianou (2008) write, "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." The human-like confusion tables found in MGR work, as well as the ambiguity between music genres, motivates evaluating MGR systems by considering as less troublesome the confusions we expect from humans (Craft et al. 2007; Craft 2007; Lippens et al. 2004; Seyerlehner et al. 2010).

In Fig. 4 we see of our systems the same kind of behaviors mentioned above: Rock is often labeled "Metal," and Metal is often labeled "Rock." We also see that no



	bl	cl	co	di	hi	ja	me	po	re	ro	Pr
bl	93.10 ±1.20	0.00 ±0.00	2.50 ±0.53	2.30 ±0.67	0.00 ±0.00	1.00 ±0.47	0.30 ±0.48	0.80 ±0.42	1.80 ±0.63	5.20 ±1.14	87.02 ±0.90
cl	0.00 ±0.00	96.80 ±0.42	0.00 ±0.00	0.20 ±0.42	0.00 ±0.00	5.30 ±0.48	0.00 ±0.00	0.00 ±0.00	0.00 ±0.00	1.00 ±0.47	93.71 ±0.48
co	3.60 ±0.97	0.20 ±0.42	79.10 ±2.38	4.50 ±1.18	1.70 ±0.67	2.10 ±1.45	0.00 ±0.00	4.40 ±1.17	6.30 ±0.82	12.20 ±1.75	69.36 ±2.25
di	0.10 ±0.32	0.00 ±0.00	3.40 ±1.51	67.20 ±2.20	2.80 ±0.79	0.10 ±0.32	0.80 ±0.79	4.20 ±0.79	1.10 ±0.74	12.20 ±1.40	73.13 ±1.87
hi	0.30 ±0.48	0.00 ±0.00	0.00 ±0.00	3.80 ±0.92	73.80 ±1.48	0.20 ±0.42	0.90 ±0.99	1.30 ±0.48	9.50 ±1.35	2.40 ±0.84	80.10 ±2.13
ja	2.00 ±0.00	1.20 ±0.42	1.20 ±0.79	0.00 ±0.00	0.00 ±0.00	90.70 ±1.42	0.00 ±0.00	0.00 ±0.00	0.60 ±0.52	1.00 ±0.00	93.80 ±0.63
me	0.00 ±0.00	0.00 ±0.00	0.00 ±0.00	0.20 ±0.42	3.20 ±0.79	0.30 ±0.48	93.00 ±1.41	0.00 ±0.00	1.30 ±0.48	4.00 ±1.05	91.19 ±1.54
po	0.00 ±0.00	0.00 ±0.00	3.80 ±1.14	8.60 ±0.70	5.70 ±0.48	0.10 ±0.32	0.00 ±0.00	84.70 ±1.49	4.90 ±0.99	5.70 ±1.34	74.64 ±1.50
re	0.50 ±0.53	0.00 ±0.00	3.30 ±0.67	3.30 ±0.95	12.20 ±1.32	0.00 ±0.00	0.00 ±0.00	2.60 ±1.07	69.60 ±3.10	4.10 ±1.29	72.81 ±1.88
ro	0.40 ±0.52	1.80 ±0.42	6.70 ±1.95	9.90 ±2.18	0.60 ±0.52	0.20 ±0.42	5.00 ±0.82	2.00 ±0.47	4.90 ±1.20	52.20 ±1.93	62.41 ±1.68
F	89.95 ±0.56	95.23 ±0.33	73.89 ±2.00	70.03 ±1.87	76.80 ±1.21	92.22 ±0.77	92.08 ±1.19	79.35 ±1.30	71.14 ±2.23	56.82 ±1.33	80.02 ±0.55

(a) AdaBFFs with two-node trees

	bl	cl	co	di	hi	ja	me	po	re	ro	Pr
bl	95.90 ±1.52	0.00 ±0.00	1.30 ±0.48	0.10 ±0.32	0.90 ±0.74	1.10 ±0.57	0.60 ±0.52	0.00 ±0.00	3.00 ±1.05	5.50 ±1.35	88.53 ±2.12
cl	1.10 ±0.57	95.80 ±0.63	2.80 ±0.42	1.60 ±0.52	1.00 ±0.00	3.20 ±0.42	0.00 ±0.00	0.10 ±0.32	1.00 ±0.00	1.00 ±0.00	89.04 ±0.76
co	0.00 ±0.00	0.00 ±0.00	77.10 ±1.29	1.80 ±0.42	1.00 ±0.00	0.40 ±0.70	0.50 ±0.53	1.60 ±0.97	1.10 ±0.88	8.00 ±0.94	84.28 ±1.92
di	0.20 ±0.42	0.00 ±0.00	2.50 ±0.97	70.20 ±2.15	3.30 ±1.34	0.30 ±0.48	0.20 ±0.42	3.00 ±0.67	5.10 ±1.10	5.70 ±1.06	77.61 ±2.27
hi	0.90 ±0.88	0.00 ±0.00	0.10 ±0.32	4.90 ±0.88	85.20 ±1.62	1.00 ±0.67	0.10 ±0.32	1.40 ±0.97	6.00 ±1.25	0.00 ±0.00	85.55 ±1.40
ja	0.40 ±0.52	0.30 ±0.48	2.00 ±0.82	0.00 ±0.00	0.90 ±0.32	92.20 ±0.92	0.00 ±0.00	0.70 ±0.67	1.40 ±0.52	0.00 ±0.00	94.20 ±1.49
me	0.40 ±0.70	1.40 ±0.52	0.10 ±0.32	0.60 ±0.52	2.60 ±0.52	1.00 ±0.00	96.80 ±0.79	2.00 ±0.82	1.10 ±0.32	15.30 ±2.06	79.83 ±1.66
po	0.00 ±0.00	0.00 ±0.00	5.50 ±0.53	5.60 ±1.51	2.00 ±0.00	0.10 ±0.32	0.00 ±0.00	87.50 ±1.43	6.10 ±1.37	1.40 ±0.84	80.90 ±1.48
re	0.90 ±0.57	0.00 ±0.00	1.40 ±0.84	8.40 ±0.84	3.10 ±0.32	0.00 ±0.00	0.00 ±0.00	3.40 ±0.84	72.40 ±1.65	1.40 ±0.84	79.57 ±1.32
ro	0.20 ±0.42	2.50 ±0.53	7.20 ±1.40	6.80 ±1.81	0.00 ±0.00	0.70 ±0.67	1.80 ±0.42	0.30 ±0.48	2.80 ±1.03	61.70 ±3.06	73.49 ±1.64
F	92.04 ±1.04	92.29 ±0.49	80.53 ±1.41	73.70 ±1.80	85.37 ±1.38	93.18 ±0.90	87.49 ±1.11	84.06 ±0.89	75.81 ±1.19	67.03 ±1.83	83.48 ±0.51

(b) SRCAM with normalized features

	bl	cl	co	di	hi	ja	me	po	re	ro	Pr
bl	90.30 ±1.06	0.00 ±0.00	0.70 ±0.67	0.00 ±0.00	0.00 ±0.00	0.90 ±0.32	0.10 ±0.32	0.00 ±0.00	2.90 ±0.74	9.00 ±0.82	86.92 ±1.31
cl	0.00 ±0.00	92.10 ±0.88	0.00 ±0.00	0.80 ±0.42	0.00 ±0.00	2.20 ±0.63	0.00 ±0.00	1.00 ±0.00	0.00 ±0.00	0.00 ±0.00	95.85 ±0.80
co	2.50 ±0.53	0.00 ±0.00	78.70 ±1.34	2.20 ±0.92	0.70 ±0.48	4.90 ±0.57	0.00 ±0.00	3.70 ±0.48	2.70 ±0.82	7.20 ±0.92	76.73 ±1.40
di	1.10 ±0.32	0.00 ±0.00	4.40 ±0.52	75.90 ±0.99	1.00 ±0.00	0.00 ±0.00	1.50 ±0.53	4.50 ±0.53	2.90 ±0.32	3.80 ±0.63	79.82 ±0.84
hi	0.00 ±0.00	0.00 ±0.00	0.90 ±0.32	5.80 ±0.63	82.30 ±1.16	0.20 ±0.42	0.00 ±0.00	1.00 ±0.00	7.60 ±0.70	0.00 ±0.00	84.15 ±0.77
ja	0.10 ±0.32	1.20 ±0.42	2.30 ±0.48	0.00 ±0.00	0.00 ±0.00	86.50 ±0.85	0.00 ±0.00	0.00 ±0.00	0.60 ±0.52	0.00 ±0.00	95.39 ±1.35
me	0.90 ±0.32	1.00 ±0.82	1.00 ±0.00	1.00 ±0.00	1.00 ±0.00	0.00 ±0.00	93.40 ±0.97	0.00 ±0.00	0.30 ±0.48	4.60 ±0.52	90.51 ±0.83
po	0.00 ±0.00	0.00 ±0.00	5.10 ±0.88	4.40 ±0.70	7.60 ±0.84	0.80 ±0.42	0.00 ±0.00	84.50 ±0.85	3.90 ±0.57	3.20 ±0.42	77.19 ±1.57
re	1.00 ±0.00	0.00 ±0.00	0.60 ±0.70	0.90 ±0.74	3.30 ±0.82	0.00 ±0.00	0.00 ±0.00	0.10 ±0.32	76.20 ±1.14	1.90 ±0.57	90.74 ±1.74
ro	4.10 ±0.74	5.70 ±0.48	6.30 ±1.64	9.00 ±0.82	4.10 ±0.88	4.50 ±0.71	5.00 ±0.67	5.20 ±0.63	2.90 ±0.57	70.30 ±1.16	60.04 ±1.26
F	88.57 ±1.01	93.93 ±0.51	77.69 ±0.93	77.81 ±0.66	83.21 ±0.79	90.72 ±0.62	91.93 ±0.74	80.67 ±1.13	82.83 ±0.98	64.77 ±1.14	83.02 ±0.38

(c) MAPsCAT with total covariance

Fig. 4 Mean confusions with standard deviations for each system (only one configuration each for lack of space). Columns are true genres, with mean precision (Pr × 100) shown in last column. Classification accuracy shown in bottom right corner. Rows are predicted genres, with mean F-measure (F × 100) in last row. Mean recalls × 100 are on *diagonal*. Classes as in Fig. 3

system labels Blues, Disco, Hip hop, Pop or Reggae as "Classical." It is thus tempting to claim that though these systems are sometimes not picking the correct labels, at least their mistakes make musical sense because even humans confuse, e.g., metal and rock, but never confuse, e.g., classical as blues or hip hop. This conclusion is unwarranted because it implicitly makes two assumptions: 1) the labels in GTZAN correspond to what we think they mean in a musical sense, e.g., that the Disco excerpts are exemplary of disco (Ammer 2004; Shapiro 2005); and 2) the systems are using cues similar, or equivalent in some way, to those used by humans when categorizing music. The first assumption is disputed by an analysis of the contents of GTZAN and how people describe them (Sturm 2013b). We see from this that the meaning of its categories are more broad than what its labels imply, e.g., many Blues excerpts are tagged on `last.fm` as "zydeco," "cajun," and "swing"; many Disco excerpts are tagged "80s," "pop," and "funk"; and many Metal excerpts are

tagged "rock," "hard rock," and "classic rock." The second assumption is disputed by work showing the inability of low-level features to capture musical information (Aucouturier and Pachet 2004; Aucouturier 2009; Marques et al. 2010), e.g., what instruments are present, and how they are being played; what rhythms are used, and what is the tempo; whether the music is for dancing or listening; whether a person is singing or rapping, and the subject matter.

4.4 Summary

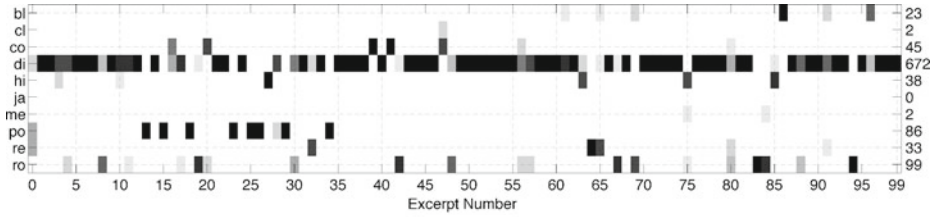
After evaluating the performance statistics of our systems using *Classify* in GTZAN, we are able to answer, in terms of classification accuracy, recall, precision, F-measure, and confusions, as well as our formal hypothesis testing, the question of how well each system can identify the labels assigned to excerpts in GTZAN. We are, however, no closer to determining the extents to which they can *recognize* from audio signals the music genres on which they are supposedly trained. One might argue, "the goal of any machine learning system is to achieve good results for a majority of cases. Furthermore, since machine learning is a methodology for building computational systems that improve their performance—whether with respect to classification accuracy, precision, or reasonable confusions—by training on examples provided by experts, then accuracy, precision, and confusion tables are reasonable indicators of the extent to which a system improves in this task. Therefore, no one can argue that an MGR system with a classification accuracy of 0.5—even for a small and problematic dataset like GTZAN—could be better than one of 0.8." Even if "better" simply means, "with respect to classification accuracy in GTZAN," there can still be question.

We find (Sturm 2013b) that classification accuracy of MAPsCAT in GTZAN using 2 fCV is more than 4 points above that of SRCAM. Using the Binomial hypothesis test above, we can reject the null hypothesis that MAPsCAT performs no better than SRCAM. Testing the same systems but with artist filtering, we find the classification accuracies of SRCAM are higher than those of MAPsCAT. Hence, even for this restricted and ultimately useless definition of "better"—for what meaning does it have to an actual user? (Schedl and Flexer 2012; Schedl et al. 2013)—we can still question the ranking of MGR systems in terms of classification accuracy. With doubt even for this limited sense of "better," how could there be less doubt with a sense that is more broad? For example, it can only be speculative to claim that a high accuracy system is "better" than another in the sense that it is doing so by recognizing genre (not confounds), or that a system will have similar performance in the real world, and so on.

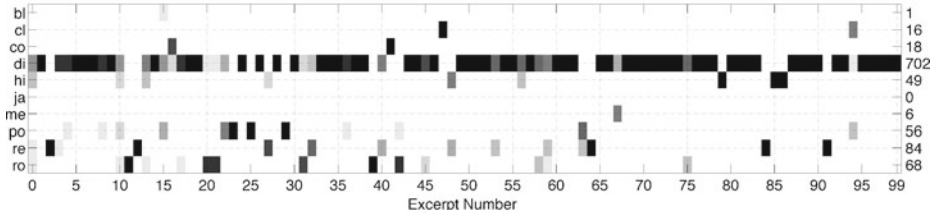
A vital point to make clear is that our aim here is not to test whether an MGR system *understands* music, or whether it is selecting labels *in a way* indistinguishable from that way used by humans (Guaus 2009), or even koi (Chase 2001), primates (McDermott and Hauser 2007), pigeons (Porter and Neuringer 1984), or sparrows (Watanabe and Sato 1999). It is unnecessary to require such things of a machine before we can accept that it is capable of merely selecting labels that are indistinguishable from those humans would choose. Our only concern here is how to reliably measure the extent to which an MGR system is *recognizing genre* and not confounds, such as bandwidth, dynamic range, etc. We take up this challenge in the next section by evaluating the behaviors of the systems.

5 Evaluating the behaviors of MGR systems

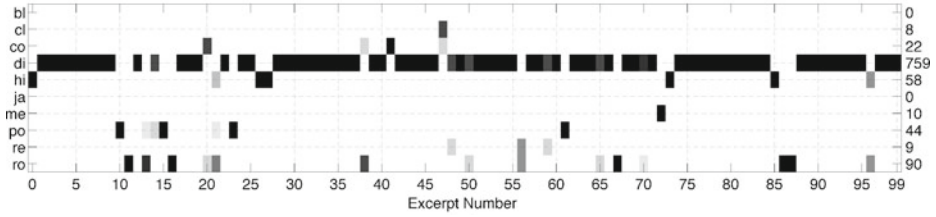
Figure 5 shows how the Disco excerpts are classified by AdaBFFs, SRCAM, and MAPsCAT. (For lack of space, we only look at excerpts labeled Disco, though all categories show similar behaviors.) Some evaluation in MGR describes particular misclassifications, and sometimes authors describe listening to confused excerpts to determine what is occurring, e.g., Deshpande et al. (2001), Lee et al. (2006), Langlois and Marques (2009), Scaringella et al. (2006). Of their experiments, Deshpande et al. (2001) 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. (2006) finds "... it is 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 when listening to the audio files."



(a) AdaBFFs with two-node decision trees



(b) SRCAM with normalized features



(c) MAPsCAT with total covariance

Fig. 5 Excerpt-specific confusions for Disco for each system (same configurations as Fig. 4) in 10 × 10 fCV, with number of classifications in each genre labeled at *right*. Classes as in Fig. 3. *Darkness of a square* indicates the number of times an excerpt is labeled a genre (*left*), with *black* as all 10 times

Unlike the confusion tables in Fig. 4, Fig. 5 shows the specific Disco excerpts that AdaBFFs most often classifies as "Pop" and "Rock," that SRCAM most often classifies as "Reggae" and "Rock," and that MAPsCAT most often classifies as "Rock" and "Hip hop." We also see that some excerpts are never classified "Disco," and/or are always classified with the same class; and that five excerpts always classified the same way by all three systems. We might see such pathological behavior as bias in a system, or perhaps an indication of an error in GTZAN; nonetheless, the pathological behaviors of our systems present themselves as teachable moments (Sturm 2012b): by considering the specific excerpts producing these consistent errors, we might learn what causes problems for some or all of the systems.

We find very few works discussing and using such behaviors of an MGR system. Lopes et al. (2010) propose training an MGR system using only the training data that is easily separable. In a similar vein, Bağci and Erzin (2007) refine their class models by only using perfectly separable training data (Sturm and Gouyon 2013, unpublished). Related is experimental work (Pampalk et al. 2005; Flexer 2007; Flexer and Schnitzer 2009) showing inflated performance of systems for MGR or music similarity when the same artists and/or albums occur in training and testing sets. Within music similarity and retrieval, a related topic is hubs (Aucouturier and Pachet 2004; Gasser et al. 2010; Flexer and Schnitzer 2010; Schnitzer et al. 2012), where some music signals appear equally close to many other music signals regardless of similarity.

We define a *consistent misclassification* (CM) when in all CV runs a system selects the same but "wrong" class for an excerpt. When in all CV runs a system selects different "wrong" classes for an excerpt, we call it a *persistent misclassification* (PM). When in all CV runs a system selects the "correct" class for an excerpt, we call it a *consistently correct classification* (C3). Table 5 summarizes these types of classification in our 10×10 fCV for AdaBFFs, SRCAM, and MAPsCAT. We see that of the Disco excerpts, AdaBFFs produced 55 C3s, 15 CMs, 5 PMs, and it consistently misclassified 11 excerpts as "Disco." In total, MAPsCAT appears to have the highest number of C3s and CMs, and AdaBFFs the lowest.

Table 5 Classification type results for each system (same configurations as Fig. 4) on GTZAN

Genre	AdaBFFs				SRCAM				MAPsCAT			
	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

The column "CM as" signifies the number of excerpts consistently misclassified as each genre (*rows*). We do not consider here the replicas and mislabelings in GTZAN (Sturm 2013b)

We can consider three different possibilities with such errors. The first is that the GTZAN label is "wrong," and the system is "right"—which can occur since GTZAN has mislabelings (Sturm 2013b). Such a result provides strong evidence that the system is recognizing genre, and means that the classification accuracy of the system we compute above is too low. The second possibility is that the label is "right," but the system almost had it "right." This suggests it is unfair to judge a system by considering only its top choice for each excerpt. For instance, perhaps the decision statistic of the system in the "correct" class is close enough to the one it picked that we can award "partial credit." The third possibility is that the label is "right," but the class selected by the system would have been selected by a human. Hence, this suggests the error is "acceptable" because the selected class is indistinguishable from what humans would choose. Such a result provides evidence that the system is recognizing genre. We now deal with each of these possibilities in order.

5.1 Evaluating behavior through the mislabelings in GTZAN

So far, our evaluation has employed descriptive and inferential statistics, looked at how specific excerpt numbers are classified, and noted pathological behaviors, but we have yet to make use the actual music embodied by any excerpts. It is quite rare to find in the MGR literature any identification of the music behind problematic classifications (Sturm 2012a). Langlois and Marques (2009) notice in their evaluation that all tracks from an album by João Gilberto are PMs for their system. They attribute this to the tracks coming from a live recording with speaking and applause. In their results, we see that the system by Lee et al. (2006) classifies as "Techno," John Denver's "Rocky Mountain High," but they do not discuss this result.

The first three columns of Table 6 list for only Disco (we see similar results for all other categories) the specific excerpts of each pathological classification of our systems (same configurations as in Fig. 4). We know that in Disco, GTZAN has six repeated excerpts, two excerpts from the same recording, seven mislabeled excerpts, and one with a problematic label (Sturm 2013b). Of the excerpts consistently

Table 6 Classification type results for each system (same configurations as Fig. 4) for GTZAN Disco excerpts considering the replicas and mislabelings in GTZAN (Sturm 2013b)

System	Classification type			
	C3	CM *excerpts*	PM *excerpts*	CM as "Disco" *label and excerpts*
AdaBFFs	49	*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	51	*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	66	*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	39	*(23), 41, 86*		*hi 00*

A *struck-though number* is a mislabeled or problematic excerpt in GTZAN. A *circled number* is an excerpt the system classified "correctly." Classes as in Fig. 3

misclassified "Disco" by our systems, six are mislabeled in GTZAN. The last column lists those excerpts in other categories that each system consistently misclassifies "Disco." A number struck-through in Table 6 is a mislabeled excerpt; and a circled number is an excerpt for which the system selects the "correct" class. It is important to note that the categories of GTZAN are actually more broad than what their titles suggest (Sturm 2013b), i.e., Disco includes music that is not considered exemplary of disco as described: "A style of dance music of the late 1970s and early 1980s ... It is characterized by a relentless 4/4 beat, instrumental breaks, and erotic lyrics or rhythmic chants" (Ammer 2004).

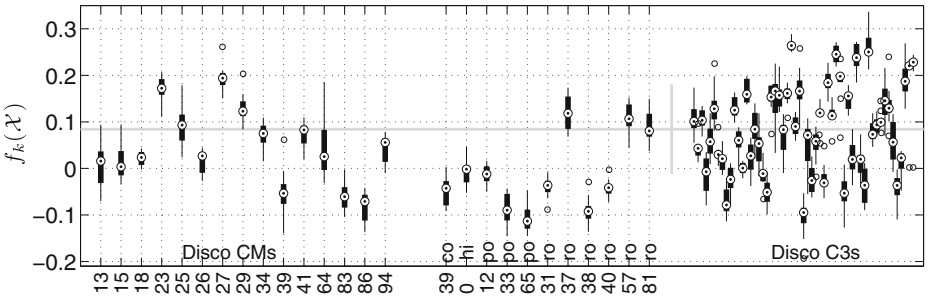
We see from Table 6 and Fig. 5, all seven Disco mislabelings in GTZAN appear in the pathological behaviors of the three systems. AdaBFFs consistently "correctly" classifies four of the six mislabelings it finds, SRCAM two of the five it finds, and MAPsCAT two of the six it finds. All systems are firm in classifying as "Pop" Disco 23 ("Playboy," Latoya Jackson). Both AdaBFFs and MAPsCAT are firm in classifying as "Hip hop" Disco 27 ("Rapper's Delight," The Sugarhill Gang), but SRCAM most often selects "Reggae." Both AdaBFFs and SRCAM are firm in classifying as "Pop" Disco 29 ("Heartless," Evelyn Thomas). While AdaBFFs is adamant in its classification as "Pop" Disco 26 ("(Baby) Do The Salsa," Latoya Jackson), MAPsCAT is "incorrect" in its firm classification as "Hip hop." Finally, both SRCAM and MAPsCAT are firm in classifying as "Rock" Disco 11 ("Can You Feel It," Billy Ocean). AdaBFFs classifies as only "Country," and MAPsCAT as only "Rock," Disco 20 ("Patches," Clarence Carter), but never as the "correct" label "Blues" (Sturm 2013b).

We now consider the excerpts consistently misclassified as "Disco" by the systems. First, AdaBFFs "correctly" classifies Pop 65 ("The Beautiful Ones," Prince), and SRCAM "correctly" classifies Pop 63 ("Ain't No Mountain High Enough," Diana Ross). However, AdaBFFs insists Country 39 ("Johnnie Can't Dance," Wayne Toups & Zydecajun) is "Disco" though the "correct" label is "Blues" (Sturm 2013b). and insists Rock 40 ("The Crunge," Led Zeppelin) is "Disco" though the "correct" labels are "Metal" and/or "Rock" (Sturm 2013b). SRCAM insists Rock 77 ("Freedom," Simply Red) is "Disco," though it should be "Pop" (Sturm 2013b). Finally, both AdaBFFs and SRCAM insist Rock 38 ("Knockin' On Heaven's Door," Guns N' Roses) is "Disco," though it should be "Metal" and/or "Rock" (Sturm 2013b).

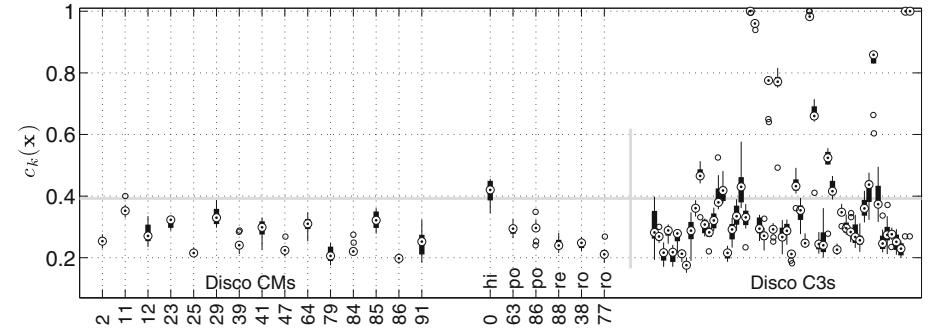
In summary, we see that each system finds some of the mislabelings in Country, Disco, Pop, and Rock. When a system "correctly" classifies such an error, it provides strong evidence that it could be recognizing genre, e.g., AdaBFFs finds and "correctly" classifies as "Pop" both excerpts of Latoya Jackson. When a system "incorrectly" classifies a mislabeling, however, it provides evidence that it is not recognizing genre, e.g., all three systems consistently misclassify as "Country" Disco 41 ("Always and Forever," Heatwave). We now consider the remaining Disco excerpts in Table 6, and investigate the decision confidences.

5.2 Evaluating behavior through the decision statistics

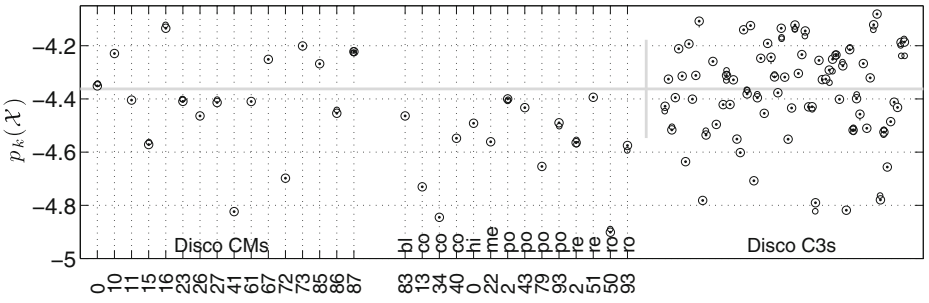
We now explore the second possibility of these pathological behaviors: whether for these kinds of misclassifications a system deserves credit because its decision statistics between the correct and selected class are close, or at least the "correct" label is ranked second. Figure 6 shows a boxplot of the decision statistics for AdaBFFs (2),



(a) Decision statistic of AdaBFFs (2) with two-node decision trees



(b) Decision statistic of SRCAM (6) with normalized features



(c) Decision statistic of MAPsCAT (8) with total covariance

Fig. 6 Boxplot of decision statistics for Disco excerpts: CMs labeled on *left*; C3s on *right*; CMs as "Disco" labeled in *center*. Excerpt numbers are shown. Mean of C3s shown as *gray line* with one standard deviation *above* and *below*. Classes as in Fig. 3

for SRCAM (6), and for MAPsCAT (8). We only consider Disco for lack of space. Figure 6a shows for AdaBFFs that its decision statistic for most Disco CMs and CMs as "Disco" lie within two standard deviations of the mean of the C3s decision statistics. This suggests AdaBFFs is as "confident" in its pathological misclassifications as it is in its C3s. We reach the same conclusion for SRCAM and MAPsCAT. We next consider each excerpt individually.

Table 7 lists the music artist and title of the CMs from Table 6 that are not struck-through, as well as statistics of decision statistics (confidence), the rank of "Disco," and the top three `last.fm` tags (from the "count" parameter) of each song or artist

(retrieved Oct. 15, 2012). We do not include tags that duplicate 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 decision statistic of the selected class with that of "Disco." This is different in Fig. 6, where we show the decision statistics and not a difference. Table 8 lists in the same way the details of the CMs as "Disco."

In Table 7 we see that AdaBFFs, SRCAM, and MAPsCAT all appear quite confident in many of their CMs; however, "Disco" appears in the penultimate position for many by AdaBFFs and MAPsCAT. Only two excerpts are CMs for all three systems: Disco 41 is always classified "Country," even though its top tags have no overlap with the top tags in Country ("country," "classic country," "oldies," etc.) (Sturm 2013b). For AdaBFFs and SRCAM, "Disco" appears around rank 5; and rank 8 for MAPsCAT (above only "Metal" and "Classical"). The other common CM is Disco 86, which is consistently misclassified by each system in radically different ways. Its tags have little overlap with the top tags of Rock (i.e., "70s"), but none with those of Blues or Hip hop (Sturm 2013b).

Table 8 show for all three systems the details of the CMs as "Disco" in Table 6, as well as their confidences, and the rank of the GTZAN category to which each belongs. Some of these appear to be satisfactory, e.g., the excerpts with top tags "pop," "70s," "80s," "soul," "funk" and "dance" have overlap with the top tags of Disco (Sturm 2013b). Others, however, appear quite unsatisfactory, such as "Knockin' on Heaven's Door" by Guns 'N Roses, for which AdaBFFs ranks its labels from penultimate: "Pop" and "Hip hop" before "Rock" or "Metal."

In summary, from our evaluation of the decision statistics involved in their pathological errors, it is hard to find evidence to support a conclusion that any of these systems is working in a way corresponding to confidence in its decision when it comes to music genre. The question thus arises: to what extent do humans show the same kinds of classification behavior as AdaBFFs, SRCAM, and MAPsCAT? We test this next using a listening experiment.

5.3 Evaluating behavior by listening tests

Listening tests for the evaluation of systems are by and large absent from the MGR literature (Sturm 2012a). The work of Gjerdingen and Perrott (2008) is a widely cited study of human music genre classification (Aucouturier and Pampalk 2008), and Krumhansl (2010) and Mace et al. (2011) extend this work. Both Ahrendt (2006) and Meng and Shawe-Taylor (2008) use listening tests to gauge the difficulty of discriminating the genres of their datasets, and to compare with the performance of their system. Lippens et al. (2004) use listening tests to produce a dataset that has excerpts more exemplary of single genres; and Craft et al. (2007) and Craft (2007) propose that evaluating MGR systems makes sense only with respect to the generic ambiguity of music. Seyerlehner et al. (2010) expands upon these works. Guaus (2009) conducts listening tests to determine the relative importance of timbre or rhythm in genre recognition. Finally, Cruz and Vidal (2003, 2008) and we (Sturm 2012b) conduct listening tests to determine if an MGR system can create music representative of a genre it has learned.

These latter works motivate the use of listening tests to circumvent the need to compare tags, or demarcate elements, necessary to argue that it is more acceptable to label an excerpt, e.g., "Pop" or "Rock," neither or both. Our hypothesis is that

Table 7 Details of Disco CMs from Table 6

	No.	Artist	Title of work	Assigned class k	$\{f_k(\mathscr{X}) - f_D(\mathscr{X})\}$ (2) mean ± st.dev.	Rank of "Disco" mean ± st.dev.	Top `last.fm` tags
AdaBFFs	13	Donna Summer	Back Off Boogaloo	Pop	0.15 ± 0.05	2.10 ± 0.32	Artist: disco, pop, 70s
	15	Heatwave	Boogie Nights	Pop	0.17 ± 0.07	3.20 ± 0.63	Disco, funk, 70s
	18	?	?	Pop	0.17 ± 0.04	2.20 ± 0.42	
	25	Evelyn Thomas	High Energy	Pop	0.14 ± 0.05	2.00 ± 0.00	Disco, 80s, dance
	34	Evelyn Thomas	Reflections	Pop	0.09 ± 0.05	2.00 ± 0.00	80s, disco; artist: disco
	39	George McCrae	I Can't Leave You Alone	Country	0.10 ± 0.07	2.00 ± 0.00	Disco, 70s, pop
	41	Heatwave	Always and Forever	Country	0.48 ± 0.07	5.80 ± 0.63	Soul, slow jams, 70s
	64	Lipps, Inc.	Funkytown	Reggae	0.20 ± 0.10	2.00 ± 0.00	Disco, 80s, 70s
	83	Rick Dees	Disco Duck	Rock	0.18 ± 0.03	3.90 ± 0.32	70s, pop, Disco
	86	Alicia Bridges	I Love the Night Life	Blues	0.22 ± 0.05	4.40 ± 1.07	70s; artist: disco, dance
	94	Bronski Beat	Why?	Rock	0.21 ± 0.05	2.00 ± 0.00	80s, new wave, synthpop
	No.	Artist	Title of work	Assigned class k	$\{c_k(\mathbf{x}) - c_D(\mathbf{x})\}$ (6) mean ± st.dev.	Rank of "Disco" mean ± st.dev.	Top `last.fm` tags
SRCAM	02	Archie Bell and the Drells	Look Back Over Your Shoulder	Reggae	0.16 ± 0.03	4.70 ± 0.67	Northern soul, soul
	12	Carl Carlton	She's a Bad Mama Jama	Reggae	0.07 ± 0.05	2.10 ± 0.32	Funk, disco, 70s
	25	Evelyn Thomas	High Energy	Pop	0.05 ± 0.01	3.00 ± 0.00	Disco, 80s, dance
	39	George McCrae	I Can't Leave You Alone	Rock	0.12 ± 0.04	3.50 ± 0.71	Disco, 70s, pop
	41	Heatwave	Always and Forever	Country	0.19 ± 0.06	4.50 ± 1.78	Soul, slow jams, 70s
	64	Lipps, Inc.	Funky Town	Reggae	0.16 ± 0.03	2.80 ± 0.79	Disco, 80s, 70s
	79	Peter Brown	Love is Just the Game	Hip hop	0.07 ± 0.04	3.30 ± 1.83	70s, disco; artist: funk
	84	?	?	Reggae	0.12 ± 0.02	4.40 ± 0.52	
	85	Tom Tom Club	Wordy Rappinghood	Hip hop	0.12 ± 0.04	2.00 ± 0.00	New wave, 80s, funk
	86	Alicia Bridges	I Love the Night Life	Hip hop	0.08 ± 0.03	4.30 ± 1.06	70s; artist: disco, dance
	91	Silver convention	Fly Robin Fly	Reggae	0.08 ± 0.08	2.60 ± 1.35	Disco, pop
	No.	Artist	Title of work	Assigned class k	$\{p_k(\mathscr{X}) - p_D(\mathscr{X})\}$ (8) mean ± st.dev. (10^{-2})	rank of "Disco" mean ± st.dev.	Top `last.fm` Tags
MAPsCAT	00	Boz Scaggs	Lowdown	Hip hop	1.68 ± 0.18	4.00 ± 0.42	70s, classic rock
	10	?	?	Pop	0.71 ± 0.13	3.00 ± 0.47	
	15	Heatwave	Boogie Nights	Pop	0.65 ± 0.16	3.00 ± 0.48	Disco, funk, 70s
	16	?	?	Rock	0.35 ± 0.14	2.00 ± 0.00	
	41	Heatwave	Always and Forever	Country	3.96 ± 0.17	8.00 ± 0.42	Soul, slow jams, 70s
	61	Anita Ward	Ring my Bell	Pop	0.75 ± 0.28	2.00 ± 0.00	Disco, 70s, dance
	67	ABBA	Dancing Queen	Rock	0.87 ± 0.17	5.00 ± 0.70	Pop, Disco, 70s
	72	ABBA	Mamma Mia	Metal	1.00 ± 0.14	4.00 ± 0.79	Pop, 70s, disco
	73	KC & Sunshine Band	I'm Your Boogie Man	Hip hop	0.40 ± 0.12	2.00 ± 0.00	Disco, 70s, funk
	85	Tom Tom Club	Wordy Rappinghood	Hip hop	0.94 ± 0.24	2.00 ± 0.00	New wave, 80s, funk
	86	Alicia Bridges	I Love the Night Life	Rock	1.21 ± 0.17	3.00 ± 0.00	70s; artist: disco, dance
	87	The Supremes	He's my Man	Rock	0.31 ± 0.17	2.00 ± 0.00	Soul, vocalization

Table 8 Details of CMs as "Disco" from Table 6

	Genre & no.	Artist	Title of work	$\{f_D(\mathscr{X}) - f_k(\mathscr{X})\}$ (2) mean ± st.dev.	Rank of "k" mean ± st.dev.	Top `last.fm` tags
AdaBFFs	co 39	Zydecajun & Wayne Toups	Johnnie Can't Dance	0.23 ± 0.06	3.50 ± 1.58	Artist: zydeco, cajun, folk
	hi 00	Afrika Bambaataa	Looking for the Perfect Beat	0.16 ± 0.07	2.30 ± 0.48	Electro, hip-hop, old school
	po 12	Aretha Franklin, et al.	You Make Me Feel Like a Natural Woman	0.16 ± 0.04	3.60 ± 0.52	Pop, Ballad; artist: soul
	po 33	Britney Spears	Pepsi Now and Then	0.19 ± 0.04	3.30 ± 0.48	Artist: pop, dance
	po 65	Prince	The Beautiful Ones	0.21 ± 0.04	5.60 ± 0.70	80s, funk, pop
	ro 31	The Rolling Stones	Honky Tonk Women	0.15 ± 0.04	2.40 ± 0.70	Classic rock, rock, 60s
	ro 37	The Rolling Stones	Brown Sugar	0.08 ± 0.05	2.00 ± 0.00	Classic rock, rock, 70s
	ro 38	Guns 'N Roses	Knockin' on Heaven's door	0.19 ± 0.04	4.10 ± 0.74	Rock, hard rock, classic rock
	ro 40	Led Zeppelin	The Crunge	0.28 ± 0.04	5.20 ± 0.79	Classic rock, hard rock, rock
	ro 57	Sting	If You Love Somebody Set Them Free	0.23 ± 0.06	2.40 ± 0.52	Rock, 80s, pop
	ro 81	Survivor	Poor Man's Son	0.34 ± 0.05	2.60 ± 0.52	80s, rock, melodic rock
	Genre & no.	Artist	Title of work	$\{c_D(\mathbf{x}) - c_k(\mathbf{x})\}$ (6) mean ± st.dev.	Rank of "k" mean ± st.dev.	Top `last.fm` tags
SRCAM	hi 00	Afrika Bambaataa	Looking for the Perfect Beat	0.27 ± 0.04	2.00 ± 0.00	Electro, hip-hop, old school
	po 63	Diana Ross	Ain't No Mountain High Enough	0.13 ± 0.01	2.00 ± 0.00	Soul, motown, 70s
	po 86	Madonna	Cherish	0.08 ± 0.04	2.00 ± 0.00	Pop, 80s, dance
	re 88	Marcia Griffiths	Electric Boogie	0.20 ± 0.04	8.20 ± 0.79	Funk, reggae, dance
	ro 38	Guns 'N Roses	Knocking on Heaven's Door	0.08 ± 0.03	2.90 ± 0.32	Rock, hard rock, classic rock
	ro 77	Simply Red	Freedom	0.08 ± 0.04	3.70 ± 0.95	Pop, rock, easy
	Genre & no.	Artist	Title of work	$\frac{1}{100}\{p_D(\mathscr{X}) - p_k(\mathscr{X})\}$ (8) mean ± st.dev.	Rank of "k" mean ± st.dev.	Top `last.fm` tags
MAPsCAT	bl 83	Buckwheat Zydeco	?	2.67 ± 0.19	2.00 ± 0.42	
	co 13	Loretta Lynn	Let Your Love Flow	0.96 ± 0.20	3.00 ± 0.52	Country, traditional country
	co 34	Merle Haggard	Sally Let Your Bangs Hang Down	0.81 ± 0.35	2.00 ± 0.71	Artist: country, classic country, outlaw country
	co 40	Kentucky Headhunters	Dumas Walker	0.67 ± 0.29	2.00 ± 0.00	Country, outlaw country, rock
	hi 00	Afrika Bambaataa	Looking for the Perfect Beat	2.25 ± 0.16	2.00 ± 0.32	Electro, hip-hop, old school
	me 22	Ozzy Osbourne	Crazy Train	0.75 ± 0.15	2.00 ± 0.32	Heavy metal, metal, hard rock
	po 02	Mariah Carey	My All	0.26 ± 0.22	2.00 ± 0.00	Pop, rnb, soul
	po 43	Cher	Believe	1.17 ± 0.24	2.00 ± 0.00	Pop, dance, 90s
	po 79	Kate Bush	Couldbusting	1.93 ± 0.13	7.00 ± 0.57	80s, pop, alternative
	po 93	Mandy Moore	I Wanna Be with You	1.08 ± 0.17	2.00 ± 0.00	Pop, romantic, love
	re 02	Bob Marley	Could You Be Loved	0.70 ± 0.19	2.00 ± 0.32	Reggae, roots reggae, Jamaican
	re 51	?	?	2.49 ± 0.26	3.00 ± 0.48	
	ro 50	Simple Minds	See the Lights	1.21 ± 0.22	3.00 ± 0.52	Rock, 80s, new wave
	ro 93	The Stone Roses	Waterfall	0.43 ± 0.21	2.00 ± 0.00	Indie, britpop, madchester

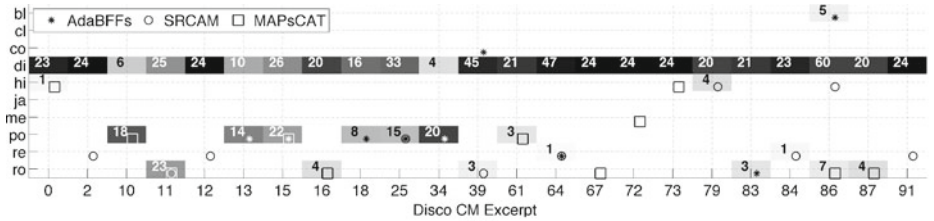
the difference between the label given by a human and that given by a system will be large enough that it is extremely clear which label is given by a human. Hence, we wish to determine the extent to which the CMs of AdaBFFs, SRCAM or MAPsCAT—of which we see the systems are confident—are acceptable in the sense that they are indistinguishable from those humans would produce.

We conduct a listening test in which a human subject must choose for each 12 s excerpt which label of two was given by a human (i.e., G. Tzanetakis); the other label was given by AdaBFFs, SRCAM or MAPsCAT. 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. (We pick "representative" excerpts by listening: 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 labels of all excerpts continues to the second part of the test, where s/he must discriminate between the human- and system-given genres for each music excerpt. For instance, the test GUI presents "Back Off Boogaloo" by Donna Summer along with the labels "Disco" and "Pop." The subject selects the one s/he thinks is given by a human before proceeding. We record the time each subject uses to listen to an excerpt before proceeding. We test all Disco CMs and CMs as "Disco" in Tables 7 and 8. Although all other classes have CMs in all three systems (Table 5), we test only these ones because of the effort required. In total, 24 test subjects completed the second part.

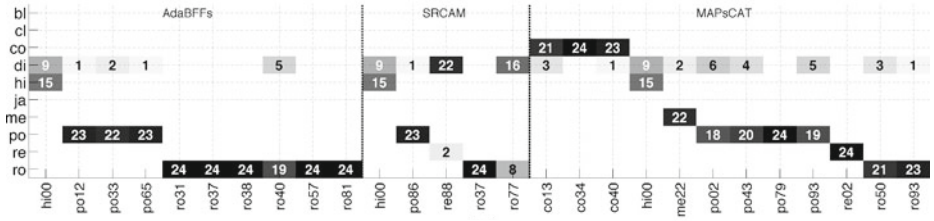
Figure 7 shows the distribution of choices made by the test subjects. Figure 7a shows that out of the 24 times Disco 10 was presented, six people selected "Disco" (the human-given label), and 18 people selected "Pop" (the class selected by MAPsCAT). Some excerpts are presented more than 24 times because they are misclassified by more than one system. We see that of the Disco CMs of the systems, for only two of nine by AdaBFFs, one of 11 by MAPsCAT, and none of 10 by SRCAM, did a majority of subjects side with the non-human class. For some excerpts, agreement with the human label is unanimous. Figure 7b shows that for Hip hop 00, "Hip hop" (the class selected by AdaBFFs) was selected by 15 subjects; and 9 subjects picked "Disco" (the human-given label). 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."

We now test the null hypothesis that the subjects are unable to recognize the difference between the label given by a human and the class selected by a system. 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 for 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]: $\hat{x}(h) = (h+1)/(N+2)$ (Song et al. 2009). The variance of this estimate is given by

$$\hat{\sigma}^2(\hat{x}) = \frac{\hat{x}(1-\hat{x})}{(N-1) + \frac{N+1}{N\hat{x}(1-\hat{x})}}. \tag{10}$$



(a) Disco CMs



(b) CMs as "Disco"

Fig. 7 Distribution of choices from listening tests. **a** Disco CMs in Table 7. The class selected by a system is marked using a *symbol* (*legend*). **b** CMs as "Disco" in Table 8. Classes as in Fig. 3

We test the null hypothesis by computing $P[T > |\hat{x} - 0.5|/\hat{o}(\hat{x})]$ 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 human-given labels ($p < 0.008$). 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 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 the human-given labels ($p < 0.012$).

So, what is it about Disco 13, 15 and 18 that makes subjects divided between the labels "Disco" and "Pop," and choose "Pop" more often than "Disco" for Disco 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 "produced." In these five cases then, we can argue that AdaBFFs and MAPsCAT are classifying acceptably. (Some subjects were also dissatisfied by some label pairs, e.g., "Metal" and "Disco" for ABBA's "Mamma Mia" while they wished to select "Pop" instead.)

In the case of Disco 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. After inspecting the duration each subject listened to Disco 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 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." In these two cases, then, we can argue that SRCAM and MAPsCAT are classifying acceptably.

5.4 Summary

In Section 4, we were concerned with quantitatively measuring the extent to which an MGR system predicts the genre labels of GTZAN. This presents a rather rosy picture of performance: all of our systems have high classification accuracies, precision and F-measures in many categories, and confusion behaviors that appear to make musical sense. Though their classification accuracies in GTZAN drop significantly when using an artist filter (Sturm 2013b), they still remains high above that of chance. Due to *Classify*, however, we cannot reasonably argue that this means they are recognizing the genres in GTZAN, or more broadly that they will perform well in the real world recognizing the same genres (Urbano et al. 2013). In this section, we have thus been concerned with evaluating the extent to which an MGR system displays the kinds of behavior we expect of a system that has capacity to recognize genre.

By inspecting the pathological errors of the systems, and taking into consideration the mislabelings in GTZAN (Sturm 2013b), we find evidence for and against the claim that any of them can recognize genre, or that any of them are better than the others. We see MAPsCAT has over one hundred more C3s than SRCAM and AdaBFFs, but AdaBFFs "correctly" classifies the most mislabeled Disco excerpts than the other two. All three systems, however, make errors that are difficult to explain if genre is what each is recognizing. We see that the confidence of these systems in their pathological errors are for the most part indistinguishable from their confidence in their C3s. While the rank of the "correct" class is often penultimate to the "wrong" one they select, there are rankings that are difficult to explain if genre is what each is recognizing. Finally, our listening test reveals that for the most part the pathological errors of these systems are readily apparent from those humans would commit. Their performance in that respect is quite poor.

6 On evaluation

While genre is an inescapable result of human communication (Frow 2005), it can also sometimes be ambiguous and subjective, e.g., Lippens et al. (2004), Ahrendt (2006), Craft et al. (2007), Craft (2007), Meng and Shawe-Taylor (2008), Gjerdingen and Perrott (2008) and Seyerlehner et al. (2010). A major conundrum in the evaluation of MGR systems is thus the formal justification of why particular labels are better than others. For instance, while we deride it above, an argument might be made that ABBA's "Mamma Mia" employs some of the same stylistic elements of metal used by Motörhead in "Ace Of Spades"—though it is difficult to imagine

the audiences of the two would agree. 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 (2007) 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." Such an approach to evaluation might thus be a poor match with the nature of genre (Frow 2005).

We have shown how evaluating the performance statistics of MGR systems using *Classify* in GTZAN is inadequate to meaningfully measure the extents to which a system is recognizing genre, or even whether it addresses the fundamental problem of MGR. Indeed, replacing GTZAN with another dataset, e.g., ISMIR2004 (ISMIR 2004), or expanding it, does not help as long as we do not control for all independent variables in a dataset. On the other hand, there is no doubt that we see systems performing with classification accuracies significantly above random in GTZAN and other datasets. Hence, something is working in the prediction of the labels in these datasets, but is that "something" *genre recognition*? One might argue, "The answer to this question is irrelevant. The 'engineering approach'—assemble a set of labeled data, extract features, and let the pattern recognition machinery learn the relevant characteristics and discriminating rules—results in performance significantly better than random. Furthermore, with a set of benchmark datasets and standard performance measures, we are able to make meaningful comparisons between systems." This might be agreeable in so far that one restricts the application domain of MGR to predicting the single labels of the music recording excerpts in the handful of datasets in which they are trained and tested. When it comes to ascertaining their success in the real world, to decide which of several MGR systems is best and which is worst, which has promise and which does not, *Classify* and classification accuracy provide no reliable or even relevant gauge.

One might argue, "accuracy, recall, precision, F-measures are standard performance measures, and this is the way it has always been done for recognition systems in machine learning." We do not advocate eliminating such measures, not using *Classify*, or even of avoiding or somehow "sanitizing" GTZAN. We build all of Section 5 upon the outcome of *Classify* in GTZAN, but with a major methodological difference from Section 4: we consider the *contents* of the categories. We use the faults of GTZAN, the decision statistics, and a listening test, to illuminate the pathological behaviors of each system. As we look more closely at their behaviors, the rosy picture of the systems evaporates, as well as our confidence that any of them is addressing the problem, that any one of them is better than the others, or even that one of them will be successful in a real-world context.

One might argue that confusion tables provide a realistic picture of system performance. However, in claiming that the confusion behavior of a system "makes musical sense," one implicitly makes 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? The faults of GTZAN, and the wide composition of its categories, obviously do not bode well for the first assumption (Sturm 2013b). The second assumption is difficult to justify, and requires one to dig deeper than the confusion behaviors, to determine how the system is encoding and using such relevant features.

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. Comparing the classification results with the tags given by a community of listeners show that some behaviors do "make musical sense," but other appear less acceptable. 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 acceptability of its choices, circumvents the need to compare tags, and gets to the heart of whether a system is producing genre labels indistinguishable from those humans would produce. Hence, we finally see by this that though our systems have classification accuracies and other statistics 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 acceptability of their classifications reveals that they are likely not *recognizing* genre at all.

If performance statistics better than random do not reflect the extents to which a system is solving a problem, then what can? The answer to this has import not just for MGR, but music information research in general. To this end, consider a man claiming his horse "Clever Hans" can add and subtract integers. We watch the owner ask Hans, "What is 2 and 3?" Then Hans taps his hoof until his ears raise after its fifth tap, at which point he is rewarded by the owner. To measure the extent to which Hans understands the addition and subtraction of integers, having the owner ask more questions in an uncontrolled environment does not add evidence. We can instead perform a variety of experiments that do. For instance, with the owner present and handling Hans, two people can whisper separate questions to Hans and the owner, with the ones whispering not knowing whether the same question is given or not. In place of real questions, we might ask Hans nonsensical questions, such as, "What is Bert and Ernie?" Then we can compare its answers with each of the questions. If this demonstrates that something other than an understanding of basic mathematics might be at play, then we must search for the mechanism by which Hans is able to correctly answer the owner's questions in an uncontrolled environment. We can, for instance, blindfold Hans to determine whether it is vision; or isolate it in a sound proof room with the owner outside to determine whether it is sound. Such a historical case is well-documented by Pfungst (1911).

Classify using datasets having many independent variables changing between classes is akin to asking Hans to answer more questions in an uncontrolled environment. What is needed is a richer and more powerful toolbox for evaluation (Urbano et al. 2013). One must search for the mechanism of correct response, which can be evaluated by, e.g., *Rules* and *Robust*. Dixon et al. (2010) use *Rules* to inspect the sanity of what their system discovers useful for discriminating different genres. We show using *Robust* (Sturm 2012b) that two high-accuracy MGR systems can classify the same excerpt of music in radically different ways when we make minor adjustments by filtering that do not affect its musical content. Akin to nonsense questions, Matityaho and Furst (1995) notice that their system classifies a zero-amplitude signal as "Classical," and white noise as "Pop." Porter and Neuringer (1984), investigating the training and generalization capabilities of pigeons in discriminating between two genres, test whether responses are due to the music itself, or to confounds such as characteristics of the playback mechanisms, and the lengths and loudness of excerpts. Chase (2001) does the same for koi, and looks at the effect of timbre as well.

Since it is as remarkable a claim that an artificial system "recognizes genre with 85 % accuracy" as a horse is able to perform mathematics, this advocates approaching an MGR system—or autotagger, or any music information system—as if it were "Clever Hans." This of course necessitates creativity in experimental design, and requires much more effort than comparing selected tags to a "ground truth." One might argue, "One of the reasons MGR is so popular is because evaluation is straightforward and easy. Your approach is less straightforward, and certainly unscalable, e.g., using the million song dataset (Bertin-Mahieux et al. 2011; Hu and Ogihara 2012; Schindler et al. 2012)." To this we can only ask: why attempt to solve very big problems with a demonstrably weak approach to evaluation, when the smaller problems have yet to be indisputably solved?

7 Conclusion

In this work, we have evaluated the performance statistics and behaviors of three MGR systems. Table 4 shows their classification accuracies are significantly higher than chance, and are among the best observed (and reproduced) for the GTZAN dataset. Figure 3 shows their recalls, precisions, and F-measures to be similarly high. Finally, Fig. 4 shows their confusions "make musical sense." Thus, one might take these as evidence that the systems are capable of recognizing some of the genres in GTZAN. The veracity of this claim is considerably challenged when we evaluate the behaviors of the systems. We see that SRCAM has just as high confidences in its consistent misclassifications as in its consistently correct classifications. We see MAPsCAT—a system with a high F-score in Metal—always mistakes the excerpt of "Mamma Mia" by ABBA as "Metal" first, "Rock" second, and "Reggae" or "Country" third. We see that all subjects of our listening test have little trouble discriminating between a label given by a human and that given by these systems. In short, though these systems have superb classification accuracy, recalls, etc., in GTZAN, they do not reliably produce genre labels indistinguishable from those humans produce.

From the very nature of *Classify* in GTZAN, we are unable to reject the hypothesis that any of these systems is not able to recognize genre, *no matter the accuracy we*

observe. In essence, "genre" is not the only independent variable changing between the excerpts of particular genres in our dataset; and *Classify* does not account for them. There is also, just to name a few, instrumentation (disco and classical may or may not use strings), loudness (metal and classical can be played at high or low volumes), 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), etc. Hence, to determine if an MGR system has a capacity to recognize any genre, one must look deeper than classification accuracy and related statistics, and from many more perspectives than just *Classify*.

Acknowledgements Many thanks to: Carla T. Sturm for her bibliographic prowess; and Geraint Wiggins, Nick Collins, Matthew Davies, Fabien Gouyon, Arthur Flexer, and Mark Plumbley for numerous and insightful conversations. Thank you to the numerous anonymous peer reviewers who contributed greatly to this article and its organization.

References

Abeßer, J., Lukashevich, H., Bräuer, P. (2012). Classification of music genres based on repetitive basslines. *Journal of New Music Research, 41*(3), 239–257.

Ahrendt, P. (2006). *Music genre classification systems—A computational approach*. Ph.D. thesis, Technical University of Denmark.

Ammer, C. (2004). *Dictionary of music* (4th ed.). New York: The Facts on File, Inc.

Andén, J., & Mallat, S. (2011). Multiscale scattering for audio classification. In *Proc. International Society for Music Information Retrieval* (pp. 657–662).

Andén, J., & Mallat, S. (2012). *Scatterbox v. 1.02*. http://www.cmap.polytechnique.fr/scattering/. Accessed 15 Oct 2012.

Anglade, A., Benetos, E., Mauch, M., Dixon, S. (2010). Improving music genre classification using automatically induced harmony rules. *Journal of New Music Research, 39*(4), 349–361.

Ariyaratne, H.B., & Zhang, D. (2012). A novel automatic hierachical approach to music genre classification. In *Proc. IEEE International Conference on Multimedia & Expo* (pp. 564–569).

Aucouturier, J.J. (2009). Sounds like teen spirit: Computational insights into the grounding of everyday musical terms. In J. Minett, & W. Wang (Eds.), *Language, evolution and the brain: Frontiers in linguistic series* (pp. 35–64). Academia Sinica Press.

Aucouturier, J.-J. & Bigand, E. (2013). Seven problems that keep MIR from attracting the interest of cognition and neuroscience. *Journal of Intelligent Information Systems*. doi:10.1007/s10844-013-0251-x.

Aucouturier, J.J., & Pachet, F. (2003). Representing music genre: A state of the art. *Journal of New Music Research, 32*(1), 83–93.

Aucouturier, J.J., & Pachet, F. (2004). Improving timbre similarity: How high is the sky? *Journal of Negative Results in Speech and Audio Sciences, 1*(1), 1–13.

Aucouturier, J.J., & Pampalk, E. (2008). Introduction—from genres to tags: A little epistemology of music information retrieval research. *Journal of New Music Research, 37*(2), 87–92.

Bağci, U., & Erzin, E. (2007). Automatic classification of musical genres using inter-genre similarity. *IEEE Signal Processing Letters, 14*(8), 521–524.

Barbedo, J.G.A., & Lopes, A. (2007). Automatic genre classification of musical signals. *EURASIP Journal on Advances in Signal Processing*. doi:10.1155/2007/64960.

Barbedo, J.G.A., & Lopes, A. (2008). Automatic musical genre classification using a flexible approach. *Journal of the Audio Engineering Society, 56*(7/8), 560–568.
Benbouzid, D., Busa-Fekete, R., Casagrande, N., Collin, F.D., Kégl, B. (2012). Multiboost: A multi-purpose boosting package. *Journal of Machine Learning Research, 13*, 549–553.
Benetos, E., & Kotropoulos, C. (2008). A tensor-based approach for automatic music genre classification. In *Proc. European Signal Processing Conference*.
Benetos, E., & Kotropoulos, C. (2010). Non-negative tensor factorization applied to music genre classification. *IEEE Transactions on Audio, Speech, and Language Processing, 18*(8), 1955–1967.
Berenzweig, A., Logan, B., Ellis, D.P.W., Whitman, B. (2004). A large-scale evaluation of acoustic and subjective music-similarity measures. *Computer Music Journal, 28*(2), 63–76.
Bergstra, J., Casagrande, N., Erhan, D., Eck, D., Kégl, B. (2006a). Aggregate features and AdaBoost for music classification. *Machine Learning, 65*(2–3), 473–484.
Bergstra, J., Lacoste, A., Eck, D. (2006b). Predicting genre labels for artist using FreeDB. In *Proc. International Society for Music Information Retrieval* (pp. 85–88).
Bergstra, J., Mandel, M., Eck, D. (2010). Scalable genre and tag prediction with spectral covariance. In *Proc. International Society for Music Information Retrieval* (pp. 507–512).
Bertin-Mahieux, T., Eck, D., Mandel, M. (2010). Automatic tagging of audio: The state-of-the-art. In W. Wang (Ed.), *Machine audition: Principles, algorithms and systems* (pp. 334–352). IGI Publishing.
Bertin-Mahieux, T., Ellis, D.P., Whitman, B., Lamere, P. (2011). The million song dataset. In *Proc. International Society for Music Information Retrieval* (pp. 591–596). http://labrosa.ee.columbia.edu/millionsong/.
Burred, J.J., & Lerch, A. (2004). Hierarchical automatic audio signal classification. *Journal of the Audio Engineering Society, 52*(7), 724–739.
Chai, W., & Vercoe, B. (2001). Folk music classification using hidden Markov models. In *Proc. International Conference on Artificial Intelligence*
Chang, K., Jang, J.S.R., Iliopoulos, C.S. (2010). Music genre classification via compressive sampling. In *Proc. International Society for Music Information Retrieval* (pp. 387–392).
Chase, A. (2001). Music discriminations by carp "Cyprinus carpio". *Learning & Behavior, 29*, 336–353.
Chen, S.H., & Chen, S.H. (2009). Content-based music genre classification using timbral feature vectors and support vector machine. In *Proc. International Conference on Interaction Sciences: Information Technology, Culture and Human* (pp. 1095–1101).
Collins, N. (2012). Influence in early electronic dance music: An audio content analysis investigation. In *Proc. International Society for Music Information Retrieval* (pp. 1–6).
Craft, A. (2007). *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.
Craft, A., Wiggins, G.A., Crawford, T. (2007). How many beans make five? The consensus problem in music-genre classification and a new evaluation method for single-genre categorisation systems. In *Proc. International Society for Music Information Retrieval* (pp. 73–76).
Cruz, P., & Vidal, E. (2003). Modeling musical style using grammatical inference techniques: a tool for classifying and generating melodies. In *Proc. Web Delivering of Music* (pp. 77–84). doi:10.1109/WDM.2003.1233878.
Cruz, P., & Vidal, E. (2008). Two grammatical inference applications in music processing. *Applied Artificial Intelligence, 22*(1/2), 53–76.
DeCoro, C., Barutcuoglu, S., Fiebrink, R. (2007). Bayesian aggregation for hierarchical genre classification. In *Proc. International Society for Music Information Retrieval* (pp. 77–80).
Deshpande, H., Singh, R., Nam, U. (2001). Classification of music signals in the visual domain. In *Proc. Digital Audio Effects*. Limerick, Ireland.
Dietterich, T. (1996). *Statistical tests for comparing supervised learning algorithms*. Tech. Rep., Oregon State University, Corvallis, OR.
Dixon, S., Mauch, M., Anglade, A. (2010). Probabilistic and logic-based modelling of harmony. In *Proc. Computer Music Modeling and Retrieval* (pp. 1–19).
Fabbri, F. (1982). A theory of musical genres: Two applications. In P. Tagg & D. Horn (Eds.), *Popular music perspectives* (pp. 55–59). Gothenburg and Exeter.
Flexer, A. (2006). Statistical evaluation of music information retrieval experiments. *Journal of New Music Research, 35*(2), 113–120.
Flexer, A. (2007). A closer look on artist filters for musical genre classification. In *Proc. International Society for Music Information Retrieval* (pp. 341–344).

Flexer, A., & Schnitzer, D. (2009). Album and artist effects for audio similarity at the scale of the web. In *Proc. Sound and Music Computing* (pp. 59–64).
Flexer, A., & Schnitzer, D. (2010). Effects of album and artist filters in audio similarity computed for very large music databases. *Computer Music Journal, 34*(3), 20–28.
Freund, Y., & Schapire, R.E. (1997). A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of Computer and System Sciences, 55*, 119–139.
Frow, J. (2005). *Genre*. New York: Routledge.
Fu, Z., Lu, G., Ting, K.M., Zhang, D. (2011). A survey of audio-based music classification and annotation. *IEEE Transactions on Multimedia, 13*(2), 303–319.
García, J., Hernández, E., Meng, A., Hansen, L.K., Larsen, J. (2007). Discovering music structure via similarity fusion. In *Proc. NIPS workshop on music, brain & cognition: Learning the structure of music and its effects on the brain*.
Gasser, M., Flexer, A., Schnitzer, D. (2010). Hubs and orphans—an explorative approach. In *Proc. Sound and Music Computing*.
Gjerdingen, R.O., & Perrott, D. (2008). Scanning the dial: The rapid recognition of music genres. *Journal of New Music Research, 37*(2), 93–100.
Goto, M., Hashiguchi, H., Nishimura, T., Oka, R. (2003). RWC music database: Music genre database and musical instrument sound database. In *Proc. International Society for Music Information Retrieval* (pp. 229–230).
Gouyon, F., & Dixon, S. (2004). Dance music classification: A tempo-based approach. In *Proc. International Society for Music Information Retrieval* (pp. 501–504).
Guaus, E. (2009). *Audio content processing for automatic music genre classification: Descriptors, databases, and classifiers*. Ph.D. thesis, Universitat Pompeu Fabra, Barcelona, Spain.
Holzapfel, A., & Stylianou, Y. (2008). Musical genre classification using nonnegative matrix factorization-based features. *IEEE Transactions on Audio, Speech, and Language Processing, 16*(2), 424–434.
Homburg, H., Mierswa, I., Möller, B., Morik, K., Wurst, M. (2005). A benchmark dataset for audio classification and clustering. In *Proc. International Society for Music Information Retrieval* (pp. 528–531).
Hu, Y., & Ogihara, M. (2012). Genre classification for million song dataset using confidence-based classifiers combination. In *Proc. ACM Special Interest Group on Information Retrieval* (pp. 1083–1084).
Humphrey, E.J., Bello, J.P., LeCun, Y. (2013). Feature learning and deep architectures: New directions for music informatics. *Journal of Intelligent Information Systems*. doi:10.1007/s10844-013-0248-5.
ISMIR (2004). *Genre results*. http://ismir2004.ismir.net/genre_contest/index.htm. Accessed 15 Oct 2012.
Krumhansl, C.L. (2010). Plink: "Thin slices" of music. *Music Perception: An Interdisciplinary Journal, 27*(5), 337–354.
Langlois, T., & Marques, G. (2009). Automatic music genre classification using a hierarchical clustering and a language model approach. In *Proc. International Conference on Advances in Multimedia* (pp. 188–193).
Law, E. (2011). Human computation for music classification. In T. Li, M. Ogihara, G. Tzanetakis (Eds.), *Music data mining* (pp. 281–301). Boca Raton, FL: CRC Press.
Lee, J.W., Park, S.B., Kim, S.K. (2006). 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). Berlin/Heidelberg: Springer. doi:10.1007/11760023_27.
Lerch, A. (2012). *An introduction to audio content analysis: Applications in signal processing and music informatics*. New York, Hoboken: Wiley/IEEE Press.
Li, T., & Ogihara, M. (2004). Music artist style identification by semi-supervised learning from both lyrics and contents. In *Proc. ACM Multimedia* (pp. 364–367).
Lin, C.R., Liu, N.H., Wu, Y.H., Chen, A. (2004). 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). Berlin/Heidelberg: Springer.
Lippens, S., Martens, J., De Mulder, T. (2004). A comparison of human and automatic musical genre classification. In *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing* (pp. 233–236).
Lopes, M., Gouyon, F., Koerich, A., Oliveira, L.E.S. (2010). Selection of training instances for music genre classification. In *Proc. International Conference on Pattern Recognition* (pp. 4569–4572).

Lukashevich, H., Abeßer, J., Dittmar, C., Großmann, H. (2009). From multi-labeling to multi-domain-labeling: A novel two-dimensional approach to music genre classification. In *International Society for Music Information Retrieval* (pp. 459–464).
Mace, S.T., Wagoner, C.L., Teachout, D.J., Hodges, D.A. (2011). Genre identification of very brief musical excerpts. *Psychology of Music, 40*(1), 112–128.
Mallat, S. (2012). Group invariant scattering. *Communications on Pure and Applied Mathematics, 65*(10), 1331–1398.
Marques, G., Domingues, M., Langlois, T., Gouyon, F. (2011a). Three current issues in music autotagging. In *Proc. International Society for Music Information Retrieval* (pp. 795–800).
Marques, G., Langlois, T., Gouyon, F., Lopes, M., Sordo, M. (2011b). Short-term feature space and music genre classification. *Journal of New Music Research, 40*(2), 127–137.
Marques, G., Lopes, M., Sordo, M., Langlois, T., Gouyon, F. (2010). Additional evidence that common low-level features of individual audio frames are not representative of music genres. In *Proc. Sound and Music Computing*.
Matityaho, B., & Furst, M. (1995). Neural network based model for classification of music type. In *Proc. Convention of Electrical and Electronics Engineers in Israel* (pp. 1–5). doi:10.1109/EEIS.1995.514161.
McDermott, J., & Hauser, M.D. (2007). Nonhuman primates prefer slow tempos but dislike music overall. *Cognition, 104*(3), 654–668. doi:10.1016/j.cognition.2006.07.011.
McKay, C. (2004). *Automatic genre classification of MIDI recordings*. Ph.D. thesis, McGill University, Montréal, Canada.
McKay, C., & Fujinaga, I. (2005). Automatic music classification and the importance of instrument identification. In *Proc. Conference on Interdisciplinary Musicology*.
McKay, C., & Fujinaga, I. (2006). Music genre classification: Is it worth pursuing and how can it be improved? In *Proc. International Society for Music Information Retrieval* (pp. 101–106).
Meng, A., Ahrendt, P., Larsen, J. (2005). Improving music genre classification by short-time feature integration. In *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing* (pp. 497–500).
Meng, A., & Shawe-Taylor, J. (2008). An investigation of feature models for music genre classification using the support vector classifier. In *Proc. International Society for Music Information Retrieval* (pp. 604–609).
MIREX (2005). *Genre results*. http://www.music-ir.org/mirex/wiki/2005:MIREX2005_Results. Accessed 15 Oct 2012.
Moerchen, F., Mierswa, I., Ultsch, A. (2006). Understandable models of music collections based on exhaustive feature generation with temporal statistics. In *Int. Conference on Knowledge Discovery and Data Mining* (pp. 882–891).
Pampalk, E. (2006). *Computational models of music similarity and their application in music information retrieval*. Ph.D. thesis, Vienna University of Tech., Vienna, Austria.
Pampalk, E., Flexer, A., Widmer, G. (2005). Improvements of audio-based music similarity and genre classification. In *Proc. International Society for Music Information Retrieval* (pp. 628–233).
Panagakis, Y., & Kotropoulos, C. (2010). Music genre classification via topology preserving non-negative tensor factorization and sparse representations. In *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing* (pp. 249–252).
Panagakis, Y., Kotropoulos, C., Arce, G.R. (2009a). Music genre classification using locality preserving non-negative tensor factorization and sparse representations. In *Proc. International Society for Music Information Retrieval* (pp. 249–254).
Panagakis, Y., Kotropoulos, C., Arce, G.R. (2009b). Music genre classification via sparse representations of auditory temporal modulations. In *Proc. European Signal Processing Conference* (pp. 1–5).
Panagakis, Y., Kotropoulos, C., Arce, G.R. (2010a). Non-negative multilinear principal component analysis of auditory temporal modulations for music genre classification. *IEEE Transactions on Audio, Speech, and Language Processing, 18*(3), 576–588.
Panagakis, Y., Kotropoulos, C., Arce, G.R. (2010b). Sparse multi-label linear embedding nonnegative tensor factorization for automatic music tagging. In *Proc. European Signal Processing Conference* (pp. 492–496).
Pfungst, O. (translated by C.L. Rahn) (1911). *Clever hans (The horse of Mr. Von Osten): A contribution to experimental animal and human psychology*. New York: Henry Holt.
Porter, D., & Neuringer, A. (1984). Music discriminations by pigeons. *Experimental Psychology: Animal Behavior Processes, 10*(2), 138–148.

Ren, J.M., & Jang, J.S.R. (2011). Time-constrained sequential pattern discovery for music genre classification. In *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing* (pp. 173–176).
Ren, J.M., & Jang, J.S.R. (2012). Discovering time-constrained sequential patterns for music genre classification. *IEEE Transactions on Audio, Speech, and Language Processing, 20*(4), 1134–1144.
Rizzi, A., Buccino, N.M., Panella, M., Uncini, A. (2008). Genre classification of compressed audio data. In *Proc. International Workshop on Multimedia Signal Processing* (pp. 654–659).
Salzberg, S.L. (1997). On comparing classifiers: Pitfalls to avoid and a recommended approach. *Data Mining and Knowledge Discovery, 1*, 317–328.
Sanden, C. (2010). *An empirical evaluation of computational and perceptual multi-label genre classification on music*. Master's thesis, University of Lethbridge.
Sanden, C., & Zhang, J.Z. (2011a). Algorithmic multi-genre classification of music: An empirical study. In *Proc. International Computer Music Conference* (pp. 559–566).
Sanden, C., & Zhang, J.Z. (2011b). Enhancing multi-label music genre classification through ensemble techniques. In *Proc. ACM Special Interest Group on Information Retrieval* (pp. 705–714).
Scaringella, N., Zoia, G., Mlynek, D. (2006). Automatic genre classification of music content: A survey. *IEEE Signal Processing Magazine, 23*(2), 133–141.
Schapire, R., & Singer, Y. (1999). Improved boosting algorithms using confidence-rated predictions. *Machine Learning, 37*(3), 297–336.
Schedl, M., & Flexer, A. (2012). Putting the user in the center of music information retrieval. In *Proc. International Society for Music Information Retrieval* (pp. 385–390).
Schedl, M., Flexer, A., Urbano, J. (2013). The neglected user in music information retrieval research. *Journal of Intelligent Information Systems*. doi:10.1007/s10844-013-0247-6.
Schindler, A., Mayer, R., Rauber, A. (2012). Facilitating comprehensive benchmarking experiments on the million song dataset. In *Proc. International Society for Music Information Retrieval* (pp. 469–474).
Schindler, A., & Rauber, A. (2012). Capturing the temporal domain in echonest features for improved classification effectiveness. In *Proc. Adaptive Multimedia Retrieval*.
Schnitzer, D., Flexer, A., Schedl, M., Widmer, G. (2012). Local and global scaling reduce hubs in space. *Journal of Machine Learning Research, 13*, 2871–2902.
Seyerlehner, K., Schedl, M., Sonnleitner, R., Hauger, D., Ionescu, B. (2012). From improved auto-taggers to improved music similarity measures. In *Proc. Adaptive Multimedia Retrieval*.
Seyerlehner, K., Widmer, G., Knees, P. (2010). A comparison of human, automatic and collaborative music genre classification and user centric evaluation of genre classification systems. In *Proc. Adaptive Multimedia Retrieval* (pp. 118–131).
Shapiro, P. (2005). *Turn the beat around: The secret history of disco*. London, UK: Faber & Faber.
Silla, C.N., Koerich, A.L., Kaestner, C.A.A. (2008). The Latin music database. In: *Proc. International Society for Music Information Retrieval* (pp. 451–456).
Slaney, M. (1998). *Auditory toolbox*. Tech. Rep., Interval Research Corporation.
Smith, J.B.L., Burgoyne, J.A., Fujinaga, I., Roure, D.D., Downie, J.S. (2011). Design and creation of a large-scale database of structural annotations. In *Proc. International Society for Music Information Retrieval* (pp. 555–560).
Song, W., Chang, C.J., Liou, S. (2009). Improved confidence intervals on the Bernoulli parameter. *Communications and Statistics Theory and Methods, 38*(19), 3544–3560.
Sturm, B.L. (2012a). A survey of evaluation in music genre recognition. In *Proc. Adaptive Multimedia Retrieval*.
Sturm, B.L. (2012b). Two systems for automatic music genre recognition: What are they really recognizing? In *Proc. ACM Workshop on Music Information Retrieval with User-Centered and Multimodal Strategies* (pp. 69–74).
Sturm, B.L. (2013a). On music genre classification via compressive sampling. In *Proc. IEEE International Conference on Multimedia & Expo*.
Sturm, B.L. (2013b). The GTZAN dataset: Its contents, its faults, their effects on evaluation, and its future use. http://arxiv.org/abs/1306.1461.
Sturm, B.L., & Noorzad, P. (2012). On automatic music genre recognition by sparse representation classification using auditory temporal modulations. In *Proc. Computer Music Modeling and Retrieval*.
Sundaram, S., & Narayanan, S. (2007). Experiments in automatic genre classification of full-length music tracks using audio activity rate. In *Proc. Workshop on Multimedia Signal Processing* (pp. 98–102).

Tacchini, E., & Damiani, E. (2011). What is a "musical world"? An affinity propagation approach. In *Proc. ACM Workshop on Music Information Retrieval with User-Centered and Multimodal Strategies* (pp. 57–62).

Theodoridis, S., & Koutroumbas, K. (2009). *Pattern Recognition* (4th ed.). Amsterdam, The Netherlands: Academic Press, Elsevier.

Turnbull, D., & Elkan, C. (2005). Fast recognition of musical genres using RBF networks. *IEEE Transactions on Knowledge and Data Engineering, 17*(4), 580–584.

Tzanetakis, G., & Cook, P. (2002). Musical genre classification of audio signals. *IEEE Transactions on Speech and Audio Processing, 10*(5), 293–302.

Tzanetakis, G., Ermolinskyi, A., Cook, P. (2003). Pitch histograms in audio and symbolic music information retrieval. *Journal of New Music Research, 32*(2), 143–152.

Umapathy, K., Krishnan, S., Jimaa, S. (2005). Multigroup classification of audio signals using time-frequency parameters. *IEEE Transactions on Multimedia, 7*(2), 308–315.

Urbano, J., Schedl, M., Serra, X. (2013). Evaluation in music information retrieval. *Journal of Intelligent Information Systems*. doi:10.1007/s10844-013-0249-4.

van den Berg, E., & Friedlander, M.P. (2008). Probing the Pareto frontier for basis pursuit solutions. *SIAM Journal on Scientific Computing, 31*(2), 890–912.

Vatolkin, I. (2012). 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). Berlin: Springer.

Wang, F., Wang, X., Shao, B., Li, T., Ogihara, M. (2009). Tag integrated multi-label music style classification with hypergraph. In *Proc. International Society for Music Information Retrieval* (pp. 363–368).

Watanabe, S., & Sato, K. (1999). Discriminative stimulus properties of music in java sparrows. *Behavioural Processes, 47*(1), 53–57.

Wiggins, G.A. (2009). Semantic gap?? Schemantic schmap!! Methodological considerations in the scientific study of music. In *Proc. IEEE International Symposium on Multimedia* (pp. 477–482).

Wright, J., Yang, A.Y., Ganesh, A., Sastry, S.S., Ma, Y. (2009). Robust face recognition via sparse representation. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 31*(2), 210–227.

Wu, M.J., Chen, Z.S., Jang, J.S.R., Ren, J.M. (2011). Combining visual and acoustic features for music genre classification. In *Proc. International Conference on Machine Learning and Applications and Workshops* (pp. 124–129).

Yao, Q., Li, H., Sun, J., Ma, L. (2010). Visualized feature fusion and style evaluation for musical genre analysis. In *Proc. International Conference on Pervasive Computing, Signal Processing and Applications* (pp. 883–886).